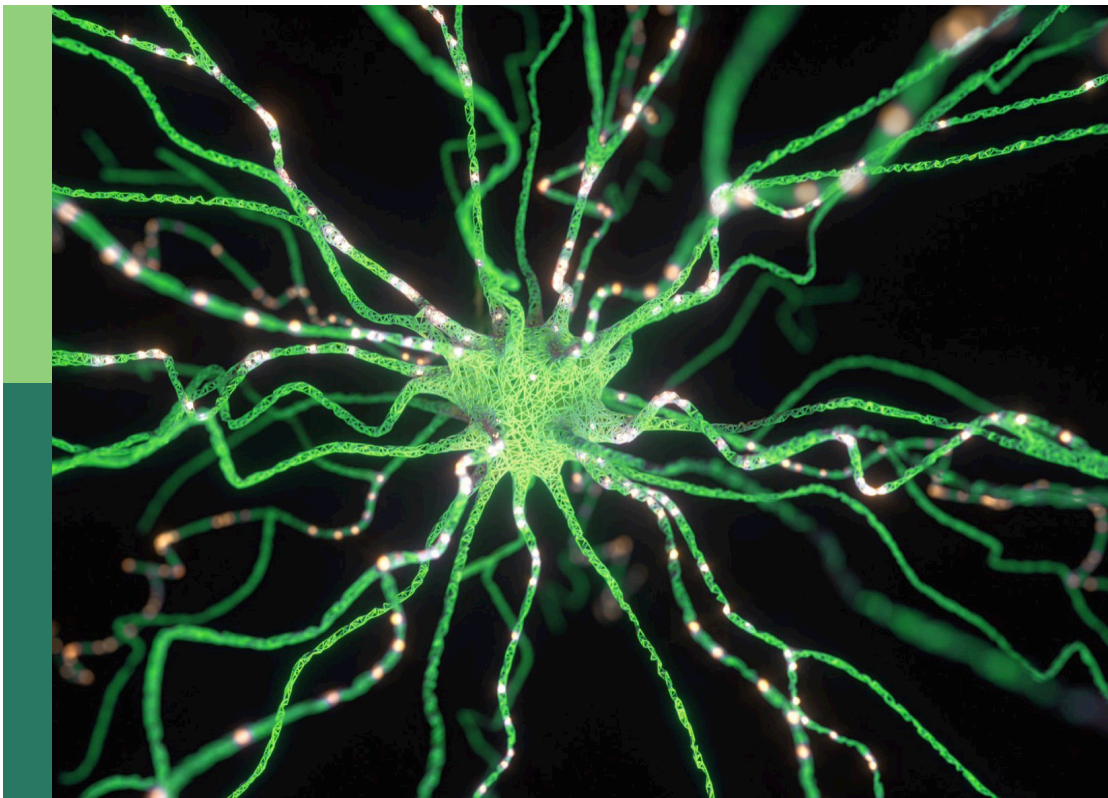
frontiers | Research Topics

Bridging the gap between machine learning and affective computing

Edited by
Zhen Cui, Abhinav Dhall, Xiaopeng Hong, Yong Li,
Wenming Zheng and Yuan Zong

Published in
Frontiers in Neurorobotics
Frontiers in Psychology
Frontiers in Computer Science



ISSN 1664-8714
ISBN 978-2-83250-379-9
DOI 10.3389/978-2-83250-379-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Bridging the gap between machine learning and affective computing

Topic editors

Zhen Cui — Nanjing University of Science and Technology, China
Abhinav Dhall — Monash University, Australia
Xiaopeng Hong — Harbin Institute of Technology, China
Yong Li — Nanjing University of Science and Technology, China
Wenming Zheng — Southeast University, China
Yuan Zong — Southeast University, China

Citation

Cui, Z., Dhall, A., Hong, X., Li, Y., Zheng, W., Zong, Y., eds. (2023). *Bridging the gap between machine learning and affective computing*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-379-9

Table of contents

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 10 August 2021
doi: 10.3389/fnbot.2021.704226

A Novel sEMG-Based Gait Phase-Kinematics-Coupled Predictor and Its Interaction With Exoskeletons

Baichun Wei[1,2], *Zhen Ding*[3], *Chunzhi Yi*[3], *Hao Guo*[1,2], *Zhipeng Wang*[3], *Jianfei Zhu*[3] and *Feng Jiang*[1,2]*

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Pengcheng Laboratory, Shenzhen, China, [3] School of Mechatronics Engineering, Harbin Institute of Technology, Harbin, China

The interaction between human and exoskeletons increasingly relies on the precise decoding of human motion. One main issue of the current motion decoding algorithms is that seldom algorithms provide both discrete motion patterns (e.g., gait phases) and continuous motion parameters (e.g., kinematics). In this paper, we propose a novel algorithm that uses the surface electromyography (sEMG) signals that are generated prior to their corresponding motions to perform both gait phase recognition and lower-limb kinematics prediction. Particularly, we first propose an end-to-end architecture that uses the gait phase and EMG signals as the priori of the kinematics predictor. In so doing, the prediction of kinematics can be enhanced by the ahead-of-motion property of sEMG and quasi-periodicity of gait phases. Second, we propose to select the optimal muscle set and reduce the number of sensors according to the muscle effects in a gait cycle. Finally, we experimentally investigate how the assistance of exoskeletons can affect the motion intent predictor, and we propose a novel paradigm to make the predictor adapt to the change of data distribution caused by the exoskeleton assistance. The experiments on 10 subjects demonstrate the effectiveness of our algorithm and reveal the interaction between assistance and the kinematics predictor. This study would aid the design of exoskeleton-oriented motion-decoding and human–machine interaction methods.

Keywords: electromyography, motion decoding algorithm, kinematics prediction, gait recognition, long short-term memory

OPEN ACCESS

Edited by:
Yuan Zong,
Southeast University, China

Reviewed by:
Deepak Joshi,
Indian Institutes of Technology (IIT), India
Ye Ma,
Ningbo University, China

***Correspondence:**
Feng Jiang
fjiang@hit.edu.cn

Received: *02 May 2021*
Accepted: *15 July 2021*
Published: *10 August 2021*

Citation:
Wei B, Ding Z, Yi C, Guo H, Wang Z, Zhu J and Jiang F (2021) A Novel sEMG-Based Gait Phase-Kinematics-Coupled Predictor and Its Interaction With Exoskeletons. Front. Neurorobot. 15:704226. doi: 10.3389/fnbot.2021.704226

INTRODUCTION

For the past few decades, with the development of human–machine interaction and human motion-decoding methods, an advanced technology was developed to bridge the gap between the human and robots (Bonato, 2010). This robotic technology, known as the wearable robot, directly interacts with the human body to enhance the mobility of healthy people (exoskeletons), to treat muscles or skeletal parts which are injured or after the operation (orthosis), or to replace the missing limbs of disabled people (prostheses) (Viteckova et al., 2013; Chadwell et al., 2020).

As an important branch of wearable robots, the lower-limb exoskeletons run in parallel to the human lower-limbs, with representative applications to daily assistance, medical rehabilitation, and other areas (Kazerooni, 2008; Sankai, 2010; Awad et al., 2017). In recent years, with the

development of human-machine interaction technology and advanced wearable sensors, the exoskeletons have been able to decode the human motions based on physiological or kinematic signals, meanwhile autonomously and promptly assist the user's locomotion at the critical timing, which has enhanced the initiative and intelligence of the system (Yan et al., 2015).

Surface electromyography (sEMG), one of the commonly used neural signals for motion-decoding, integrates the spatial and temporal information of the muscles (Joshi et al., 2013). The amplitude of sEMG is highly related to the level of muscle activation, owing to which sEMG is widely used in control strategies of exoskeletons (Yang et al., 2008; Fan and Yin, 2009). The traditional and practical control strategy for exoskeletons and prostheses is known as the 'direct myoelectric control' approach. The strategy collects the sEMG signals to control the motors of the mechanical joints (Williams, 1990). Although this control strategy has achieved considerable reliability, it becomes non-intuitive when the number of mechanical joints increases. The user training process also tends to be quite time-consuming and cumbersome (Resnik et al., 2018).

As a potential solution to the problem, sEMG-based pattern recognition methods have been developed for motion-decoding and myoelectric control, which seeks the synergistic relationship between muscles based on multichannel sEMG signals, and then matches it with the defined patterns (Scheme and Englehart, 2011). For lower-limb exoskeletons, the motion pattern that is necessary for achieving the mode switching of the control system is the gait phase, which may help to provide a more proper assistant force on human movement (Vu et al., 2018). One of the commonly used gait phase definitions for exoskeletons is shown in **Figure 1**, which segments the gait cycle based on several significant events, such as the initial contact or the toe off (Taborri et al., 2016).

As a general rule, the sEMG-based phase classification process includes extracting the temporal or spatial-temporal features from window-segmented sEMG signals, followed by a classifier to align the features to the pre-defined phases (Novak and Riener, 2015). Compared with the non-stationary raw sEMG signals, the feature-extraction process maximally separates the desired output classes, with an impressive performance in pattern recognition (Hudgins et al., 1993). However, the feature representation will lead to the increased dimension of data, which may increase the burden to the limited computing equipment of the exoskeleton. Dimension-reduction plays an important role in the related research, with representative methods such as principal component analysis (PCA) (Englehart et al., 2001), linear discriminant analysis (LDA) (Chu et al., 2007), and profile likelihood maximization (Naik et al., 2018). Although various methods were proposed to deal with the 'curse of dimension' problem in the feature space, few studies focused on the source data space, i.e., the selected muscles in the studies. Dealing with the muscle redundancy problem, i.e., removing the muscles that have less effect on phase recognition, will reduce not only the dimension of the input data but also the number of sensors.

Due to the motion continuity, the kinematics of the lower-limb joints is time-varying in a gait phase. In addition, the mode switching of the control system may diminish the continuity and smoothness of assistance during the transition of different phases (Kim et al., 2019). Thus, continuous decoding of lower-limb kinematics is beneficial to provide additional knowledge for more precise exoskeleton control. So far, extensive work has been done to estimate the joint kinematics, such as joint angles (Ngeo et al., 2014) of trajectories (Xia et al., 2018). However, when considering the limited computing power of the exoskeletons, the application of these methods may cause a time delay between the estimated kinematics and the actual occurrence of the motion event, which may reduce the effectiveness of the exoskeleton and even cause a potential injury to the subject (Tanghe et al., 2020). In order to compensate for this time delay, our previous work achieved the ahead-of-time prediction of kinematics (Yi et al., 2021). However, the study did not consider the simultaneous classification of the gait phases, which would be beneficial for kinematics prediction because of the common quasi-periodicity.

For exoskeletons, there exists another problem in applications of sEMG-based motion-decoding methods. According to Sylos-Labini et al. (2014), the assistive forces provided by an exoskeleton may result in a change of the muscle coordination manners (i.e., muscle synergies). Similar conclusions were also given by the related studies that investigated the effect of robotic gait assistance on the muscle function of the subjects (Moreno et al., 2013; Li et al., 2019). The altered muscle functions would cause an unknown distribution change of sEMG, therefore cause adverse effects on the sEMG-based motion-decoding methods. However, there is still a lack of investigation of how the exoskeleton affects the sEMG-based motion decoding methods, which matters a lot for the applications of the methods to exoskeletons.

In this study, we propose a novel motion-decoding method that combines the recognition of gait phases and the prediction of lower-limb joint angles. The main contributions of this paper are integrated as follows:

- We propose an sEMG and gait phase-based continuous lower-limb kinematics predictor, which leverages not only the ahead-of-motion property of sEMG but also the quasi-periodicity of gait phases to present the ahead-of-time joint kinematics prediction.
- We propose a muscle selection scheme in view of the effects of muscles on the classification of gait phases.
- We experimentally quantify how the assistance of an ankle exoskeleton affects the motion-decoding methods and propose a fine-tuning scheme to adapt to the performance degradation caused by exoskeleton assistance.

The structure of the paper is as follows. In Related Works section, the related works are briefly described. Materials and Methods section details the data acquisition process, the experimental design, and the structure of the proposed motion-decoding method. The evaluation metrics validating the effectiveness of our method are also described in this section. The experimental results are detailed in Results section and analyzed in Discussion section. The conclusion underlines the performance of the proposed method in Conclusion section.

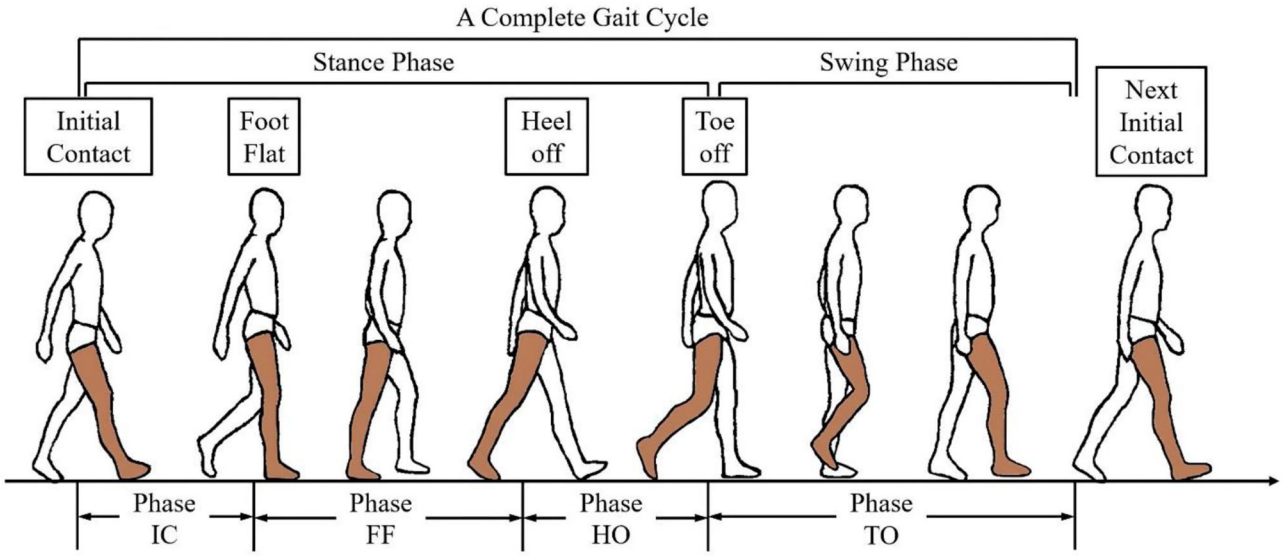


FIGURE 1 | The gait phase definition in a gait cycle.

RELATED WORKS

Phase Recognition and Dimension Reduction

The gait phase recognition is a non-trivial problem for exoskeleton and prosthesis, which is used to permit the control system to work with more initiative and precision (Ferris et al., 2007). Joshi et al. presented a method that combined the Bayesian information criterion and LDA to recognize eight phases based on four-channel EMG signals (Joshi et al., 2013). With an average accuracy of 76.12%, the recognized phases were applied in an exoskeleton orthosis. The study in (Ryu and Kim, 2014) implemented fractal analysis to analyze the change of vibroathrographic signals. Based on four-channel EMG signals, the support vector machine (SVM) classifier could recognize four phases with an average accuracy of 91%.

In recent years, deep learning has revolutionized the fields correlated with machine learning and pattern recognition (LeCun et al., 2015). Compared with other machine learning methods, deep learning is better at searching for the relations of the source data with the labels. In addition, the change of the gait phase is quasi-periodic, which means the temporal-contextual data is beneficial for phase recognition. Because of the reasons described above, we adopted the Long-Short Term Memory (LSTM) to design the phase classifier.

For exoskeleton systems, the motion-decoding algorithms usually run on an onboard microcomputer, which means the source data need to be carefully selected to avoid the control system hysteresis caused by high computational complexity. Moreover, the feature extraction process increases the dimension of the input data by multiples, which may add another layer of complexity. Thus, dimension reduction usually plays an important role in exoskeleton systems. The study in Chu et al. (2007) compared different feature projection methods, such as LDA and PCA, and evaluated through Sammon's stress and Fisher's index. A study by Naik et al. (2018) introduced a screen-plot-based statistical technique for feature reduction. With the implementation of the Fisher score, the method reduced the feature dimension from 28 to 13.

Although various dimension reduction methods have been proposed to avoid the model overfitting and reduce computational complexity, few studies have analyzed the selected muscles. In their works, the muscles were mostly determined by related works or experiences. In this study, we propose a muscle selection scheme that analyzes the effects of muscles on phase classification. Through this scheme, the redundant muscles will be discarded in order to both reduces the dimension of the data and the number of the sensors.

Continuous Decoding of Joint Kinematics

Because most of the lower-limb exoskeletons are located at the joints, such as the knee or the ankle, it is beneficial to obtain the kinematic parameters of the joints, which provide more continuous and detailed knowledge for smooth control. See et al. solved the joint axis using the numerical optimization method, established the limb coordinate system, and calculated the lower limb joint angle based on the IMU signals (Seel et al., 2014). Ameri et al. proposed a real-time upper limb wrist joint trajectory decoding method based on support vector regression (SVR). They implemented this method for proportional control based on EMG signals (Ameri et al., 2014). In the study of Xia et al. (2018), a deep architecture-based model was proposed to estimate the limb trajectory, which combined the convolutional neural network (CNN) and recurrent neural networks (RNN). The results showed that the accuracy and robustness of the proposed method are much higher than those of SVR and CNN.

Although the above studies have shown considerable performance, the time delay in control hinders their application

to exoskeleton systems. In order to deal with this problem and enhance the control of exoskeleton, the future joint kinematics are required. Kevin et al. proposed a probabilistic model to present the future prediction of the current kinematics and gait events, which leveraged the quasi-periodicity of lower-limb motions (Tanghe et al., 2020). The method presented a pioneer frame for kinematic prediction, and it can be enhanced by physiological knowledge. According to the previous studies, there exists a time delay between the onset of the sEMG and the occurrence of the movement (Hioki and Kawasaki, 2012). This phenomenon, known as the electromechanical delay (EMD), can be helpful for the ahead-of-time prediction of kinematics. Thus, we propose an LSTM-based lower-limb kinematic predictor, which leverages the quasi-periodicity of phase and EMD to present the ahead-of-time lower-limb joint angles.

Effects of Exoskeletons on Muscle Functions

How the exoskeletons affect the muscle functions of the subjects have been investigated for many years (Steele et al., 2017). Prior studies have revealed that external forces that were provided by the exoskeletons would alter the activity-level and recruitment patterns of the muscle groups (Sylos-Labini et al., 2014; Li et al., 2019). The study (Sylos-Labini et al., 2014) recorded the sEMG activity of six healthy individuals during overground walking with a lower-limb exoskeleton. The result revealed that the activity of some muscles increased in the exoskeleton-assisted condition compared with the normal walking condition, while the other muscles did not change significantly. Pearson correlation coefficients were implemented as another metric to compare the sEMG waveforms in these two conditions, and a significant difference was found. In Steele et al. (2017), muscle synergy and muscle activity were implemented to evaluate the changes in muscle recruitment and coordination patterns. The result revealed that the subjects could selectively modulate the activity of individual muscles and were not constrained to synergistic patterns of muscle coordination.

The related studies designed complete experiments to investigate the effects of exoskeleton on muscle functions, and concluded that exoskeletons could alter the muscle recruitment patterns (Li et al., 2019). However, there is still a lack of research on the investigation of the exoskeleton effect on sEMG-based motion-decoding methods. Such effect is worthy of investigation since sEMG has obvious advantages in application to exoskeletons, such as the EMD and information of kinematics, dynamics, and personal identity. Thus, we experimentally quantified the effect of an ankle exoskeleton on the proposed motion-decoding model. Also, we implemented a fine-tuning scheme to allow the model to adapt to the change of data distribution caused by exoskeletons' assistance.

MATERIALS AND METHODS

Data Acquisition and Experimental Protocol

This study was conducted under the approval of the Chinese Ethics Committee of Registering Clinical Trials, and all the subjects signed the consent form corresponding to the experiments, who could decide to stop the experiment at any time. The subjects include 10 healthy males with an average height of 178 $\pm$ 5 cm and an average weight of 77.6 $\pm$ 10 kg. The data collection was performed using the EMG acquisition equipment (Delsys Trigno, IM type and Avanti type), a designed foot pressure acquisition device, and an optical motion capture system (VICON). At the beginning of the data collecting process, the signals from various acquisition devices were synchronized by a trigger device.

In this study, we constructed two datasets for the experimental protocol. In the first dataset, ten subjects were involved to performed the level-walking on a treadmill with a constant walking speed of 4.5 km/h. As shown in **Figure 2**, nine quadrupolar EMG electrodes were mounted on the lower-limb muscles, some of which have proved the validity in lower-limb motion decoding, corresponding to rectus femoris (RF), vastus lateralis (VL), vastus medialis (VM), tibialis anterior (TA), soleus (SL), biceps femoris (BF), semitendinosus (ST), gastrocnemius medial head (GM) and gastrocnemius lateral head (GL), with a sampling frequency of 1111.11 Hz. In order to decode the lower-limb kinematics of the subjects, 16 reflective markers were attached to the lower-limb, following the experimental scheme of the VICON user guide, and the lower-limb joint angles were collected with a sampling frequency of 100 Hz. In addition, two FSR sensors were attached to the heel and first metatarsal bone of the subject for phase labeling, foot pressure signals were collected with a sampling rate of 500 Hz.

In the second dataset, four of the ten subjects were recruited to participate in the experiments. With an ankle exoskeleton, the subjects performed the level-walking on a treadmill with a speed of 4.5 km/h. Based on the proposed muscle selection scheme described in Ankle Exoskeleton Frame section, a subset was selected from nine muscles to collect the EMG signals. The attachment of VICON markers and FSR sensors are the same as the first dataset. In both datasets, each subject was instructed to complete at least two trials of level-walking. Each trial lasted for 8 mins, and 15-min rest followed with each trial to avoid muscle fatigue.

Ankle Exoskeleton Frame

In this study, an ankle exoskeleton was implemented, which was shown in **Figure 3**. The designed ankle exoskeleton comprised a waist textile belt, two thigh textile belts, a shank textile belt, and an ankle end-effector mounted on the boot. The exoskeleton was actuated by a powerful motor, with the mechanical power transmitted through a flexible Boden cable tether which terminated at the heel.

The electronic control strategy of the exoskeleton was compiled in LabVIEW software and deployed to the Sbrio-9636 controller through a shared local area network, which was a single task mode control. At the event of heel-off, the motor pulled up on the end-effector through the Borden cable to provides an upward force of 100 N beneath the subject's heel, which assisted in reducing the plantarflexion forces provided by the subjects.

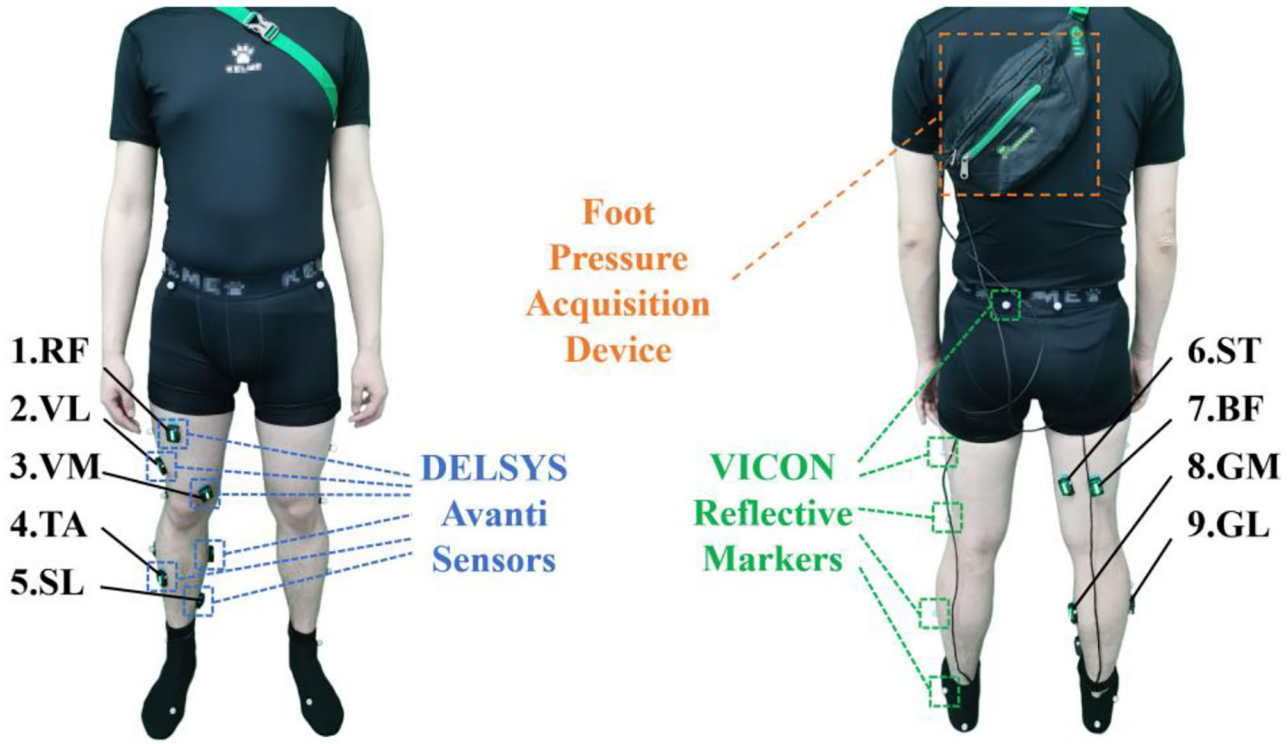


FIGURE 2 | Sensor attachment of the subjects.

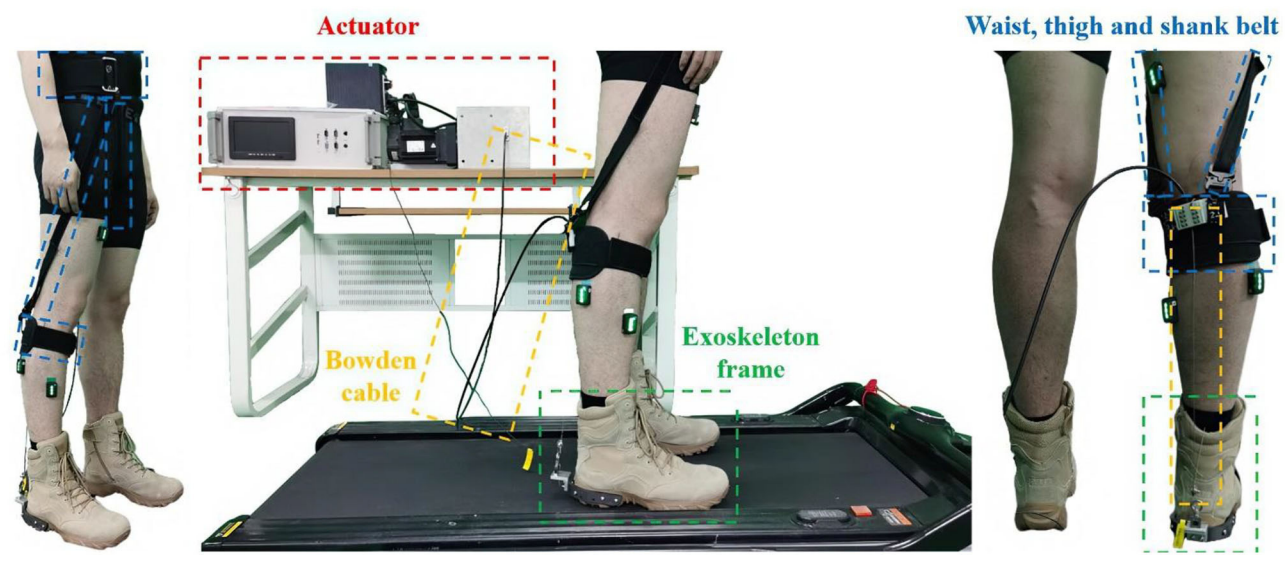


FIGURE 3 | The exoskeleton frame implemented in the second dataset.

Muscle Subset Selection

sEMG signals are generated by nerve signals stimulating muscle activation, which contain massive human motion information. The amplitude and pulse duration of sEMG is highly correlated with the extent and duration of muscle activation, which varies in different phases. **Figure 4** shows the sEMG amplitude of the tibialis anterior from different subjects, which was magnified 10,000 times. From the figure, a phenomenon can be found that the tibialis anterior is mainly activated in the fourth phase among the three subjects, which means that the muscles may not play a

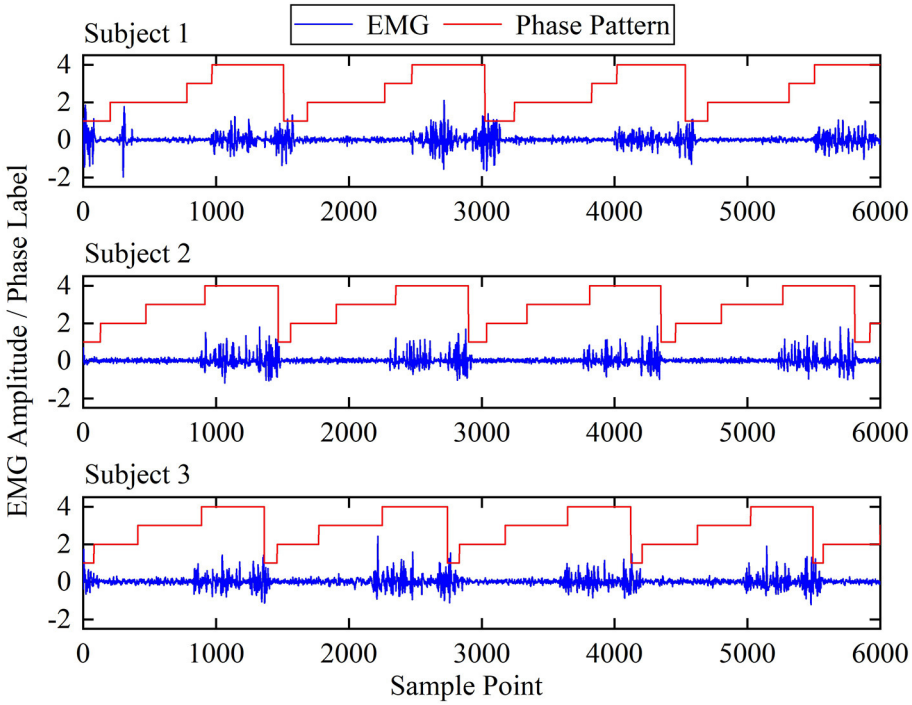


FIGURE 4 | The EMG amplitude of tibialis anterior among different subjects.

role in walking all the time. Instead, they activate at a certain time of the gait cycle. Moreover, although the sEMG amplitude and vibration frequency are different among the subjects, the timings of sEMG pulses in a gait cycle are roughly the same, which means different subjects may share a similar pattern of muscle activation (Chvatal et al., 2011).

Based on the assumption, a muscle-activation-based muscle selection scheme was proposed, which evaluated the effects of the muscles on phase recognition. Firstly, a standard manipulation was implemented to remove the motion artifact and other interference (Ngeo et al., 2014). Then, the signals were processed by full-wave rectification and normalized by dividing by the peak rectified EMG. A low-pass filter was carried out for the processed signal, as the frequency of muscle activation was much lower than that of EMG signals (Ding et al., 2011).

After the above manipulation, the neural activation u(i) of the ith processed EMG sample e(i) with TE sampling interval was calculated as follows:

$$u\,(i) = \alpha \times e\left(i - \frac{d}{T_E}\right) - \beta_1 \times u\,(i-1) \; - \beta_2 \times u\,(i-1) \qquad (1)$$

where α, β_1 and β_2 are the recursive coefficients that maintain the stability of u(i), d is the time delay. Based on the neural activation derived from sEMG signal, the corresponding muscle activation a(t) was calculated by a simplified model (Lloyd and Besier, 2003). In equation (2), A is the nonlinear shape factor that varies between −3 and 0, with −3 represents highly exponential and 0 represents a linear relationship. This factor and the recursive coefficients can be determined by minimizing a mean-square error cost function (Ngeo et al., 2014). In this study, A is equal to −2.

$$a\,(t) = \frac{e^{Au(t)} - 1}{e^{A} - 1} \qquad (2)$$

Muscle activation sequence was calculated from EMG signals of each channel. Then, data of a gait cycle was extracted and segmented by different phases. After that, the average area Ai of muscle activation a(t) in phase i was calculated by:

$$A_i = \int_{t_p}^{T_p} a\,(t)\,dt \quad i = 1, 2, 3, 4 \qquad (3)$$

Through the above calculation, A_i corresponding to four phases was obtained. In order to compare the activations of muscles in different phase more intuitively, a normalization operation was implemented to obtain the effect E_i of muscle to the ith phase. We would then evaluate the muscles based on the muscle effects, following the rule that at least four muscles should be selected, which have the highest activation in the corresponding four phases, and the muscles with similar activations in at least three phases would be discarded.

$$E_i = \frac{A_i}{max(A)} \quad i = 1, 2, 3, 4 \qquad (4)$$

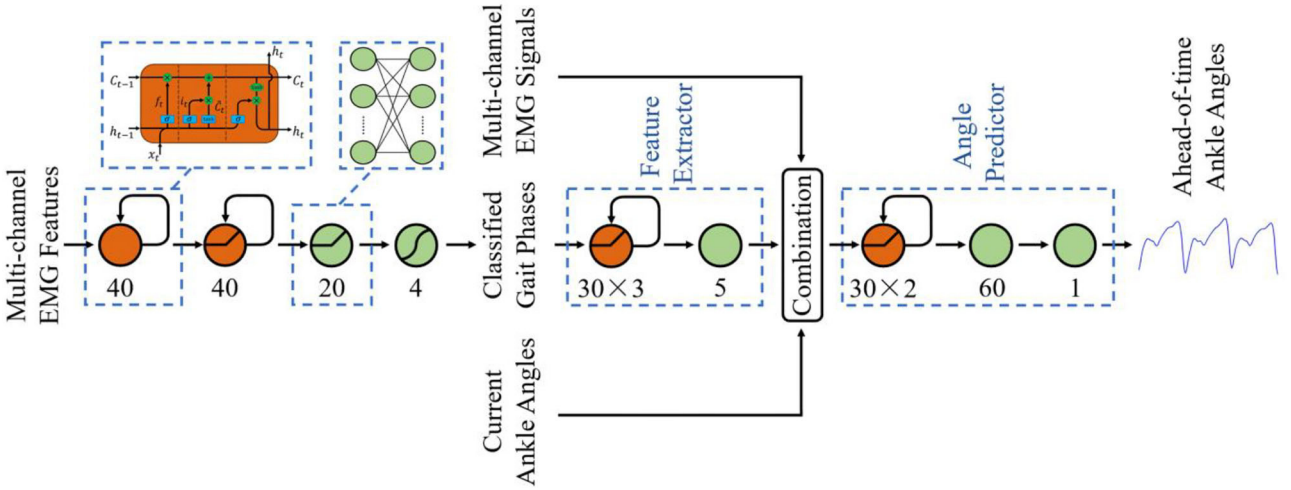


FIGURE 5 | The structure of the phase classifier and ankle angle predictor.

Data Processing

After the optimal muscle subset was determined, the multi-channel EMG and joint angle streams were segmented by a continuous sliding window scheme, with a window length of 180 ms and a window increment of 40 ms (Englehart and Hudgins, 2003). In order to facilitate the phase classification and consider the time efficiency, the following time-domain features were extracted from each EMG segment, which were mean absolute value (MAV), zero crossing (ZC), slope signal change (SSC), and waveform length (WL). The effectiveness and the real-time capability of these features had already been verified in the related studies (He et al., 2011; Zhang et al., 2020). The feature vector x of a sliding window with the dimension of 4n is presented in the form of equation (5), where n represents the number of muscles and f denotes the extracted features from each muscle.

$$x = [f_1, f_2, \cdots, f_n] \tag{5}$$

Phase Classifier and Angle Predictor

For the classification of gait phases, options abound of machine learning, such as HMM (Evans and Arvind, 2014), LDA (Joshi et al., 2013), SVM (Ryu and Kim, 2014), etc. However, despite the verified effectiveness of these classifiers, they did not utilize the previous context of the gait phase, which was also an important element because of the quasi-periodicity of the changing phase state. Thus, we designed an LSTM-based phase classifier. The structure of the classifier was shown in **Figure 5**, consisting of an input layer with the dimension equal to the input features, two LSTM hidden layers of 40, a fully connected layer of 20, and a softmax layer of four corresponding to the gait phases. ReLU activation function was used to connect the LSTM layer, the fully connected layer, and the output layer. In order to prevent model overfitting, dropout regularization was applied after every fully connected hidden layer with probabilities of 0.5.

For the ahead-of-time prediction of ankle angles, LSTM was also implemented to leverage the quasi-periodicity of the changing ankle angles and gait phases. Different from the studies of phase classification, few feature extraction methods have been verified to be efficient for angle regression. Thus, as shown in **Figure 5**, a deep structure was designed, which combined a four-layer LSTM-based feature extractor (30-30-30-5) and a three-layer LSTM-based angle predictor (30-30-60-1). ReLU activation function was also implemented to connect the LSTM layer and the fully connected layer.

The models were tested on Nvidia Xavier Module Interface, with the overall running time for a time window was <30 ms. As the window increment was 40 ms, the prediction time of the angle predictor was set to 40 ms to compensate for the time delay and match the kinematics with the next incoming data stream.

Evaluation Metrics

Several quantitative metrics were used to evaluate the performance of our method. The motion-decoding method we proposed is subject-specific. Thus, to improve the reliability of classification results while avoiding the problem of cross-subject, a modified leave-one-out cross-validation was carried out. Each time, one trial from a subject (defined in Data Acquisition and Experimental Protocol section) was regarded as the testing data, and the other trial from the same subject with all trials from other subjects were regarded as the training data. The procedure continued until each trial from each subject was tested. For all the evaluation processes, one-way ANOVA was implemented to validate the significant effect of a single variable on the results.

In order to verify the performance of the proposed classifier, the SVM classifier with the radial basis function kernel and the LDA classifier with the singular value decomposition solver were compared, which were implemented from the scikit-learn library. The feasibility of these classifiers have already been proved in the

related research (He et al., 2011; Naik et al., 2018). In addition, classification accuracy (ACC) was used for the visualization of the performance.

SVR has been implemented to estimate the simultaneous DOFs of the joints in the related study, and outperformed ANN in myoelectric control tasks (Ameri et al., 2014). Thus, to verify the effectiveness of the proposed method, SVR was also implemented for angle prediction tasks in this study. The output of the predictor was a continuous time series of joint angles. Thus, the Pearson correlation coefficient (R-value) was implemented to quantify the linear relationship between the predicted and reference ankle angles:

$$R = \frac{cov(\theta_{pre}, \theta_{ref})}{\sigma_{pre}\,\sigma_{ref}} \tag{6}$$

where θ_{pre} and θ_{ref} are defined as the predicted knee angles and reference knee angles, respectively. σ is the standard deviation, and *cov* represents the covariance. In addition to the similarity evaluation of the signals, the deviation and residual variance between the predicted and reference angles were estimated by the root mean square error (RMSE) and the normalized RMSE (NRMSE), where n denotes the total number of sampled data, and θ of equation (8) represents the predicted knee angles.

$$RMSE = \sqrt{\frac{1}{n}\sum\left(\theta_{pre} - \theta_{ref}\right)^2} \tag{7}$$

$$NRMSE = \frac{RMSE}{\theta_{max} - \theta_{min}} \tag{8}$$

RESULTS

To begin with, the effect of each muscle described in Ankle Exoskeleton Frame section was calculated, which was shown in **Table 1**. In order to avoid the error caused by abnormal phases, the whole procedure was repeated three times, and the corresponding E_i were averaged to obtain the result. Based on the muscle selection scheme, RF, TA, ST, GM, and GL were selected since they contained the muscles with the highest activation level in different phases, and each of them also had a discriminative activation level in another phase (shown in bold values), which might be beneficial for the phase classification task.

Based on the selected muscles, the phase classification accuracy is shown in **Figure 6**, where MA represents the proposed muscle selection scheme. In order to verify the validity of the proposed method, the exhaustive method (EX) was compared. This method searched for optimal muscle subsets based on the classification accuracy, which was a time-consuming way. The result of nine muscles (ALL) was also presented to quantify the loss of information caused by muscle selection. In the figure, the average accuracy of MA (93.15% of LSTM) was a little lower than that of nine muscles (93.59% of LSTM), which meant that the excluded muscles contained some effective

TABLE 1 | Effects of nine muscles on different phase patterns.

Muscle	Muscle effects on different gait phases			
	IC	FF	HO	TO
RF	**1**	**0.49**	0.25	0.29
VM	1	0.35	0.27	0.33
VL	1	0.54	0.56	0.53
TA	**0.81**	0.18	0.21	**1**
SL	0.21	0.94	1	0.21
ST	**0.74**	0.24	0.23	**1**
BF	1	0.88	0.96	0.87
GM	0.16	**0.61**	**1**	0.16
GL	**0.44**	**0.78**	**1**	**0.18**

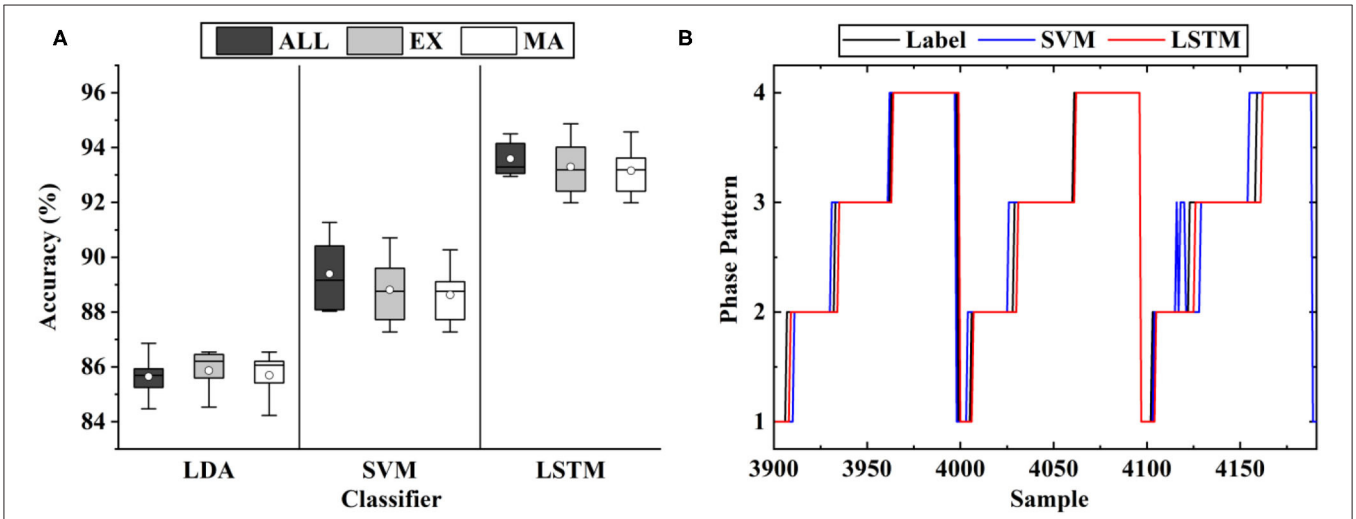


FIGURE 6 | The classification results of gait phases: **(A)** The results of the different muscle sets, where MA represents the proposed muscle selection scheme, EX represents the exhaustive method and ALL represents all the nine muscles; **(B)** The representative results of the three classifiers.

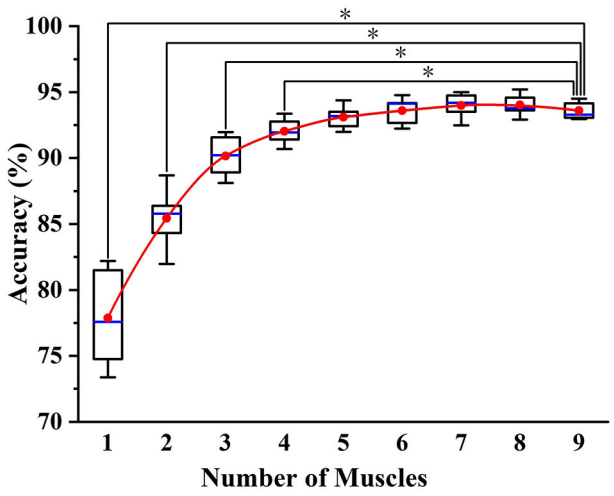


FIGURE 7 | The classification results of the optimal muscle subsets based on different muscle numbers. * indicates a statistically significant difference (one-way ANOVA, $P < 0.05$).

information, but no statistically significant difference was found ($P > 0.05$). In addition, the accuracy of MA was almost the same as that of EX (93.29% of LSTM, $P > 0.05$). When comparing the muscle subsets obtained by MA and EX, we discovered that the muscle subsets of six subjects were the same, while those of the other four were a little different, largely due to the error caused by muscle palpation and sensor location.

As shown in **Figure 6B**, the error mostly occurred in the transition of the phases, largely due to the ambiguity of the phase boundaries. As shown in the figure, the average classification accuracy of LSTM (93.15%) was significantly higher than that of SVM (88.63%) and LDA (85.69%). The same inference was also given when comparing the phase boundaries deduced by the classifiers. Thus, LSTM was implemented as the classifier in the following experiments.

Figure 7 shows the classification results of the optimal muscle subsets based on different muscle numbers. The result of five muscles was based on the muscle selection scheme, while the results of other muscle numbers were based on the exhaustive method. From the figure, a phenomenon could be found that the average accuracy of nine muscles (93.59%) was lower than that of eight muscles (94.02%) and seven muscles (93.98%), which was largely due to the muscle redundancy. In addition, the result of five muscles (93.15%) was not significantly different from that of nine muscles ($P > 0.05$), while that of four muscles was the opposite ($P < 0.05$). It meant that the selected muscle subset contained the minimum number of muscles while retaining the classification accuracy as much as possible.

Figure 8 depicts the representative results of the angle predictor based on different data inputs. In the figure, Angle Only represent the inputs of one-channel current angles, while EMG and Phase-based represents those of five-channel sEMG, one-channel phases and one-channel current angles. As SVR is not able to extract features from sEMG, the feature set of **Muscle Subset Selection** section was implemented. As shown in the figure, the proposed LSTM-based predictor outperformed SVR in both Angle Only and EMG and Phase-based conditions.

In **Figure 9**, the results of different data inputs were evaluated by the three metrics, where EMG-based represents the inputs of five-channel sEMG and one-channel current angles. As shown in the figure, the predicted angles of LSTM were significantly better than those of SVR (RMSE, 1.89° versus 6.51°; NRMSE, 20.07 versus 5.83%; *R*-value, 0.97 versus 0.41). For LSTM, it is shown that the results of EMG and Phase-based outperformed those of EMG-based, and a significant difference was found in the comparison of the results ($P < 0.05$). Thus, the data stream of sEMG and phases, and LSTM-based predictor were implemented in the following experiments.

The effects of exoskeletons on phases have been quantified in **Figure 10**, where wo to w Exo represents that the model was trained in wo Exo (without exoskeleton) condition and tested in w Exo (with exoskeleton) condition. When the classifier was trained and tested in a single condition, the accuracy is quite high and stable, exhibiting that the muscle recruitment pattern of w Exo is as stationary as that of wo Exo. However, when the training and testing sets came from different conditions, the accuracy declined significantly. The most influenced phases were the IC (92.63–56.61%) and HO (93.91–77.12%), which corresponded to the difference in phase duration. A possible reason for this significant decline is that the altered muscle function significantly affects the distribution of sEMG, which have been reported in the related studies (Sylos-Labini et al., 2014; Li et al., 2019).

As shown in **Figure 11**, the results of angle prediction also supported the above view. In order to control the number of variables, the input phases of the predictor were the labels. Similar to the phase classifier, the angle predictor performed quite well in the single condition, but the accuracy declined significantly when the training and testing set came from different conditions (RMSE, 1.89°-5.68°; NRMSE, 5.83–17.52%; *R*-value, 0.97–0.84).

In order to investigate the difference in muscle function in the two conditions and pursue a potential solution to the decline of accuracy, we adopted the fine-tuning method to update the classifier. Each time, 1 min w Exo data was added to update the model, which had already been trained by wo Exo data. The rest of the w Exo data was regarded as the testing set. As shown in **Figure 12**, the accuracies of phase IC and HO significantly increased (IC, 55.61–79.81%; HO, 77.12–87.13%) when the model was updated by 1-min data. In addition, the accuracy gradually stabilized when 4-min data was added, and the accuracy was roughly the same as that in the single w Exo condition (IC, 92.87 versus 92.63%; FF, 93.45 versus 93.91%).

The results of fine-tuning-based angle prediction are shown in **Figure 13**. The accuracy was also significantly increased when 1-min data was added (RMSE, 5.68°-3.05°; NRMSE, 17.41–9.33%; *R*-value, 0.84–0.90), and gradually stabilized when 2-min data was added. Although the performance was not as good as that of only w Exo condition (RMSE, 2.52° versus 1.89°; NRMSE, 7.51 versus 5.83%; *R*-value, 0.95 versus 0.97), it was accurate enough to perform the ahead-of-time angle prediction.

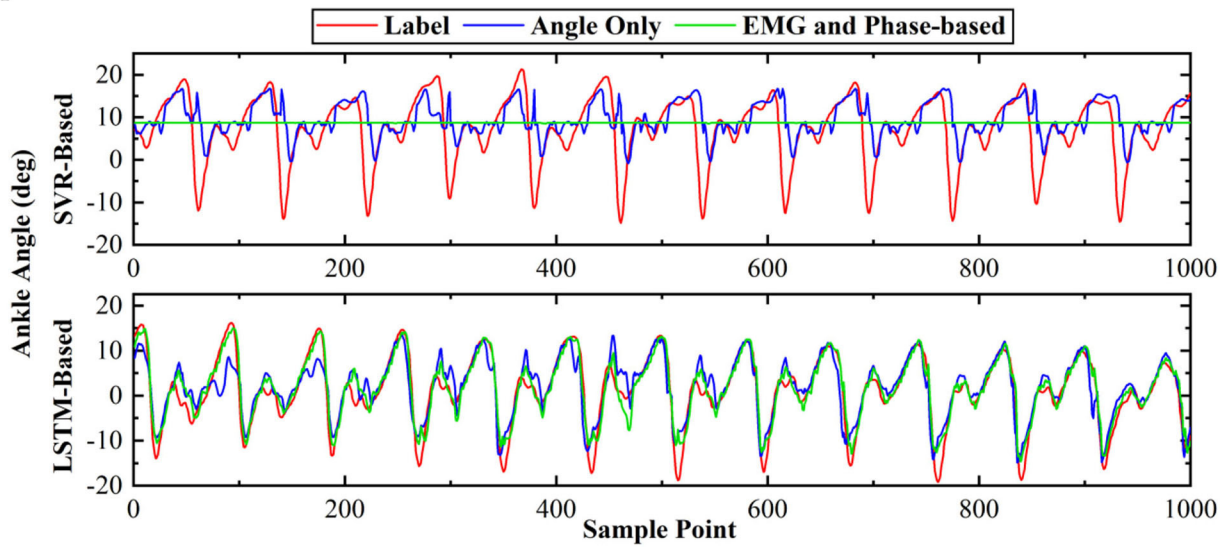


FIGURE 8 | The representative results of different ankle angle predictors.

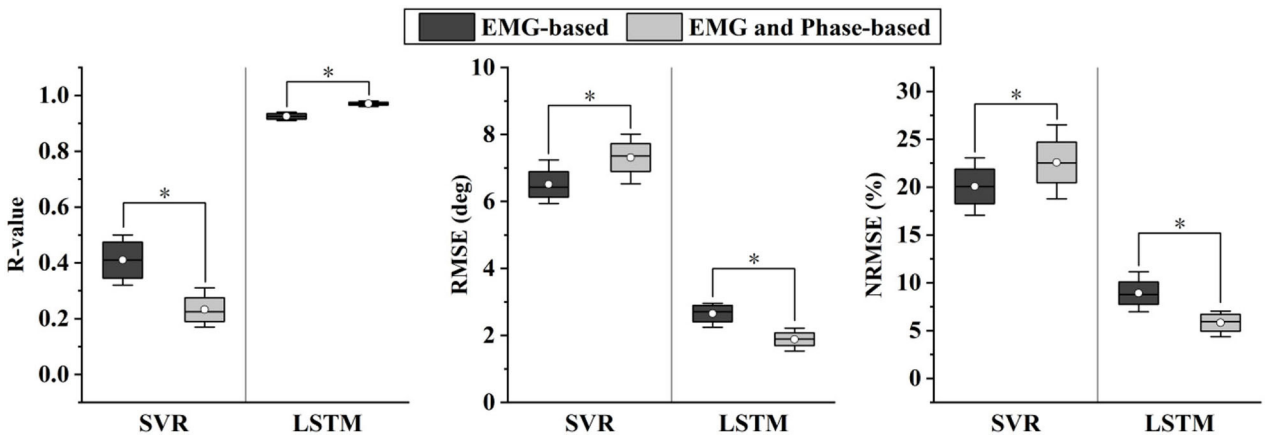


FIGURE 9 | Comparison of different angle predictors based on two evaluation metrics. * indicates a statistically significant difference (one-way ANOVA, $P < 0.05$).

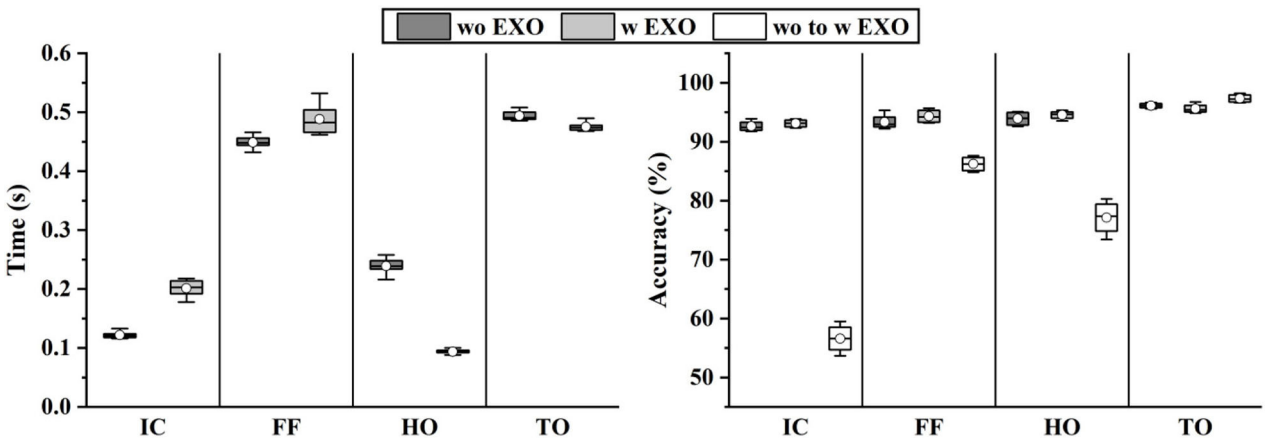


FIGURE 10 | The duration and phase classification results involving the exoskeleton.

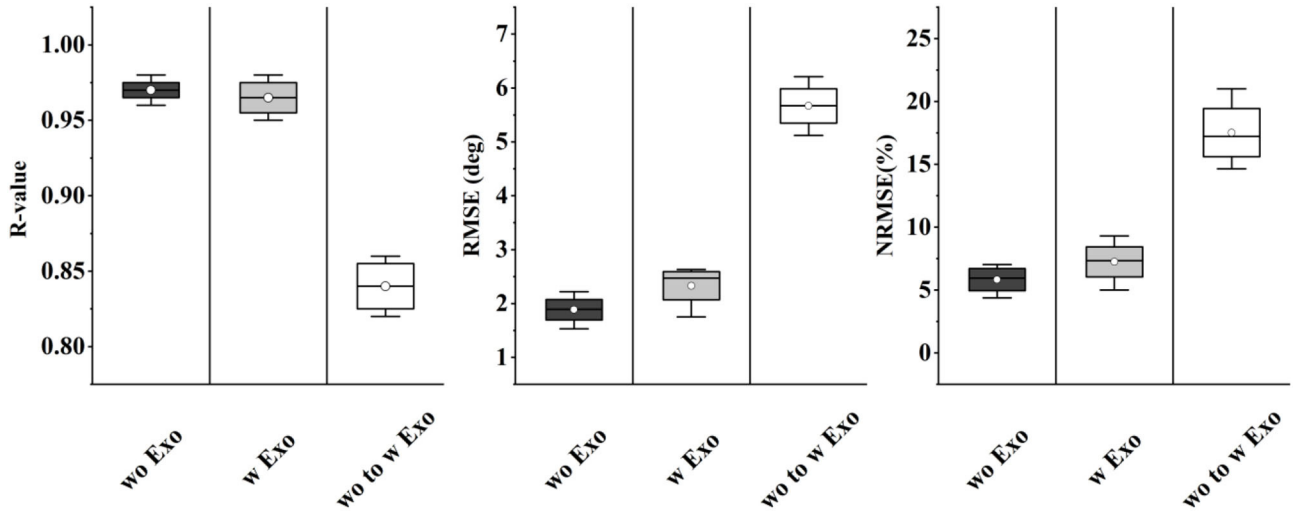


FIGURE 11 | The prediction results involving the exoskeleton.

DISCUSSION

As noted in the study, we proposed a novel sEMG and phase-based angle predictor and compared the contributions akin to ours. Through the muscle selection scheme, we reduced the number of muscles from nine to five, and the changes have little effect on the accuracy. The proposed method, which combined phase recognition and ankle angle prediction, significantly outperformed the related methods. In addition, through the fine-tuning scheme, the feasibility of the method was also verified in the exoskeleton condition, which effectively counteracted the signal distribution changes caused by exoskeleton assistance.

For data dimension reduction tasks, related studies either directly projected the data to the lower-dimensional space or selected the features that would best discriminate various movements via source estimates (Chu et al., 2007; Naik et al., 2018). Based on evaluating the muscle effects in each gait phase, we both reduced the dimension of data and the number of sensors. In addition, a surprising result is shown in **Figure 7**, exhibiting that the accuracy of nine muscles is slightly lower than that of eight and seven muscles. It indicated that some muscles might be not beneficial or even adverse to phase classification. In general, a viewpoint can be summarized that for phase classification, it is preferable to construct a muscle set with the activation of the muscles that are discriminative in different phases, rather than add as many muscles as possible to allow the classifiers to search for a complete muscle-phase relationship.

Various studies have been proposed for motion-decoding tasks, such as the discrete locomotion and gait phase recognition (Godiyal et al., 2018a,b), or continuous kinematic and dynamic estimation (Lloyd and Besier, 2003; Yi et al., 2018). However, the control of an exoskeleton can be enhanced if information in the future is available. Recently, Tanghe et al. proposed an IMU-based kinematics predictor, which was oriented to exoskeletons (Tanghe et al., 2020). Compared with their work, we leveraged the prior knowledge of the gait phase and EMD for kinematics prediction. In addition, transfer learning can be easily applied to the proposed data-driven method, especially when data distribution changes due to the intervention of exoskeletons. As shown in **Figure 8**, the results of sEMG and phase-based were significantly better than those of angle-based. The possible reason is twofold. Firstly, the EMD property of sEMG provides the ahead-of-time information for the prediction of the incoming ankle angles, which has been extracted by the deep LSTM-based feature extractor. Secondly, the joint training process both optimize the feature extractor and angle predictor, and reinforces the correlation between sEMG signals, phases, and ankle angles. In addition, the effect of phase priori for angle prediction was

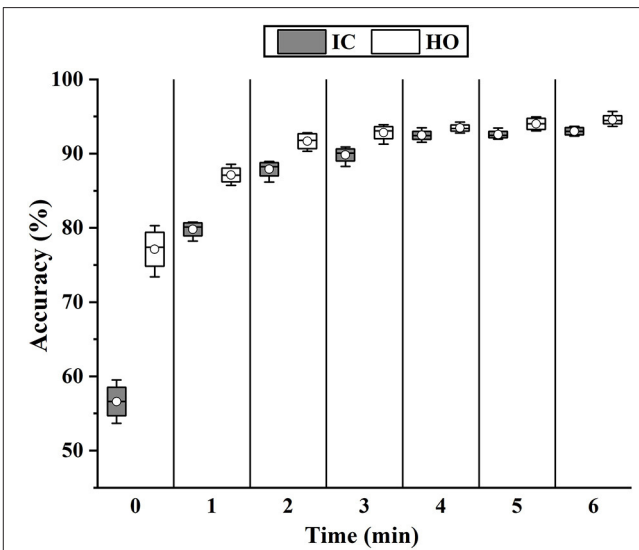


FIGURE 12 | The change of classification accuracy when more data of w Exo was added to update the fine-tuning-based classifier.

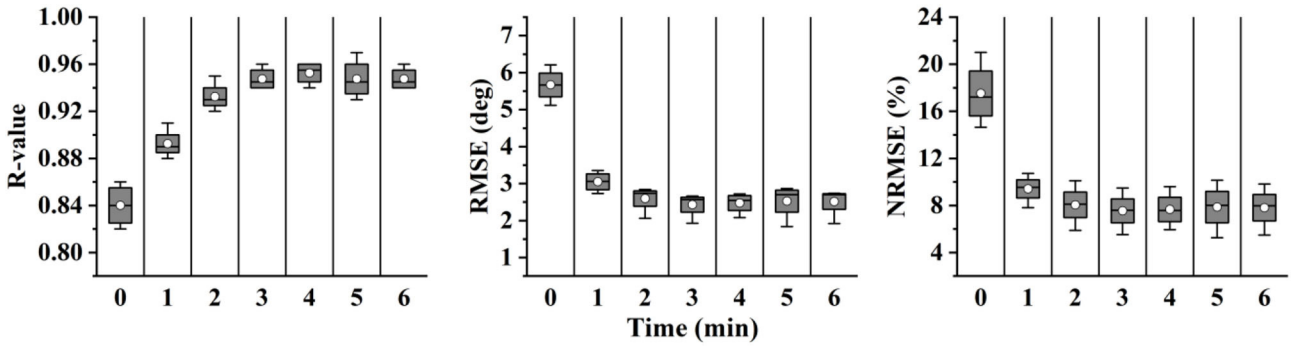


FIGURE 13 | The change of prediction performance when more data of w Exo was added to update the fine-tuning-based predictor.

also tested. The results shown in **Figure 9** suggested that besides sEMG, the gait phase also provided the prior knowledge for angle prediction, thus further improved the prediction accuracy.

In this study, we investigated and quantified how the exoskeleton affected the sEMG-based motion decoding methods. As shown in **Figures 10**, **11**, the results significantly declined when the training and testing data came from different conditions. The interference of the exoskeleton was considered to be the main reason for this phenomenon. First, the exoskeleton disturbed the walking patterns, which was shown in **Figure 10A** and was also reported in Tanghe et al. (2020). Second, the exoskeleton altered the muscle recruitment patterns, exhibiting that the muscles were not restricted to the fixed synergistic patterns. They will selectively modulate the activity given the external interference, instead (Steele et al., 2017).

In order to seek a potential solution to this problem, the fine-tuning scheme was implemented to update the model. As shown in **Figures 12**, **13**, when adding 1-min w Exo data into the training set, the performance of the phase classifier and angle predictor significantly increased. This phenomenon suggested that a correlation might exist between the altered muscle synergy and the original one, thus enabling the fine-tuning of the model with a small size of data. In addition, the performance of the models that were updated through 4-min data was close to that of the models based on whole data of the trials with the exoskeleton, which validated the feasibility of the proposed scheme.

Despite the LSTM-based angle predictor achieved good performance in ahead-of-time ankle angle prediction, there is still room to improve the validity of the method. Since the phases were inputs of the angle predictor, the error caused by phase misclassification would affect the performance of angle prediction. Therefore, the proper post-processing procedure is beneficial to reduce the occurrence of the accumulated error. In addition, even though the fine-tuning scheme was validated to be efficient for the accuracy decline of the model caused by exoskeleton interference, the need for data of trials with the exoskeleton is still inconvenient. The adaption of motion-decoding methods from normal walking to exoskeleton-involved walking would be an important pointer for future research, which necessitates a larger dataset with sufficient subjects and more investigation of the effects of the exoskeleton on muscle functions.

CONCLUSION

In this study, we proposed a novel ankle angle predictor, which presented the prediction of kinematics. First of all, a reduced set of muscles was selected by the proposed muscle selection scheme, which was meant to reduce the data dimension in the muscle level. Secondly, An LSTM-based phase classifier was designed to assign the sEMG to four phases. Finally, with the aid of sEMG and phases, the proposed angle predictor presents the ahead-of-time prediction based on the measured ankle angles.

In order to investigate the perturbance of the exoskeleton to the proposed method, the method was trained on a dataset for normal walking and tested on a dataset for walking with an exoskeleton. From the result, we showed that the method is effective for both phase classification and angle prediction on the training set, while the accuracy significantly declined on the testing set. In order to compensate for the decline of accuracy, a fine-tuning scheme was implemented. After the model update manipulation, the accuracy of phase classification and angle prediction on the testing set had significantly increased and close to that on the training set.

The method enabled the quantitative compensation for the time delay of the exoskeletons, which offers opportunities to achieve a more accurate and smooth control system. In addition, the study enabled us to comprehend the inherent limitations for the applications of the motion-decoding method to the exoskeletons. Being cognizant of these factors, our future work objective is to explore the physiological mechanism of human-exoskeleton interaction and seek for a solution to allow the exoskeletons to adapt to a new subject without the pretraining procedure.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Chinese Ethics Committee of Registering Clinical Trials. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BW and CY: conceptualization and writing–original draft preparation. BW and ZD: methodology. BW: software and visualization. BW and FJ: validation and investigation. BW and HG: formal analysis and data curation. FJ: resources, supervision, project administration, and funding acquisition. HG, ZW, and JZ: writing–review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research is partially funded by National Key Research and Development Program of China via Grant 2018YFC0806802 and 2018YFC0832105.

ACKNOWLEDGMENTS

The authors would like to thank all subjects who participated in experiments and the members of HIT human motion perception team.

REFERENCES

Ameri, A., Kamavuako, E. N., Scheme, E. J., Englehart, K. B., and Parker, P. A. (2014). Support vector regression for improved real-time, simultaneous myoelectric control. *IEEE Trans. Neural Syst. Rehabil. Eng.* 22, 1198–1209. doi: 10.1109/TNSRE.2014.2323576

Awad, L. N., Bae, J., O'Donnell, K., De Rossi, S. M. M., Hendron, K., Sloot, L. H., et al. (2017). A soft robotic exosuit improves walking in patients after stroke. *Sci. Transl. Med.* 9:eaai9084. doi: 10.1126/scitranslmed.aai9084

Bonato, P. (2010). Wearable sensors and systems. *IEEE Eng. Med. Biol. Mag.* 29, 25–36. doi: 10.1109/MEMB.2010.936554

Chadwell, A., Diment, L., Micó-Amigo, M., Morgado Ramírez, D. Z., Dickinson, A., Granat, M., et al. (2020). Technology for monitoring everyday prosthesis use: a systematic review. *J. NeuroEng. Rehabil.* 17:93. doi: 10.1186/s12984-020-00711-4

Chu, J.-U., Moon, I., Lee, Y.-J., Kim, S.-K., and Mun, M.-S. (2007). A supervised feature-projection-based real-time emg pattern recognition for multifunction myoelectric hand control. *IEEE/ASME Trans. Mechatron.* 12, 282–290. doi: 10.1109/TMECH.2007.897262

Chvatal, S. A., Torres-Oviedo, G., Safavynia, S. A., and Ting, L. H. (2011). Common muscle synergies for control of center of mass and force in nonstepping and stepping postural behaviors. *J. Neurophysiol.* 106, 999–1015. doi: 10.1152/jn.00549.2010

Ding, Q. C., Xiong, A. B., Zhao, X. G., and Han, J. D. (2011). "A novel EMG-driven state space model for the estimation of continuous joint movements," in *IEEE,* 2891–2897. doi: 10.1109/ICSMC.2011.6084104

Englehart, K., Hudgin, B., and Parker, P. A. (2001). A wavelet-based continuous classification scheme for multifunction myoelectric control. *IEEE Trans. Biomed. Eng.* 48, 302–311. doi: 10.1109/10.914793

Englehart, K., and Hudgins, B. (2003). A robust, real-time control scheme for multifunction myoelectric control. *IEEE Trans. Biomed. Eng* 50, 848–854. doi: 10.1109/TBME.2003.813539

Evans, R. L., and Arvind, D. K. (2014). "Detection of gait phases using orient specks for mobile clinical gait analysis," in *2014 11th International Conference on Wearable and Implantable Body Sensor Networks* (Zuirch, Switzerland: IEEE), 149–154.

Fan, Y., and Yin, Y. (2009). "Mechanism design and motion control of a parallel ankle joint for rehabilitation robotic exoskeleton," in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (Guilin, China: IEEE), 2527–2532.

Ferris, D. P., Sawicki, G. S., and Daley, M. A. (2007). A physiologist's perspective on robotic exoskeletons for human locomotion. *Int. J. Human. Robot.* 04, 507–528. doi: 10.1142/S0219843607001138

Godiyal, A. K., Mondal, M., Joshi, S. D., and Joshi, D. (2018a). Force myography based novel strategy for locomotion classification. *IEEE Trans. Human-Mach. Syst.* 48, 648–657. doi: 10.1109/THMS.2018.2860598

Godiyal, A. K., Verma, H. K., Khanna, N., and Joshi, D. (2018b). A force myography-based system for gait event detection in overground and ramp walking. *IEEE Trans. Instrum. Meas.* 67, 2314–2323. doi: 10.1109/TIM.2018.2816799

He, H.uang, Fan, Z.hang, Hargrove, L. J., Zhi, D.ou, Rogers, D. R., and Englehart, K. B. (2011). Continuous locomotion-mode identification for prosthetic legs based on neuromuscular–mechanical fusion. *IEEE Trans. Biomed. Eng.* 58, 2867–2875. doi: 10.1109/TBME.2011.2161671

Hioki, M., and Kawasaki, H. (2012). Estimation of finger joint angles from sEMG using a neural network including time delay factor and recurrent structure. *ISRN Rehabi.* 2012, 1–13. doi: 10.5402/2012/604314

Hudgins, B., Parker, P., and Scott, R. N. (1993). A new strategy for multifunction myoelectric control. *IEEE Trans. Biomed. Eng.* 40, 82–94. doi: 10.1109/10.204774

Joshi, C. D., Lahiri, U., and Thakor, N. V. (2013). "Classification of gait phases from lower limb EMG: application to exoskeleton orthosis," in *2013 IEEE Point-of-Care Healthcare Technologies (PHT)* (Bangalore, India: IEEE), 228–231. doi: 10.1109/PHT.2013.6461326

Kazerooni, H. (2008). "Exoskeletons for human performance augmentation," in *Springer Handbook of Robotics*, eds B. Siciliano, and O. Khatib (Berlin, Heidelberg: Springer), 773–793. doi: 10.1007/978-3-540-30301-5_34

Kim, J., Lee, G., Heimgartner, R., Arumukhom Revi, D., Karavas, N., Nathanson, D., et al. (2019). Reducing the metabolic rate of walking and running with a versatile, portable exosuit. *Science* 365, 668–672. doi: 10.1126/science.aav7536

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, Z., Liu, H., Yin, Z., and Chen, K. (2019). Muscle synergy alteration of human during walking with lower limb exoskeleton. *Front. Neurosci.* 12:1050. doi: 10.3389/fnins.2018.01050

Lloyd, D. G., and Besier, T. F. (2003). An EMG-driven musculoskeletal model to estimate muscle forces and knee joint moments in vivo. *J. Biomech.* 36, 765–776. doi: 10.1016/S0021-9290(03)00010-1

Moreno, J. C., Barroso, F., Farina, D., Gizzi, L., Santos, C., Molinari, M., et al. (2013). Effects of robotic guidance on the coordination of locomotion. *J. NeuroEng. Rehabil.* 10:79. doi: 10.1186/1743-0003-10-79

Naik, G. R., Selvan, S. E., Arjunan, S. P., Acharyya, A., Kumar, D. K., Ramanujam, A., et al. (2018). An ICA-EBM-based sEMG classifier for recognizing lower limb movements in individuals with and without knee pathology. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 675–686. doi: 10.1109/TNSRE.2018.2796070

Ngeo, J. G., Tamei, T., and Shibata, T. (2014). Continuous and simultaneous estimation of finger kinematics using inputs from an EMG-to-muscle activation model. *J. Neuroeng. Rehabil.* 11:122. doi: 10.1186/1743-0003-11-122

Novak, D., and Riener, R. (2015). A survey of sensor fusion methods in wearable robotics. *Rob. Auton. Syst.* 73, 155–170. doi: 10.1016/j.robot.2014.08.012

Resnik, L., Huang, H., Winslow, A., Crouch, D. L., Zhang, F., and Wolk, N. (2018). Evaluation of EMG pattern recognition for upper limb prosthesis control: a case study in comparison with direct myoelectric control. *J. NeuroEng. Rehabil.* 15:23. doi: 10.1186/s12984-018-0361-3

Ryu, J. H., and Kim, D. H. (2014). "Multiple gait phase recognition using boosted classifiers based on sEMG signal and classification matrix," in *International Conference on Ubiquitous Information Management and Communication,* 1–4.

Sankai, Y. (2010). "HAL: Hybrid Assistive Limb Based on Cybernics," in *Robotics Research*, eds M. Kaneko and Y. Nakamura (Berlin, Heidelberg: Springer), 25–34. doi: 10.1007/978-3-642-14743-2_3

Scheme, E., and Englehart, K. (2011). Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use. *JRRD* 48:643. doi: 10.1682/JRRD.2010.09.0177

Seel, T., Raisch, J., and Schauer, T. (2014). IMU-based joint angle measurement for gait analysis. *Sensors* 14, 6891–6909. doi: 10.3390/s140406891

Steele, K. M., Jackson, R. W., Shuman, B. R., and Collins, S. H. (2017). Muscle recruitment and coordination with an ankle exoskeleton. *J. Biomech.* 59, 50–58. doi: 10.1016/j.jbiomech.2017.05.010

Sylos-Labini, F., La Scaleia, V., d'Avella, A., Pisotta, I., Tamburella, F., Scivoletto, G., et al. (2014). EMG patterns during assisted walking in the exoskeleton. *Front. Hum. Neurosci.* 8, 1–12. doi: 10.3389/fnhum.2014.00423

Taborri, J., Palermo, E., Rossi, S., and Cappa, P. (2016). Gait partitioning methods: a systematic review. *Sensors* 16:66. doi: 10.3390/s16010066

Tanghe, K., De Groote, F., Lefeber, D., De Schutter, J., and Aertbelien, E. (2020). Gait trajectory and event prediction from state estimation for exoskeletons during gait. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 211–220. doi: 10.1109/TNSRE.2019.2950309

Viteckova, S., Kutilek, P., and Jirina, M. (2013). Wearable lower limb robotics: A review. *Biocybern. Biomedi. Eng.* 33, 96–105. doi: 10.1016/j.bbe.2013.03.005

Vu, H., Gomez, F., Cherelle, P., Lefeber, D., Nowé, A., and Vanderborght, B. (2018). ED-FNN: a new deep learning algorithm to detect percentage of the gait cycle for powered prostheses. *Sensors* 18:2389. doi: 10.3390/s18072389

Williams, T. W. (1990). Practical methods for controlling powered upper-extremity prostheses. *Assis. Technol.* 2, 3–18. doi: 10.1080/10400435.1990.10132142

Xia, P., Hu, J., and Peng, Y. (2018). EMG-based estimation of limb movement using deep learning with recurrent convolutional neural networks: EMG-based estimation of limb movement. *Artif. Organs* 42, E67–E77. doi: 10.1111/aor.13004

Yan, T., Cempini, M., Oddo, C. M., and Vitiello, N. (2015). Review of assistive strategies in powered lower-limb orthoses and exoskeletons. *Rob. Auton. Syst.* 64, 120–136. doi: 10.1016/j.robot.2014.09.032

Yang, X., Lihua, G., Yang, Z., and Gu, W. (2008). "Lower extreme carrying exoskeleton robot adative control using wavelet neural networks," in *IEEE,* 399–403. doi: 10.1109/ICNC.2008.754

Yi, C., Jiang, F., Zhang, S., Guo, H., Yang, C., Ding, Z., et al. (2021). Continuous prediction of lower-limb kinematics from multi-modal biomedical signals. *IEEE Trans. Circuits Syst. Video Technol.* 1–1. doi: 10.1109/TCSVT.2021.3071461

Yi, C., Ma, J., Guo, H., Han, J., Gao, H., Jiang, F., et al. (2018). Estimating three-dimensional body orientation based on an improved complementary filter for human motion tracking. *Sensors* 18:3765. doi: 10.3390/s18113765

Zhang, K., Wang, J., de Silva, C. W., and Fu, C. (2020). Unsupervised Cross-Subject Adaptation for Predicting Human Locomotion Intent. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 646–657. doi: 10.1109/TNSRE.2020.2966749

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 26 August 2021
doi: 10.3389/fnbot.2021.684037

Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech

Yan Zhao[1], *Zhenlin Liang*[1], *Jing Du*[1], *Li Zhang*[2,3], *Chengyu Liu*[4] *and Li Zhao*[1]*

[1] Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, China, [2] Computational Intelligence Group, Northumbria University, Newcastle upon Tyne, United Kingdom, [3] National Subsea Centre, Robert Gordon University, Aberdeen, United Kingdom, [4] School of Instrument Science and Engineering, Southeast University, Nanjing, China

Depression is a mental disorder that threatens the health and normal life of people. Hence, it is essential to provide an effective way to detect depression. However, research on depression detection mainly focuses on utilizing different parallel features from audio, video, and text for performance enhancement regardless of making full usage of the inherent information from speech. To focus on more emotionally salient regions of depression speech, in this research, we propose a multi-head time-dimension attention-based long short-term memory (LSTM) model. We first extract frame-level features to store the original temporal relationship of a speech sequence and then analyze their difference between speeches of depression and those of health status. Then, we study the performance of various features and use a modified feature set as the input of the LSTM layer. Instead of using the output of the traditional LSTM, multi-head time-dimension attention is employed to obtain more key time information related to depression detection by projecting the output into different subspaces. The experimental results show the proposed model leads to improvements of 2.3 and 10.3% over the LSTM model on the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) and the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) corpus, respectively.

Keywords: depression, LSTM, multi-head attention, frame-level feature, deep learning

OPEN ACCESS

Edited by:
Zhen Cui,
Nanjing University of Science and Technology, China

Reviewed by:
Smitha Kavallur Pisharath Gopi,
Nanyang Technological University, Singapore
Yong Li,
Nanjing University of Science and Technology, China
Tong Zhang,
Nanjing University of Science and Technology, China

***Correspondence:**
Li Zhao
zhaoli@seu.edu.cn

Received: *22 March 2021*
Accepted: *19 July 2021*
Published: *26 August 2021*

Citation:
Zhao Y, Liang Z, Du J, Zhang L, Liu C and Zhao L (2021) Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech. Front. Neurorobot. 15:684037. doi: 10.3389/fnbot.2021.684037

1. INTRODUCTION

Depression is a prevalent mental disorder, affecting millions of human beings all over the world (Organization, 2017). Depression not only makes patients bear psychological pain, pessimism and, self-accusation but also leads to a high possibility of disability and death (Hawton et al., 2013). It can bring a severe burden on individuals and families. Moreover, the particularity of mental disorders makes them difficult to diagnose. Most people with depression do not seek medical advice or even ignore it. Its diagnosis mainly relies on the self-report of patient or explicit severe mental disorder symptoms (Hamilton, 1960; Zung, 1965). There are also other evaluations, such as the 9–item Patient Health Questionnaire (PHQ–9) (Kroenke and Spitzer, 2002), the PHQ–8 (Kroenke et al., 2009), and so on. Influenced by subjective factors, such methods have some limitations. Therefore, providing an effective and objective method, as an auxiliary standard, for detecting depression, is of vital significance.

In recent years, myriad models have been proposed for automatic depression detection. Senoussaoui et al. (2014) showed that an i-vector-based representation of short-term acoustic

features, which contains 20 static Mel Frequency Cepstral Coefficients (MFCC) and 40 dynamic MFCC coefficients, is effective for depression classification based on different regression models. Yang et al. (2017) proposed a Deep Convolutional Neural Network (DCNN) with the text, video, and audio descriptors for detecting depression. Rodrigues Makiuchi et al. (2019) proposed a multimodal fusion of speech and linguistic representations for depression detection. By parallel employing the textual, audio, and visual models, the acquired features compose the input features of the full connection layer. Jan et al. (2017) proposed a Convolutional Neural Network (CNN) architecture for automatic depression prediction. Various frame-level features were extracted to obtain distinctive expression information. Yin et al. (2019) proposed a Hierarchical Bidirectional LSTM with text, video, and audio features for depression prediction. Li et al. (2019a) employed CNN for mild depression recognition based on electroencephalography. We observe that most of the proposed models (Senoussaoui et al., 2014; Jan et al., 2017; Yang et al., 2017; Rodrigues Makiuchi et al., 2019; Yin et al., 2019) rely on multimodal calculation, instead of focusing on the internal relation of the speech signal. We believe that making full use of the emotional information at all times is the key to provide an effective model for depression classification.

Therefore, to emphasize the key information of speech signals, an improved attention-based LSTM model is proposed for automatic depression detection in this research. First, we apply frame-level features for LSTM. The frame-level features keep the inherent emotional information of the speech sequences. Moreover, its variable length is suitable for LSTM. Second, we apply multi-head time-dimension attention for LSTM output to utilize the critical inherent information. Besides, the multi-head attention helps linearly project the LSTM output into different subspaces for various context vectors with reduced dimensions. To indicate the model efficiency, we evaluate the proposed model on the DAIC-WOZ and MODMA corpora.

The rest of the study is organized as follows. Section 2 describes related studies. Section 3 Analysis introduces the frame-level features and the selection. The proposed attention-based LSTM model is introduced in section 4. The databases and experiment results are provided in section 5. Section 6 discusses the experiment results. Section 7 concludes this study.

2. RELATED WORK

2.1. Deep Learning Models

For depression detecting, the machine learning algorithms were initially utilized, such as support vector machine (SVM) (Long et al., 2017; Jiang et al., 2018) and Gaussian mixture model (GMM) (Jiang et al., 2018). In recent years, deep neural networks have been widely used for detecting depression (Jan et al., 2017; Yang et al., 2017; Li et al., 2019a; Rodrigues Makiuchi et al., 2019; Yin et al., 2019). Previous studies such as Yang et al. (2017) and Jan et al. (2017) employed CNN as the classification model with multiple features for depression prediction. Making full use of the multimodality features is the key success of their models. Yin et al. (2019) used a Hierarchical Bidirectional LSTM network for the processed sequence information to predict depression. Besides utilizing multimodality features, their work focused on extracting time sequence information to inform prediction. Various methods are developed for the classification of speech emotions (Tiwari et al., 2020; Abbaschian et al., 2021). In addition, studies by (Li et al., 2019b; Xie et al., 2019; Zhao et al., 2019) has proved that the LSTM network is effective for processing sequential signals. Since the existing studies lack exploring the inherent relationships of the speech signals, we proposed a multi-head time-dimension attention LSTM model for depression classification. The proposed method is utilized for emphasizing the information of emotional salient regions to boost the classification performance for depression detection.

2.2. Attention Mechanism

Recently, the attention mechanism has achieved great success in computer vision. Xiao et al. (2015) applied visual attention to deep neural network for fine-grained classification tasks. Zhao et al. (2017) proposed a diversified visual attention network for object classification. The core idea is that the attention of a person depicts different priorities for various parts of an image. Inspired by such a strategy, the attention mechanism is introduced into speech emotion recognition. Mirsamadi et al. (2017) proposed local attention using recurrent neural networks for speech emotion recognition. Xie et al. (2019) used both time and feature dimension attention mechanism to achieve better performance for speech emotion recognition. Li et al. (2019b) explored the effectiveness of the self-attention mechanisms and multitask learning for speech emotion recognition. Specifically, previous studies by Mirsamadi et al. (2017) and Xie et al. (2019) have mainly focused on calculating different attention weightings for different parts of speech waveforms.

With the widely use of attention mechanism, a multi-head attention scheme has been proposed Vaswani et al. (2017) and introduced to many areas. Jiang et al. (2019) used Bidirectional

TABLE 1 | Frame-level speech features.

Acoustic features	Description
F0	Pitch frequency
Jitter	The average absolute difference between the consecutive periods
Shimmer	The average absolute difference between the interpolated peak amplitudes of consecutive periods
Loudness	The loudness and delta regression of loudness
MFCC	MFCC and delta regression of MFCC
Pcm_Mag	Mel spectral
Lpc	Linear predictive coding coefficients
LspFreq	Line spectral pair frequency
voiceProb	The voicing probability
harmonicERMS	Harmonic component root mean square energy
noiseERMS	Noise component root mean square energy
HNR	Log harmonics-to-noise ratio

Encoder Representations from Transformers (BERT) as the encoder for unsupervised pre training. Lian et al. (2019) proposed a multi-head attention framework, fusing the context, the emotional information of speech and speakers, to reach better performance for speech emotion classification. The earlier literatures (Mirsamadi et al., 2017; Lian et al., 2019; Li et al., 2019b; Xie et al., 2019; Abbaschian et al., 2021) indicate that the attention mechanism is effective for mining the inherent emotional information from speech. Hence, it is suitable for the study to apply such an attention mechanism for depression speech detection.

3. ACOUSTIC FEATURES ANALYSIS

The depression prediction with respect to speech comprises speech processing and classification methods based on the extracted features. The performance rate of a classifier largely relies on the type of extracted features. Many hand–crafted features have been discovered and used for improving prediction performances. These include prosodic features (Yang et al., 2017), spectral features (Senoussaoui et al., 2014; Yang et al., 2017; Rodrigues Makiuchi et al., 2019; Yin et al., 2019), and energy related features (Yang et al., 2017), e.g., Previous studies indicate that speech emotions have an inherent relationship with depression detection. In this study, we evaluate the widely used ComParE openSMILE features (Schuller et al., 2016; Jassim et al., 2017) and adopt some speech features as acoustic descriptors for depression detection. **Table 1** describes the frame-level speech features.

To evaluate and visualize the impact of features on detection, samples from DAIC-WOZ and MODMA corpora are taken for comparison. For each feature, we calculate the mean value of speech segments and sort them in ascending order. Outliers cause an excessive gradient. To identify the effectiveness of features for prediction, we take speech samples from DAIC-WOZ and MODMA corpora and calculate the mean value of the features over timeframes. **Figures 1**, **2** exhibit the mean values of four features. The x-axis represents the sample numbers and the y-axis represents the amplitude. They show that the HNR feature has the largest distinction among the four features. For voiceProb, it has many overlaps for samples on DAIC-WOZ corpus, which

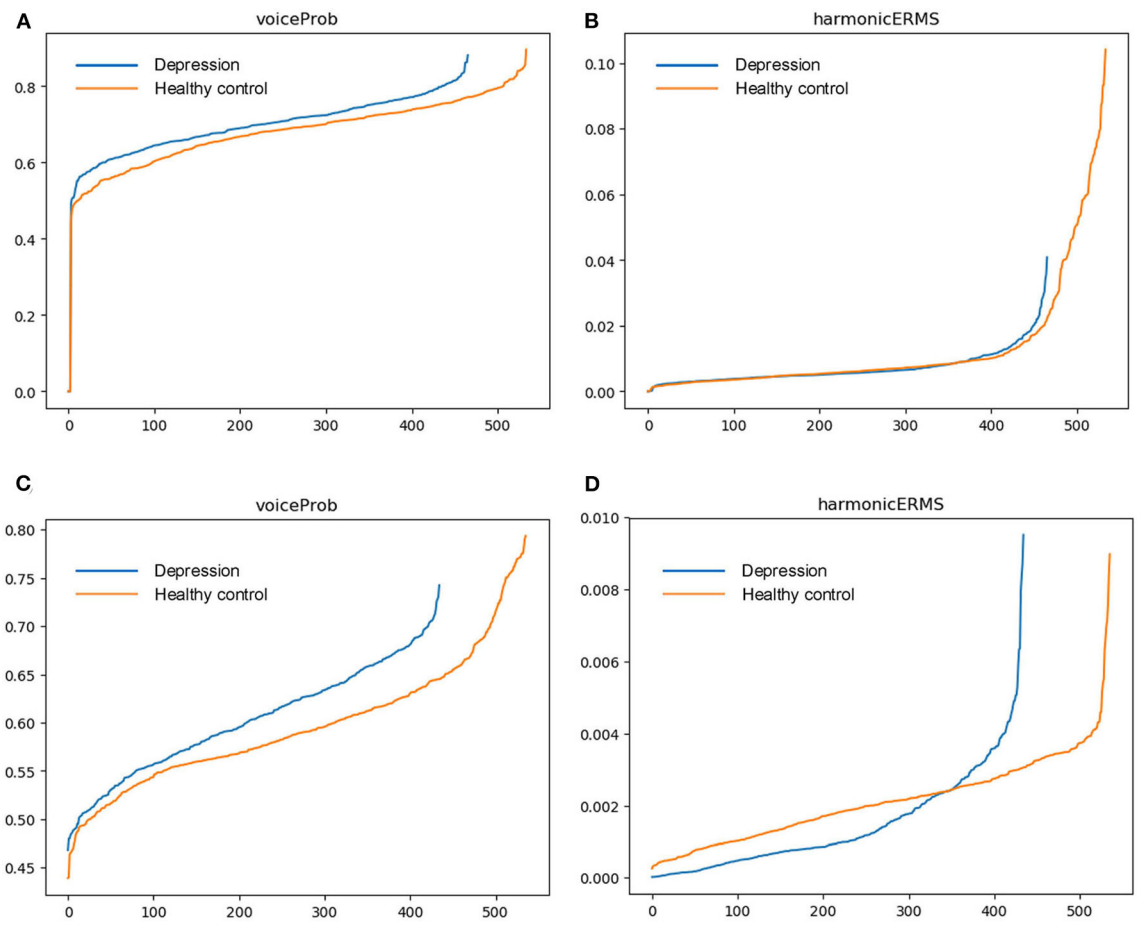


FIGURE 1 | Feature (voiceProb and harmonticERMS) comparison. **(A,B)** are features mean values on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) corpus, while **(C,D)** are features mean values on the MODMA corpus.

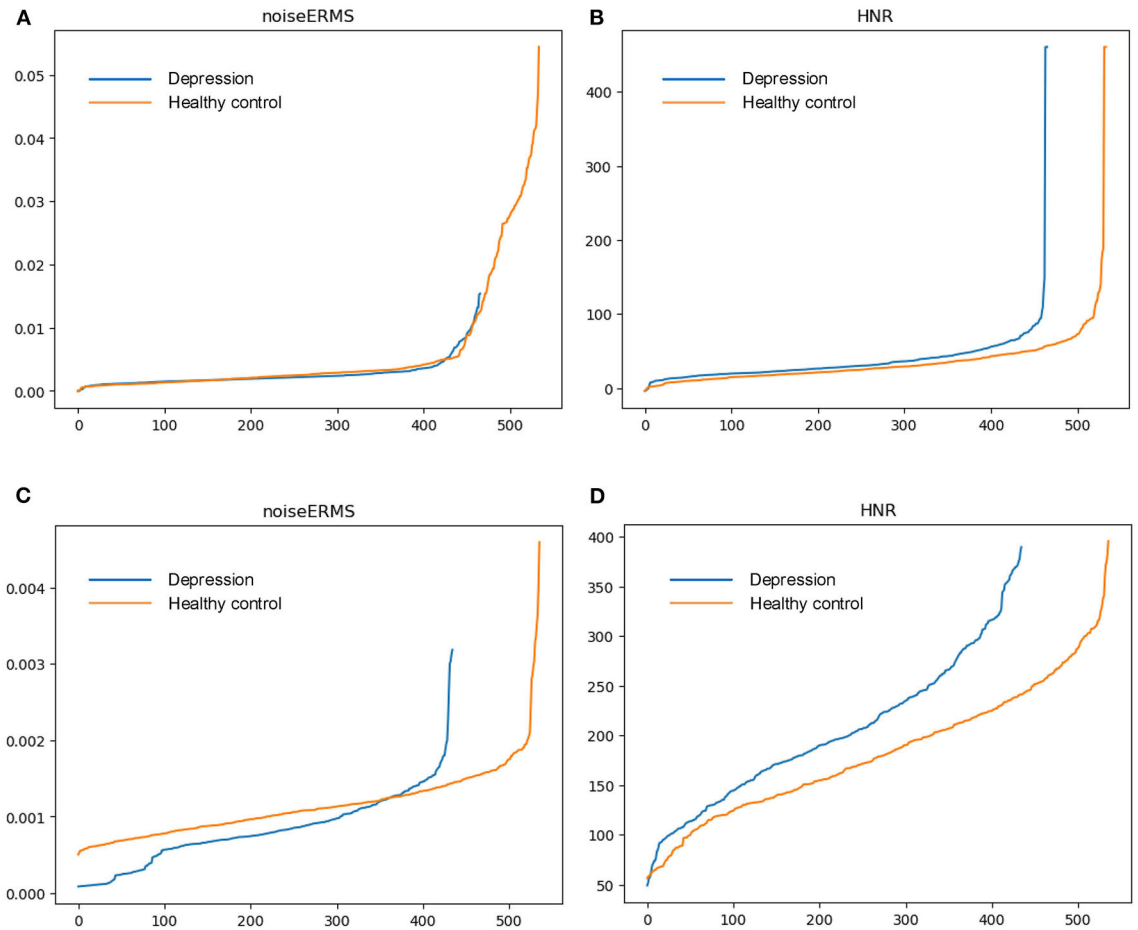


FIGURE 2 | Feature (noiseERMS and HNR) comparison. **(A,B)** are feature mean values on the DAIC-WOZ corpus, while **(C,D)** are feature mean values on the MODMA corpus.

means that it may not be effective for depression as a single feature. The same situation is observed on harmonticERMS and noiseERMS on the DAIC-WOZ database.

Furthermore, we conduct cluster analysis on DAIC-WOZ and MODMA corpora respectively. The mean values of the features over timeframes are calculated as before. The distributions of samples under different feature combinations are shown in **Figure 3**. The cluster results reveal the differences between the depression and normal samples. In **Figures 3A,C**, most of the depression samples tend to be lower on harmonicERMS and higher on MFCC, while the distributions of the two types of samples are roughly the same in terms of pcm_loudness_sma_de, which is consistent with the previous results. The second combination is voiceProb, noiseERMS, and the delta regression of MFCC. According to the previous analysis, there is significant overlap on voiceProb and noiseERMS on the DAIC-WOZ corpus. However, it can be seen from **Figures 3B,D** that there are also two distinct cluster centers despite more overlapping parts compared to **Figures 3A,C** both on DAIC-WOZ and MODMA corpora. This phenomenon indicates that a combination of two or more features can improve the ability to distinguish depression. It also demonstrates the effectiveness of the frame-level features in the identification of depression. Finding an effective model to expand the gap between depression and normal samples is right way to go.

4. MULTI-HEAD ATTENTION-BASED LSTM

The attention mechanism has been introduced to many areas successfully (Xiao et al., 2015; Mirsamadi et al., 2017; Vaswani et al., 2017; Zhao et al., 2017; Jiang et al., 2019; Lian et al., 2019; Xie et al., 2019). The main idea of the attention mechanism is to pay more attention to a certain weight distinction. In the previous study, Xie et al. (2019) studied the effectiveness of frame-level speech features, which include temporal information as well as feature-level information. The final representations multiplied by the attention layer helps model to improve the performance. In this study, for mining the multiple representations with more emotional information, we introduce the multi-head attention mechanism to depression detection and further develop the attention-based LSTM model.

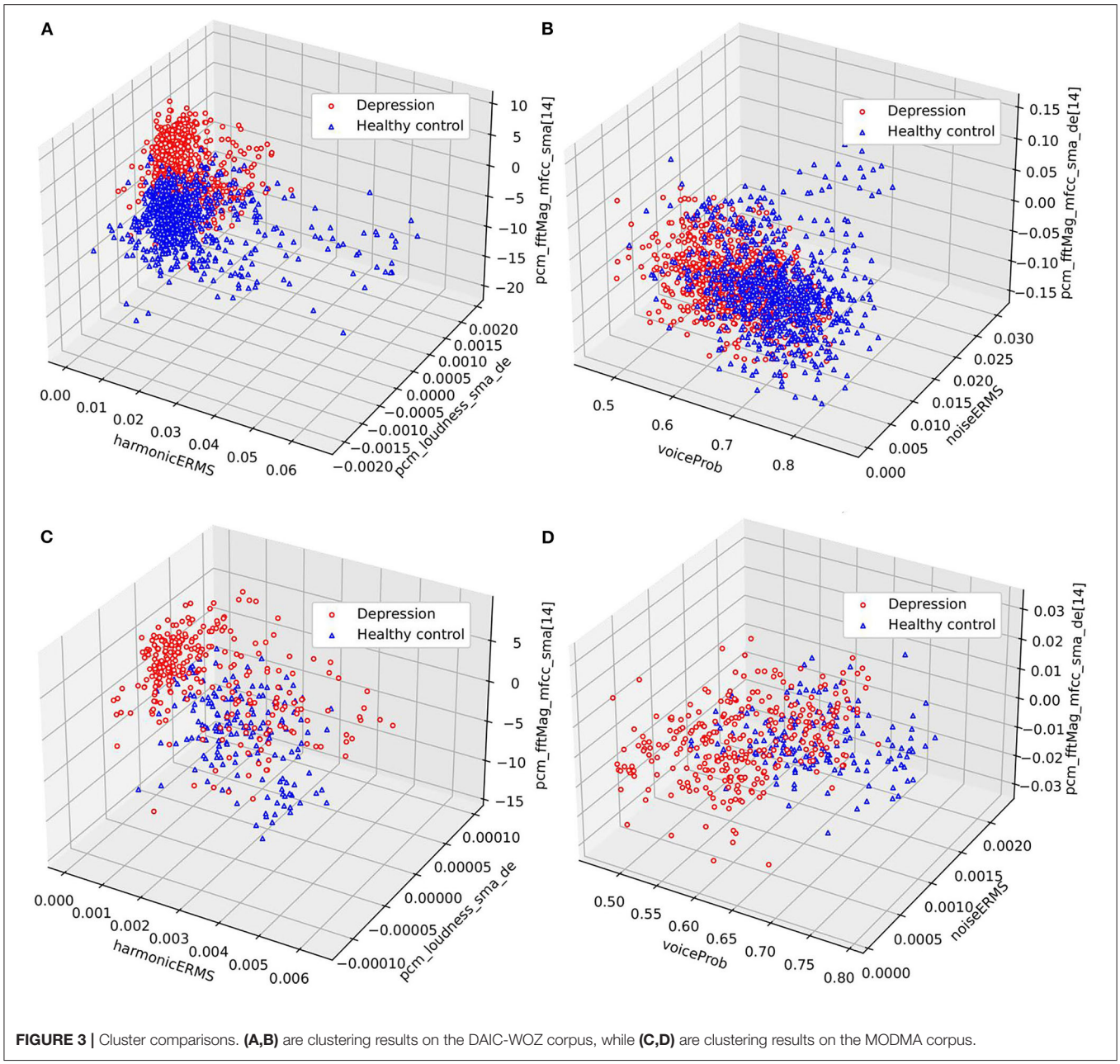


FIGURE 3 | Cluster comparisons. **(A,B)** are clustering results on the DAIC-WOZ corpus, while **(C,D)** are clustering results on the MODMA corpus.

4.1. LSTM Model

Hochreiter and Schmidhuber (1997) first proposed LSTM. Gers et al. (2000) added the forgetting gate for LSTM and proved its effectiveness. In an LSTM cell, the forgetting gate is used for discarding the useless information of the previous moment and updating the cell state. The previously hidden layer output and the current moment input are used in the updating algorithm. Multiple structures have been proposed for improving the LSTM performance, e.g., the forgetting gate (Gers et al., 2000) and peepholes (Gers and Schmidhuber, 2000). In the previous work, Xie et al. (2019) proposed an attention gate for LSTM to reduce the number of training calculations. The experiments indicate that the attention gate can help improve the effectiveness of LSTM model training. Hence, in the study, we use the modified LSTM (Xie et al., 2019) as the baseline.

4.2. Multi-Head Attention

Vaswani et al. (2017) first proposed the multi-head attention scheme. By taking an attention layer as a function, which maps a query and a set of key-value pairs to the output, their study found that it is beneficial to employ multi-head attention for the queries, values, and keys. By linearly projecting the context vectors into different subspaces, the multi-head attention layer computes the hidden information, which shows better performance than that of single-head attention. Inspired by Vaswani et al. (2017), we

calculate the output by weighted values, which are computed by queries and the corresponding keys.

Xie et al. (2019) has presented the time-dimension calculation for attention weighting:

$$s_t = softmax\left(o_{last} \times (o_{all} \times W_t)^H\right), o_{last} \in R^{B,1,Z} \quad (1)$$

$$o_t = s_t \times o_{all}, o_{all} \in R^{B,T,Z}, s_t \in R^{B,1,T} \quad (2)$$

where s_t donates the attention score of the time dimension, o_{last} represents the last time output and o_{all} is the all-time output. B represents the batch size, and T represents the number of time steps, while Z represents the feature dimension. The parameter 1 represents the last time step. H represents the transpose operator, and W_t represents the parameter matrix, while o_t donates the output of the time-dimension attention layer.

Formulas 1 and 2 are the single-head attention calculation. We only use two types of LSTM output for attention. The output of all time is essential because it contains all LSTM output information. The reason to choose the last time step output is that it includes the most redundant information among all time steps. For multi-head time-dimension attention computing, we also choose the two types of output to calculate the queries, keys, and values:

$$K_i = W_{i,k} \times o_{all} + b_{i,k}, K_i \in R^{B,T,\frac{Z}{n}}, W_{i,k} \in R^{Z,\frac{Z}{n}}, b_{i,k} \in R^{\frac{Z}{n}} \quad (3)$$

$$V_i = W_{i,v} \times o_{all} + b_{i,v}, V_i \in R^{B,T,\frac{Z}{n}}, W_{i,v} \in R^{Z,\frac{Z}{n}}, b_{i,v} \in R^{\frac{Z}{n}} \quad (4)$$

$$Q_i = W_{i,q} \times o_{last} + b_{i,q}, Q_i \in R^{B,1\frac{Z}{n}}, W_{i,q} \in R^{Z,\frac{Z}{n}}, b_{i,q} \in R^{\frac{Z}{n}} \quad (5)$$

where K,V,Q donate the value, key, and query. n is the number of attention heads and b means bias.

The multi-head attention scores and context vectors are calculated as follows:

$$s_i = softmax\left(Q_i \times K_i^H\right), s_i \in R^{B,1,T} \quad (6)$$

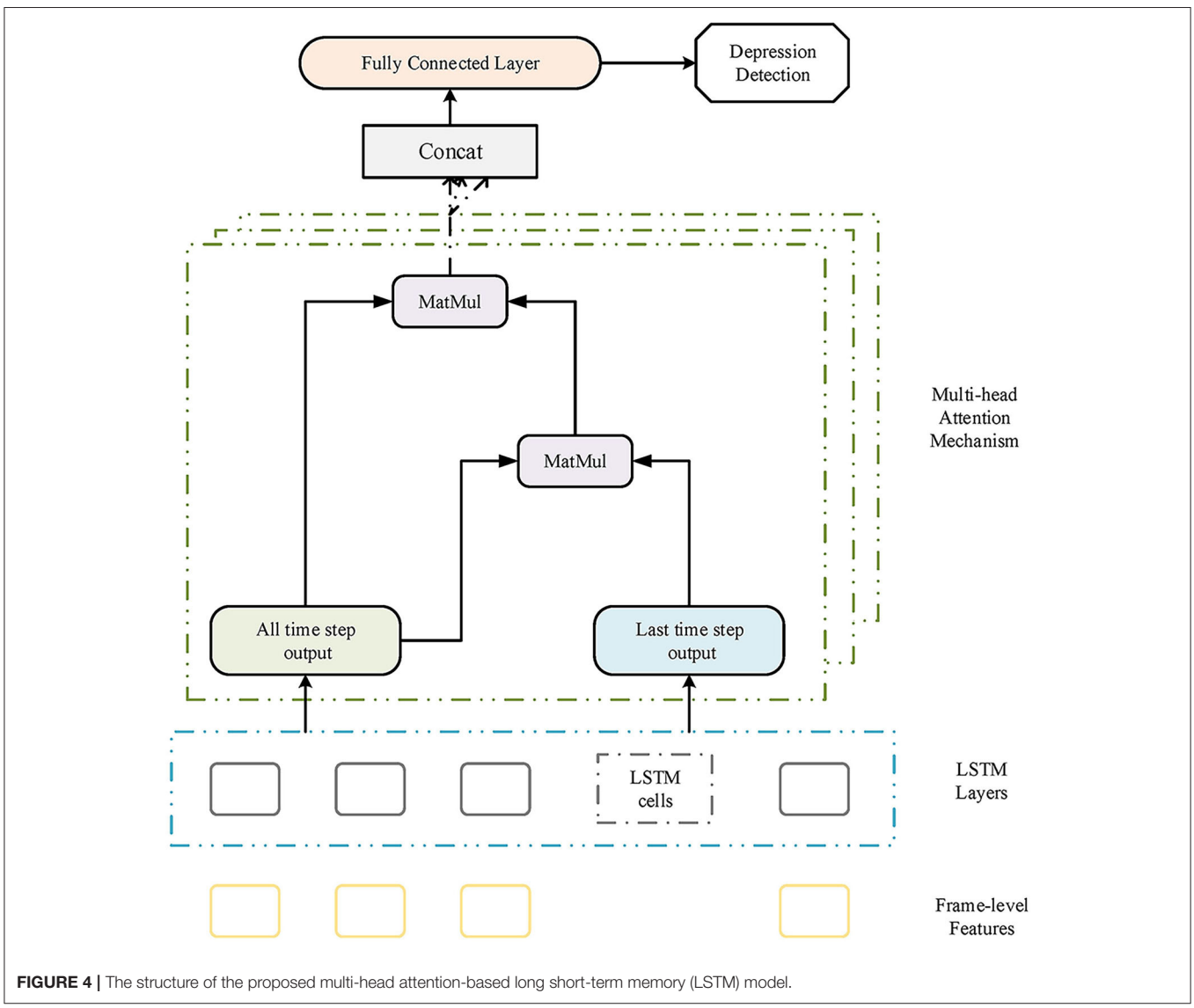


FIGURE 4 | The structure of the proposed multi-head attention-based long short-term memory (LSTM) model.

$$context_i = s_i \times V_i, context_i \in R^{B,1,\frac{Z}{n}} \quad (7)$$

$$CV = Concat\left([context_1, \ldots, context_n]\right), CV \in R^{B,1,Z} \quad (8)$$

where s_i represents the multi-head time-dimension attention score and $context_i$ represents the reduced-dimension context vectors from each subspace. The overall structure of multi-head time-dimension attention is described in **Figure 4**. Next, the context vector is put into the full connection layer. The output is then sent to the softmax layer for final prediction.

5. EXPERIMENT AND RESULTS

5.1. Datasets

In this research, we evaluate the proposed model on DAIC-WOZ (Gratch et al., 2014) and MODMA (Cai et al., 2020) corpora. The DAIC data corpus contains clinical interviews designed to support the diagnosis of psychological distress conditions. The sampling rate is 16,000 Hz. The numbers of depression and healthy control samples randomly selected are 42 and 47, respectively. Then, we divide them into segments, which makes feature extraction more convenient. We obtain 2,156 depression segments and 2,245 healthy control segments from the selected samples. To ensure the effectiveness of the fragments, abnormal segments, which are <3 s with litter information or larger than 20 s, are discarded in this research. Finally, we utilize 3,401 and 1,000 audio segments, which are randomly sorted by the software, as the train set and the test set, respectively.

The database contains 52 samples on the MODMA database, with 23 depression and 29 healthy control samples. We also divide them into sentences. Compared with samples in the DAIC-WOZ corpus, samples of MODMA contains much more information with an average duration of over 10 s. We also discard the abnormal segments, which are much larger than other segments. At last, we several 1,321 segments. We randomly split them into two different sets (train set and test set). The train set includes 971 segments while the test set contains 350 segments. Both of the corpora are grouped into two categories (depression and healthy control).

5.2. Multi-Head Time-Dimension LSTM

We utilize the attention mechanism to capture the key information from the depression speech. In the previous study, Xie et al. (2019) used single-head attention for emphasizing the reverent key information related to the task. In this study, we proposed multi-head time-dimension attention for depression detection. To prove its validity, we conduct experiments for comparison with LSTM models. We use three types of LSTM models and evaluate them on DAIC-WOZ and MODMA corpora. The models are: (1) LSTM. (2) LSTM+T, which is time-dimension attention LSTM (Xie et al., 2019). (3) LSTM+nT, which is the proposed multi-head time-dimension LSTM, and n represents the head number. The proposed models, including the LSTM and multi-head time-dimension-based LSTM, are composed of two LSTM layers. The number of hidden units for the first LSTM layer is 512 and that of the second LSTM layer is 256. The size of the fully connected layer is [128, 12]. The learning rate is set as 0.0001, and the batch size is 64. We extract the acoustic features mentioned above by openSMILE Eyben et al. (2010) and use them as the input of the proposed model. Instead of pretraining on other databases, we train the models directly on DAIC-WOZ and MODMA corpora. **Table 2** shows the experimental results.

As described in **Table 2**, the LSTM+T model has better results than those of the LSTM model, while LSTM+nT models

TABLE 2 | Unweighted average recalls (UARs) of different models on DAIC-WOZ and MODMA corpora.

Model	UAR	
	DAIC-WOZ(%)	MODMA(%)
LSTM	91.2	88.6
LSTM+T	92.1	96.6
LSTM+2T	92.9	**98.9**
LSTM+4T	**93.5**	98.3
LSTM+8T	92.5	98.0

Bold values represent the best results in the comparison.

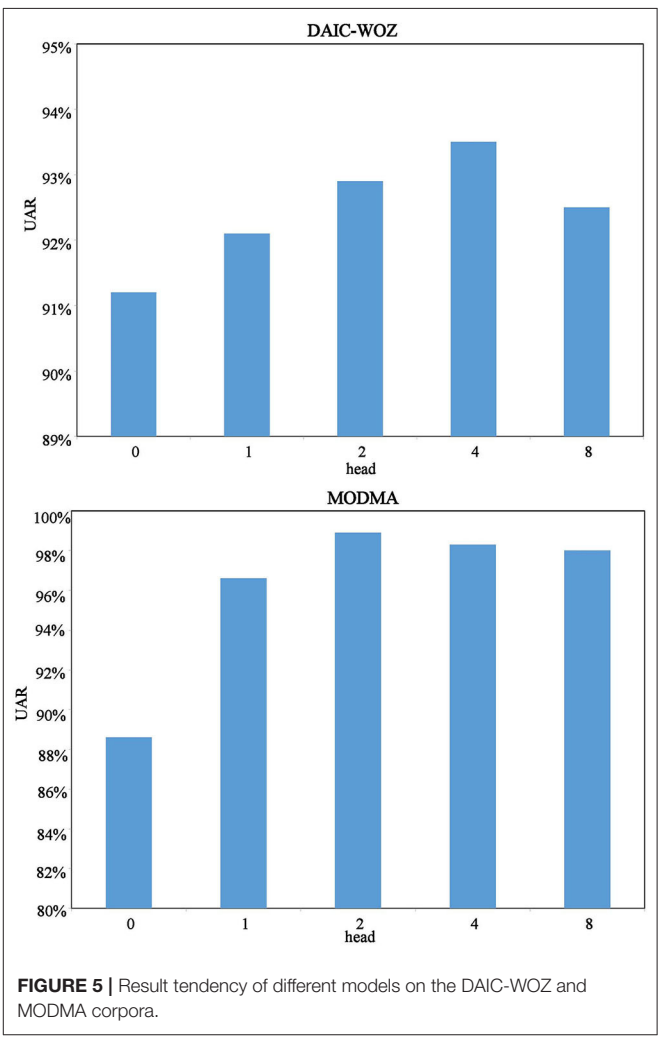


FIGURE 5 | Result tendency of different models on the DAIC-WOZ and MODMA corpora.

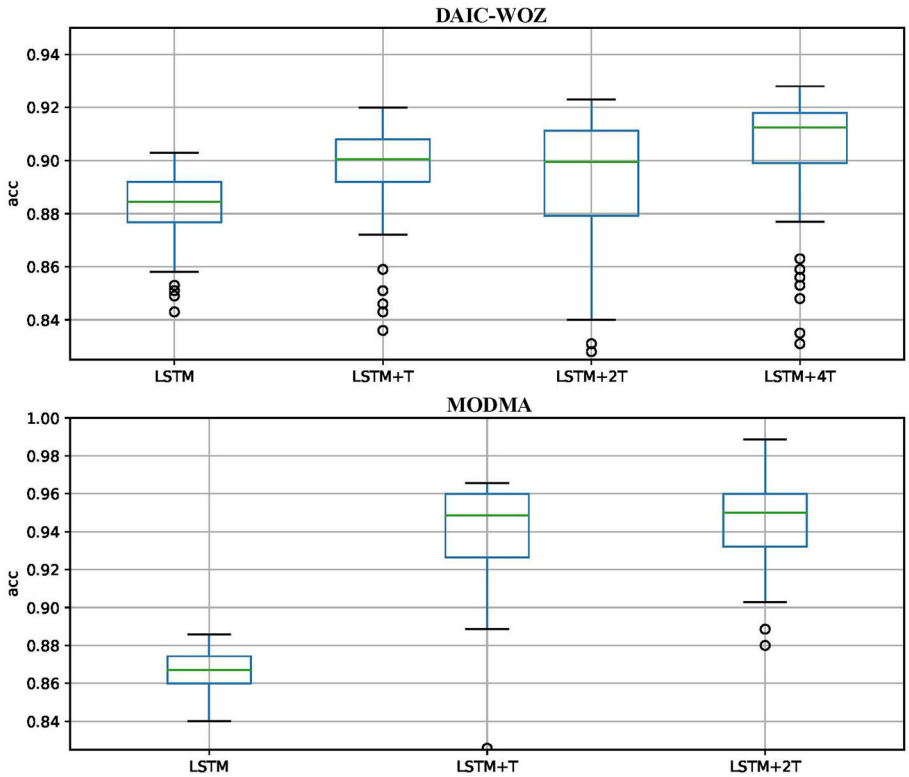


FIGURE 6 | Stability of models on test sets.

acquire the best performance on both DAIC-WOZ and MODMA corpora. We choose unweighted average recall (UAR) to evaluate the effectiveness of the two feature sets for different databases. UAR is defined as: UAR=$\frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{n_i}$, where c_i represents the correctly classified sample number of i category, n_i represents sample number of i category and N represents categories. The time-dimension attention shows its reliability for depression detection, by improving 0.9 and 8.0% on DAIC-WOZ and MODMA corpora, respectively. The LSTM+nT models achieve the best UARs (93.5% on DAIC-WOZ and 98.9% on MODMA) in the experiments. The model UAR is 93.5% on DAIC-WOZ with 4-head and it is up to 98.9% on MODMA with 2-head.

Figure 5 shows the tendency of the models. Zero-head means LSTM while 1-head denotes the LSTM+T model. For experiments on DAIC-WOZ, we observe that the multi-head time-dimension attention models tend increasing first and decreasing subsequently. We believe it is normal for multi-head attention calculation to illustrate such a performance behavior. The reason is that the increase of head number cannot always help the model obtain better performance. There must be a boundary for it. Since the tendency proves the boundary, we believe the multi-head time-dimension attention LSTM has achieved the best UAR with 4-head. For the MODMA corpus, we can see that attention is effective. All models with attention have a high UAR of over 95%. The phenomenon could be caused by distinguishing features on the MODMA corpus, which can be proved on the feature comparison of the frame-level speech feature section. The 2-head time dimension achieves the best result of 98.9%. If we put the single-head attention into consideration, it still tends increasing first and then decreasing. The experiment results prove the effectiveness of the proposed multi-head time-dimension attention.

Figure 6 shows the stability of models on the test set. The y-axis represents accuracy (UAR), and the x-axis represents the models. We exhibit the results from LSTM to the best model on the DAIC-WOZ and MODMA corpora. The blue rectangular box height indicates the stability of the model, and the lines inside the box are the stable UAR. On the test set, we could obtain thousands of results when the model converges. The stability in this study means most test results are inside the box range. The stable UAR means the median of results. The two lines outside the box mean the highest and lowest UARs. Circles represent outliers. As shown in the figures, the LSTM+nT model achieves higher stable UAR than those of LSTM and LSTM+T models on both test corpora. The overall performance indicates the LSTM+nT models are more reliable than other models.

For a better understanding of the depression patterns in speech signals, we draw the speech waveform as well as its corresponding attention score, which is shown in **Figure 7**. We experiment on one audio clip, using the 4-head attention LSTM model, and visualize one of the attention scores. What can be seen from **Figure 7** is that the multi-head attention mechanism

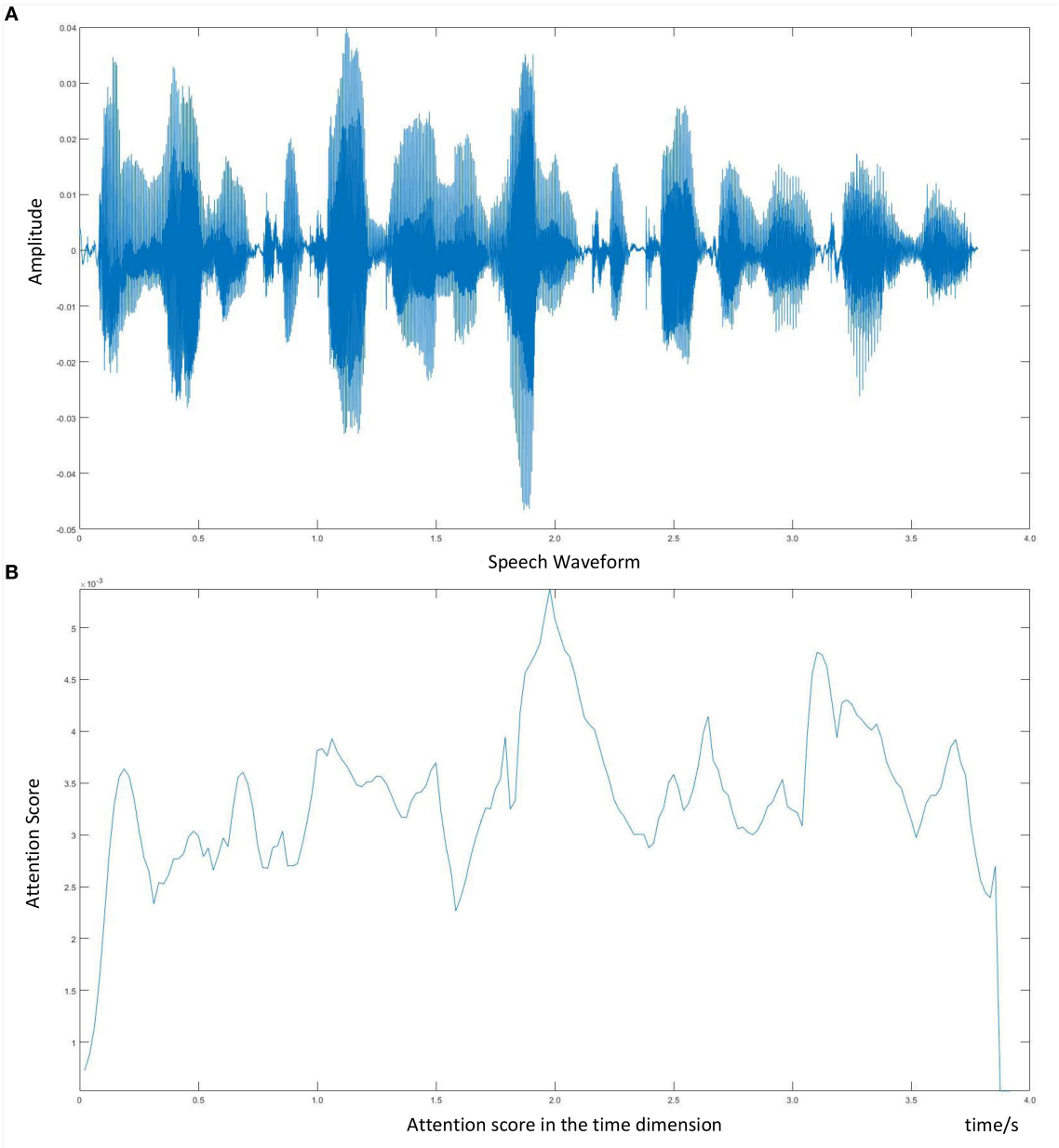


FIGURE 7 | Visualization of speech waveform and corresponding attention score. **(A)** Speech waveform. **(B)** Attention score in the time dimension.

endows diverse weights for the salient regions. For example, the attention score changes with the fluctuation of the audio clip and achieves a peak around 2 s. Moreover, with the amplification of the emotional part, we can pay more attention to the negative regions, which is beneficial for depression detection from speech.

6. DISCUSSION

In this study, we extract frame-level features to detect depression. In the previous study, Long et al. (2017) and Jiang et al. (2018) studied the speech features using different classifiers. The developments of Long et al. (2017) and Jiang et al. (2018) prove the effectiveness of MFCC, loudness, and F0 features. Therefore, we adopt those widely used features as parts of this study. To evaluate the effectiveness of features, we conduct a comparison between depression and normal samples to visualize the impact of features on detection. The results indicate that enhancing the emotional region of speech is a fundamental part of better depression classification.

Table 2 exhibits the results of LSTM and multi-head time-dimension. We could easily find that LSTM obtains the worst

TABLE 3 | Comparison between long short-term memory (LSTM) and the proposed model.

	DAIC-WOZ			MODMA		
Model	**Precision(%)**	**Recall(%)**	**F1 score**	**Precision**	**Recall**	**F1 score**
LSTM	89.7	91.6	0.907	94.6	78.7	0.859
LSTM+T	91.6	91.4	0.915	95.5	96.8	0.962
LSTM+2T	91.2	**93.8**	0.925	**99.3**	98.1	**0.987**
LSTM+4T	92.4	**93.8**	**0.931**	96.9	**99.4**	0.981
LSTM+8T	**93.0**	90.8	0.919	98.1	97.4	0.977
Zhao et al. (2019)	91.2	92.0	0.916	92.9	93.5	0.932
Li et al. (2019b)	82.2	89.1	0.855	93.5	90.1	0.918

Bold values represent the best results in the comparison.

results on both DAIC-WOZ and MODMA corpora. The best LSTM+nT model improves by 2.3 and 10.3% on DAIC-WOZ and MODMA, respectively. It indicates that the multi-head attention mechanism helps the model to emphasize the key time information of sequence. Besides that, we find that the best results of the multi-head time-dimension attention-based LSTM model achieve the 1.4 and 2.3% improvement than those of a single-head attention-based LSTM model on the DAIC-WOZ and MODMA corpora, respectively. This phenomenon proves that linear projections have a significant influence on the attention mechanism. Linearly projecting the LSTM output into different subspaces and then computing the reduced-dimension context vectors of various subspaces provides more key information than single-head attention.

Table 3 exhibits the results of LSTM and the proposed model. Besides, we also make comparisons with other models mentioned above (Li et al., 2019b; Zhao et al., 2019). We follow the model structure and keep all layer parameters the same to reimplement the models for depression detection. Audios are processed into spectrogram as input features. We use precision, recall, and F1 score as standards for comparison. TP represents the correctly classified number of samples for positive cases. *FP* represents the incorrectly classified number of samples that are misclassified as positive cases and *FN* represents the incorrectly classified number of samples that are misclassified as negative. The calculation of precision and recall is defined as: precision=$TP/(TP+FP)$, recall=$TP/(TP+FN)$, F1=2(precision$\times$recall)/(precision+recall). We use the F1 score as the harmonic mean of precision and recall. The proposed models exceed the LSTM model and the deeper models, 2-D CNN LSTM (Zhao et al., 2019) and CNN LSTM with self-attention mechanism (Li et al., 2019b), in all standards. For the DAIC-WOZ, database, the LSTM+4T model achieves the best F1 score of 0.931 while the LSTM and LSTM+T only achieve 0.907 and 0.915, respectively. For the MODMA database, the LSTM+2T model shows the best performance. It has improvements of 4.7 and 19.4% on precision and recall, respectively, in comparison with those of the LSTM model. The F1 score also increases from 0.859 to 0.987, which indicates the proposed model is effective for depression prediction. Based on the experimental results on the DAIC-WOZ and MODMA corpora, the proposed strategy shows a significant impact on depression detection.

7. CONCLUSION

In this research, an improved attention-based LSTM network is proposed for depression detection. We first study the speech features for depression detection on the DAIC-WOZ and MODMA corpora. By applying the multi-head time-dimension attention weighting, the proposed model emphasizes the key temporal information. We evaluate the proposed model on both DAIC-WOZ and MODMA corpora. It achieves great superiority over other models for depression classification.

In further directions, first, we may explore other effective speech features for depression detection. Moreover, experiments will be conducted in the future to indicate the efficiency of the modified LSTM model for other time-series predictions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The datasets (DAIC-WOZ and MODMA) for this study can be found here: https://dcapswoz.ict.usc.edu/; http://modma.lzu.edu.cn/data/index.

AUTHOR CONTRIBUTIONS

YZ designed the overall architecture and wrote the manuscript. ZL conducted the experiments to evaluate the performance of the proposed method. CL and LZ critically refined the manuscript. All authors have contributed to the summary of the presented results and the discussion points and contributed to the review and editing of the manuscript.

FUNDING

This research was founded in part by the Distinguished Young Scholars of Jiangsu Province (BK20190014), and the Natural Science Foundation of China (No.81871444, 61673108, 61571106, 61633013).

REFERENCES

Abbaschian, B. J., Sierra-Sosa, D., and Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21:1249. doi: 10.3390/s21041249

Cai, H., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., et al. (2020). Modma dataset: a multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv*:2002.09283.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia* (Firenze), 1459–1462.

Gers, F., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: continual prediction with lstm. *Neural Comput*. 12, 2451–2471. doi: 10.1162/089976600300015015

Gers, F. A., and Schmidhuber, J. (2000). "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, Vol. 3* (Como: IEEE), 189–194.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., et al. (2014). "The distress analysis interview corpus of human and computer interviews," in *LREC* (Reykjavik), 3123–3128.

Hamilton, M. (1960). A rating scale for depression. *J. Neurol Neurosurg. Psychiatry* 23:56. doi: 10.1136/jnnp.23.1.56

Hawton, K., i Comabella, C. C., Haw, C., and Saunders, K. (2013). Risk factors for suicide in individuals with depression: a systematic review. *J. Affect. Disord*. 147, 17–28. doi: 10.1016/j.jad.2013.01.004

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Jan, A., Meng, H., Gaus, Y. F. B. A., and Zhang, F. (2017). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Trans. Cogn. Dev. Syst*. 10, 668–680. doi: 10.1109/TCDS.2017.2721552

Jassim, W. A., Paramesran, R., and Harte, N. (2017). Speech emotion classification using combined neurogram and interspeech 2010 paralinguistic challenge features. *IET Signal Proc*. 11, 587–595. doi: 10.1049/iet-spr.2016.0336

Jiang, D., Lei, X., Li, W., Luo, N., Hu, Y., Zou, W., et al. (2019). Improving transformer-based speech recognition using unsupervised pre-training. *arXiv preprint arXiv*:1910.09932.

Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X., et al. (2018). Detecting depression using an ensemble logistic regression model based on multiple speech features. *Comput. Math. Methods Med*. 2018:6508319. doi: 10.1155/2018/6508319

Kroenke, K., and Spitzer, R. L. (2002). The phq-9: a new depression diagnostic and severity measure. *Psychiatr Ann*. 32, 509–515. doi: 10.3928/0048-5713-20020901-06

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *J. Affect. Disord*. 114, 163–173. doi: 10.1016/j.jad.2008.06.026

Li, X., La, R., Wang, Y., Niu, J., Zeng, S., Sun, S., et al. (2019a). Eeg-based mild depression recognition using convolutional neural network. *Med. Biol. Eng. Comput*. 57, 1341–1352. doi: 10.1007/s11517-019-01959-2

Li, Y., Zhao, T., and Kawahara, T. (2019b). "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech* (Graz), 2803–2807.

Lian, Z., Tao, J., Liu, B., and Huang, J. (2019). Conversational emotion analysis via attention mechanisms. *ArXiv* abs/1910.11263. doi: 10.21437/Interspeech.2019-1577

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., and Cai, H. (2017). "Detecting depression in speech: comparison and combination between different speech types," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Kansas City, MO: IEEE), 1052–1058.

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA: IEEE), 2227–2231.

Organization, W. H. (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. Technical report, World Health Organization.

Rodrigues Makiuchi, M., Warnita, T., Uto, K., and Shinoda, K. (2019). "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice), 55–63.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., et al. (2016). "The interspeech 2016 computational paralinguistics challenge: deception, sincerity native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016)* (San Francisco, CA), *Vols 1-5*, 2001–2005.

Senoussaoui, M., Sarria-Paja, M., Santos, J. F., and Falk, T. H. (2014). "Model fusion for multimodal depression classification and level detection." in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (Orlando, FL), 57–63.

Tiwari, U., Soni, M., Chakraborty, R., Panda, A., and Kopparapu, S. K. (2020). "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 7194–7198.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, CA: Curran Associates Inc.), 6000–6010.

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., and Zhang, Z. (2015). "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 842–850.

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., and Schuller, B. (2019). Speech emotion classification using attention-based lstm. *IEEE/ACM Trans. Audio Speech Lang. Proc.* 27, 1675–1685. doi: 10.1109/TASLP.2019.2925934

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., and Sahli, H. (2017). "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (Mountain View, CA), 53–59.

Yin, S., Liang, C., Ding, H., and Wang, S. (2019). "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice), 65–71.

Zhao, B., Wu, X., Feng, J., Peng, Q., and Yan, S. (2017). Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimedia* 19, 1245–1256. doi: 10.1109/TMM.2017.2648498

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d 2d cnn lstm networks. *Biomed. Signal Process Control* 47, 312–323. doi: 10.1016/j.bspc.2018.08.035

Zung, W. W. (1965). A self-rating depression scale. *Arch. Gen. Psychiatry* 12, 63–70. doi: 10.1001/archpsyc.1965.01720310065008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neurorobotics

METHODS
published: 25 October 2021
doi: 10.3389/fnbot.2021.763100

TFE: A Transformer Architecture for Occlusion Aware Facial Expression Recognition

Jixun Gao[1]* *and Yuanyuan Zhao*[2]

[1] Department of Computer Science, Henan University of Engineering, Zhengzhou, China, [2] Department of Computer Science, Zhengzhou University of Technology, Zhengzhou, China

Facial expression recognition (FER) in uncontrolled environment is challenging due to various un-constrained conditions. Although existing deep learning-based FER approaches have been quite promising in recognizing frontal faces, they still struggle to accurately identify the facial expressions on the faces that are partly occluded in unconstrained scenarios. To mitigate this issue, we propose a transformer-based FER method (TFE) that is capable of adaptatively focusing on the most important and unoccluded facial regions. TFE is based on the multi-head self-attention mechanism that can flexibly attend to a sequence of image patches to encode the critical cues for FER. Compared with traditional transformer, the novelty of TFE is two-fold: (i) To effectively select the discriminative facial regions, we integrate all the attention weights in various transformer layers into an attention map to guide the network to perceive the important facial regions. (ii) Given an input occluded facial image, we use a decoder to reconstruct the corresponding non-occluded face. Thus, TFE is capable of inferring the occluded regions to better recognize the facial expressions. We evaluate the proposed TFE on the two prevalent in-the-wild facial expression datasets (AffectNet and RAF-DB) and the their modifications with artificial occlusions. Experimental results show that TFE improves the recognition accuracy on both the non-occluded faces and occluded faces. Compared with other state-of-the-art FE methods, TFE obtains consistent improvements. Visualization results show TFE is capable of automatically focusing on the discriminative and non-occluded facial regions for robust FER.

Keywords: affective computing, facial expression recognition, occlusion, transformer, deep learning

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Tong Zhang,
Nanjing University of Science and Technology, China
Yuan Zong,
Southeast University, China

***Correspondence:**
Jixun Gao
gaojixun@haue.edu.cn

Received: *23 August 2021*
Accepted: *13 September 2021*
Published: *25 October 2021*

Citation:
Gao J and Zhao Y (2021) TFE: A Transformer Architecture for Occlusion Aware Facial Expression Recognition. Front. Neurorobot. 15:763100. doi: 10.3389/fnbot.2021.763100

1. INTRODUCTION

Facial expressions are the most natural way for humans to express emotions. Facial expression recognition (FER) has received significant interest from psychologists and computer scientists as it facilitates a number of practical applications, such as human-computer interaction, pain estimation, and affect analysis. Although current FER systems have obtained promising accuracy when recognizing facial images captured in controlled scenarios, these FER systems usually suffer from considerable performance degradation when recognizing expressions in the wild conditions. To fill the gap between the FER accuracy on the controlled faces and in-the-wild faces, researchers start to collect large-scale facial expression databases in uncontrolled environment (Li et al., 2017; Mollahosseini et al., 2017). Despite the usage of face images in the uncontrolled scenario, FER is still challenging due to the existence of facial occlusions. It is non-trivial to solve the occlusion problem

because facial occlusions are various and abundant. These facial occlusions may appear in many forms, such as breathing masks, hands, drinks, fruits, and other objects that might appear in front of the human faces in our daily life. The facial occlusions may block any other part of the face, and the variability of occlusions would inevitably induce the decreased FER performance.

Previous studies usually handled FER under occlusion with sub-region-based features (Kotsia et al., 2008; Li et al., 2018a,b; Wang et al., 2020b), e.g., Kotsia et al. (2008) presented a detailed analysis on occluded FER and conclude that FER will suffer from more decreased performance with occluded mouth than the occluded eyes. With the popularity of the data-driven convolutional neural network (CNN) techniques, a number of recent efforts on FER have been made on the collection of large-scale facial expression databases and exploit CNN to enhance the performance of FER. Li et al. (2018a) proposed to decompose facial regions in the convolutional feature maps with the manually defined facial landmarks and fused the local and global facial representations *via* attention mechanism. However, the recent CNN-based FER methods lack the ability to learn global interactions and relations between distant facial parts. These methods are not capable of flexibly attending to distinctive facial regions for precise FER under occlusions.

Inspired by the observation (Naseer et al., 2021) that transformers are robust to occlusions, perturbations, and domain shifts, we propose a **T**ransformer Architecture for **F**acial **E**xpression Recognition (TFE) under occlusions. Currently, vision transformers (Dosovitskiy et al., 2020; Li et al., 2021) have demonstrated impressive performance across numerous machine vision tasks. These models are based on multi-head self-attention mechanisms that can flexibly attend to a sequence of image patches to encode contextual cues. The self-attention in the transformers has been shown to effectively learn global interactions and relations between distant object parts. A number of following studies on downstream tasks such as object detection (Carion et al., 2020), segmentation (Jin et al., 2021), and video processing (Girdhar et al., 2019; Fang et al., 2020) have verified the feasibility of the transformers. Given the content-dependent long-range interaction modeling capabilities, transformers can flexibly adjust their receptive field to cope with occlusions in data and enhance the discriminability of the representations.

Intuitively, human perceives the facial expressions *via* several critical facial regions, e.g., eyes, eyebrows, and corners of the mouth. If some facial patches are occluded, human may judge the expression according to the other highly informative regions. To mimic the way that human recognizes the facial expression, we propose a region selection unit (RS-Unit) that is capable of focusing on the important facial regions. To be specific, RS-Unit selects the discriminative facial regions and removes the redundant or occluded facial parts. We then combine the global classification token with the selected part tokens as the facial expression representation. With the proposed RS-Unit, TFE is able to adaptively perceive the distinctive and unobstructed regions in facial images. To further enhance the discriminability of the representation, we exploit an auxiliary decoder to reconstruct the corresponding non-occluded face. Thus, TFE is capable of inferring the occluded facial regions *via* the unoccluded parts to better recognize the facial expressions. **Figure 1** illustrates the attention map of TFE on some facial images. It is clear that TFE is capable of focusing on the critical and unoccluded facial parts for robust FER. More visual examples and explanations can be seen in section 4.2.1.

The contributions of this study can be summarized as follows:

1. We propose a transformer architecture to recognize facial expressions (TFE) from partially occluded faces. TFE consists of a region selection unit (RS-Unit) that automatically perceives and selects the critical facial regions for robust FER. TFE is deployed to focus on the most important and unoccluded facial regions.
2. To further enhance the discriminability of the facial expression representation, TFE contains an auxiliary image decoder to reconstruct the corresponding non-occluded face. The image decoder is merely exploited during the training process and incorporates no extra computation burden at inference time.
3. Qualitative experimental results show the benefits and the advantages of the proposed TFE over other state-of-the-art approaches on two prevalent in-the-wild facial expression databases. Visualization results additionally show that TFE is superior in perceiving the informative facial regions.

2. RELATED WORK

We discuss the previous literatures that are related to our proposed TFE, i.e., FER with occlusions and the vision transformer.

2.1. Methods for FEE Under Occlusion

For FER tasks, occlusion is one of the inevitable challenges in real-world scenarios. We just classify previous FER methods into two classes: handcrafted features-based methods and deep learning-based approaches.

Early FER under occlusion methods typically encode handcrafted features from face samples, and then learn classifiers based on the encoded features (Rudovic et al., 2012; Zhang et al., 2014). Liu et al. (2013) proposed a novel FER method to mitigate the partial occlusion issue *via* fusing Gabor multi-orientation representations and local Gabor binary pattern histogram sequence. Cotter (2010) introduced to use sparse representation for FER. Especially, Kotsia et al. (2008) analyzed how partial occlusions affect FER performance and found that FER suffers more from mouth occlusion than the equivalent eyes occlusion.

Over the recent years, Convolution Neural Network (CNN) has shown exemplary performance on many computer vision tasks (Schroff et al., 2015; Krizhevsky et al., 2017; Li et al., 2020). The promising learning ability of deep CNN can be attributed to the use of hierarchical feature extraction stages that can adaptively learn the features from the data in an end-to-end fashion. There are many CNN-based FER works (Levi and Hassner, 2015; Ding et al., 2017; Meng et al., 2017; Zeng et al., 2018; Zhang et al., 2018; Li et al., 2019; Jiang et al., 2020). For FER under occlusion, Li et al. (2018a) proposed a CNN

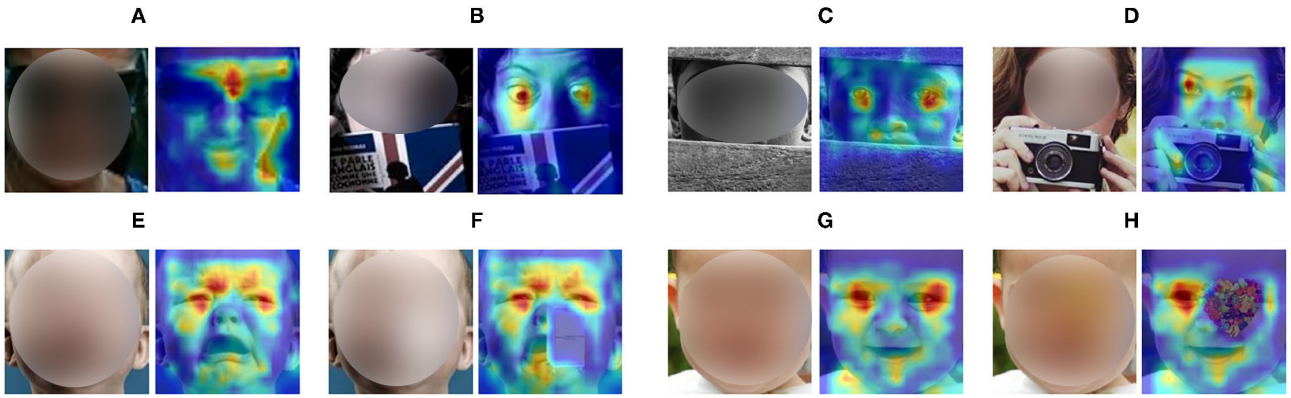


FIGURE 1 | Attention maps of several facial images with real (**A–D** in top row) or synthesized (**E–H** in bottom row) occlusions. Our proposed TFE is capable of perceiving the important facial regions for robust FER. A deep red means high attention. Better viewed in color and zoom in.

with attention mechanism (ACNN) to perceive facial expressions from unoccluded or partially occluded faces. ACNN crops facial patches from the area of important facial features, e.g., mouth, eyes, nose, and so on. The selected multiple facial patches are encoded as a weighed representation *via* a PG-Unit. The PG-Unit calculates the weight of each facial patch according to its obstructed-ness *via* an attention net. Based on this work, Wang et al. (2020b) proposed to randomly crop relative large facial patches instead of small fixed facial parts and refine the attention weights by a region bias loss function and relation-attention module. Ding et al. (2020) proposed an occlusion-adaptive deep network with a landmark-assisted attention branch network to perceive and drop the corrupted local features. Pan et al. (2019) introduced to train two CNNs from non-occluded facial images and occluded faces, respectively. Subsequently, they constrain the distribution of the encoded facial representations from two CNNs to be close *via* adversarial learning.

Our proposed TFE differs from previous CNN-based methods in two ways. One, TFE does not rely on facial landmarks for regional feature extraction. It is because the facial landmarks may show considerable misalignments under severe occlusions. Under this condition, the encoded facial parts are not part-aligned or semantic meaningful. Two, TFE is a transformer-based and the self-attention mechanism in the transformer that can flexibly attend to a sequence of image patches to encode the contextual cues. TFE consists of a region selection unit (RS-Unit) that automatically perceives and selects the critical facial regions for robust FER. TFE is potentially to obtain higher FER accuracy on both non-occluded and occluded faces. We will verify this in section 4.

2.2. Vision Transformer

Transformer models have largely facilitated research in machine translation and natural language processing (NLP) (Waswani et al., 2017). Transformer models have become the outstanding standard for NLP tasks. The main idea of the original transformer is to calculate the self-attention by comparing a representation to all other representations in the input sequence. In detail, features are first encoded to obtain memory [including value (V) and key (K)] and query (Q) embedding by linear projections. The product of the query Q with keys K is used as the attention weights for value V. A position embedding is also exploited and added to these representations to introduce the positional information in such a non-convolutional paradigm. Transformers are especially good at modeling long-range dependencies between elements of a sequence.

Inspired by the success of the transformer models, many recent studies try to use transformers in computer vision applications (Dosovitskiy et al., 2020; Li et al., 2021). Among them, Dosovitskiy et al. (2020) applied a pure transformer encoder for image classification. To obtain the input token representations, they crop the input image into 16×16 small patches and linearly map the patches to the input dimension of the encoder. Since then, ViTs are gaining rapid interest in various computer vision tasks because they offer a self-attention-based noval mechanism that can effectively capture long-range dependencies. Touvron et al. (2021) showed that ViT models can achieve competitive accuracy on ImageNet with stronger data augmentation and more regularization. Subsequently, transformer models are applied to other popular tasks such as object detection (Carion et al., 2020), segmentation (Jin et al., 2021), and video processing (Girdhar et al., 2019; Fang et al., 2020). In this study, we extend ViT to FER under occlusion and show its effectiveness.

3. METHOD

Figure 2 illustrates the main idea of the proposed TFE. Given an input face image, TFE encodes its convolutional feature maps *via* a commonly used backbone network such as ResNet-18 (He et al., 2016). Then, TFE encodes the robust facial expression representation *via* the vision transformer and the proposed RS-Unit. During the training stage, the encoded convolutional feature maps are decoded to reconstruct the unoccluded facial image. Below, we present the details of each of them.

3.1. Network Architecture

Following ViT (Dosovitskiy et al., 2020), we first preprocess the input image into a sequence of flattened image patches. However, the conventional split approach merely cuts the images into

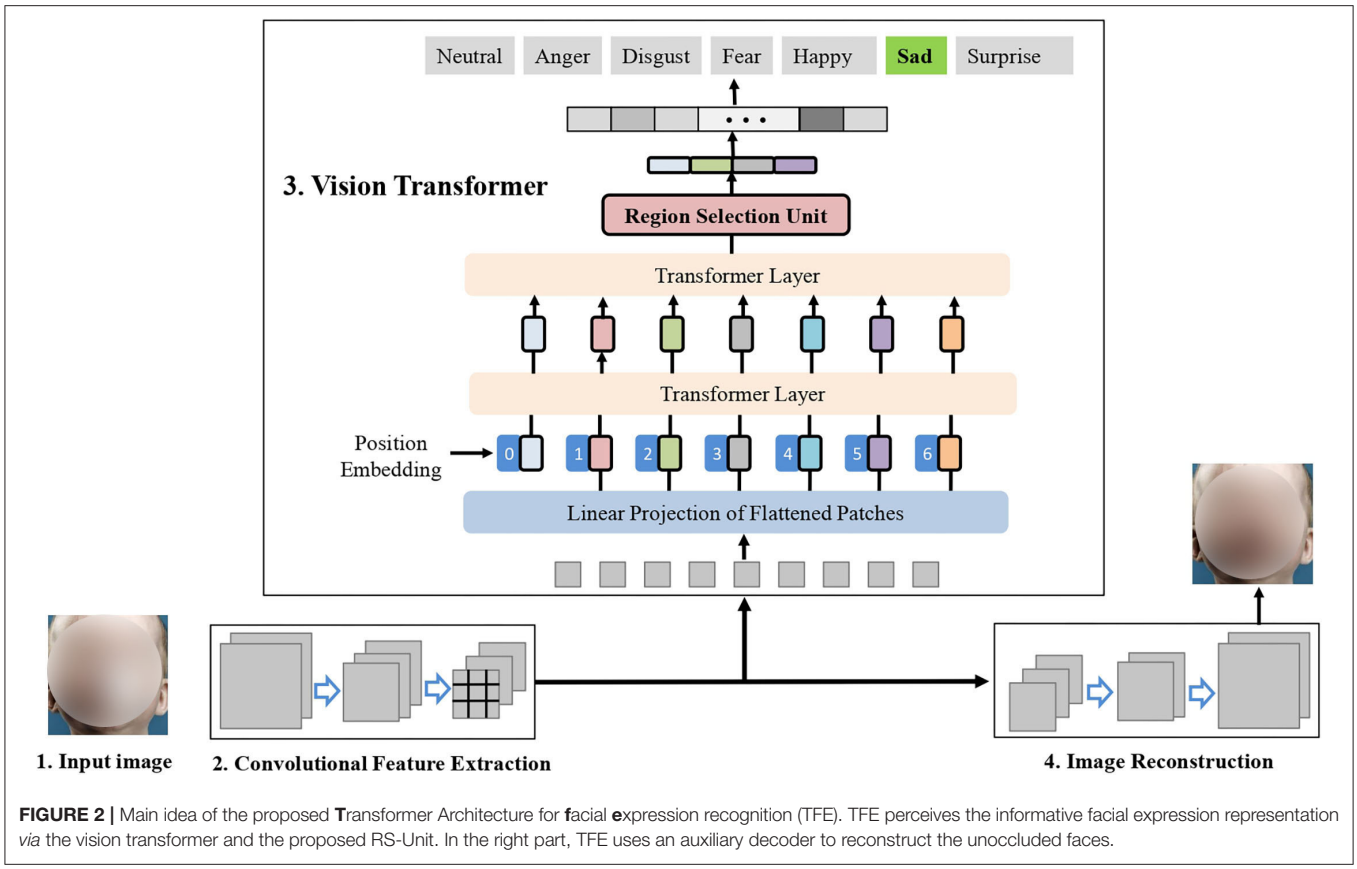


FIGURE 2 | Main idea of the proposed **T**ransformer Architecture for **f**acial **e**xpression recognition (TFE). TFE perceives the informative facial expression representation *via* the vision transformer and the proposed RS-Unit. In the right part, TFE uses an auxiliary decoder to reconstruct the unoccluded faces.

overlapping or non-overlapping patches, which harms the local neighboring structures and shows substandard optimizability (Xiao et al., 2021). Inspired by Xiao et al. (2021) that exploits a few number of stacked 3×3 convolutions for image sequentialization, we adopt the popular ResNet-based backbone (He et al., 2016) to encode the input facial image $\mathbf{I}$. A typical ResNet usually has four stages (Li et al., 2021), and we use the output of the S-th stage as the encoded feature maps $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ feature maps; thus, we get a total of $N = H \times W$ image tokens, each token $\mathcal{X}_i$ with a feature dimension of C. As H equals W, here we use $P = H = W$ for brevity. In our proposed TFE, the image tokens have the spatial size 1×1, the input sequence is obtained by: (i) flattening the spatial dimensions of the feature map and (ii) projecting the flattend tokens to the target transformer dimension.

We map the flattend image token $\mathbf{X}_i$ into a latent D-dimensional feature space *via* a learnable fully connected neural layer. With the sliced image token $\mathbf{X}_i \in \mathbb{R}^{P^2 \times D}, 1 \leq i \leq N$, a trainable position embedding is plused to the token embeddings to retain positional information as follows:

$$\mathbf{Z}_0 = [\mathbf{X}_{class}; \mathbf{X}_1\mathbf{E}; \mathbf{X}_2\mathbf{E}; \mathbf{X}_N\mathbf{E}] + \mathbf{E}_{pos}, \quad (1)$$

$$\mathbf{Z}_l{}' = MSA(LN(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1}, \quad l \in 1, 2, \cdots, L \quad (2)$$

$$\mathbf{Z}_l = MLP(LN(\mathbf{Z}_l{}')) + \mathbf{Z}_l', \quad l \in 1, 2, \cdots, L, \quad (3)$$

where N means the number of the image tokens, $\mathbf{E}$ is the token embedding projection, and $\mathbf{E}_{pos}$ means the position embedding. L means the number of layers of the multi-head self-attention (MSA) and the multi-layer perceptron (MLP) blocks. The transformer encoder includes alternating layers of multi-head self-attention (MSA) and multilayer perceptron (MLP) blocks. We also add a layernorm (LN) layer before every block and residual connections after every block. Besides, the MLP consists of two fully connected neural layers with a GELU non-linearity. $\mathbf{X}_{class}$ is a classification token that consists of an embedding attached to the sequence of embedded patches. After L transformer layers, a classification head is attached to $\mathbf{Z}_L^0$. We implemented the classification with a MLP that consists of one hidden layer at the training and testing phase.

3.2. Vision Transformer With RS-Unit

One of the most important problems in FER under occlusion is to precisely perceive the discriminative facial regions that represent subtle facial deformations caused by facial expressions. To this end, we proposed a RS-Unit to automatically select the critical facial parts for robust FER under occlusions. Different with previous methods that use facial landmarks for facial region decomposition (Li et al., 2018a; Ding et al., 2020; Wang et al., 2020b), RS-Unit does not need auxiliary annotation and merely adopts the pre-computed multi-head attention information.

Suppose the model consists of M self-attention heads and the hidden features, outputs of the last transformer layer are denoted as $\mathbf{Z}_L = [Z_L^0, Z_L^1, Z_L^2, \cdots, Z_L^N]$. To better utilize the attention

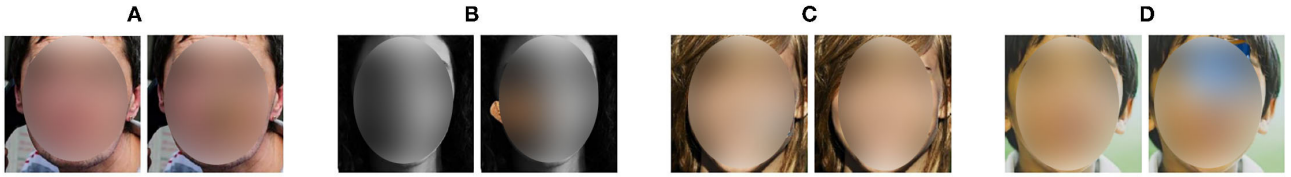


FIGURE 3 | Examples of the synthesized occluded images. The occluders are various in shape, color, and facial positions. **(A)** Anger, **(B)** neutral, **(C)** happy, and **(D)** sad.

information, the input to the final classification layer is changed. In detail, the raw attention weights are obtained *via* recursive matrix multiplication in all the layers:

$$\mathbf{a}_{total} = \sum_{l=0}^{L} \mathbf{a}_l. \tag{4}$$

As $\mathbf{a}_{total}$ spots how information propagates from the preceding transformer layer to the features in the later transformer layers, $\mathbf{a}_{total}$ should be a promising choice to capture the important local facial regions for FER (He et al., 2021). Thus, we can choose the positions of the maximum values with regard to the M different attention heads in $\mathbf{a}_{total}$. We then choose the indexes of the maximum values $A_1, A_2, \cdots, A_M$ w.r.t the M different attention heads in $\mathbf{a}_{total}$. These indexes are exploited as positions for RS-Unit to select the corresponding tokens in $\mathbf{Z}_L$. At last, we combine the classification token with the selected tokens along as the final representation:

$$\mathbf{Z}_{select} = \text{Concat}[Z_L^0, Z_L^{A_1}, Z_L^{A_2}, \cdots, Z_L^{A_M}]. \tag{5}$$

By utilizing the entire input sequence with tokens tightly related to discriminative facial regions and combine the classification token as input to the classification layer, our proposed TFE is capable of utilizing the global facial information but also the local facial regions that contain critical subtle facial deformations induced by facial expressions. Thus, our proposed TFE is expected to perceive the discriminative facial regions for robust FER under occlusions.

3.3. Image Reconstruction

Since the facial expression is a subtle deformation of faces that can be inferred from multiple facial regions, it is beneficial to explicitly infer the occluded facial parts from the unoccluded regions. In the image inpainting process, the model is tasked to precisely perceive the fine-grained facial action units to infer their co-occurrence (Li et al., 2018a).

Inspired by this, we propose to reconstruct the facial image with an auxiliary decoder. To this end, we synthesize the occluded face images by manually collecting abundant masks for generating the occluders. We show some randomly selected occluded images in **Figure 3**. With the occluded faces $\mathbf{I}_{occ}$ and the corresponding original images $\mathbf{I}_{ori}$, we are capable of reconstructing the images as follows,

$$\mathcal{L}_{rec} = \|\mathbf{I}_{ori} - Dec(Enc(\mathbf{I}_{occ}))\|_1, \tag{6}$$

where *Enc* means the convolutional feature extraction operation shown in **Figure 2**, *Dec* denotes the image decoding process.

3.4. Overall Objective

Transformer-based FER method is trained in an end-to-end fashion by minimizing the integration of the FER loss and the image reconstruction loss in Equation (6). We integrate the two goals and obtain the full objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{rec}, \tag{7}$$

where hyper-parameter λ controls the importance of the image reconstruction term.

4. EXPERIMENT

4.1. Implementation Details

We adopted ResNet-18 (He et al., 2016) as the backbone network for TFE due to its elegant structure and excellent performance in image classification. We used the output of the third stage as the convolutional feature maps: $\mathbf{X} \in \mathbb{R}^{14\times14\times1024}$. Thus, the token size is $N = 14 \times 14$. We set $L = 4$, $D = 768$, and $M = 12$. We initialized the backbone of TFE with the pre-trained model based on ImageNet dataset. We mixed all the facial expression datasets with their modifications with artificial facial occlusions with the ratio of 1:1. TFE was optimized *via* a batch-based stochastic gradient descent manner. We actually set the batch size as 128 and the base learning rate as 0.001. The weight decay was set as 0.0005 and the momentum was set as 0.9. The optimal setting for the loss weight between the FER and image reconstruction term was set as $1:1$ by grid search.

4.1.1. Datasets

We evaluated the methods on two facial expression datasets [RAF-DB (Li et al., 2017) and AffectNet (Mollahosseini et al., 2017)]. We additionally evaluate our proposed TFE on FED-RO dataset (Li et al., 2018a). **RAF-DB** consists of about 30,000 facial images annotated with compound or basic expressions by 40 trained human. We merely used the images with seven basic expressions. We obtained totally 12,271 images for training data and 3,068 images for evaluation. **AffectNet** is currently the largest dataset with annotated facial expressions. AffectNet consists of approximately 400,000 images manually annotated. We merely utilized the images with six basic and neutral expressions, We obtained about 280,000 images for training and 3,500 images for evaluation. **FED-RO** (Li et al., 2018a) is a facial expression

TABLE 1 | Test set accuracy on RAF-DB dataset.

Method	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	ACC (Overall/Ave)
AlexNet (Li et al., 2017)	60.15	58.64	21.87	39.19	86.16	60.88	62.31	–/55.60
VGG16 (Li et al., 2017)	59.88	68.52	27.50	35.13	85.32	64.85	66.32	80.96/58.22
DLP-CNN (Li et al., 2017)	80.29	71.60	52.15	62.16	92.83	80.13	81.16	80.89/74.20
gACNN (Li et al., 2018a)	84.30	78.42	53.11	55.39	93.17	82.88	86.27	85.07/76.22
TAE (Li et al., 2020)	62.80	58.01	45.03	58.12	76.03	45.85	64.44	81.68/58.61
TFE (Ours)	86.76	79.01	64.38	66.22	95.61	87.03	90.27	88.49/81.33

TABLE 2 | Validation set accuracy on AffectNet dataset.

Method	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	ACC (Overall/Ave)
AlexNet (Mollahosseini et al., 2017)*	–	–	–	–	–	–	–	47.00/47.00
RAN-ResNet18 (Wang et al., 2020b)*	–	–	–	–	–	–	–	52.90/52.90
VGG16 (Simonyan and Zisserman, 2014)	**89.61**	53.42	20.61	32.03	90.03	35.01	37.22	51.13/51.13
FAB-Net (Wiles et al., 2018)	38.64	30.62	**48.42**	32.14	82.25	35.61	51.42	45.59/45.59
TAE (Li et al., 2020)	44.42	38.63	46.84	40.39	78.01	40.81	**54.41**	49.07/49.07
gACNN (Li et al., 2018a)	73.42	66.18	32.59	46.22	93.81	55.82	43.43	58.78/58.78
OADN (Ding et al., 2020)	–	–	–	–	–	–	–	61.90/61.90
SCN (Wang et al., 2020a)	–	–	–	–	–	–	–	60.23/60.23
TFE (Ours)	76.03	**68.09**	46.83	**47.03**	**94.12**	**57.32**	53.90	**63.33/63.33**

*The bold values denotes the best results. *Means the values are reported in the original papers.*

database with real-world occlusions. Each face has real occlusions in uncontrolled environment. There are totally 400 images in FED-RO dataset annotated with seven expressions. We train the proposed TFE on the joint training data of AffectNet and RAF dataset, following the protocol suggested in Li et al. (2018a).

Following (Li et al., 2018a), we manually collected approximately 4 k images as masks for generating the occluders. These occluders were discovered and saved from search engine *via* more than 50 keywords, such as hair, hat, book, beer, apple, cabinet, computer, orange, etc. The height H and width W of the occluders S satisfy $H \in [96, 128]$ and $W \in [96, 128]$. **Figure 3** shows some occluded faces. It is evident that the artificial occluded facial images are diverse in occlusion patterns.

4.1.2. Evaluation Metric

We report FER performance on both the occluded and non-occluded images of all the datasets. We used the overall and the overall and average accuracy on seven facial expression categories (i.e., six prototypical plus neutral categories) as a performance metric. Besides, we also report some confusion matrixes on RAF-DB dataset to show the discrepancies between the expressions.

4.2. FER Experimental Results

We compare the proposed TFE with the state-of-the-art FER methods, including DLP-CNN (Li et al., 2017), gACNN (Li et al., 2018a), FAB-Net (Wiles et al., 2018), TAE (Li et al., 2020), OADN (Ding et al., 2020), and SCN (Wang et al., 2020a). The comparison results are shown in **Tables 1**–**3**.

Table 1 shows the FER results of our method and previous studies on RAF-DB dataset. Our TFE achieves 81.33% in the average accuracy on seven facial expression categories. Compared with DLP-CNN (Li et al., 2017), TFE obtains 7.13% improvements in the average accuracy. Compared with the strongest competing method in the same setting gACNN (Li et al., 2018a), TFE surpasses it by 5.61%. The benefits of TFE over other methods can be explained in two-fold. First, TFE explicitly utilizes transformer layers in the network structure. The self-attention in the transformers has been shown to effectively learn local to global interactions and relations between distant facial parts. Besides, the RS-Unit on top of the transformer layers in our proposed TFE helps perceive the critical facial regions. Thus, TFE is capable of spotting the local subtle facial deformations induced by facial expressions. Second, TFE explicitly reconstructs the unoccluded facial images with an auxiliary decoder, which facilitates the backbone CNN in TFE to learn to infer the occluded facial parts *via* the important facial regions.

TABLE 3 | Test set accuracy on FED-RO dataset.

Method	ResNet18	RAN	DLP-CNN	gACNN	OADN	TFE
ACC (AVE)	64.25	67.98	60.31	66.50	68.11	**70.60**

The bold values denotes the best results.

Table 2 shows the comparisons of our TFE and other state-of-the-art FER methods on AffectNet dataset. TFE achieves 63.33% in the average accuracy on seven facial expression categories. Compared with RAN-ResNet-18 (Wang et al., 2020b) that use multiple crops of facial images as input and learns adaptive weights for each input image, TFE obtains 10.43% improvements

TABLE 4 | Ablation study on RAF-DB dataset.

Method	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	ACC (Overall/Ave)
Original test set of RAF-DB dataset								
TFE (*w/o D, w/o T*)	83.97	79.01	60.63	60.81	94.51	85.56	86.32	85.91 /79.69
TFE (*w/ D, w/o T*)	85.15	**83.33**	**65.63**	64.86	**95.78**	**87.03**	84.80	86.20/80.94
TFE	**86.76**	79.01	64.38	**66.22**	95.61	**87.03**	**90.27**	**88.64/81.33**
Synthesized occluded test set of RAF-DB dataset								
TFE (*w/o D, w/o T*)	79.41	**76.54**	53.12	54.05	91.90	81.80	80.85	83.68/73.95
TFE (*w/ D, w/o T*)	81.47	75.93	55.62	59.46	93.42	**84.73**	80.55	84.00/75.88
TFE	**83.53**	72.84	**60.00**	**67.57**	**93.50**	82.85	**85.41**	**85.12/77.96**

The bold values denotes the best results.

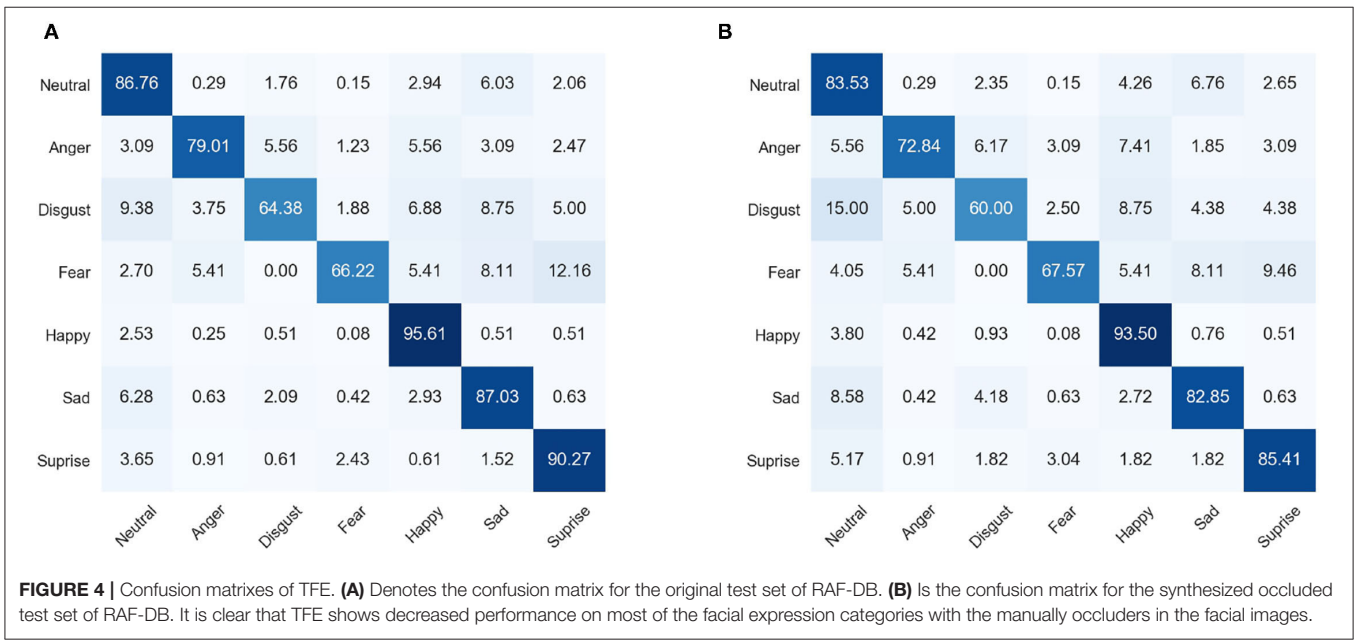


FIGURE 4 | Confusion matrixes of TFE. **(A)** Denotes the confusion matrix for the original test set of RAF-DB. **(B)** Is the confusion matrix for the synthesized occluded test set of RAF-DB. It is clear that TFE shows decreased performance on most of the facial expression categories with the manually occluders in the facial images.

in the average accuracy. Compared with the self-supervised methods FAB-Net (Wiles et al., 2018) and TAE (Li et al., 2020), TFE shows its success in almost each facial expression category. Among the state-of-the-art FER methods, gACNN (Li et al., 2018a) and OADN (Ding et al., 2020) both exploit the 24 facial landmarks for facial region decomposition and learn the path-specific representation to better capture the local details of the input facial image. However, their FER performance still lags behind our proposed TFE, as illustrated in **Table 2**. This is because the transformer layers in TFE naturally encode the patch-specific face representation by tokenizing the input convolutional feature maps. TFE does not rely on facial landmarks to extract the local representations and avoids the negative influence induced by the misalignments of the facial landmarks. We additionally show the FER performance comparison on FED-RO dataset in **Table 3**. FED-RO dataset is the first facial expression dataset with real occlusions. TFE achieves 70.60% in the average accuracy and outperforms other compared methods with no exception. In summary, the experimental results in **Tables 1**–**3** verify the superiority of the proposed TFE for robust facial expression recognition.

4.2.1. Ablation Study

Both the transformer layers and auxiliary decoder help TFE obtain improvements on FER. We performed a quantitative study of these two parts in order to better understand the benefits of TFE.

We show the FER performance of TFE without auxiliary image reconstruction decoder and without the transformer layers (as well as RS-Unit) [TFE (*w/o D, w/o T*)], and TFE with the auxiliary image reconstruction decoder but without transformer layers and RS-Unit [TFE (*w/ D, w/o T*)] in **Table 4**. It is clear that TFE (*w/o D, w/o T*) shows decreased FER performance on both the original and synthesized occluded face images. With the auxiliary image reconstruction decoder, TFE (*w/ D, w/o T*) illustrates improved FER performance in many facial expression categories. The comparisons between TFE (*w/o T, w/o D*) and TFE (*w/ T, w/o D*) demonstrate the effectiveness of the auxiliary image reconstruction decoder. With the transformer layers and the auxiliary image decoder, TFE obtains the best FER performance. As illustrated in **Table 4**, TFE shows its benefits in *Neutral*, *Fear*, *Surprise* and obtains comparable accuracy in *Disgust*, *Happy*, *Sad*.

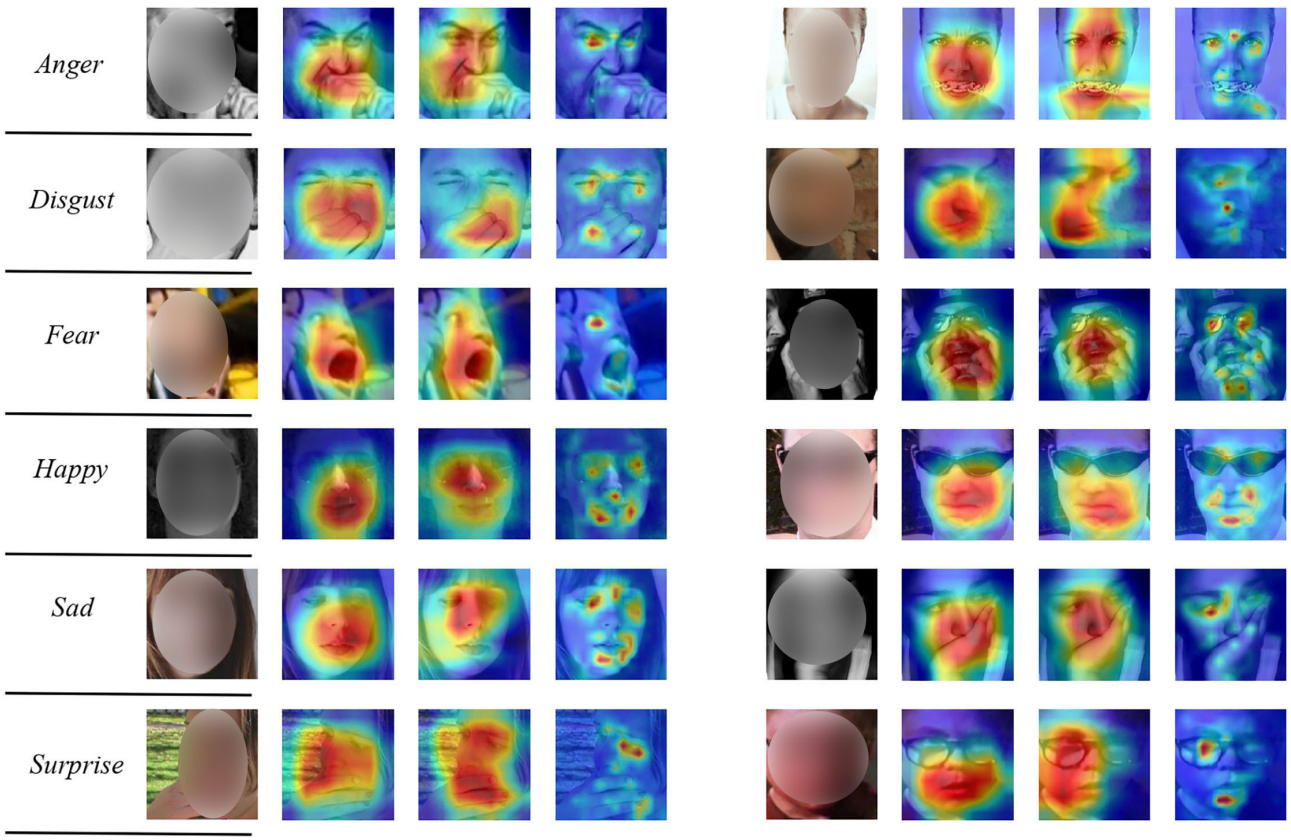


FIGURE 5 | Attention maps of several facial images with occlusions. For each input face image, the first, second, and third column, respectively, show the attention map of TFE (*w/o D, w/o T*), TFE (*w/ D, w/o T*), and TFE. Our proposed TFE is capable of perceiving the important facial regions for robust FER. A deep red denotes low attention. A deep red means high attention. Better viewed in color and zoom in.

We additionally show the confusion matrixes of our proposed TFE on both the original and synthesized occluded test set of RAF-DB dataset in **Figure 4**. It is clear that TFE shows degraded performance on most of the facial expression categories when the facial images are occluded in **Figure 4B**. Besides, TFE shows the lowest FER accuracy on *Disgust* category and highest accuracy on *Happpy* category. Easily confused expression categories are *disgust* and *sad*, *fear* and *surprise*, and *fear* and *sad*. Our above observations are consistent with the conclusions in Li et al. (2018a).

We show the attention maps of the TFE and its variants in **Figure 5**. For each input face, the first, second, and third column, respectively, show the attention map of TFE (*w/o D, w/o T*), TFE (*w/ D, w/o T*), and our proposed TFE. It is evident that TFE is capable of shifting attention from the occluded facial patches to other unobstructed regions. As a comparison, TFE (*w/o T, w/o D*) and TFE (*w/ D, w/o T*) are not capable of precisely focusing on the important and unobstructed facial parts. Taking facial images labeled with *Happy* in the fourth row for example, TFE perceives the eyes and the corner or the mouth precisely, irrespective of the facial occlusions. The visualization results show the benefits of the proposed RS-Unit and the auxiliary decoder for robust FER under occlusions.

5. CONCLUSIONS

In this study, we propose a transformer-based FER method (TFE) that is capable of adaptatively focusing on the most important and unoccluded facial regions. Considering that facial expression is represented by several specific facial parts, we propose a RS-Unit to automatically perceive the critical facial parts so as to explicitly perceive the important facial regions for robust FER. To better perceive the fine-grained facial deformations and infer the co-occurrence of different facial action units, TFE consists of an auxiliary decoder to reconstruct the facial image. Quantitative and qualitative experiments have verified the feasibility of our proposed TFE. TFE also outperforms other state-of-the-art FER approaches. Ablation and visualization analyses show TFE is capable of shifting attention from the occluded facial regions to other important ones. Currently, TFE exploits

the fixed patch size as the input to the transformer layer while larger facial patch size might be a better choice for the heavily occluded facial images. We will explore this in the future work. Besides, we will also explore how to reduce the computation overhead and make TFE suit for mobile deployment.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

JG and YZ cooperatively led the method design and experiment implementation. JG wrote the sections of the manuscript. YZ provided result review, theoretical guidance, and paper revision. Both authors have read and approved the final manuscript.

FUNDING

This publication of this paper was supported by the Henan key R & D and promotion projects (Grant: 212102310551) and the Key Scientific Research Project Plan of Henan Province colleges and universities (19A520008, 20A413002).

REFERENCES

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *European Conference on Computer Vision* (Glasgow), 213–229. doi: 10.1007/978-3-030-58452-8_13

Cotter, S. F. (2010). "Sparse representation for accurate classification of corrupted and occluded facial expressions," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (Dallas,TX), 838–841. doi: 10.1109/ICASSP.2010.5494903

Ding, H., Zhou, P., and Chellappa, R. (2020). "Occlusion-adaptive deep network for robust facial expression recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)* (Houston, TX), 1–9. doi: 10.1109/IJCB48548.2020.9304923

Ding, H., Zhou, S. K., and Chellappa, R. (2017). "Facenet2expnet: regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC), 118–126. doi: 10.1109/FG.2017.23

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Available online at: https://arxiv.org/pdf/2010.11929v1.pdf

Fang, Y., Gao, S., Li, J., Luo, W., He, L., and Hu, B. (2020). Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting. *Neurocomputing* 392, 98–107. doi: 10.1016/j.neucom.2020.01.087

Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. (2019). "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*(Long Beach, CA, USA), 244–253. doi: 10.1109/CVPR.2019.00033

He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., et al. (2021). *Transfg: A Transformer Architecture for Fifine-Grained Recognition*. Available online at: https://arxiv.org/abs/2103.07976v1

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., et al. (2020). "DFEW: a large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA), 2881–2889. doi: 10.1145/3394171.3413620

Jin, Y., Han, D., and Ko, H. (2021). TRSEG: transformer for semantic segmentation. *Pattern Recogn. Lett.* 148, 29–35. doi: 10.1016/j.patrec.2021.04.024

Kotsia, I., Buciu, I., and Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image Vis. Comput.* 26, 1052–1067. doi: 10.1016/j.imavis.2007.11.004

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Levi, G., and Hassner, T. (2015). "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 503–510. doi: 10.1145/2818346.2830587

Li, S., Deng, W., and Du, J. (2017). "Reliable crowdsourcing and deep locality preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 2852–2861. doi: 10.1109/CVPR.2017.277

Li, Y., Sun, Y., Cui, Z., Shan, S., and Yang, J. (2021). Learning fair face representation with progressive cross transformer. *arXiv preprint arXiv:2108.04983*.

Li, Y., Zeng, J., and Shan, S. (2020). Learning representations for facialactions from unlabeled videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 99, 1–1. doi: 10.1109/TPAMI.2020.3011063

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018a). Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Process.* 28, 2439–2450. doi: 10.1109/TIP.2018.2886767

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018b). "Patch-gated CNN for occlusion aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)* (Beijing), 2209–2214. doi: 10.1109/ICPR.2018.8545853

Li, Y., Zeng, J., Shan, S., and Chen, X. (2019). "Self-supervised representation learning from videos for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 10924–10933. doi: 10.1109/CVPR.2019.01118

Liu, S.-S., Zhang, Y., Liu, K.-P., and Li, Y. (2013). "Facial expression recognition under partial occlusion based on gabor multi-orientation features fusion and local gabor binary pattern histogram sequence," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (Beijing), 218–222. doi: 10.1109/IIH-MSP.2013.63

Meng, Z., Liu, P., Cai, J., Han, S., and Tong, Y. (2017). "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition* (Washington, DC), 558–565. doi: 10.1109/FG.2017.140

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* 10, 18–31. doi: 10.1109/TAFFC.2017.2740923

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F. S., and Yang, M. H. (2021). *Intriguing Properties of Vision Transformers*. Available online at: https://arxiv.org/abs/2105.10497

Pan, B., Wang, S., and Xia, B. (2019). "Occluded facial expression recognition enhanced through privileged information," in *Proceedings of the 27th ACM International Conference on Multimedia* (Nice), 566–573. doi: 10.1145/3343031.3351049

Rudovic, O., Pantic, M., and Patras, I. (2012). Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1357–1369. doi: 10.1109/TPAMI.2012.233

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). "Facenet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 815–823. doi: 10.1109/CVPR.2015.7298682

Simonyan, K., and Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Available online at: https://export.arxiv.org/abs/1409.1556

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*(Virtual Event), 10347–10357.

Wang, K., Peng, X., Yang, J., Lu, S., and Qiao, Y. (2020a). "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6897–6906. doi: 10.1109/CVPR42600.2020.00693

Wang, K., Peng, X., Yang, J., Meng, D., and Qiao, Y. (2020b). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* 29, 4057–4069. doi: 10.1109/TIP.2019.2956143

Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al. (2017). "Attention is all you need," in *NIPS*.

Wiles, O., Koepke, A., and Zisserman, A. (2018). Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*. doi: 10.1109/ICCVW.2019.00364

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., and Girshick, R. (2021). *Early Convolutions Help Transformers See Better*. Available online at: https://arxiv.org/pdf/2106.14881.pdf

Zeng, J., Shan, S., and Chen, X. (2018). "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision* (Munich), 222–237. doi: 10.1007/978-3-030-01261-8_14

Zhang, L., Tjondronegoro, D., and Chandran, V. (2014). Random gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing* 145, 451–464. doi: 10.1016/j.neucom.2014.05.008

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2018). Spatial-temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* 49, 839–847. doi: 10.1109/TCYB.2017.2788081

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neurorobotics

REVIEW
published: 29 November 2021
doi: 10.3389/fnbot.2021.784514

Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives

Shiqing Zhang[1], *Ruixin Liu*[1,2], *Xin Tao*[1] *and Xiaoming Zhao*[1]*

[1] Institute of Intelligence Information Processing, Taizhou University, Zhejiang, China, [2] School of Sugon Big Data Science, Zhejiang University of Science and Technology, Zhejiang, China

Automatic speech emotion recognition (SER) is a challenging component of human-computer interaction (HCI). Existing literatures mainly focus on evaluating the SER performance by means of training and testing on a single corpus with a single language setting. However, in many practical applications, there are great differences between the training corpus and testing corpus. Due to the diversity of different speech emotional corpus or languages, most previous SER methods do not perform well when applied in real-world cross-corpus or cross-language scenarios. Inspired by the powerful feature learning ability of recently-emerged deep learning techniques, various advanced deep learning models have increasingly been adopted for cross-corpus SER. This paper aims to provide an up-to-date and comprehensive survey of cross-corpus SER, especially for various deep learning techniques associated with supervised, unsupervised and semi-supervised learning in this area. In addition, this paper also highlights different challenges and opportunities on cross-corpus SER tasks, and points out its future trends.

Keywords: speech emotion recognition, cross-corpus, deep learning, feature learning, survey

OPEN ACCESS

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Xiaoya Zhang,
Nanjing University of Science and Technology, China
Dongmei Jiang,
Northwestern Polytechnical University, China
Ziping Zhao,
Tianjin Normal University, China

***Correspondence:**
Xiaoming Zhao
tzxyzxm@163.com

Received: *28 September 2021*
Accepted: *08 November 2021*
Published: *29 November 2021*

Citation:
Zhang S, Liu R, Tao X and Zhao X (2021) Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives. Front. Neurorobot. 15:784514. doi: 10.3389/fnbot.2021.784514

INTRODUCTION

Emotion recognition is an important direction in psychology, biology, and computer science, and has recently received extensive attention from the engineering research field. One of the starting points for emotion recognition is to assist in designing more humane human-computer interaction (HCI) methods, since emotion plays a key role in the fields of HCI, artificial intelligence (Cowie et al., 2001; Ramakrishnan and El Emary, 2013; Feng and Chaspari, 2020).

Traditional HCI is mainly carried out through keyboard, mouse, screen, etc. It only pursues convenience and accuracy, and cannot understand and adapt to people's emotions or mood. And if the computer lacks the ability to understand and express emotions, it is difficult to expect the computer to have the same intelligence as human beings. Moreover, it is also difficult to expect HCI to be truly harmonious and natural. Since the communications and exchanges between humans are natural and emotional, people naturally expect computers to have emotional capabilities in the procedure of HCI. The purpose of affective computing (Picard, 2010) is to endow computers the ability to observe, understand, and generate various emotional features similar to humans, and ultimately enable computers to interact naturally, cordially, and vividly like humans.

Emotion recognition is one of the most basic and important research subjects in the field of affective computing. Speech signals convey human emotional information most naturally. At present, speech emotion recognition (SER), which aims to classify

human emotions from affective speech signals, has become a hot research topic in the fields of signal processing, pattern recognition, artificial intelligence, HCI, etc. Studying on SER has been going on for more than two decades (Schuller, 2018) and it has been applied to HCI (Cowie et al., 2001; Fragopanagos and Taylor, 2005), affective robots (Samani and Saadatian, 2012; Zhang et al., 2013), call-centers (Morrison et al., 2007), e-learning system (Li et al., 2007), computer games (Yildirim et al., 2011), depression severity classification (Harati et al., 2018), detection of autism spectrum disorder (ASD) (Lin et al., 2020), and so on.

During the past two decades, tremendous efforts have been made to focus on SER. Several survey related to SER can be found in El Ayadi et al. (2011), Anagnostopoulos et al. (2015), and Akçay and Oguz (2020). Note that the majority of existing SER systems are trained and evaluated on a single corpus and a single language setting. However, in many practical applications, there are great differences between training corpus and testing corpus. For example, the training and testing corpora come from two (or more) different languages, cultures, distribution modes, data scales, and so on. These differences across corpora result in significant idiosyncratic variations impeding the generalization of current SER techniques, thereby yielding an active research subject called cross-corpus SER in the field of SER.

Generally, in a basic cross-corpus SER system there are two crucial steps: emotion classifier and domain-invariant feature extraction. In the following, we will introduce these two steps of cross-corpus SER in brief.

As for emotion classifier, various traditional machine learning methods can be utilized for cross-corpus SER. The representative emotion classification methods contain linear discriminant classifier (LDC) (Banse and Scherer, 1996; Dellaert et al., 1996), K-Nearest Neighbor (Dellaert et al., 1996), artificial neural network (ANN) (Nicholson et al., 2000), support vector machines (SVM) (Kwon et al., 2003), hidden Markov models (HMM) (Nwe et al., 2003), Gaussian mixture models (GMM) (Ververidis and Kotropoulos, 2005), sparse representation classification (SRC) (Zhao and Zhang, 2015) and so on. Nevertheless, each classifier has its own advantages and disadvantages. The classifier combination method integrating the advantages of multiple classifiers (Morrison et al., 2007; Albornoz et al., 2011) began to draw researchers' attention.

Domain-invariant feature extraction, which aims to learn generalized feature representations of affective speech that are invariant across corpora, is another critical step in a cross-corpus SER system. So far, a variety of domain-invariant feature extraction methods have been explored for cross-corpus SER. According to the fact that the used data label information is whether included or not, existing domain-invariant feature extraction techniques for cross-corpus SER can be divided into three categories: supervised learning, semi-supervised learning, and unsupervised learning. Supervised learning is defined by its use of labeled sample data. In terms of labeled inputs and outputs, the used algorithm could measure its performance over time. In contrast, unsupervised learning aims to discover the inherent structure of unlabeled sample data without the demand for human intervention. Semi-supervised learning characterizes a type of the learning algorithms which try to learn from unlabeled and labeled sample data, generally supposing that the samples come from the same or similar distribution.

In the early cross-corpus SER literatures, to alleviate the problem of corpus-specific discrepancy for generalization, a variety of supervised, unsupervised, and semi-supervised techniques have been already developed on the basis of several typical hand-crafted low-level descriptors (LLDs), such as prosodic features, voice quality features and spectral features (Luengo et al., 2010; Zhang and Zhao, 2013), the INTERSPEECH-2009 emotion challenge (384 parameters) (Schuller et al., 2009b), the INTERSPEECH-2010 paralinguistic challenge (1,582 parameters) (Schuller et al., 2010a), the INTERSPEECH-2013 computational paralingusitics challengE (ComParE) set (6,373 parameters) (Schuller et al., 2013), the Geneva minimalistic acoustic parameter set (GeMAPS) (88 parameters) (Eyben et al., 2016), and so on. In particular, after extracting hand-crafted LLDs, for simply eliminating differences of cross-corpus acoustic features, corpus-based normalization in a supervised (Schuller et al., 2010b) or unsupervised manner (Zhang et al., 2011) was presented. In addition, several more sophisticated methods were also developed to learn common feature representations from the extracted hand-crafted LLDs, by means of supervised-based (Song et al., 2016b) or semi-supervised based matrix factorization (Luo and Han, 2019), supervised-based (Mao et al., 2017), or unsupervised-based domain adaption (Deng et al., 2017), etc. In recent years, the current state-of-art technique is to employ an adversarial learning scheme in an unsupervised (Abdelwahab and Busso, 2018) or semi-supervised (Latif et al., 2020) manner for learning a domain-invariant acoustic feature representation on cross corpus SER tasks.

Although the above-mentioned hand-crafted acoustic features associated with supervised, unsupervised, and semi-supervised learning approaches can produce good domain-invariant features for cross-corpus SER, they are still low-level and not highly discriminative. It is thus desirous to obtain high-level domain-invariant feature representations for cross-corpus SER.

To achieve high-level domain-invariant feature representations for cross-corpus SER, the recently-emerged deep learning (LeCun et al., 2015) methods may present a possible solution. The representative deep leaning techniques contain deep belief networks (DBNs) (Hinton and Salakhutdinov, 2006), convolutional neural networks (CNNs) (Krizhevsky et al., 2012), recurrent neural networks (RNNs) (Elman, 1990) and its variant called long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), autoendcoders (AEs) (Ballard, 1987; Schmidhuber, 2015) and so on. So far, deep learning methods have shown good performance on object detection and classification (Wu et al., 2020), natural language processing (Otter et al., 2020), speech signal processing (Purwins et al., 2019), multimodal emotion recognition (Zhou et al., 2021), and so on, due to its strong feature learning ability.

Inspired by the lack of summarizing recent advances in various deep learning techniques for cross-corpus SER, this paper aims to present an up-to-date and comprehensive survey of cross-corpus SER, especially for various deep learning techniques associated with supervised, unsupervised and semi-supervised

learning in this area. In addition, this paper highlights different challenges and opportunities on cross-corpus SER tasks, and point out its future trends. To the best of our knowledge, we are the first attempt to provide such a review for deep cross-corpus SER.

The organization of this paper is as follows. A review of speech emotion databases is presented at first. Then, we simply review supervised, unsupervised, and semi-supervised learning in details. Next, we review traditional methods for cross-corpus SER. We show recent advances of the applications of deep learning techniques incorporated with supervised, unsupervised and semi-supervised learning for cross-corpus SER. Next, we give a summary of open challenge and future directions. Finally, concluding remarks are provided.

SPEECH EMOTION DATABASES

For cross-corpus SER, a variety of speech emotion databases have been developed. **Table 1** presents a brief summary of existing speech emotion databases. In this section, we describe briefly these existing speech emotion databases, as described below.

DES

The Danish Emotional Speech (DES) (Engberg et al., 1997) dataset contains 5,200 audio utterances, simulated by four professional actors (2 females, 2 males). The simulated utterances consist of five emotional states: anger, happiness, neutral, sadness, and surprise. The audio recordings from each actor are composed of two isolated words, nine sentences and two passages of fluent speech materials. The whole audio utterances last about 30 min in duration. For a listening test, 20 listeners were employed.

SUSAS

The Speech Under Simulated and Actual Stress (SUSAS) (Hansen and Bou-Ghazale, 1997) dataset is a speech under stress corpus including five kinds of stress and feelings. It contains a highly confused collection of 35 aircraft communication vocabulary words. The researchers invited 32 speakers (13 females, 19 males) to produce more than 16,000 utterances. Simulated speech under stress is composed of ten stress styles such as speaking style, single tracking task, and Lombard effect domain.

SmartKom

The SmartKom (Steininger et al., 2002) dataset is a multimodal corpus consisting of Wizard-Of-Oz dialogues in German and English from 70 subjects (31 males and 39 females). This dataset includes several audio tracks and two video tracks (face, side of body). The main purpose of this dataset is to conduct empirical researches on human-computer interaction in a variety of tasks and technological settings. This dataset contains several sessions, each of which has a one-person recording of about 4.5 min. All the collected 3,823 utterances were annotated with seven emotional states: neutral, joy, anger, helplessness, contemplation, surprise.

FAU-AIBO

The FAU-AIBO (Batliner et al., 2004) corpus was collected from the recordings of children interacting with the Aibo pet robot. This dataset consists of spontaneous German speech. The children were made to believe that Aibo was reacting to their orders, while the robot was effectively controlled by a human operator. This dataset were obtained from 51 children (21 males, 30 females) ranging from 10 to 13 years old. The audio was recorded by using a DAT recorder (16-bit, 16 kHz). The audio recording is automatically segmented into "tums" using a 1 s pause. Five annotators were asked to listen to the tums in order and label each word individually as neutral (default) or the other ten categories. For annotation, the majority voting (MV) was employed. Finally, the utterance number for MV is 4,525, and contains 10 affective states: happy, surprise, stressed, helplessness, sensitivity, irritation, anger, mother, boredom, and condemnation.

EMO-DB

The Berlin emotional speech database (EMO-DB) (Burkhardt et al., 2005), covers seven emotional states: anger, boredom, disgust, fear, happiness, neutral, and sadness. Verbal contents come from 10 German (5 males and 5 females) pre-defined neutral utterances. Ten professional actors were invited to speak each utterance in all seven emotional states. EMO-DB consists of approximately 535 sentences from seven emotions. The audio files were recorded with a sampling rate of 16 kHz and a 16-bit resolution and mono channel. The duration for all audio files are average 3 s.

MASC

The Mandarin affective speech corpus (MASC) (Wu et al., 2006) consists of 68 native speakers (23 women, 45 man) and five affective states: neutral, anger, pride, panic and sadness. Each participant reads 5 phrases and 10 sentences for 3 times for every emotion, thereby yielding 25,636 utterances. These sentences involves in all the phonemes in Chinese language. The purpose of this corpus is to investigate the prosody and linguistic information of affective expressions in Chinese. Additionally, prosody feature analysis and speaker identification baseline experiments were also carried out.

eNTERFACE05

The eNTERFACE05 (Martin et al., 2006) corpus is an audio-visual video database which includes six elicited emotions: anger, disgust, fear, joy, sadness, and surprise. It is composed of 1,277 audio-visual video samples from 42 participants (8 females) with 14 different countries. Every participant was demanded to listen to six consecutive short tales, which were designed to invoke a particular feeling. Two experts were asked to determine whether the induced emotional response clearly characterizes the expected emotion.

SAL

The Belfast Sensitive Artificial Listener (SAL) (Douglas-Cowie et al., 2007) corpus is a subset of the developed HUMAINE database. The used SAL subset (Wöllmer et al., 2008) includes

TABLE 1 | A brief summary of speech emotion databases.

Corpus/References	Language	Year	Categories	Size	Speakers	Recordings	Modalities
DES/ (Engberg et al., 1997)	Danish	1997	Neutral, surprise, anger, happiness, sadness	5,200	4 (2f)	Acted	Audio
SUSAS/ (Hansen and Bou-Ghazale, 1997)	English	1997	Four states of speech under stress: neutral, angry, loud, Lombard	16,000	32 (13f)	Natural	Audio
SmartKom/ (Steininger et al., 2002)	German	2002	Neutral, joy, anger, helplessness, contemplation, surprise	3,823	70 (39f)	Natural	Audio
FAU-AIBO/ (Batliner et al., 2004)	German	2004	Anger, bored, emphatic, helpless, joyful, motherese, neutral	4,525	51 (30f)	Natural	Audio
EMO-DB/ (Burkhardt et al., 2005)	German	2005	Anger, boredom, disgust, fear, happiness, sadness, neutral	535	10 (5f)	Acted	Audio
eNTERFACE05/ (Martin et al., 2006)	English	2006	Anger, disgust, fear, happiness, sadness, surprise	1,277	42 (8f)	Elicited	Audiovisual
MASC/ (Wu et al., 2006)	Mandarin	2006	Neutral, anger, pride, panic, sadness	25,636	68 (23f)	acted	Audio
SAL/ (Douglas-Cowie et al., 2007)	English	2007	Anger, sadness, happiness, fear, neutral	1,692	4 (2f)	Natural	Audiovisual
ABC/ (Schuller et al., 2007)	German	2007	Aggressive, cheer, intoxicated, nervous, neutral, tire	431	8 (4f)	Elicited	audiovisual
CASIA/ (Zhang and Jia, 2008)	Mandarin	2008	Surprise, happiness, sadness, anger, fear, neutral	9,600	4 (2f)	Acted	Audio
VAM/ (Grimm et al., 2008)	German	2008	Dimension emotions (valence, arousal, dominance)	946	47 (32f)	Natural	audiovisual
IEMOCAP/ (Busso et al., 2008)	English	2008	Happiness, anger, sadness, frustration, neutral	1,150	10 (5f)	Elicited	Audiovisual
AVIC/ (Schuller et al., 2009a)	German	2009	Breathing, consent, garbage, hesitation, laughter	996	21 (10f)	Natural	Audiovisual
Polish/ (Staroniewicz and Majewski, 2009)	Polish	2009	Anger, sadness, happiness, fear, disgust, surprise, neutral	2,351	13 (7f)	Acted	audiovisual
IITKGPSEHSC/ (Koolagudi et al., 2011)	Hindi	2011	Happy, sad, angry, sarcastic, fear, neutral, disgust, surprise	1,200	10 (5f)	Acted	Audio
EMOVO/ (Costantini et al., 2014)	Italian	2014	disgust, fear, anger, joy, surprise, sadness	588	6 (3f)	Acted	Audiovisual
SAVEE/ (Jackson and Haq, 2014)	English	2014	Anger, sadness, fear, disgust neutral, joy, surprise	480	4 (-)	Acted	Audiovisual
AFEW/ (Dhall et al., 2015)	English	2015	Anger, disgust, fear, joy, neutral, sadness, surprise	1,645	330 (-)	Natural	Audiovisual
BAUM-1/ (Zhalehpour et al., 2016)	Turkish	2016	Happiness, anger, sadness, disgust, fear, surprise, boredom	1,222	31 (13f)	Natural	Audiovisual
MSP-IMPROV/ (Busso et al., 2017)	English	2017	Happiness, anger, sadness, neutral	8,438	12 (6f)	acted	Audiovisual
CHEAVD/ (Li et al., 2017)	Mandarin	2017	Anger, anxious, disgust, happiness, neutral, sadness, surprise, worried	2,852	238 (125f)	Natural	Audiovisual
NNIME/ (Chou et al., 2017)	Mandarin	2017	Discrete emotions (angry, happy, sad, neutral, frustration, surprise) and dimension emotions (valence, arousal, dominance)	102	44 (22f)	Acted	Multimodal
URDU/ (Latif et al., 2018a)	Urdu	2018	angry, sad, neutral, happy	400	38 (11f)	Natural	Audiovisual
RAVDESS/ (Livingstone and Russo, 2018)	English	2018	Calm, happy, sad, angry, fearful, surprise, disgust	7,356	24(12f)	Acted	Audiovisual
MSP-PODCAST/ (Lotfian and Busso, 2019)	English	2019	Discrete emotions (anger, sadness, happiness, surprise, fear, disgust, contempt and neutral) and dimension emotions (valence, arousal, dominance)	2,317	197 (87f)	Natural	Audio

25 recording sessions from 4 speakers (2 men and 2 women). The average duration of each session is 20 min. These audio-visual recordings in this dataset were collected from natural man-machine sessions developed by a SAL interaction. Four annotators were employed to continually mark the real-time data based on the Feeltrace tool (Cowie et al., 2000). These 25 recording sessions were divided into turns in terms of energy-based voice activity detection, yielding a total of 1,692 turns.

ABC

The Airplane Behavioral Corpus (ABC) (Schuller et al., 2007) is an audio-visual emotional database, which is designed for particular applications to public transportation. In order to elicit a certain emotion, a script was utilized to make the subject enter into the context of the guided storyline. The selected public transportation contains holiday flights with return flights related to serving of wrong food, tumultuous currents, falling asleep, talking to neighbors and so on. Eight gender-balanced participants between the ages of 25–48 years were invited to take part in the audio recording with the German language. After pre-segmentation by three experienced male annotators, a total of 11.5 h of video with 431 clips was collected. The mean duration of all 431 video clips is 8.4 s.

VAM

The VAM (Vera-Am-Mittag) corpus (Grimm et al., 2008) contains audio-visual transcripts collected from the German television talk show, which was recorded in unscripted and spontaneous discussions. This dataset consists of 946 utterances collected from 47 guests (15 males and 32 females) of talk show. The discussion themes were related to private problems, including friendship crises, fatherhood, or happy events. To annotate speech data, the audio recordings were segmented into the utterance-level, making each utterance include at least one phrase. A certain number of human annotators were employed for labeling data (17 annotators for half of all the data, 6 annotators for the others).

CASIA

The CASIA corpus (Zhang and Jia, 2008), developed by the institute of Automation, Chinese Academy of Science, consists of 9,600 audio files in total. This dataset contains six emotional states: happiness, sadness, anger, surprise, fear, and neutral. Four professional actors (two males and two females) were asked to simulate these emotions.

IEMOCAP

The Interactive Emotive Binary Motion Capture Database (IEMOCAP) (Busso et al., 2008) was developed by the team of speech analysis and interpretation laboratory (SAIL) from the University of Southern California (USC). This dataset contains five sessions lasting around 12 h, and 1,150 utterances in total. They were collected from 10 professional actors in dyadic sessions whose faces, heads, and hands were marked in scripted and natural verbal interaction scenarios. The actors performed chosen affective scripts and elicited five emotions (happiness, anger, sadness, frustration, and neutral states) under the designed imaginary settings.

AVIC

The Audio-Visual interest corpus (AVIC) (Schuller et al., 2009a) is an audio-visual emotional dataset designed for commercial applications. In this commercial scenario, the product demonstrator leads one of 21 subjects (10 women) by means of an English business presentation. The level of interest was annotated for each sub-speaker. In addition, the conversation content and non-verbal vocalization were also annotated in the AVIC collection. Finally, only 996 phrases with high inter-annotator agreement were obtained.

Polish

The Polish (Staroniewicz and Majewski, 2009) corpus is a spontaneous emotional speech dataset with six affective states: anger, sadness, happiness, fear, disgust, surprise and neutral. This dataset was recorded by three groups of speakers: professional actors, amateur actors and amateurs. A total of 2,351 utterances were recorded in which 1,168 with female and 1,183 with male voice. The average duration of all utterances was about 1 s. Then, 202 listeners were invited to attend the listening tests, in which 33 of them were musically educated and 27 foreigners did not know the Polish language.

IITKGP-SEHSC

The Indian Institute of Technology Kharagpur Simulated Emotional Hindi Speech Corpus (IITKGP-SEHSC) (Koolagudi et al., 2011) is an affective song and spoken corpus for the Hindi language. This dataset comprises of 10 participants (5 males, 5 females), each of which speaks 15 utterances in 10 sessions. It contains 1,200 audio files from 8 emotions: joy, sadness, anger, sarcasm, fear, neutral, disgust, surprise.

EMOVO

The EMOVO (Costantini et al., 2014) corpus is the first affective dataset for the Italian language. This dataset was established by six professional actors who speak 14 sentences to simulate seven affective states: disgust, fear, anger, joy, surprise sadness, and neutral. These utterances were recorded with specialized facilities in the Ugo Bordoni laboratory. This corpus also presents a subjective verification test based on the emotion-classification of two sentences conducted by two different groups of 24 listeners.

SAVEE

The Surrey audio-visual expression of emotion (SAVEE) (Jackson and Haq, 2014) corpus is a multimodal acted affective dataset with the British English language. It contains a total of 480 utterances with seven different emotions: neutral, happy, sad, angry, surprise, fear, and disgust. These utterances produced by four professional male actors. To keep the good quality of the affective acting, all the recordings in this dataset were verified by ten different evaluators under audio, visual and audio-visual condition. The scripts in these recordings were chosen from the conventional TIMIT corpus (Garofolo et al., 1993).

AFEW

The Acted Facial Expressions in the Wild (AFEW) is a natural audio-visual affective video corpus which is provided for emotion recognition in the wild (EmotiW) challenge. There have been various recently-developed versions of AFEW datasets (Kossaifi et al., 2017). One of the popular AFEW datasets is AFEW5.0 (Dhall et al., 2015) collected from 330 speakers in 2015. AFEW5.0 consists of seven affective states: anger, disgust, fear, joy, neutral, sadness and surprise, evaluated by 3 annotators. AFEW5.0 contains 1,645 utterances in total and is split into three parts: the training set (723 samples), the validation set (383 samples), and the testing set (539 samples).

BAUM-1

The BAUM-1 (Zhalehpour et al., 2016) audio-visual corpus is a spontaneous emotional dataset containing eight emotions (joy, anger, sadness, disgust, fear, surprise, boredom, and contempt), and four mental states (unsure, thinking, concentrating, and bothered). This dataset consists of 1,222 audio-visual samples from 31 Turkish participants (17 female, 14 males). The average duration of the whole samples is about 3 s. Five annotators were invited to label each sample by means of a majority voting.

MSP-IMPROV

The MSP-IMPROV (Busso et al., 2017) acted database is an audio-visual affective dataset that records the English interaction of 12 actors (6 males, 6 females) in binary conversations. Each conversation is manually split into speech turns. It consists of 8,438 emotion sentences over 9 h from four emotions: happiness, anger, sadness, and neutral. At least 50,000 evaluators were recruited by using crowdsourcing to annotate these emotional contents. The audio recording rate was 48 kHz.

CHEAVD

The CASIA Natural Emotion Audiovisual Data (CHEAVD) (Li et al., 2017) contains 2,852 natural emotional clips with 140 min extracted from 238 speakers (113 males, 125 females). This dataset is collected from 34 films, 2 television series, and 4 other television programs. This dataset is divided into three parts: the training set (1981), validation set (243) and testing set (628). The average duration of the whole samples is 3.3 s. It consists of eight emotional categories, such as angry, happy, sad, worried, anxious, surprise, disgust, and neutral. The sampling rate of audio files is 41 kHz.

NNIME

The NTHU-NTUA Chinese Interactive Emotion Corpus (NNIME) (Chou et al., 2017) is a multimodal spontaneous emotional database, collected from 44 speakers (22 females, 22 males), involved in spontaneous dyadic spoken interactions. This dataset contains 102 dyadic interaction sessions with ∼11 h of audio-video data. These participants come from the Department of Drama at National Taiwan University of Arts. Another 49 annotators were invited to implement a rich set of emotion annotations on discrete and dimensional annotation (valence, arousal, dominance). For discrete emotions, there are six categories: angry, happy, sad, neutral, frustration, surprise. The sample rate of audio recordings is 44.1 kHz.

URDU

The URDU corpus (Latif et al., 2018a) is an unscripted and natural emotional spoken dataset with the first URDU language. It consists of 400 audio samples in four affective states (angry, happy, sad and neutral). In this dataset, the audio recordings were collected from the conversations of 38 participants (27 males and 11 females) on the Urdu television talk shows. Four different annotators were requested to make annotations for all the audio recordings based on the audio-visual condition.

RAVDESS

The RAVDESS dataset (Livingstone and Russo, 2018) is a multimodal corpus of affective speech and songs. This dataset is gender-balanced and comprises 24 specialized actors (12 males, 12 females) who produce speech and song samples in a neutral North American pronunciation. For affective speech, it consists of calm, joy, sadness, anger, fear, surprise, and disgust. For affective songs, it consists of calm, joy, sadness, anger, fear, surprise, and disgust and fear. Every expression is generated at two levels of affective intensity with an additional neutral expression. The final collection of 7,356 recordings was individually rated for 10 times on these aspects of affective validity, intensity, and genuineness. For these ratings, 247 untrained research subjects from North America were employed.

MSP-PODCAST

The MSP-PODCAST (Lotfian and Busso, 2019) natural corpus contains 2,317 utterances collected from 403 podcasts. These utterances come from 197 speakers' (110 males, 87 females) spontaneous English speech in the Creative Commons authorized recording downloaded from the audio sharing websites. These podcasts are evaluated by using crowdsourcing to be dimensional emotions (valence, arousal, dominance) and discrete emotions including anger, sadness, happiness, surprise, fear, disgust, contempt, and neutral. In total, 278 different workers are invited to evaluate these utterances. Audio recordings have a sampling rate of 8 kHz.

REVIEW OF SUPERVISED, UNSUPERVISED, AND SEMI-SUPERVISED LEARNING

In this section, we will simply review the concept and typical supervised, unsupervised, and semi-supervised learning techniques, as described below.

Supervised Learning

Supervised learning usually requires a large number of labeled samples to carefully train the model for achieving better model generalization ability (Cunningham et al., 2008). At the same time, due to the problem of dimension disaster, when processing high-dimensional data, the number of labeled samples required to train a good supervised model will further show an exponential explosion trend. This makes it difficult for traditional supervised

learning to be applied to some tasks that lack training samples. Nevertheless, supervised learning methods are usually simpler than unsupervised learning methods. Therefore, when training a supervised model, how to reduce the demand for labeled samples and improve the performance of model learning has become an important research problem (Alloghani et al., 2020).

Supervised learning can be further grouped into classification and regression. A classification problem is to deal with categorical outputs, whereas a regression problem is to process continuous outputs. The typical supervised learning methods contains ANN, SVM, HMM, GMM, random forest, Bayesian networks, decision tree, linear regression, logistic regression, and so on (Kotsiantis et al., 2007; Sen et al., 2020).

Unsupervised Learning

Unlike supervised learning with labeled data, unsupervised learning aims to extract inherent feature representations from unlabeled sample data. Therefore, unsupervised learning mainly relies on previously learned knowledge to distinguish likely classes within unlabeled sample data. As a result, unsupervised learning is very appropriate for feature learning tasks (Alloghani et al., 2020).

In general, unsupervised learning methods can be divided into three categories (Usama et al., 2019): hierarchical learning, data clustering, and dimensionality reduction. Hierarchical learning aims to learn complicated feature representations from a hierarchy of multiple linear and non-linear activation operations. Autoencoders (AEs) (Ballard, 1987; Schmidhuber, 2015) are one of the earliest unsupervised hierarchical learning algorithms. Data clustering is a well-known unsupervised learning task that concentrates on seeking hidden patterns from input unlabeled sample data in the form of clusters. Data clustering methods can be grouped into three categories (Saxena et al., 2017): hierarchical clustering, Bayesian clustering, and partitional clustering. One of the widely-used data clustering approaches is k-means clustering (Likas et al., 2003) which belongs to partitional clustering. Dimensionality reduction (also called subspace learning) aims to seek the hidden pattern of the underlying data by means of extracting intrinsic low-dimensional structure. Dimensionality reduction can be categorized into two types: linear and non-linear methods (Van Der Maaten et al., 2009). Principal component analysis (PCA) (Wold et al., 1987) and non-negative matrix factorization (NMF) (Lee and Seung, 1999) are two popular linear dimensionality reduction methods.

Semi-supervised Learning

In order to make full use of the advantages of unsupervised learning and supervised learning, semi-supervised learning aims to combine a small number of labeled data and a large number of unlabeled data for performing certain learning tasks. The main goal of semi-supervised learning is to harness unlabeled data for constructing better learning procedures. For example, for a classification problem, additional sample data without label information can be utilized to aid in the classification process for performance improvement.

Semi-supervised learning can be divided into two main types (van Engelen and Hoos, 2020): inductive and transductive methods. Inductive methods aim to construct a classification model that can be utilized to predict the label of previously unseen sample data. In this case, unlabelled data may be employed when training this classification model. The representative inductive methods (Ligthart et al., 2021) contain self-training, co-training, multi-view learning, generative models, and so on. Different from inductive methods, transductive methods do not need to build a classifier for the whole input space. The typical transductive methods are graph-based semi-supervised learning algorithms (Chong et al., 2020) in which they attempt to transfer the label information of a small set of labeled data to the remaining large unlabeled data with the aid of a graph. The popular graph-based semi-supervised learning algorithms include the graph Laplacian (Fergus et al., 2009), graph-based semi-supervised neural network models (Alam et al., 2018) like graph convolutional networks (Chen et al., 2020).

TRADITIONAL METHODS FOR CROSS-CORPUS SER

From the view of point of supervised, unsupervised, and semi-supervised learning, in this section we will introduce traditional methods for cross-corpus SER, as described below.

Supervised Learning for Traditional Methods

On supervised cross-corpus SER tasks, researchers usually combine one or more databases as training sets and testify the performance on each labeled database as a testing set in a cross-validation scheme. In early supervised cross-corpus SER, the typical hand-crafted acoustic features and conventional classifiers were employed in a supervised learning manner. For instance, in Schuller et al. (2010b), they extracted 93 LLD features such as prosody, voice quality and articulatory features and performed speaker-corpus normalization so as to deal with the differences among corpora. Then, the linear SVM was used to conduct cross-corpus evaluation experiments. They adopted different combinations of training and testing sets on all used labeled databases for cross-corpus experiments. In Feraru et al. (2015), 1,941 LLD acoustic features like prosody, voice quality and spectral features were derived, then the linear SVM was employed for cross-corpus SER. A post-processing of the trained SVM models was performed by rule-based model inversion to cope with the difference among corpora. For cross-corpus experiments, they trained and tested each used labeled database against each. Based on the extracted INTERSPEECH-2010 Paralinguistic Challenge feature set with 1,582 LLDs, a new method of transfer non-negative matrix factorization (TNMF) (Song et al., 2016b), in which the non-negative matrix factorization (NMF) and the maximum mean discrepancy (MMD) algorithms were combined, was developed for cross-corpus SER. They also trained and tested each other for all used labeled database. They showed that the performance of the proposed TNMF was much better than the baseline method with the linear SVM. A domain adaptation based approach,

named emotion-discriminative and domain-invariant feature learning method (EDFLM) (Mao et al., 2017), was presented for cross-corpus SER. Training and testing each other for all used labeled database was implemented. In this method, domain discrepancy was minimized, whereas emotion-discrimination was employed to produce emotion-discriminative and domain-invariant features, followed by the linear SVM for SER. They extracted the INTERSPEECH-2009 Emotion Challenge feature set as inputs of EDFLM. In Kaya and Karpov (2018), they provided a cascaded normalization method, integrating linear speaker level, non-linear value level and feature vector level normalization, and then employed an extreme learning machine (ELM) classifier for cross-corpus SER. Here, they extracted the ComParE feature set with 6,373 LLDs. They conducted cross-corpus experiments in two settings: single corpus training (one-vs.-one), and multiple corpus training via leave-one-corpus-out (LOCO) setting. A non-negative matrix factorization based transfer subspace learning method (NMFTSL) (Luo and Han, 2020), in which the knowledge of the source data could be transferred to the target data, was developed to seek a shared feature subspace for the source and target corpus on cross-corpus SER tasks. They extracted the INTERSPEECH-2010 Paralinguistic Challenge feature set and then adopted the linear SVM for cross-corpus SER. Based on all the used databases, they constructed 30 cross-corpus SER schemes by using multiple combinations for source and target corpus on cross-corpus SER task.

Unsupervised Learning for Traditional Methods

For unsupervised cross-corpus SER tasks, researchers tried to investigate how agglomeration of unlabeled data. For instance, in Zhang et al. (2011) they extracted 39 functionals of 56 acoustic LLDs, yielding 6,552 features in total, and then employed the linear SVM to conduct a cross-corpus LOCO strategy for experiments. To evaluate the effectiveness of normalization techniques before data agglomeration, they investigated the performance of centering, normalization and standardization for per corpus normalization. Experiment results on multiple databases showed that adding unlabelled emotional samples to agglomerated multi-corpus training sets could improve SER recognition performance. To mitigate the different feature distributions between the source and target speech signals, a domain-adaptive subspace learning (DoSL) approach (Liu et al., 2018) was presented to learn a project matrix for yielding similar feature distributions. They utilized the INTERSPEECH-2009 feature set with 384 features and adopted the linear SVM for cross-corpus LOCO SER experiments. Likewise, to reduce the disparity of source and target feature distributions, a transfer subspace learning (TRaSL) (Liu et al., 2021) was also proposed for cross-corpus SER. The proposed TRaSL aimed to find a projection matrix which transformed the source and target speech signals into a common feature subspace. Finally, they adopted the INTERSPEECH-2009 feature set and the linear SVM for cross-corpus LOCO SER experiments.

Semi-supervised Learning for Traditional Methods

For semi-supervised cross-corpus SER, some recent literatures have focused on the combination of unlabeled and labeled sample data for performance improvement. In particular, a new transfer learning technique, namely transfer semi-supervised linear discriminant analysis (TSDA) (Song et al., 2016a), was provided to produce corpus-invariant discriminative feature representations on cross-corpus SER tasks. They obtained the INTERSPEECH-2010 Paralinguistic Challenge feature set, and then performed cross-corpus SER with the linear SVM. They conducted cross-corpus experiments with a LOCO scheme, and showed that TSDA outperformed other methods. A semi-supervised adaptation regularized transfer non-negative matrix factorization (SATNMF) (Luo and Han, 2019) was presented to extract common features for cross-corpus SER. The proposed SATNMF method aimed to integrate the label information of training data with NMF, and found a latent low-rank feature space to minimize simultaneously the marginal and conditional distribution differences among several language datasets. They employed the ComParE feature set and the linear SVM for LOCO SER experiments.

In summary, **Table 2** presents a summary of the above-mentioned supervised, unsupervised, and semi-supervised learning literatures for traditional methods on cross-corpus SER tasks.

DEEP LEARNING METHODS FOR CROSS-CORPUS SER

From the view of point of supervised, unsupervised, and semi-supervised learning, in this section we will introduce deep learning methods for cross-corpus SER, as described below.

Supervised Learning for Deep Learning Methods

For supervised cross-corpus SER with labeled databases, the typical CNN, LSTM, DBN, and its combinations in a hybrid way, associated with the transfer learning strategy, have been recently adopted. Specially, in Marczewski et al. (2017), to alleviate the different distributions of features and labels across domains, they proposed a deep learning network architecture composed of two uni-dimensional convolutional layers, one LSTM layer, and two FC layers for cross-corpus SER. The used CNN layers aimed to derive spatial features of varying abstract levels, whereas the LSTM layer was used to learn temporal information related to emotion evolution over time. In this case, they jointly exploited CNNs to extract domain-shared features and LSTMs to identify emotions with domain specific features. All the samples data from all databases were used for training and testing by using a 5-fold cross validation scheme. Experiments showed that they could learn transferable features to enable model adaptation from multiple source domains. In Latif et al. (2018b), considering the fact that DBNs have a strong generalization power, this study presented a transfer learning technique based on DBNs to improve the performance of SER in cross-language and

TABLE 2 | A brief summary of traditional cross-corpus SER literatures.

References	Category	Input features	Methods for cross-corpus	Datasets
Schuller et al. (2010b)	Supervised	93 LLDs	speaker-corpus normalization	DES/, EMO-DB, SUSAS, AVIC, SmartKom, eNTERFACE05
Feraru et al. (2015)	Supervised	1,941 LLDs	rule-based model inversion	EMO-DB, DES, eNTERFACE05
Song et al. (2016b)	Supervised	INTERSPEECH-2010	TNMF	FAU-AIBO, eNTERFACE05, EMO-DB
Mao et al. (2017)	Supervised	INTERSPEECH-2009	EDFLM	ABC, EMO-DB, FAU-AIBO
Kaya and Karpov (2018)	Supervised	ComParE	cascaded normalization	EMO-DB, DES, eNTERFACE05
Luo and Han (2020)	Supervised	INTERSPEECH-2010	NMFTSL	CASIA, SAVEE, EMO-DB, IEMOCAP, eNTERFACE05
Zhang et al. (2011)	Unsupervised	6,552 LLDs	corpus normalization	ABC, AVIC, DES, VAM, SAL, eNTERFACE05
Liu et al. (2018)	Unsupervised	INTERSPEECH-2009	DoSL	EMO-DB, eNTERFACE05
Liu et al. (2021)	Unsupervised	INTERSPEECH-2009	TRaSL	EMO-DB,eNTERFACE05, IEMOCAP
Song et al. (2016a)	Semi-supervised	INTERSPEECH-2010	TSDA	EMO-DB, eNTERFACE05
Luo and Han (2019)	Semi-supervised	ComParE	SATNMF	CASIA, EMO-DB, eNTERFACE05

cross-corpus scenarios. The used DBNs consisted of three RBM layers, in which the first two RBMs contained 1,000 hidden neurons, and the third RBM included 2,000 hidden neurons. The simple variant (eGeMAPS) of typical GeMAPS feature set, including 88 LLDs like pitch, energy, spectral, and so on, was employed as inputs of DBNs. For all used databases, a LOCO scheme was used for cross-corpus SER experiments. Experiment result demonstrated that DBNs provided better performance on cross-corpus SER tasks, compared with a SAE and the linear SVM. In Parry et al. (2019), after extracting 40 Mel filterbank coefficients, they presented a comparative analysis of the generalization capability of deep learning models like CNNs, LSTMs, and CNN-LSTM. The used CNNs were composed of one-dimension convolutional layer, and one max-pooling layer. The used LSTMs were two-layer bi-directional LSTMs. The used CNN-LSTM contained three CNNs and two LSTMs above-mentioned. This study indicated that the CNN and CNN-LSTM models gave very close performance, but better than LSTM. For cross-corpus experiments, all corpora were combined together, thereby producing 11 h 45 min for training, 1 h 30 min each for validation and testing. In Rehman et al. (2020), to develop a more adaptable SER in adversarial conditions, they presented a hybrid neural network framework for cross-corpus SER. The hybrid neural network consisted of two-layer LSTMs and a ramification layer. LSTMs aimed to learn temporal sequence data in the one-hot input matrices, yielded by the latter ramification layer. The ramification layer comprised of multiple embedding units and split the input MFCCs into subsequent one-hot output. They validated the performance of different methods by means of training deep models on two of the used databases and then testing it on the third database. Experiments showed the effectiveness of the proposed method on cross-corpus SER tasks.

Unsupervised Learning for Deep Learning Methods

For unsupervised cross-corpus SER tasks by leveraging unlabeled data, the popular unsupervised autoendcoder (Ballard, 1987; Schmidhuber, 2015) and its variants have been widely employed. For instance, to address the discrepancy between training and testing data, an adaptive denoising autoencoder (A-DAE) based an unsupervised domain adaptation approach (Deng et al., 2014b) was developed for cross-corpus SER. In this method, the prior knowledge learned from a target set was utilized to regularize the training on a source set. When obtaining the INTERSPEECH-2009 Emotion Challenge feature set, A-DAE was employed to learn a common representation across training and test samples, followed by the linear SVM for cross-corpus SER. They conducted cross-corpus SER experiments by using a LOCO corpus scheme. In Deng et al. (2017), an end-to-end domain adaptation method, named universum autoencoder (U-AE), which retained feature representation ability to discover the intrinsic structure in input data, was presented for cross-corpus SER. The proposed U-AE aimed to enable the unsupervised learning autoencoder to have supervised learning ability, thereby improving the performance of cross-corpus LOCO SER. The standard INTERSPEECH-2009 Emotion Challenge feature set was employed as inputs of the proposed U-AE. This study indicated that the proposed U-AE outperformed other domain adaptation methods such as kernel mean matching (Gretton et al., 2009), and shared-hidden-layer autoencoders (Deng et al., 2014a). In Neumann and Vu (2019), they investigated how unsupervised representation learning on additional unlabeled data could be used to promote SER performance. More specially, they integrated feature representations learnt by using an unsupervised autoencoder into an attentive CNN-based emotion classifier so as to improve recognition performance on cross-corpus LOCO SER tasks. In detail, they firstly trained a recurrent sequence-to-sequence autoencoder on unlabeled data and then adopted it to produce feature representations for labeled target data. These produced feature representations were then incorporated as additional source information for emotion identification during the training process of the used attentive CNN.

In recent years, several advanced unsupervised learning methods such as adversarial learning (Goodfellow et al.,

2014) and attentive learning have also been used for cross-corpus SER. Specially, in Abdelwahab and Busso (2018), a domain adversarial neural network (DANN), consisting of three parts: a feature representation layer, a task classification layer, and a domain classification layer, was employed to learn a common feature representation between training and testing data. DANN was trained by using labeled sample data in the source domain and unlabeled sample data in the target domain. The extracted acoustic features were the ComParE feature set as inputs of DANN. They conducted cross-corpus experiments by using single corpus training (one-vs.-one), and multiple corpus training via a LOCO scheme. This study demonstrated that adversarial training on the basis of unlabeled training data yielded an obvious performance improvement compared with training with the source data. In Ocquaye et al. (2021), a deep learning framework including three attentive asymmetric CNNs was presented to emotion identification for cross-lingual and cross-corpus speech signals in an unsupervised manner. They implemented cross-corpus SER experiments by using a LOCO corpus scheme. The proposed approach employed jointly supervised learning incorporated with softmax loss and center loss in order to learn high-level discriminative feature representations for target domain data with the aid of pseudo-labeled data. Evaluation results indicated that the proposed method outperformed a SAE and DBNs with three RBMs.

Semi-supervised Learning for Deep Learning Methods

For semi-supervised cross-corpus SER by leveraging unlabeled and labeled data, adversarial learning (Goodfellow et al., 2014) was usually taken as a generative model for. For instance, in Chang and Scherer (2017), they explored a semi-supervised learning approach, called a multitask deep convolutional generative adversarial network (DCGAN), to improve cross-corpus performance. DCGAN was utilized to learn strong feature representation from the computed spectrograms on unlabeled data. For multitask learning, the proposed multitask model took emotional valence as a primary target and emotional activation as a secondary target. For evaluation, they combined unlabeled data from all used databases and testified the performance on one labeled database. Experiment results found that unsupervised learning presented significant improvements for cross-corpus SER. In Deng et al. (2018), to take advantage of the available unlabeled speech data, they proposed a semi-supervised autoencoder to improve the performance of cross-corpus SER. The proposed method extended a typical unsupervised autoencoder by means of adjoining the supervised learning objective of a deep feed forward network. The extracted acoustic features were the INTERSPEECH-2009 Emotion Challenge feature set. Cross-corpus experiments were implemented by using multiple corpus training via a LOCO scheme. Experimental results showed that

TABLE 3 | A brief summary of existing deep cross-corpus SER literatures.

References	Category	Input features	Methods for cross-corpus	Datasets
Marczewski et al. (2017)	Supervised	54,000 dimensional data points	CNN, LSTM	AFEW, EMO-DB, EMOVO, eNTERFACE05, IEMOCAP
Latif et al. (2018b)	Supervised	eGeMAPS	DBNs	FAU-AIBO, IEMOCAP, EMO-DB, SAVEE, EMOVO
Parry et al. (2019)	Supervised	Mel filterbank coefficients	CNN, LSTM, CNN-LSTM	IEMOCAP, EMOVO, EMO-DB, RAVDESS, SAVEE
Rehman et al. (2020)	Supervised	13 MFCCs	LSTMs, a ramification layer	IEMOCAP, RAVDESS, EMO-DB
Deng et al. (2014b)	Unsupervised	INTERSPEECH-2009	A-DAE	FAU-AIBO, ABC, SUSAS
Deng et al. (2017)	Unsupervised	INTERSPEECH-2009	U-AE	ABC, EMO-DB, SUSAS
Abdelwahab and Busso (2018)	Unsupervised	INTERSPEECH-2013	DANN	IEMOCAP, MSP-IMPROV, MSP-PODCAST
Neumann and Vu (2019)	Unsupervised	128 Mel frequency bands	unsupervised autoencoder and ACNN	IEMOCAP, MSP-IMPROV
Ocquaye et al. (2021)	Unsupervised	spectrogram	three attentive asymmetric CNNs	SAVEE, IEMOCAP, EMO-DB,FAU-AIBO, EMOVO
Chang and Scherer (2017)	Semi-supervised	spectrogram	DCGAN	AMI, IEMOCAP
Deng et al. (2018)	Semi-supervised	INTERSPEECH-2009	Unsupervised autoencoder	FAU-AIBO, ABC, EMO-DB, SUSAS
Gideon et al. (2019)	Semi-supervised	40 dimensional Mel-filter banks	ADDoG	IEMOCAP, MSP-IMPROV
Latif et al. (2020)	Semi-supervised	spectrogram	AAE	IEMOCAP, MSP-IMPROV
Parthasarathy and Busso (2020)	Semi-supervised	INTERSPEECH-2013	ladder network	MSP-PODCAST, IEMOCAP, MSP-IMPROV

the proposed approach obtained promising performance with a very small number of labeled data. In Gideon et al. (2019), the extracted 40 dimensional Mel-filter banks were passed into an adversarial discriminative domain generalization (ADDoG) algorithm to learn more generalized feature representations for cross-corpus SER. Based on the idea of GANs (Goodfellow et al., 2014), ADDoG could make full use of the unlabeled test data to generalize the intermediate feature representation across different datasets. They combined multiple corpora for training and testified the performance of different methods on other corpora via a LOCO scheme. Experiment results showed that ADDoG performed better than CNNs. In Latif et al. (2020), a multi-task semi-supervised adversarial autoencoding (AAE) method was provided for cross-corpus SER. The proposed AAE was a two-step approach. First, semi-supervised learning was implemented in an adversarial autoencoder to generate latent representation. Then, a multi-task learning framework, which considered emotion, speaker and gender identification as auxiliary tasks incorporating with semi-supervised adversarial autoencoding, was built to improve the performance of primary SER task. The spectrograms achieved by a short time Fourier transform (STFT) were employed as inputs of the proposed AAE. They performed cross-corpus experiments with a LOCO scheme on all the used databases. Experiment results demonstrated that the proposed AAE outperformed CNN, CNN+LSTM, as well as DBN.

In recent years, researchers explored ladder network (Valpola, 2015) based semi-supervised methods (Huang et al., 2018; Tao et al., 2019; Parthasarathy and Busso, 2020) for cross-corpus SER and had shown superior results to supervised methods. Here, a ladder network is regarded as an unsupervised DAE trained along with a supervised classification or regression problem. For instance, in Parthasarathy and Busso (2020), a ladder network based semi-supervised method, incorporating with an unsupervised auxiliary task, was presented to reduce the diversity between the source and target domains on cross-corpus SER tasks. The primary task aimed to predict dimensional emotional attributes. The auxiliary task aimed to produce the reconstruction of intermediate feature representations with a DAE. This auxiliary task was trained on a large amount unlabeled data from the target domain in a semi-supervised manner. The ComParE feature set was fed into the ladder network. They conducted cross-corpus experiments with a LOCO scheme. This study indicated that the proposed method achieved superior performance to fully supervised single-task learning (STL) and multi-task learning (MTL) baselines.

In summary, **Table 3** presents a summary of the above-mentioned supervised, unsupervised and semi-supervised learning literatures for deep learning methods on cross-corpus SER tasks.

OPEN CHALLENGES

Although deep learning based cross-corpus SER has made great progress in recent years as mentioned above, there exist still several open challenges that should be addressed in future. In the following, we will discuss these open challenges, and show its potential trends.

One of the most important problems for cross-corpus SER is the generation of natural emotional speech data (Cao et al., 2015). As shown in **Table 1**, we can see that the majority of emotional databases for cross-corpus SER are acted and recorded in specific silent labs. However, in the real-world sceneries, the collected emotional speech data is usually noisy. In addition, there are also legal and ethical issues when recording the true natural speech emotions. Most existing utterances from natural datasets are collected from talk-shows, call-center recordings, and similar conditions in which the involved participants are informed of the recording. In this case, these natural datasets do not include all emotion categories and may not reflect the true emotions that are felt. Moreover, there is a scarcity for speech emotional datasets in numbers. Considering that deep cross-corpus SER is a data-driven task based on deep learning models with high hyper-parameters, a great number of training data is needed for training sufficiently deep models. Hence, another main challenge for deep cross-corpus SER is the scarcity of enough large emotional datasets.

The second challenge is to integrate more modalities characterized by human emotion expression for cross-corpus emotion recognition (Tzirakis et al., 2021). It is well-known that the typical bimodalities (audio, visual) (Zhang et al., 2017; Zhou et al., 2021), triple modalities (audio, visual, text) (Shoumy et al., 2020), user's physiological responses like electroencephalogram (EEG) and electrocardiogram (ECG) signals (Katsigiannis and Ramzan, 2017; Li et al., 2021), and so on, are highly correlated with human emotion expression. To further improve emotion recognition, it is thus interesting to combine speech clues with other modalities such as visual, text, and physiological clues for multimodal cross-corpus.

Another challenge is the inherent limitation of deep learning techniques. First, although various deep leaning techniques have been successfully employed to capture high-level feature representations for cross-corpus SER, most of deep learning techniques have a large number of network parameters. This makes deep learning techniques usually have very large computation complexity, resulting in its training which demands for large data. To alleviate this problem, it is a promising direction to investigate the application of deep compression and acceleration (Han et al., 2016; Choudhary et al., 2020) techniques such as pruning, trained quantization, and so on, for real-world cross-corpus SER. Additionally, deep learning is a the black-box technique. In particular, due to the used multilayer non-linear architecture, deep learning algorithms are frequently criticized to be non-transparent, and non-explainable. Therefore, it is also a promising research subject to investigate how to understand the explainability and interpretability of deep learning techniques (Fellous et al., 2019; Langer et al., 2021) for cross-corpus SER. In addition, it is also interesting to investigate the performance of recently-developed transformer (Vaswani et al., 2017; Lian et al.,

2021) method incorporating with deep learning techniques for cross-corpus SER in our future work.

CONCLUSIONS

This paper has presented an up-to-date and comprehensive review of cross-corpus SER techniques, exhibiting recent advances and perspectives in this area. It has summarized the related speech emotional databases and the applications of deep leaning techniques associated with supervised, unsupervised, semi-supervised learning for cross-corpus SER in recent years. In addition, it highlights several challenging research directions to further improve the performance of cross-corpus SER in future.

AUTHOR CONTRIBUTIONS

SZ contributed to the writing and drafted this article. RL and XT contributed to the collection and analysis of existing literatures. XZ contributed to the conception and design of this work and revised this article. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Zhejiang Provincial National Science Foundation of China and National Science Foundation of China (NSFC) under Grant Nos. LZ20F020002, LQ21F020002, and 61976149.

REFERENCES

Abdelwahab, M., and Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 2423–2435. doi: 10.1109/TASLP.2018.2867099

Akçay, M. B., and Oguz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* 116, 56–76. doi: 10.1016/j.specom.2019.12.001

Alam, F., Joty, S., and Imran, M. (2018). "Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets," in *Twelfth International AAAI Conference on Web and Social Media*. (Palo Alto, CA), 556–559.

Albornoz, E. M., Milone, D. H., and Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Comput. Speech Lang.* 25, 556–570. doi: 10.1016/j.csl.2010.10.001

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., and Aljaaf, A. J. (2020). "A systematic review on supervised and unsupervised machine learning algorithms for data science," in *Supervised unsupervised Learn Data Sci.* 3–21. doi: 10.1007/978-3-030-22475-2_1

Anagnostopoulos, C.-N., Iliou, T., and Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artif. Intell. Rev.* 43, 155–177. doi: 10.1007/s10462-012-9368-5

Ballard, D. H. (1987). "Modular learning in neural networks," in *AAAI* (Seattle, WA), 279–284.

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636. doi: 10.1037/0022-3514.70.3.614

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. J., et al. (2004). "You Stupid Tin Box"-Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus," in *Lrec*. 171–174.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). "A database of German emotional speech," in *Ninth European Conference on Speech Communication and Technology* (Lisbon). doi: 10.21437/Interspeech.2005-446

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2017). MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput.* 8, 67–80. doi: 10.1109/TAFFC.2016.2515617

Cao, H., Verma, R., and Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: studies on acted and spontaneous speech. *Comput. Speech Lang.* 29, 186–202. doi: 10.1016/j.csl.2014.01.003

Chang, J., and Scherer, S. (2017). "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA), 2746–2750. doi: 10.1109/ICASSP.2017.7952656

Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. (2020). "Simple and deep graph convolutional networks," in: *International Conference on Machine Learning* (Long Beach, CA), 1725–1735.

Chong, Y., Ding, Y., Yan, Q., and Pan, S. (2020). Graph-based semi-supervised learning: a review. *Neurocomputing* 408, 216–230. doi: 10.1016/j.neucom.2019.12.130

Chou, H., Lin, W., Chang, L., Li, C., Ma, H., and Lee, C. (2017). "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII))* (San Antonio, TX), 292–298. doi: 10.1109/ACII.2017.8273615

Choudhary, T., Mishra, V., Goswami, A., and Sarangapani, J. (2020). A comprehensive survey on model compression and acceleration. *Artif. Intell. Rev.* 53, 5113–5155. doi: 10.1007/s10462-020-09816-7

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). "EMOVO corpus: an Italian emotional speech database," in *International Conference on Language Resources and Evaluation (LREC 2014)*: European Language Resources Association *(ELRA))* 3501–3504.

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., and Schröder, M. (2000). "FEELTRACE': an instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (Northern Ireland).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* 18, 32–80. doi: 10.1109/79.911197

Cunningham, P., Cord, M., and Delany, S. J. (2008). "Supervised learning," in *Machine Learning Techniques for Multimedia* eds. Cord, M. and Cunningham, P (Berlin, Heidelberg: Springer; Cognitive Technologies), p. 21–49. doi: 10.1007/978-3-540-75171-7_2

Dellaert, F., Polzin, T., and Waibel, A. (1996). "Recognizing emotion in speech," in: *4th International Conference on Spoken Language Processing (ICSLP'96)* (Philadelphia, PA: ISCA), p. 1970–3. doi: 10.1109/ICSLP.1996.608022

Deng, J., Xia, R., Zhang, Z., Liu, Y., and Schuller, B. (2014a). "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). p. 4818–22. doi: 10.1109/ICASSP.2014.6854517

Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. (2017). Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* 24, 500–504. doi: 10.1109/LSP.2017.2672753

Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. (2018). Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 31–43. doi: 10.1109/TASLP.2017.2759338

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014b). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* 21, 1068–1072. doi: 10.1109/LSP.2014.2324759

Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., and Gedeon, T. (2015). "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. (Seattle, WA), p. 423–426.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., et al. (2007). "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *International Conference on Affective Computing and Intelligent Interaction* (Lisbon: Springer), p. 488–500. doi: 10.1007/978-3-540-74889-2_43

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognit.* 44, 572–587. doi: 10.1016/j.patcog.2010.09.020

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). "Design, recording and verification of a Danish emotional speech database," in *Fifth European Conference on Speech Communication and Technology*) (Rhodes).

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., and Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: behavioral neurostimulation. *Front. Neurosci.* 13:1346. doi: 10.3389/fnins.2019.01346

Feng, K., and Chaspari, T. (2020). A review of generalizable transfer learning in automatic emotion recognition. *Front. Comput. Sci.* 2:9. doi: 10.3389/fcomp.2020.00009

Feraru, S. M., Schuller, D., and Schuller, B. (2015). "Cross-language acoustic emotion recognition: an overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction* (Xi'an: ACII), p. 125–131. doi: 10.1109/ACII.2015.7344561

Fergus, R., Weiss, Y., and Torralba, A. (2009). "Semi-Supervised Learning in Gigantic Image Collections," in *NIPS* (Vancouver, BC: Citeseer), p. 1–9.

Fragopanagos, N., and Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Netw.* 18, 389–405. doi: 10.1016/j.neunet.2005.03.006

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. J. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. *NIST Speech Disc.* 93:27403. doi: 10.6028/NIST.IR.4930

Gideon, J., McInnis, M., and Provost, E. M. (2019). Improving cross-corpus speech emotion recognition with Adversarial Discriminative Domain Generalization (ADDoG). *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2019.2916092. [Epub ahead of print].

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems*. (Montreal, QC).

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift Mach. Learn.* 3, 131–160. doi: 10.7551/mitpress/9780262170055.003.0008

Grimm, M., Kroschel, K., and Narayanan, S. (2008). "The Vera am Mittag German audio-visual emotional speech database," in *2008 IEEE International Conference on Multimedia and Expo* (Hannover: IEEE), p. 865–868. doi: 10.1109/ICME.2008.4607572

Han, S., Mao, H., and Dally, W. J. (2016). "Deep compression: compressing deep neural network with pruning, trained quantization and huffman coding," in *International Conference on Learning Representations (ICLR)* (Vancouver, BC, Canada).

Hansen, J. H., and Bou-Ghazale, S. E. (1997). "Getting started with SUSAS: A speech under simulated and actual stress database," in *Fifth European Conference on Speech Communication and Technology*.

Harati, S., Crowell, A., Mayberg, H., and Nemati, S. (2018). "Depression severity classification from speech emotion," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI: IEEE), p. 5763–5766. doi: 10.1109/EMBC.2018.8513610

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, J., Li, Y., Tao, J., Lian, Z., Niu, M., and Yi, J. (2018). "Speech emotion recognition using semi-supervised learning with ladder networks," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (Beijing), 1–5. doi: 10.1109/ACIIAsia.2018.8470363

Jackson, P., and Haq, S. (2014). *Surrey Audio-Visual Expressed Emotion (savee) Database*. Guildford: University of Surrey.

Katsigiannis, S., and Ramzan, N. (2017). DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* 22, 98–107. doi: 10.1109/JBHI.2017.2688239

Kaya, H., and Karpov, A. A. (2018). Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* 275, 1028–1034. doi: 10.1016/j.neucom.2017.09.049

Koolagudi, S. G., Reddy, R., Yadav, J., and Rao, K. S. (2011). "IITKGP-SEHSC: Hindi speech corpus for emotion analysis," in *2011 International Conference on Devices and Communications (ICDeCom)* (IEEE). p. 1–5. doi: 10.1109/ICDECOM.2011.5738540

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* 65, 23–36. doi: 10.1016/j.imavis.2017.02.001

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: a review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* 160, 3–24. doi: 10.1007/s10462-007-9052-3

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 1097–1105.

Kwon, O., Chan, K., Hao, J., and Lee, T. (2003). "Emotion recognition by speech signals," in *EUROSPEECH-2003* (ISCA). p. 125–128.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., et al. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* 296, 103473. doi: 10.1016/j.artint.2021.103473

Latif, S., Qayyum, A., Usman, M., and Qadir, J. (2018a). "Cross lingual speech emotion recognition: Urdu vs. western languages," in *2018 International Conference on Frontiers of Information Technology (FIT)* (IEEE). p. 88–93. doi: 10.1109/FIT.2018.00023

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., and Schuller, B. W. (2020). Multi-Task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Trans. Affect. Comput.* 1–1. doi: 10.1109/TAFFC.2020.2983669

Latif, S., Rana, R., Younis, S., Qadir, J., and Epps, J. (2018b). Transfer learning for improving speech emotion classification accuracy. *arXiv preprint arXiv* 1801.06353. doi: 10.21437/Interspeech.2018-1625

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, W., Huan, W., Hou, B., Tian, Y., Zhang, Z., and Song, A. (2021). Can emotion be transferred?–A review on transfer learning for EEG-Based Emotion Recognition. *IEEE Trans. Cogn. Dev. Syst.* doi: 10.1109/TCDS.2021.3098842

Li, W., Zhang, Y., and Fu, Y. (2007). "Speech emotion recognition in e-learning system based on affective computing," in *Third International Conference on Natural Computation (ICNC-2007)* (Haikou: IEEE), 809–813. doi: 10.1109/ICNC.2007.677

Li, Y., Tao, J., Chao, L., Bao, W., and Liu, Y. (2017). CHEAVD: a Chinese natural emotional audio–visual database. *J. Ambient Intell. Humaniz. Comput.* 8, 913–924. doi: 10.1007/s12652-016-0406-z

Lian, Z., Liu, B., and Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29, 985–1000. doi: 10.1109/TASLP.2021.3049898

Ligthart, A., Catal, C., and Tekinerdogan, B. (2021). Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. *Appl. Soft Comput.* 101:107023. doi: 10.1016/j.asoc.2020.107023

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognit.* 36, 451–461. doi: 10.1016/S0031-3203(02)00060-2

Lin, Y., Gau, S. S., and Lee, C. (2020). A multimodal interlocutor-modulated attentional BLSTM for classifying autism subgroups during clinical interviews.

IEEE J. Sel. Top. Signal Process. 14, 299–311. doi: 10.1109/JSTSP.2020.2970578

Liu, N., Zhang, B., Liu, B., Shi, J., Yang, L., Li, Z., et al. (2021). Transfer subspace learning for unsupervised cross-corpus speech emotion recognition. *IEEE Access* 9, 95925–95937. doi: 10.1109/ACCESS.2021.3094355

Liu, N., Zong, Y., Zhang, B., Liu, L., Chen, J., Zhao, G., et al. (2018). "Unsupervised Cross-Corpus Speech Emotion Recognition Using Domain-Adaptive Subspace Learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5144–5148. doi: 10.1109/ICASSP.2018.8461848

Livingstone, S. R., and Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* 13:e0196391. doi: 10.1371/journal.pone.0196391

Lotfian, R., and Busso, C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.* 10, 471–483. doi: 10.1109/TAFFC.2017.2736999

Luengo, I., Navas, E., and Hernáez, I. (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Trans. Multimedia* 12, 490–501. doi: 10.1109/TMM.2010.2051872

Luo, H., and Han, J. (2019). "Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization," in *INTERSPEECH*. 3247–3251. doi: 10.21437/Interspeech.2019-2041

Luo, H., and Han, J. (2020). Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 28, 2047–2060. doi: 10.1109/TASLP.2020.3006331

Mao, Q., Xu, G., Xue, W., Gou, J., and Zhan, Y. (2017). Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Commun.* 93, 1–10. doi: 10.1016/j.specom.2017.06.006

Marczewski, A., Veloso, A., and Ziviani, N. (2017). "Learning transferable features for speech emotion recognition," in *Proceedings of the on Thematic Workshops of ACM Multimedia* (Mountain View, CA), 529–536. doi: 10.1145/3126686.3126735

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)* (IEEE). p. 8-8. doi: 10.1109/ICDEW.2006.145

Morrison, D., Wang, R., and De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun.* 49, 98–112. doi: 10.1016/j.specom.2006.11.004

Neumann, M., and Vu, N. T. (2019). "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton), 7390–7394. doi: 10.1109/ICASSP.2019.8682541

Nicholson, J., Takahashi, K., and Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing Appl.* 9, 290–296. doi: 10.1007/s005210070006

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Commun.* 41, 603–623. doi: 10.1016/S0167-6393(03)00099-2

Ocquaye, E. N., Mao, Q., Xue, Y., and Song, H. (2021). Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *Int. J. Intelligent Syst.* 36, 53–71. doi: 10.1002/int.22291

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transact. Neural Netw. Learn. Syst.* 32, 604–624. doi: 10.1109/TNNLS.2020.2979670

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., et al. (2019). "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Interspeech-2019* (Graz), 1656–1660. doi: 10.21437/Interspeech.2019-2753

Parthasarathy, S., and Busso, C. (2020). Semi-supervised speech emotion recognition with ladder networks. *IEEE/ACM Transact. Audio Speech Language Proc.* 28, 2697–2709. doi: 10.1109/TASLP.2020.3023632

Picard, R. W. (2010). Affective computing: from laughter to IEEE. *IEEE Transact. Affect. Computing* 1, 11–17. doi: 10.1109/T-AFFC.2010.10

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE J. Sel. Top. Signal Process.* 13, 206–219. doi: 10.1109/JSTSP.2019.2908700

Ramakrishnan, S., and El Emary, I. M. (2013). Speech emotion recognition approaches in human computer interaction. *Telecommun. Syst.* 52, 1467–1478. doi: 10.1007/s11235-011-9624-z

Rehman, A., Liu, Z. T., Li, D. Y., and Wu, B. H. (2020). "Cross-corpus speech emotion recognition based on hybrid neural networks," in *2020 39th Chinese Control Conference (CCC)* (Shenyang), 7464–7468. doi: 10.23919/CCC50068.2020.9189368

Samani, H. A., and Saadatian, E. (2012). A multidisciplinary artificial intelligence model of an affective robot. *Int. J. Advanced Robotic Syst.* 9, 1–11. doi: 10.5772/45662

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., et al. (2017). A review of clustering techniques and developments. *Neurocomputing* 267, 664–681. doi: 10.1016/j.neucom.2017.06.053

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schuller, B., Arsic, D., Rigoll, G., Wimmer, M., and Radig, B. (2007). "Audiovisual behavior modeling by combined feature spaces," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Honolulu, HI: IEEE), p. II-733-II-736. doi: 10.1109/ICASSP.2007.366340

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., et al. (2009a). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image Vision Computing* 27, 1760–1774. doi: 10.1016/j.imavis.2009.02.013

Schuller, B., Steidl, S., and Batliner, A. (2009b). "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association* (Brighton). doi: 10.21437/Interspeech.2009-103

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2010a). "The INTERSPEECH 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association* (Makuhari). doi: 10.21437/Interspeech.2010-739

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association* (Lyon). doi: 10.21437/Interspeech.2013-56

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., et al. (2010b). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transact. Affect. Computing* 1, 119–131. doi: 10.1109/T-AFFC.2010.8

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 90–99. doi: 10.1145/3129340

Sen, P. C., Hajra, M., and Ghosh, M. (2020). "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics* (Springer). p. 99–111. doi: 10.1007/978-981-13-7403-6_11

Shoumy, N. J., Ang, L.-M., Seng, K. P., Rahaman, D. M., and Zia, T. (2020). Multimodal big data affective analytics: a comprehensive survey using text, audio, visual and physiological signals. *J. Netw. Computer Applicat.* 149:102447. doi: 10.1016/j.jnca.2019.102447

Song, P., Zhang, X., Ou, S., Liu, J., Yu, Y., and Zheng, W. (2016a). "Cross-corpus speech emotion recognition using transfer semi-supervised discriminant analysis," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*), 1–5. doi: 10.1109/ISCSLP.2016.7918395

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., et al. (2016b). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Commun.* 83, 34–41. doi: 10.1016/j.specom.2016.07.010

Staroniewicz, P., and Majewski, W. (2009). "Polish emotional speech database–recording and preliminary validation," in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions* (Springer). p. 42–49. doi: 10.1007/978-3-642-03320-9_5

Steininger, S., Schiel, F., Dioubina, O., and Raubold, S. (2002). "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation* (Las Palmas), 33–37.

Tao, J.-H., Huang, J., Li, Y., Lian, Z., and Niu, M.-Y. (2019). Semi-supervised ladder networks for speech emotion recognition. *Int. J. Automation Comput.* 16, 437–448. doi: 10.1007/s11633-019-1175-x

Tzirakis, P., Chen, J., Zafeiriou, S., and Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion* 68, 46–53. doi: 10.1016/j.inffus.2020.10.011

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K.-L. A., Elkhatib, Y., et al. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE Access* 7, 65579–65615. doi: 10.1109/ACCESS.2019.2916648

Valpola, H. (2015). "From neural PCA to deep unsupervised learning," in *Advances in Independent Component Analysis and Learning Machines,* eds E. Bingham, S. Kaski, J. Laaksonen and J. Lampinen (Academic Press). p. 143–171. doi: 10.1016/B978-0-12-802806-3.00008-7

Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J. Mach. Learn. Res.* 10, 66–71.

van Engelen, J. E., and Hoos, H. H. (2020). A survey on semi-supervised learning. *Mach. Learn.* 109, 373–440. doi: 10.1007/s10994-019-05855-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Ververidis, D., and Kotropoulos, C. (2005). "Emotional speech classification using Gaussian mixture models," in *IEEE International Conference on Multimedia and Expo (ICME'05)* (Amsterdam), 2871–2874. doi: 10.1109/ISCAS.2005.1465226

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometr. Intelligent Lab. Syst.* 2, 37–52. doi: 10.1016/0169-7439(87)80084-9

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., et al. (2008). "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 Incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008* (Brisbane), 597–600. doi: 10.21437/Interspeech.2008-192

Wu, T., Yang, Y., Wu, Z., and Li, D. (2006). "Masc: a speech corpus in mandarin for emotion analysis and affective speaker recognition," in *2006 IEEE Odyssey-the Speaker and Language Recognition Workshop* (San Juan: IEEE), p. 1–5. doi: 10.1109/ODYSSEY.2006.248084

Wu, X., Sahoo, D., and Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing* 396, 39–64. doi: 10.1016/j.neucom.2020.01.085

Yildirim, S., Narayanan, S., and Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Comput. Speech Lang.* 25, 29–44. doi: 10.1016/j.csl.2009.12.004

Zhalehpour, S., Onder, O., Akhtar, Z., and Erdem, C. E. (2016). BAUM-1: a spontaneous audio-visual face database of affective and mental states. *IEEE Transact. Affect. Comput.* 8, 300–313. doi: 10.1109/TAFFC.2016.2553038

Zhang, J. T., and Jia, H. (2008). "Design of speech corpus for mandarin text to speech," in *The Blizzard Challenge 2008 Workshop* (Brisbane, QLD).

Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transact. Circuits Syst. Video Tech.* 28, 3030–3043. doi: 10.1109/TCSVT.2017.2719043

Zhang, S., and Zhao, X. (2013). Dimensionality reduction-based spoken emotion recognition. *Multimed. Tools Appl.* 63, 615–646. doi: 10.1007/s11042-011-0887-x

Zhang, S., Zhao, X., and Lei, B. (2013). Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. *Int. J. Adv. Robotic Syst.* 10, 1–7. doi: 10.5772/55403

Zhang, Z., Weninger, F., Wöllmer, M., and Schuller, B. (2011). "Unsupervised learning in cross-corpus acoustic emotion recognition," in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding* (Waikoloa, HI), 523–528. doi: 10.1109/ASRU.2011.6163986

Zhao, X., and Zhang, S. (2015). Spoken emotion recognition via locality-constrained kernel sparse representation. *Neural Comput. Appl.* 26, 735–744. doi: 10.1007/s00521-014-1755-1

Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q. F., and Lee, C. H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transact. Audio Speech Language Processing* 29, 2617–2629. doi: 10.1109/TASLP.2021.3096037

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fnbot.2021.746985

Micro-Expression Recognition Based on Pixel Residual Sum and Cropped Gaussian Pyramid

Yuan Zhao, Zhuang Chen and Song Luo*

School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China

Facial micro-expression(ME) recognition has great significance for the progress of human society and could find a person's true feelings. Meanwhile, ME recognition faces a huge challenge, since it is difficult to detect and easy to be disturbed by the environment. In this article, we propose two novel preprocessing methods based on Pixel Residual Sum. These methods can preprocess video clips according to the unit pixel displacement of images, resist environmental interference, and be easy to extract subtle facial features. Furthermore, we propose a Cropped Gaussian Pyramid with Overlapping(CGPO) module, which divides images of different resolutions through Gaussian pyramids and crops different resolutions images into multiple overlapping subplots. Then, we use a convolutional neural networks of progressively increasing channels based on the depthwise convolution to extract preliminary features. Finally, we fuse preliminary features and make position embedding to get the last features. Our experiments show that the proposed methods and model have better performance than the well-known methods.

Keywords: micro-expression recognition, deep learning, Gaussian pyramid, pixel residual sum, position embedding

OPEN ACCESS

Edited by:
Zhen Cui,
Nanjing University of Science and Technology, China

Reviewed by:
Sofiane Boucenna,
UMR8051 Equipes Traitement de l'Information et Systèmes (ETIS), France
Weisheng Li,
Chongqing University of Posts and Telecommunications, China

****Correspondence:***
Yuan Zhao
1290889111@qq.com

Received: *25 July 2021*
Accepted: *22 November 2021*
Published: *20 December 2021*

Citation:
Zhao Y, Chen Z and Luo S (2021) Micro-Expression Recognition Based on Pixel Residual Sum and Cropped Gaussian Pyramid. Front. Neurorobot. 15:746985. doi: 10.3389/fnbot.2021.746985

1. INTRODUCTION

Facial expression is a crucial channel for interpersonal socializing and can be used to convey inner emotions in daily life. Facial expression is divided into micro-expression(ME) and macro-expression. In past decades, macro-expression had a wide range of applications, and scholars have done a lot of research on macro-expression and facial recognition (Boucenna et al., 2014; Liu et al., 2018; Kim et al., 2019; Xie et al., 2019), but macro-expression is deceptive and can be easily hidden by human control. In contrast, ME will be unintentionally exposed as long as people intend to hide their true feeling. Hence, ME recognition has attracted much attention and has an extensive application prospect, such as clinical diagnosis, judiciary authorities, political elections, and national security.

ME has the following characteristics:

- ME is a very short facial expression and lasts between 1/25 and 1/3 (Yan et al., 2013). As a result, untrained individuals have a weaker ability to recognize ME (Lies, 1992).
- ME is an unconscious and involuntary facial expression appearing when people disguise one's emotions and can be triggered in high-risk environments and show real or hidden emotions.
- ME usually only appears in specific locations (Ekman and Friesen, 1971; Ekman, 2009b).
- ME usually needs to be analyzed in the video clip, and macro-expression can be analyzed in the image.

Due to these characteristics, it is difficult to recognize the ME artificially. Therefore, Ekman and Paul tried a lot of efforts to improve the ability of individuals to recognize the ME, and they developed a tool for ME recognition in 2002 Micro Expression Training Tool (METT) (Ekman, 2009a), which can effectively improve the individual's ability to recognize ME. However, the accuracy of relying on human recognition of ME is not high. According to reports, the accuracy of human-identified ME is only 47% (Frank et al., 2009). Therefore, it is particularly important to recognize the ME through computer vision. With the development of technology, the rise of high-speed cameras and deep learning has made it possible to accurately recognize the ME. However, the current ME recognition is mainly faced with the following problems.

- How to extract the subtle feature of the human face?
- How to overcome frame redundancy in the ME video?
- How to have stronger universality and overcome environmental changes?

The structure of the study is as follows: In Section II, the pieces of literature related to ME recognition are reviewed in detail; In Section III, a preprocessing method and network framework for ME recognition are proposed; In Section IV, we describe the experimental settings and analyze the experimental results; Finally, Section V summarizes this study with remarks. The contributions of this study are as follows.

- We propose two more effective methods of preprocessing, which combine spatio-temporal dimensionality and can extract more robust features.
- We design a module of Cropped Gaussian Pyramid with Overlapping(CGPO), which can use different scales information.
- We design a network with feature fusion, and the network structure adopts a gradual way of increasing channels.

2. RELATED WORK

2.1. Handcrafted Features

Several years before, ME recognition was mainly based on traditionally handcrafted feature descriptors. These descriptors can be divided into geometric features and appearance features.

2.1.1. Appearance-Based Features

For instance, Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) (Zhao and Pietikainen, 2007), Spatiotemporal Completed Local Quantization Patterns (STCLQP) (Huang et al., 2016), and LBP with Six Intersection Points (LBP-SIP) (Wang et al., 2014) can be considered as methods based on appearance features. These methods led that the features, dimensions are relatively high with more redundant information.

The LBP-TOP, a development of the LBP in a three-dimensional space, is a typical LBP descriptor with spatial-temporal characteristics. The LBP-TOP operator extracts LBP features on the three orthogonal planes. Next, obtained results are stitched as the final LBP-TOP feature, since the video can be regarded as a cube in the three dimensions of x, y, and t. The LBP-TOP not only considers the spatial information but also considers the information in the video sequence. After obtaining the LBP-TOP features, Zhao et al. use Support Vector Machine(SVM) for spotting and classification. Zhao et al. made good use of LBP-TOP features, and used many tricks of conventional expression analysis. As an early work, the work has achieved good results and has established the basis for the subsequent ME recognition.

The LBP-TOP has great limitations for only considering the local appearance and movement characteristics. So, Huang et al. (2016) proposed STCLQP for the ME recognition. First, three significant information, including magnitude, orientation, and sign components, are extracted by STCLQP. Second, for each component in temporal and appearance domains, Huang et al. (2016) made dense and characteristic codebooks by developing productive codebook selection and vector quantization. Finally, in terms of this codebook, Huang et al. (2016) extracted and fused spatio-temporal features, included orientation components, magnitude, and sign. Compared with LBP-TOP, the STCLQP method considers more information. Although the recognition accuracy is improved, it will inevitably lead to higher dimensions.

Furthermore, Wang et al. (2014) proposed LBP-SIP volumetric descriptor, which is based on three intersecting lines passing through a central point. The superabundance of LBP-TOP patterns is diminished by LBP-SIP. Furthermore, LBP-SIP provides a more dense and weightless characterization and reduces computational complexity. It further promotes the improvement of the accuracy of the ME recognition and has become the baseline for many subsequent works.

2.1.2. Geometric-Based Features

Optical flow, a geometric-based feature, calculates the displacement of facial feature points or the optical flow of the action area. It can extract representative motion features that are robust for the diversity of facial textures. Furthermore, the data except for RGB channels can be enhanced by optical flow (Liu et al., 2019).

Many works treat optical flow as a data preprocessing step. Liu et al. (2015) proposed an uncomplicated yet productive Main Directional Mean Optical-flow (MDMO) feature. On the ME video clips, an effective optical flow method is adopted. Meanwhile, Liu utilizes partial action units to divide the face into regions of interest (ROIs). MDMO is a normalized feature based on ROIs. It combines both spatial location and local statistic motion characteristics. MDMO has the advantage of small feature dimensions.

Some works (Liong et al., 2019; Liu et al., 2019; Zhou et al., 2019) utilized optical flow information for ME recognition and have achieved good results. For instance, Liu et al. (2019) utilized two domain adaptation methods, which include expression magnification and reduction and adversarial training. Then, he preprocessed the raw images to capture the spatio-temporal optical flow from facial movements from onset frame (the first frame in the ME video) to apex frame (the most intense frame of action in the ME video), won the championship of 2019-the second facial Micro-expressions Grand Challenge (MEGC2019) (See et al., 2019). Zhou et al. (2019) captured the TV-L1 optical

flow (Zach et al., 2007) of the onset frame and the mid-position frame, and then performs ME recognition through the Dual-Inception network. Instead of using apex frames, they use mid-position frames to cut down computation complexity. Furthermore, Liong et al. (2019) designed a STSTNet, which can be used to learn three features of optical flow, namely vertical optical flow, optical strain, and horizontal optical flow. These features are calculated by the onset frame and apex frame of ME video.

Optical flow has the advantage of small feature dimensions and the ability to capture subtle muscle movements. However, the optical flow has higher requirements on light and is easily affected by the external environment. In addition, these works only use the optical flow information of the apex frame and onset frame and lose the motion information of other frames.

2.2. Deep Neural Networks

Deep learning (LeCun et al., 2015) is universally used in various industries. Especially during the immediate past, the works on deep learning in the ME recognition field has gradually increased. In the field of deep learning, the features preprocessed by the optical flow method and LBP can be used as the input of convolution neural network (CNN). Then, CNN is usually used for feature extractors. For instance, Xia et al. (2019) proposed spatio-temporal recurrent convolutional networks based on optical flow, which extracts the optical flow information from the onset frame until the apex frame and inputs it into recurrent convolutional networks.

Furthermore, some works also use Long Short-term Memory (LSTM) to directly input ME video clips. One early work (Khor et al., 2018) proposed an Enriched Long-term Recurrent Convolutional Network (ELRCN). First, every ME frame is encoded into a feature vector by CNN modules. Then, ELRCN uses an LSTM module to pass the feature vector and predicts ME at last. ELRCN uses the feature that the information can be retained for a long time in the gating unit to detect ME in the video, and achieve good performance. Therefore, the combination of LSTM and CNN have greater advantages in recognizing ME in videos. However, due to the small changes in the ME video clips, there is frame redundancy, leading to greater computational complexity.

In conclusion, compared with traditional manual features for ME recognition, deep learning technology can extract features from ME videos and classify them with higher accuracy. However, due to frame redundancy in ME videos, the speed of the deep learning training model is greatly affected. Therefore, we propose two new ME video preprocessing methods to overcome frame redundancy in ME video and improve the recognition of ME classes.

3. METHOD

3.1. Preprocessing

As we discussed above, it is an inevitable stage to extract a discriminative and efficient feature. Therefore, this study proposes two methods based on the residual sum of image pixels to extract salient features: (1) Absolute Residual Sum (ARS) and (2) Relative Residual Sum (RRS). These methods take the frames in the ME clip at fixed intervals and consider the regional pixel displacement between frames. It not only avoids the redundancy of the ME clip but also makes full use of the ME information. The pixel-level displacement difference sum, named RS, can explain the tiny movement of the object. ARS and RRS preprocessing procedure are shown in **Figure 1**.

3.1.1. Absolute Residual Sum

Preprocessing is divided into five stages.

3.1.1.1. Select Video Clip

He et al. proposed MDMD, which used a reciprocal change from the onset frame to the offset frame to spotting ME (He et al., 2020). Therefore, we only recognize the ME from the onset frame to the apex frame. First, we select a video clip from the ME video and calculate its start and end. We select the partial video clips from the ME video clip. The onset frame is taken as the start by Equation (1), and select the end by Equation (2).

$$start = T(onset) \tag{1}$$

Where T(x) represents the frame sequence of x in the video.

$$end = \begin{cases} min(T(onset) + 10, T(offset)) & if\ T(apex) - T(onset) \\ & < 10 \\ min(T(apex), T(offset)) & else \end{cases} \tag{2}$$

Where $min(x, y)$ represents the smaller values of x and y.

3.1.1.2. Detect Feature Point

The dlib library is utilized to spotting facial feature points.

3.1.1.3. Cropping

Cropping the face through the face feature points.

3.1.1.4. Select Five Frames

Notice that, ME data is very redundant. Useful information must be mined from the data. A few other works (Li et al., 2013; Le Ngo et al., 2015, 2016) have proposed many methods to reduce frame redundancy in ME videos by using partial frames. Therefore, we require mining crucial frames from ME video clip. We define crucial frames as key-frames and define frames except for the key-frames as transition frames. Furthermore, we make two assumptions for getting rid of transition frames: (1) Transition frames are highly similar to the key-frames, and deletion does not affect the recognition accuracy. (2) Transition frames are continuously distributed, centered on key-frames.

Hence, we choose appropriate intervals by Equation (3) and select five key-frames as elements in $\mathbb{F}$ according to Equation (4).

$$gap = \lceil \frac{end - start}{N_{key} + 1} \rceil \tag{3}$$

$$\mathbb{F} = \{min(start + gap, end), min(start + gap * 2, end), ..., end\} \tag{4}$$

Where $\lceil x \rceil$ is taking the smallest integer not less than x for some scalar, and N_{key} represents the number of key frames. N_{key} is set to five in the paper.

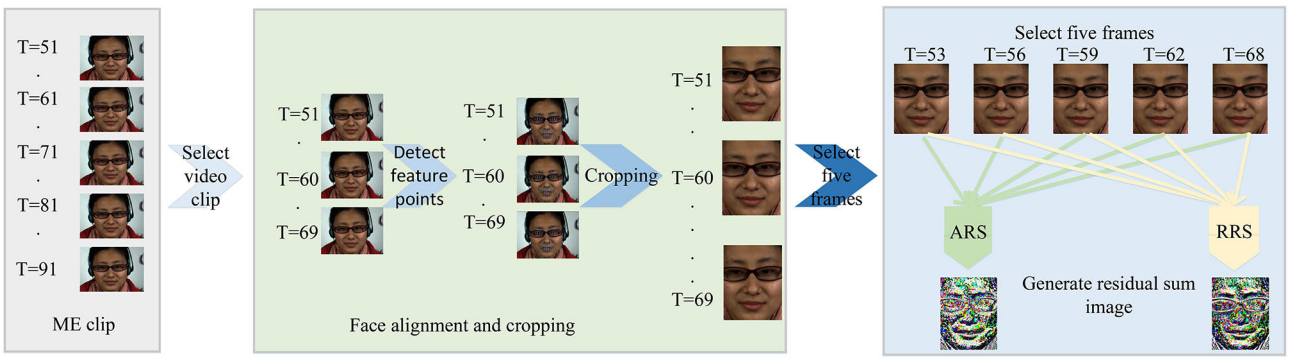


FIGURE 1 | Preprocessing Flow chart (©Xiaolan Fu).

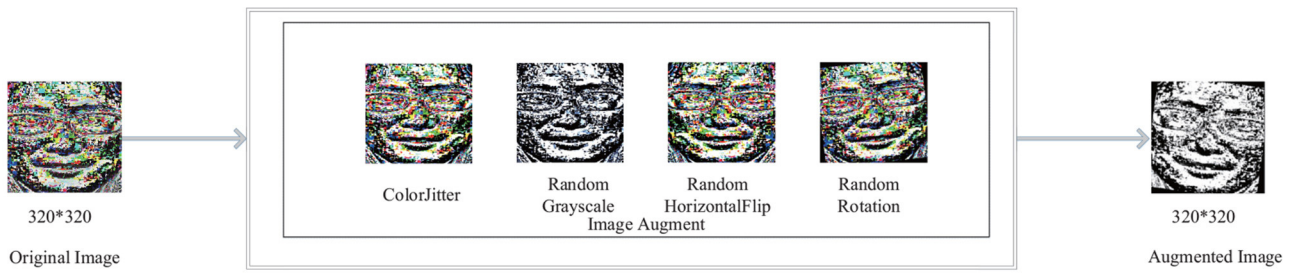


FIGURE 2 | Image augment module.

3.1.1.5. Generate Redisual Sum Image

Liu et al. (2019) took the motion difference between the onset frame and each frame to calibrate the apex frame, because the intensity relationship of ME can be indicated by the motion difference. Therefore, we cumulate the motion difference for calculating the variation trend of a single pixel. For the key frame in $\mathbb{F}$, Equation (5) is used to calculate the ARS.

$$ares(x, y, z) = (\sum_{f \in \mathbb{F}} (|Q_f(x, y, z) - Q_{start}(x, y, z)|)) \ \% \ 256 \quad (5)$$

Where $Q_f(x, y, z)$ represents the pixel value of the three-channel image (x, y, z) of the f_{th} frame and $ares(x, y, z)$ represents the pixel value of the generated ARS image.

3.1.2. Relative Residual Sum

As shown in **Figure 1**, the steps before the fifth step are the same as ARS. In the fifth step, we use Equation (6) to calculate the sum of residuals between frames. Then, we use Equation (7) to transform the range of sum to between *gmin* and *gmax*. In this experiment, $gmin = 0$ and $gmax = 255$.

$$diff(x, y, z) = (\sum_{f \in \mathbb{F}} (|Q_f(x, y, z) - Q_{start}(x, y, z)|)) \quad (6)$$

$$rres(x, y, z) = \frac{(diff(x, y, z) - min(diff))}{max(diff) - min(diff)} * (gmax - gmin) + gmin \quad (7)$$

Where $max(x, y)$, $diff(x, y, z)$, and $rres(x, y, z)$ represent the greater values of x and y, the sum of the displacement of the video frame at the three-channel image (x, y, z), and the pixel value of the generated RRS image, respectively.

3.2. Framework

CropNet, based on the depthwise convolution (Sandler et al., 2018), is used as a classification model. CropNet takes advantage of CGPO. The architecture of the CropNet is shown in **Figure 3**. Conv, BN, and FC in the figure represent Convolutional Layer, Batch Normalization Layer, and Fully Connected Layer, respectively.

3.2.1. Image Augmentation

The number of network parameters is approximately 7.6M. Image augment is essential as the network framework is slightly large. According to the characteristics of the human face, we performed the following four data augmentation in turn. (1) The image brightness, contrast, and saturation are randomly changed to [20%, 180%] of the original image brightness, and the hue offset of the image is changed to $[-0.5, 0.5]$ of the original image. (2) The picture is converted to grayscale with a probability of 20%. (3) Flipping the image horizontally with a 50% probability. (4) Rotating the image randomly clockwise $[-15,15]$ degrees. The image augment module is shown in **Figure 2**.

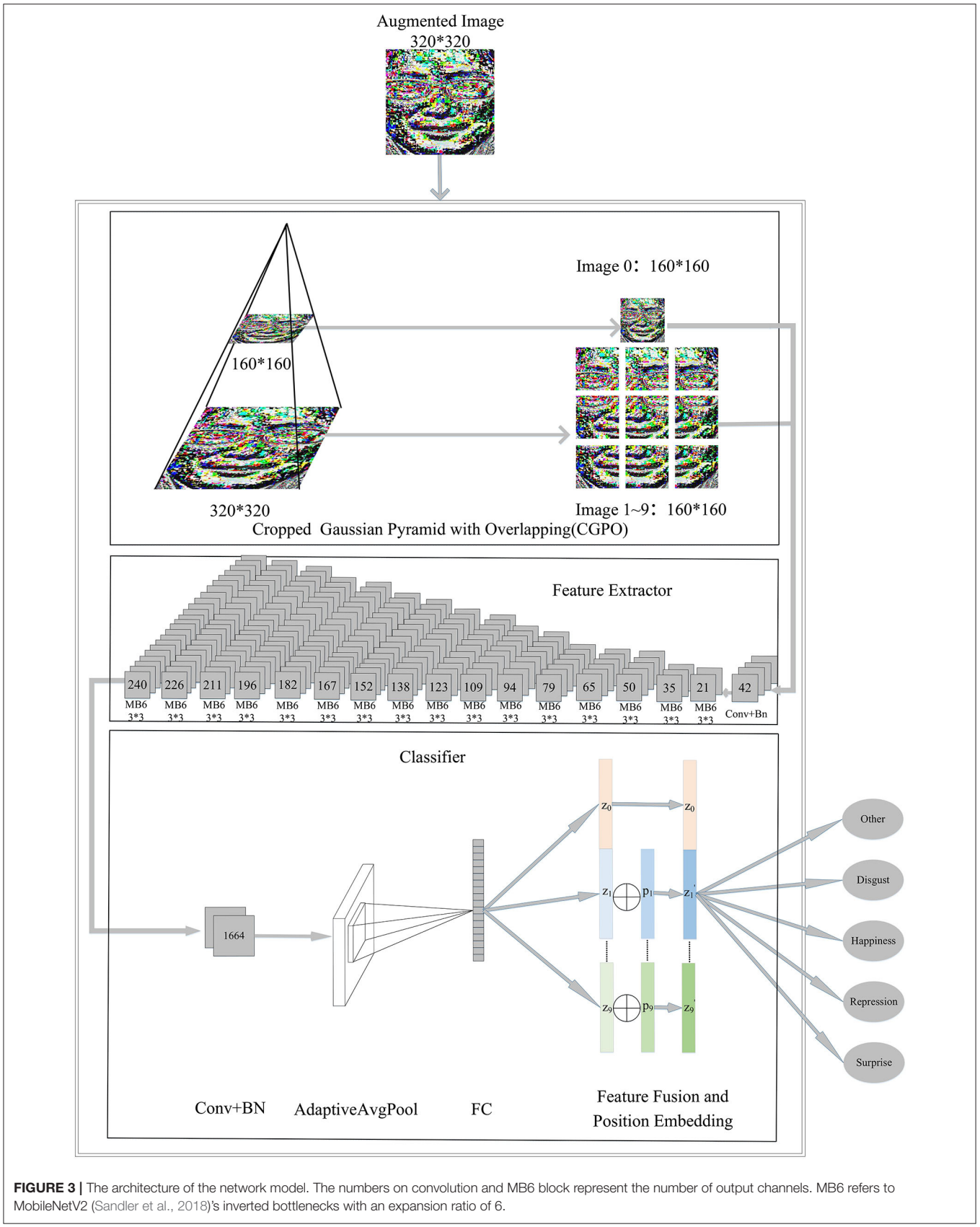


FIGURE 3 | The architecture of the network model. The numbers on convolution and MB6 block represent the number of output channels. MB6 refers to MobileNetV2 (Sandler et al., 2018)'s inverted bottlenecks with an expansion ratio of 6.

3.2.2. Cropped Gaussian Pyramid With Overlapping

Different facial areas have different importance in the production of ME. Therefore, we propose a CGPO module, which divides ME video frames with different resolutions of the image into 10 overlapping subplots. It can separate the mouth, the eyes, the nose, etc. The introduction of overlapping mechanisms can reduce the risk of separating important parts of the face. The CGPO module is shown in **Figure 3** CGPO, and its processing flow is as follows.

- First, we require 320 × 320 resolution of the image input and down-sample it to get an image with a resolution of 160 × 160.
- Second, for each image with different scale resolution, we divide them into several 160 × 160 images and introduce the overlap factor α. α is used to control the size of the overlap when crop images with different precision. In this study, α is 0.3.
- Finally, after going through the above process, images are fed CNN based with the depthwise convolution.

3.2.3. Feature Extraction

Han et al. (2020) designed ReXNet, which has achieved very good results in the ImageNet Challenge. Therefore, we use the ReXNet feature extraction module as the extractor. A network of progressively increasing channels are leveraged on the extracting feature, as shown in **Figure 3** feature extractor.

Due to the difficulties in data collection and identification of ME, there are few ME datasets. It is difficult to apply deep learning in ME recognition. Therefore, we train this module on the ImageNet datasets (Deng et al., 2009) and then apply it to the ME recognition through the transfer learning method (Pan and Yang, 2009).

3.2.4. Feature Fusion and Classifier

Feature Fusion and Classifier are shown in **Figure 3** Classifier. The features extracted in the previous module go through the Convolutional Layer, Batch Normalization Layer, Adaptive Pooling Layer, and Fully Connected Layer, in turn, and become a feature vector $z_i \in \mathbb{R}^{24}$, where i represents the order of segmented images. Since the CGOP module segmented a total of 10 images, we could obtain 10 feature vectors $\{z_0, z_1 \cdots\cdots\cdots z_9\}$.

However, because the position information after image cropping becomes blurred, the model has a hard time learning about correlations between images. We combine the location information with the feature to make the features more explanatory. Therefore, for feature vectors $\{z_1, z_2 \cdots\cdots\cdots z_9\}$ of segmented images, we introduce trainable position embedding vectors $\{p_1, p_2 \cdots\cdots\cdots p_9\}$ to learn the position information of the image, where p_i has the same dimension as z_i. The position embedding vectors are initialized to random values that follow a normal distribution. The mean of the random values is 0 and the variance is 0.2. As shown in Equation (8), we calculate the new feature vectors $\{z_1', z_2' \cdots\cdots\cdots z_9'\}$.

$$z_i' = z_i \oplus p_i \qquad 0 < i < 10 \tag{8}$$

Finally, we mix $\{z_0, z_1', z_2' \cdots\cdots\cdots z_9'\}$ by splicing and classifying ME.

4. EXPERIMENT

4.1. Datasets

Due to the characteristics of ME and its difficulty in triggering and collecting, the dataset is very scarce. As far as we know, there are three spontaneous datasets generally utilized for ME recognition: SMIC-HS (Li et al., 2013), SAMM (Davison et al., 2016), and CASME II (Yan et al., 2014a). The details of these three spontaneous datasets are shown in **Table 1**.

4.2. Experiment Settings

All experiments for this study were all carried out on Ubuntu 16.04 and Python 3.6.2 with Pytorch 1.6 on NVIDIA GTX Titan RTX GPU (24 GB). The label smoothing loss function (Lukasik et al., 2020) is leveraged as the loss function. It can better generalize the network and ultimately produce, more accurate predictions on invisible data. AdamP (Heo et al., 2021) is used as an optimizer. We use UF1 (commonly referred to as the macro average F1 score), UAR (commonly referred to as balanced accuracy), and Accuracy as our evaluation standard.

TABLE 1 | Micro-expression (ME) datasets.

Datasets	CASME II	SMIC-HS	SAMM
Particpants	26	16	29
Samples	255	157	159
Resolution	640*480	640*480	960*650
Frame rate(fps)	200	100	200
FACS coded	✓	x	✓
APEX index	✓	x	✓
Emotion	Other(99) Disgust(63) Surprise(28) Repression(27) Sadness(4) Happiness(32) Fear(2)	Negative(66) Positive(51) Surprise(40)	Other(26) Happiness(26) Disgust(9) Surprise(15) Sadness(6) Anger(57) Fear(8) Contempt(12)

TABLE 2 | Comparison of ME recognition performance in CASME II (5 classes).

Method	Accuracy
LBP-Top+AdaBoost (Le Ngo et al., 2014)	0.437
STCLQP (Huang and Zhao, 2017)	0.583
ELRCN (Khor et al., 2018)	0.524
DSSN (Khor et al., 2019)	0.707
TSCNN-I (Song et al., 2019)	0.740
SSSN (Khor et al., 2019)	0.711
TSCNN-II (Song et al., 2019)	0.810
Bi-WOOF (apex and onset) (Liong et al., 2018)	0.578
Su et al. (Su et al., 2021)	0.727
RRS+CropNet(ours)	0.790
ARS+CropNet(ours)	**0.862**

- **UF1** score can equally emphasize in a rare class. So, it is a suitable indicator in a multi-class evaluation. The calculation formula for UF1 is as follows:

$$UF1 = \frac{1}{C}\sum_{i=1}^{C}\left(\frac{2 * TP_i}{2 * TP_i + FP_i + FN_i}\right) \tag{9}$$

Where C represents the number of classes and FP_i, TP_i, and FN_i represent the false positive, the true positive, and the false negative for the i_{th} class, respectively.

- **UAR** is a more appropriate indicator instead of the standard accuracy indicator that may be partial to larger classes. The calculation formula for UAR is as follows:

$$UAR = \frac{1}{C}\sum_{i=1}^{C}\left(\frac{TP_i}{N_i}\right) \tag{10}$$

Where N_i represents the number of i_{th} class.

- **Accuracy** is commonly used as a CASME II experiment in five classes. The calculation formula for Accuracy is as follows:

$$Accuracy = \frac{TP}{N} \tag{11}$$

4.3. Experiment With Five Classes of ME in the CASME II

We choose the CASME II as the evaluation dataset. Only five classes (Others, Disgust, Happiness, Repression, and Surprise) are considered, since the fear and sadness samples are very scarce. In this experiment, Leave-One-Subject-Out (LOSO) cross validation is utilized for evaluation protocol. LOSO cross

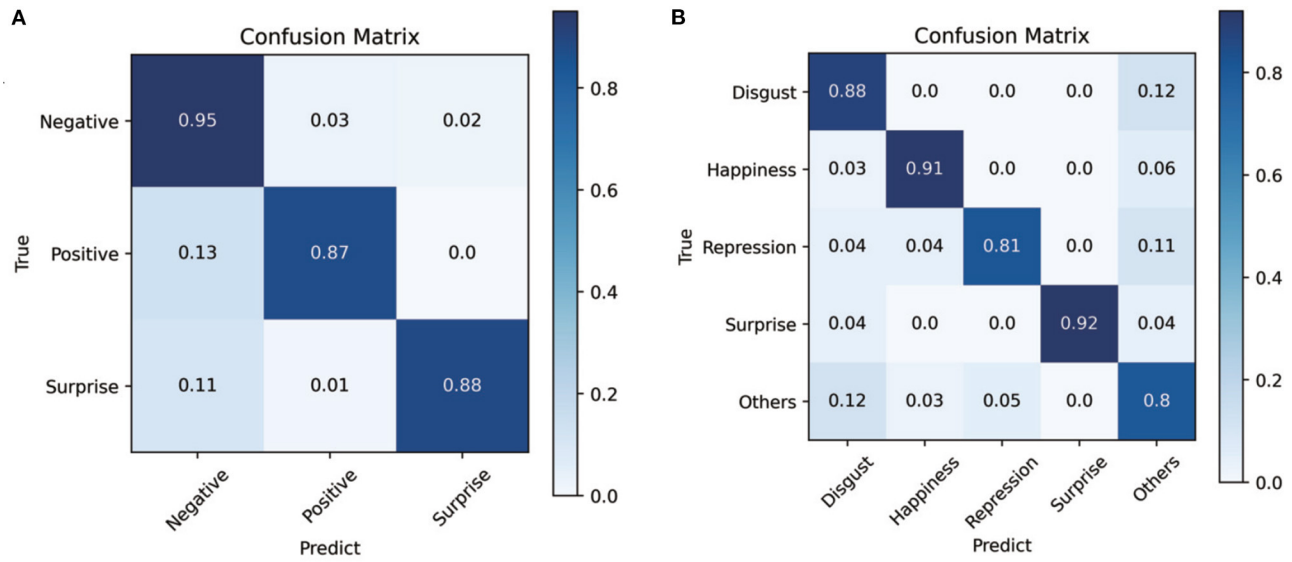


FIGURE 4 | (A) is the confusion matrix of composite datasets (SMIC-HS, CASME II, and SAMM) in the absolute residual sum (ARS) and the CropNet. **(B)** is the confusion matrix of CASME II in the ARS and the CropNet.

TABLE 3 | Comparison of ME recognition performance composite datasets.

Method	Composite		SMIC-HS		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP (Zhao and Pietikainen, 2007)	0.588	0.578	0.200	0.528	0.702	0.742	0.395	0.410
Bi-WOOF (Liong et al., 2018)	0.629	0.622	0.572	0.582	0.780	0.802	0.521	0.512
CapsuleNet (Van Quang et al., 2019)	0.652	0.650	0.582	0.587	0.706	0.701	0.620	0.598
OFF-ApexNet (Gan et al., 2019)	0.719	0.709	0.681	0.669	0.876	0.868	0.540	0.539
Dual-Inception (Zhou et al., 2019)	0.732	0.727	0.664	0.672	0.862	0.856	0.586	0.566
STSTNet (Liong et al., 2019)	0.735	0.760	0.680	0.701	0.838	0.868	0.658	0.681
ELTRCN (Khor et al., 2018)	0.788	0.782	0.746	0.753	0.829	0.820	0.775	0.715
RCN-S (Xia et al., 2020)	0.746	0.710	0.651	0.657	0.836	0.791	0.764	0.656
STSTNet+GA (Liu et al., 2021)	0.836	0.836	0.814	0.812	0.882	0.891	0.800	0.790
RRS+CropNet(ours)	0.875	0.877	0.813	0.819	0.972	0.969	0.842	0.827
ARS+CropNet(ours)	**0.911**	**0.904**	**0.855**	**0.851**	**0.974**	**0.979**	**0.912**	**0.893**

validation refers to using the samples of one subject as the test set, and the rest as the training set in each fold. It can prevent the test set and the training set from having the same sample, thereby avoiding data leakage. Recognition Accuracy can be calculated by the LOSO cross validation evaluation protocol. In the same evaluation standard, we compare with a variety of methods. The result is shown in **Table 2**.

The confusion matrix obtained by applying the ARS and the CropNet is shown in **Figure 4B**. Through the confusion matrix, the overall recognition rate is very high. The proposed method has great performance for all classes.

4.4. Composite Datasets Evaluation (CDE)

Composite datasets evaluation is a very effective evaluation method in cross-database ME recognition. In this experiment, we use the MEGC2019 standard. According to MEGC2019 standards, we combined all samples of the datasets (SAMM, CASME II, and SMIC-HS) into a composite dataset by unifying the number of ME class. ME are divided into three classes: negative, surprised, and positive. Disgust, contempt, fear, sadness, and anger is regarded as the negative class. Surprise is still regarded as surprise class. Happiness is regarded as the positive class. LOSO cross validation is utilized to split the training set and test set. **Table 3** compares the performance of proposed methods against a number of recent study. The methods in **Table 3** were all compared in the same datasets and at the same evaluation standard. The confusion matrix obtained by applying the ARS and the CropNet is shown in **Figure 4A**. It shows that three classes have similar performance, and the proposed method also has a good fit for unbalanced data.

Note that, the apex frame spotting is indispensable for ME recognition since the apex frame of the SMIC-HS dataset is not labeled. In recent years, there are a lot of apex frames spotting works (Yan et al., 2014b; Li et al., 2018; Peng et al., 2019; Zhou et al., 2019). In fact, apex frame spotting is a very difficult work. Therefore, this experiment considers a trade-off between efficiency and effectiveness. The middle frame of the video in the SMIC-HS dataset is used as the apex frame.

TABLE 4 | Ablation experiments in CASME II (5 classes).

Ablation module	Ablation method	UF1	Accuracy
paper method	**CropNet+ARS**	**0.863**	**0.862**
Preprocessing Method	CropNet+RRS	0.803	0.790
	CropNet+Optical FLow(Farneback)	0.661	0.625
	CropNet+Optical FLow(TV-L1)	0.697	0.669
Model architect	CropNet without GCOP +ARS	0.841	0.813

4.5. Ablation Experiments

We performed two ablation experiments on the CASME II dataset to verify the effectiveness of the module.

- We performed ablation experiments on preprocessing methods for comparing the effectiveness of the four preprocessing methods ARS, RRS, Farneback optical flow (Farnebäck, 2003), and TV-L1 optical flow.

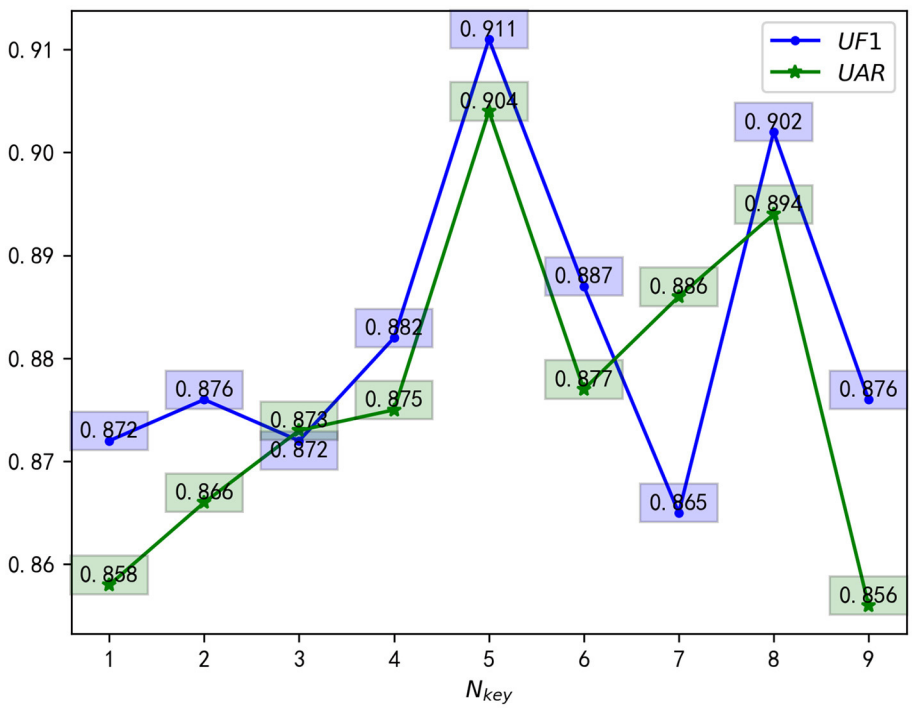


FIGURE 5 | N_{key} hyperparameter's ablation experiments.

- We performed ablation experiments on model architect for verifying the effect of the GCOP module.

As shown in **Table 4**, ARS stands out among the four preprocessing methods. It can extract more reliable spatio-temporal features and improve the UF1 value of ME recognition. RRS also achieves very good results. There are significant differences between these two methods. RRS pays more attention to areas with greater displacement by relative displacement change between unit pixels, while is not too sensitive to small displacement areas. ARS considers the trade-off between displacement regions of different scales, which can focus on both small displacement areas and large displacement areas. Therefore, subtle displacement can be captured. At the same time, for areas with frequent displacement, ARS ignores the displacement of unit pixels and pays attention to regional displacement. But in our experimental environment, Farneback optical flow and TV-L1 optical flow are far less effective than the proposed methods in this study.

The Cropped Gaussian Pyramid with Overlapping module focuses on different areas of the face, extracts features for each area, and then stitches the obtained features to classify them. Through the ablation experiment in **Table 4**, it is easy to find the efficiency of the CGPO module and the ARS method.

Furthermore, we conducted hyperparameter's ablation experiments in MEGC2019 composite datasets for verifying the effectiveness of the hyperparameters N_{key}. The experimental results are shown in **Figure 5**, which can be concluded that there is greater universality when N_{key} is set to five. Therefore, in all experiments, we only select five key-frames at equal intervals in the ME video clip.

4.6. Visualization Experiments

We use T-SNE (Van der Maaten and Hinton, 2008) to visualize the preprocessed image for better comparing the effects of the proposed preprocessing methods. **Figure 6** shows the feature distribution of images preprocessed by various methods. In

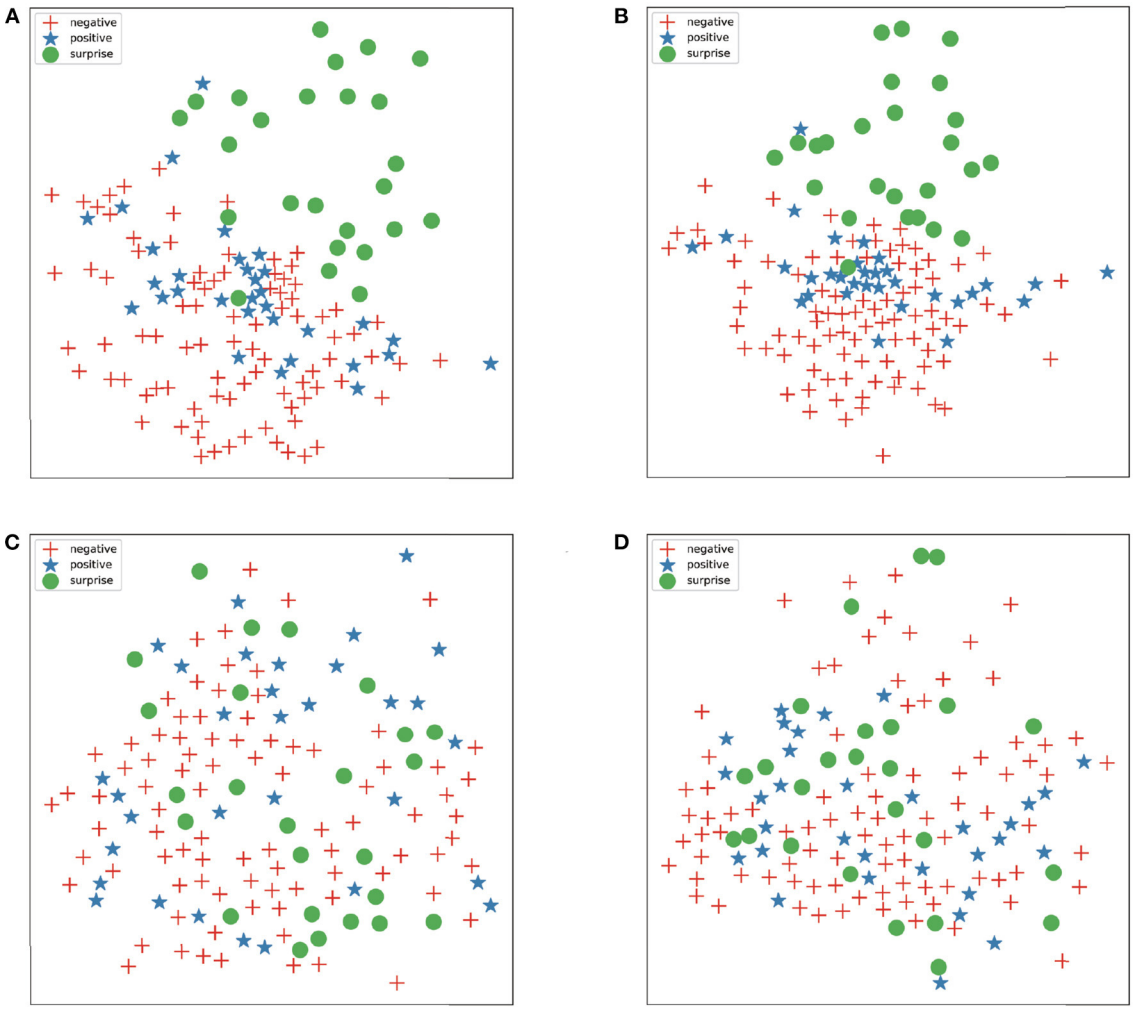


FIGURE 6 | (A–D) represent preprocessing images by ARS, relative residual sum (RRS), farneback optical flow and TV-L1 optical flow, respectively.

this experiment, we use three classes (negative, positive, and surprised) of CASME II.

The features extracted using Farneback optical flow and TV-L1 optical flow are disorganized, but the image extracted by residual sum methods can already distinguish many features. For example, surprise ME is easy to distinguish from other expressions. After preprocessing by the residual sum method, the features become initially orderly, but some of the ME are still mixed together. Therefore, further extraction of features through CNN can enhance the validity of features.

5. CONCLUSION

In this study, we propose two novel preprocessing methods to solve ME recognition tasks with spatial-temporal feature extraction. These methods use the displacement residual sum of the unit pixels of the ME clip to extract a subtle motion feature. Through our experiment, it responds well to environmental change and subtle displacement. In addition, we propose a CGPO module, which divides the image into partial overlapping pictures of different precision and extracts features from different pictures. Hence, the model can focus on each facial local area, and then recognize the subtle movements of specific locations. Furthermore, we design CropNet which have a gradual way of increasing channels, features fusion module, and position embedding function.

In the experiment, we test the proposed two preprocessing methods and the designed network on the mixed dataset of MEGC2019 and five classes of ME on CASME II. The traditional manual method based on optical flow is labor-expensive and time-consuming, while the RRS and ARS preprocessing methods greatly improve the situation of frame redundancy and improve the recognition accuracy of each ME. In addition, the CGPO module can separate key parts of a person's face for more subtle feature extraction. In general, the method proposed in the study has better performance than the well-known method.

However, the proposed model does not belong to an end-to-end model, because it must go through the preprocessing method, which takes a certain amount of time to detect key points, align faces, crop, and calculate RRS and ARS. Therefore, in the future improvement, we will improve the method and model in this study into an end-to-end model.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: this paper involves three databases (CASMEII, SMIC and SAMM). As each database involves human facial expressions, you need to apply for access. Requests to access these datasets should be directed to SMIC: Xiaobai.Li@oulu.fi, SAMM: M.Yap@mmu.ac.uk, CASMEII: eagan-ywj@foxmail.com.

AUTHOR CONTRIBUTIONS

YZ led the method design and experiment implementation. YZ and SL wrote sections of the manuscript. SL and ZC provided theoretical guidance, result review, and paper revision. All authors read and approved the final manuscript.

FUNDING

This publication of this paper was supported by the National Natural Science Foundation of China (no. 61872051), the Scientific and Technological Research Program of Chongqing Municipal Education Commission of China (no. KJ1600932), and the Graduate Innovation Fund of Chongqing University of Technology (no. clgycx20203123).

REFERENCES

Boucenna, S., Gaussier, P., Andry, P., and Hafemeister, L. (2014). A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game. *Int. J. Soc. Rob.* 6, 633–652. doi: 10.1007/s12369-014-0245-z

Davison, A. K., Lansley, C., Costen, N., Tan, K., and Yap, M. H. (2016). Samm: a spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput*. 9, 116–129. doi: 10.1109/TAFFC.2016.2573832

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255.

Ekman, P. (2009a). Lie catching and microexpressions. *Philos. Decept*. 1, 5. doi: 10.1093/acprof:oso/9780195327939.003.0008

Ekman, P. (2009b). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. New York, NY: WW Norton & Company.

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol*. 17, 124. doi: 10.1037/h0030377

Farnebäck, G. (2003). "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis* (Berlin; Heidelberg: Springer).

Frank, M., Herbasz, M., Sinuk, K., Keller, A., and Nolan, C. (2009). "I see how you feel: training laypeople and professionals to recognize fleeting emotions," in *The Annual Meeting of the International Communication Association* (New York, NY: Sheraton New Yorkpages), 1–35.

Gan, Y., Liong, S.-T., Yau, W.-C., Huang, Y.-C., and Tan, L.-K. (2019). Off-apexnet on micro-expression recognition system. *Signal Proc. Image Commun*. 74, 129–139. doi: 10.1016/j.image.2019.02.005

Han, D., Yun, S., Heo, B., and Yoo, Y. (2020). Rexnet: Diminishing representational bottleneck on convolutional neural network. *arXiv preprint* arXiv:2007.00992.

He, Y., Wang, S. J., Li J., and Yap, H. M. (2020). "Spotting macro-and micro-expression intervals in long video sequences," in *15th IEEE International Conference on Automatic Face and Gesture Recognition* (Buenos Aires: IEEE), 742–748.

Heo, B., Chun, S., Oh, S. J., Han, D., Yun, S., Kim, G., et al. (2021). "Adamp: slowing down the slowdown for momentum optimizers on scale-invariant weights," in *International Conference on Learning Representations, Vol*. 6, 1–5.

Huang, X., and Zhao, G. (2017). "Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)* (Xi'an: IEEE), 159–164.

Huang, X., Zhao, G., Hong, X., Zheng, W., and Pietikäinen, M. (2016). Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing* 175, 564–578. doi: 10.1016/j.neucom.2015.10.096

Khor, H.-Q., See, J., Liong, S.-T., Phan, R. C., and Lin, W. (2019). "Dual-stream shallow networks for facial micro-expression recognition," in *2019 IEEE International Conference on Image Processing (ICIP)* (Taipei: IEEE), 36–40.

Khor, H.-Q., See, J., Phan, R. C. W., and Lin, W. (2018). "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (Xi'an: IEEE), 667–674.

Kim, J.-H., Kim, B.-G., Roy, P. P., and Jeong, D.-M. (2019). Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access* 7, 41273–41285. doi: 10.1109/ACCESS.2019.2907327

Le Ngo, A. C., Liong, S.-T., See, J., and Phan, R. C.-W. (2015). "Are subtle expressions too sparse to recognize?" in *2015 IEEE International Conference on Digital Signal Processing (DSP)* (Singapore: IEEE), 1246–1250.

Le Ngo, A. C., Phan, R. C.-W., and See, J. (2014). "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Asian Conference on Computer Vision* (Singapore: Springer), 33–48.

Le Ngo, A. C., See, J., and Phan, R. C.-W. (2016). Sparsity in dynamics of spontaneous subtle emotions: analysis and application. *IEEE Trans. Affect. Comput.* 8, 396–411. doi: 10.1109/TAFFC.2016.2523996

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, X., Pfister, T., Huang, X., Zhao, G., and Pietikäinen, M. (2013). "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (fg)* (Shanghai: IEEE), 1–6.

Li, Y., Huang, X., and Zhao, G. (2018). "Can micro-expression be recognized based on single apex frame?" in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens: IEEE), 3094–3098.

Lies, T. (1992). *Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY: Norton.

Liong, S.-T., Gan, Y., See, J., Khor, H.-Q., and Huang, Y.-C. (2019). "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (Lille: IEEE), 1–5.

Liong, S. T., See, J. S. Y., Wong, K. S., and Phan, R. C. W. (2018). Less is more: micro-expression recognition from video using apex frame. *Signal Proc. Image Commun.* 62:82–92. doi: 10.1016/j.image.2017.11.006

Liu, K.-H., Jin, Q.-S., Xu, H.-C., Gan, Y.-S., and Liong, S.-T. (2021). Micro-expression recognition using advanced genetic algorithm. *Signal Proc. Image Commun.* 93:116153. doi: 10.1016/j.image.2021.116153

Liu, Y., Du, H., Zheng, L., and Gedeon, T. (2019). "A neural micro-expression recognizer," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (Lille: IEEE), 1–4.

Liu, Y., Yuan, X., Gong, X., Xie, Z., Fang, F., and Luo, Z. (2018). Conditional convolution neural network enhanced random forest for facial expression recognition. *Pattern Recognit.* 84, 251–261. doi: 10.1016/j.patcog.2018.07.016

Liu, Y.-J., Zhang, J.-K., Yan, W.-J., Wang, S.-J., Zhao, G., and Fu, X. (2015). A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* 7, 299–310. doi: 10.1109/TAFFC.2015.2485205

Lukasik, M., Bhojanapalli, S., Menon, A., and Kumar, S. (2020). "Does label smoothing mitigate label noise?" in *International Conference on Machine Learning* (PMLR), 6448–6458.

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Peng, M., Wang, C., Bi, T., Shi, Y., Zhou, X., and Chen, T. (2019). "A novel apex-time network for cross-dataset micro-expression recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (Cambridge, UK: IEEE), 1–6.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4510–4520.

See, J., Yap, M. H., Li, J., Hong, X., and Wang, S.-J. (2019). "Megc 2019- the second facial micro-expressions grand challenge," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (Lille: IEEE), 1–5.

Song, B., Li, K., Zong, Y., Zhu, J., Zheng, W., Shi, J., et al. (2019). Recognizing spontaneous micro-expression using a three-stream convolutional neural network. *IEEE Access* 7, 184537–184551. doi: 10.1109/ACCESS.2019.2960629

Su, Y., Zhang, J., Liu, J., and Zhai, G. (2021). "Key facial components guided micro-expression recognition based on first amp; second-order motion," in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (Shenzhen), 1–6.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.

Van Quang, N., Chun, J., and Tokuyama, T. (2019). "Capsulenet for micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (Lille: IEEE), 1–7.

Wang, Y., See, J., Phan, R. C.-W., and Oh, Y.-H. (2014). "Lbp with six intersection points: reducing redundant information in lbp-top for micro-expression recognition," in *Asian Conference on Computer Vision* (Singapore: Springer International Publishing), 525–537.

Xia, Z., Hong, X., Gao, X., Feng, X., and Zhao, G. (2019). Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Trans. Multimedia* 22, 626–640. doi: 10.1109/TMM.2019.2931351

Xia, Z., Peng, W., Khor, H.-Q., Feng, X., and Zhao, G. (2020). Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans. Image Proc.* 29, 8590–8605. doi: 10.1109/TIP.2020.3018222

Xie, S., Hu, H., and Wu, Y. (2019). Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognit.* 92, 177–191. doi: 10.1016/j.patcog.2019.03.019

Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., et al. (2014a). Casme ii: an improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* 9:e86041. doi: 10.1371/journal.pone.0086041

Yan, W.-J., Wang, S.-J., Chen, Y.-H., Zhao, G., and Fu, X. (2014b). "Quantifying micro-expressions with constraint local model and local binary pattern," in *European Conference on Computer Vision* (Zurich: Springer), 296–305.

Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., and Fu, X. (2013). How fast are the leaked facial expressions: the duration of micro-expressions. *J. Nonverbal. Behav.* 37, 217–230. doi: 10.1007/s10919-013-0159-8

Zach, C., Pock, T., and Bischof, H. (2007). "A duality based approach for realtime tv-l 1 optical flow," in *Joint Pattern Recognition Symposium* (Heidelberg: Springer), 214–223.

Zhao, G., and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 915–928. doi: 10.1109/TPAMI.2007.1110

Zhou, L., Mao, Q., and Xue, L. (2019). "Dual-inception network for cross-database micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (Lille: IEEE), 1–5.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fcomp.2021.786964

Singular Learning of Deep Multilayer Perceptrons for EEG-Based Emotion Recognition

Weili Guo[1,2], *Guangyu Li*[1,2]*, *Jianfeng Lu*[2] *and Jian Yang*[1,2]*

[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology, Nanjing, China, [2]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

Human emotion recognition is an important issue in human–computer interactions, and electroencephalograph (EEG) has been widely applied to emotion recognition due to its high reliability. In recent years, methods based on deep learning technology have reached the state-of-the-art performance in EEG-based emotion recognition. However, there exist singularities in the parameter space of deep neural networks, which may dramatically slow down the training process. It is very worthy to investigate the specific influence of singularities when applying deep neural networks to EEG-based emotion recognition. In this paper, we mainly focus on this problem, and analyze the singular learning dynamics of deep multilayer perceptrons theoretically and numerically. The results can help us to design better algorithms to overcome the serious influence of singularities in deep neural networks for EEG-based emotion recognition.

Keywords: emotion recognition, EEG, deep multilayer perceptrons, singular learning, theoretical and numerical analysis

OPEN ACCESS

Edited by:
Xiaopeng Hong,
Xi'an Jiaotong University, China

Reviewed by:
Lianyong Qi,
Qufu Normal University, China
Liangfei Zhang,
University of St Andrews,
United Kingdom

***Correspondence:**
Guangyu Li
guangyu.li2017@njust.edu.cn
Jian Yang
csjyang@njust.edu.cn

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *30 September 2021*
Accepted: *05 November 2021*
Published: *21 December 2021*

Citation:
Guo W, Li G, Lu J and Yang J (2021) Singular Learning of Deep Multilayer Perceptrons for EEG-Based Emotion Recognition. Front. Comput. Sci. 3:786964. doi: 10.3389/fcomp.2021.786964

1 INTRODUCTION

Emotion recognition is a fundamental task in affective computing and has attracted many researchers' attention in recent years (Mauss and Robinson, 2009). Human emotion can be expressed through external signals and internal signals, where external signals usually include facial expressions, body actions, and speeches, and electroencephalograph (EEG) and galvanic skin response (GSR) are typical internal signals. EEG is the method to measure electrical activities of the brain by using electrodes along the scalp skin and it is rather reliable; therefore, EEG has played a more significant role in investigating human emotion recognition problem in recent years (Yin et al., 2021).

For the emotion recognition problem based on EEG signals, researchers mainly investigate this issue from two aspects: how to extract better features from EEG signals and how to construct a model with better performance. For aspect 1, researchers have investigated the feature extraction methods of EEG signals from a time domain, frequency domain, and time–frequency domain, respectively, and a series of results have been given previously (Fang et al., 2020; Nawa et al., 2020). In this paper, we mainly focus on aspect 2, i.e., the computational model problem, and researchers have proposed many models to recognize emotions through EEG signals (Zong et al., 2016; Yang et al., 2018a; Zhang et al., 2019; Cui et al., 2020). In recent years, deep learning technology has achieved great success in many fields (Yang et al., 2018b; Yang et al., 2019; Basodi et al., 2020; Zhu and Zhang, 2021), and many works are devoted to addressing the EEG emotion recognition issue by applying deep neural networks (DNNs) (Cao et al., 2020; Natarajan et al., 2021), where the performances based on deep

learning also show significant superiority of conventional methods (Ng et al., 2015; Tzirakis et al., 2017; Hassan et al., 2019). However, the learning dynamics of DNNs, including deep multilayer perceptrons (MLPs), deep belief networks and deep convolution neural networks, are often affected by singularities, which exist in the parameter space of DNNs (Nitta, 2016).

Due to the influence of singularities, the training of DNNs often becomes very slow and the plateau phenomenon can often be observed. When the DNNs are applied to EEG-based emotion recognition, the severe negative effect of singularities on the learning process of DNNs is also inevitable, where the efficiency and performance of networks can also not be guaranteed. However, up to now, there are rarely literatures investigating this problem. In this paper, we mainly concern this problem. The main contribution of this paper is to take the theoretical and numerical analysis of singular learning in DNNs for EEG-based emotion recognition. We choose deep MLPs as the learning machine, where deep MLPs are of typical DNNs and the results are also representative for other DNNs. The types of singularities in parameter space are analyzed and the specific influence of the singularities is clearly shown. Based on the obtained results in this paper, we can further design the related algorithms to overcome this issue.

The rest of this paper is organized as follows. A brief review of related work is presented in **Section 2**. In **Section 3**, theoretical analysis of singularities in deep MLPs for EEG-based emotion recognition is taken and then the learning dynamics near singularities are numerically analyzed in **Section 4**. **Section 5** states conclusion and discussion.

2 RELATED WORK

In this section, we provide a brief overview of previous work on EEG-based emotion recognition and singular learning of DNNs.

In recent years, due to the high accuracy and stabilization of EEG signals, EEG-based algorithms have attracted ever-increasing attention in emotion recognition field. To extract better features of EEG signals, researchers have proposed various feature extraction models (Zheng et al., 2014; Zheng, 2017; Tao et al., 2020; Zhao et al., 2021), such as power spectral density (PSD), differential entropy (DE), and differential asymmetry (DASM). By using PSD and DE to extract dimension reduced features of EEG signals, Fang et al. (2020) chose the original features and dimension reduced features as the multi-feature input and verified the validity of the proposed method in the experiment part. Li et al. (2020) integrated psychoacoustic knowledge and raw waveform embedding within an augmented feature space. Song et al. (2020) employed an additional branch to characterize the intrinsic dynamic relationships between different EEG channels and a type of sparse graphic representation was presented to extract more discriminative features. Besides the feature extraction methods, more attention is paid to study the emotion classification. Given that the deep learning technology has excellent capabilities, various types of DNNs have been widely used in emotion classification (Li et al., 2018; Li et al., 2019; Ma et al., 2019; Atmaja and Akagi, 2020; Cui et al., 2020; Zhong et al., 2020), including deep convolution neural networks, deep MLPs, long short term memory (LSTM)-based recurrent neural networks, and graph neural networks. The obtained results show that these DNN models can provide superior performance compared to previous models (Yang et al., 2021a; Yang et al., 2021b).

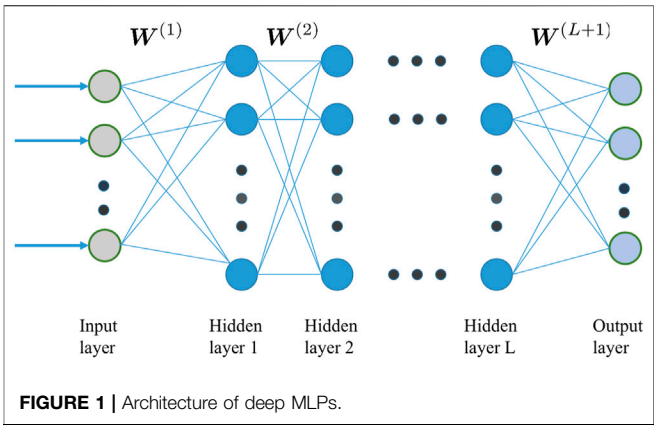


FIGURE 1 | Architecture of deep MLPs.

As mentioned above, various DNNs have been widely used in EEG-based emotion recognition; however, the training processes of DNNs often encounter many difficulties. Even if numerous research studies have been developed to conduct explanatory research, it is still very far to revealing the mechanism. As there are singularities in the parameter space of DNNs where the Fisher information matrix is singular, the singular learning dynamics of DNNs have been studied and have attracted more and more attention. As the basis of DNNs, traditional neural networks often suffer from the serious influence of various singularities (Amari et al., 2006; Guo et al., 2018; Guo et al., 2019), and the learning dynamics of DNNs are also easy to be influenced by the singularities. Nitta (2016, 2018) analyzed the types of singularity in DNNs and deep complex-value neural networks. Ainsworth and Shin (2020) investigated the plateau phenomenon in Relu-based neural networks. By using the spectral information of Fisher information matrix, Liao et al. (2020) proposed an algorithm to accelerate the training process of DNNs.

In view of the serious influence of singularities to DNNs, the training processes of DNNs will also encounter difficulties when applying DNNs to EEG-based emotion recognition. Thus, it is necessary to take the theoretical and numerical analysis to reveal the mechanism and propose related algorithms to overcome the influence of singularities.

3 THEORETICAL ANALYSIS OF SINGULAR LEARNING DYNAMICS OF DEEP MULTILAYER PERCEPTRONS

In this section, we theoretically analyze the learning dynamics near singularities of deep MLPs for the EEG-based emotion recognition.

3.1 Learning Paradigm of Deep Multilayer Perceptrons

Firstly, we introduce a typical learning paradigm of deep MLPs. For a typical deep multilayer perceptrons with L hidden layers (the architecture of the networks is shown in **Figure 1**), assuming M_i is the neuron number of hidden layer i, M_0 is the dimension of the input layer and M_{L+1} is the dimension of the output layer, we denote that: $W_{jk}^{(i)}$ represents the weight connecting from the jth node of the previous layer to the kth node of hidden layer i, and $W_{pq}^{(L+1)}$ represents the weight connecting from the pth node of hidden layer L to the qth node of output layer for $1 \le i \le L$, $1 \le j \le M_{i-1}$, $1 \le k \le M_i$, $1 \le p \le M_L$, and $1 \le q \le M_{L+1}$. Then $\boldsymbol{\theta} = \{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(L+1)}\}$ represents all the parameters of the networks, where $\boldsymbol{W}^{(i)} = [\boldsymbol{W}_1^{(i)}, \boldsymbol{W}_2^{(i)}, \ldots, \boldsymbol{W}_{M_i}^{(i)}]$ and $\boldsymbol{W}_j^{(i)} = [W_{1j}^{(i)}, W_{2j}^{(i)}, \ldots, W_{M_{(i-1)}j}^{(i)}]^T$ for $1 \le i \le L+1$ and $1 \le j \le M_i$. In this paper, the widely used log-sigmoid function $\phi(x) = \frac{1}{1+e^{-x}}$ is adopted as the activation of hidden layers and the purelin function $\psi(x) = x$ is adopted as the activation function of output layer, then for the input $\boldsymbol{x} \in \mathbb{R}^{M_0}$, by denoting the input to hidden layer k as $\boldsymbol{X}^{(k-1)}$ for $1 \le k \le L$ and the input to output layer as $\boldsymbol{X}^{(L)}$, the mathematical model of the networks can be described as follows:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \psi((\boldsymbol{W}^{(L+1)})^T \boldsymbol{X}^{(L)}) = (\boldsymbol{W}^{(L+1)})^T \boldsymbol{X}^{(L)}. \tag{1}$$

For $1 \le k \le L$, $\boldsymbol{X}^{(k)}$ can be computed as $\boldsymbol{X}^{(k)} = \phi(\boldsymbol{X}^{(k-1)}, \boldsymbol{W}^{(k)}) = \phi((\boldsymbol{W}^{(k)})^T \boldsymbol{X}^{(k-1)})$ and $\boldsymbol{X}^{(0)}$ is the input $\boldsymbol{x}$.

We choose the square loss function to measure the error:

$$l(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2}(y - f(\boldsymbol{x}, \boldsymbol{\theta}))^2, \tag{2}$$

and use the gradient descent method to minimize the loss:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\partial l(y, \boldsymbol{x}, \boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}, \tag{3}$$

where η is the learning rate.

3.2 Singularities of Deep Multilayer Perceptrons in Electroencephalograph-Based Emotion Recognition

In this paper, we mainly focus on the mechanism of singular learning dynamics of deep MLPs applied to EEG-based emotion recognition domain, not seeking the best performance; therefore, the size of the networks need not to be very large, and an appropriate size that can capture the essence of singular learning dynamics can satisfy the requirement. Without loss of generality, we choose the deep MLPs with two hidden layers and a single output neuron, i.e., $L = 2$ and $M_3 = 1$, i.e., $\boldsymbol{W}^{(3)} = \boldsymbol{W}_1^{(3)} = [W_{11}^{(3)}, W_{21}^{(3)}, \ldots, W_{M_2 1}^{(3)}]^T$, we simply denoted as $\boldsymbol{W}^{(3)} = [W_1^{(3)}, W_2^{(3)}, \ldots, W_{M_2}^{(3)}]$. Then, the deep MLPs can be rewritten as:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\theta}) &= (\boldsymbol{W}^{(3)})^T \phi(\phi(\boldsymbol{x}, \boldsymbol{W}^{(1)}), \boldsymbol{W}^2) \\ &= \sum_{j=1}^{M_2} W_j^{(3)} \phi(\phi(\boldsymbol{x}, \boldsymbol{W}^{(1)}), \boldsymbol{W}_j^{(2)}). \end{aligned} \tag{4}$$

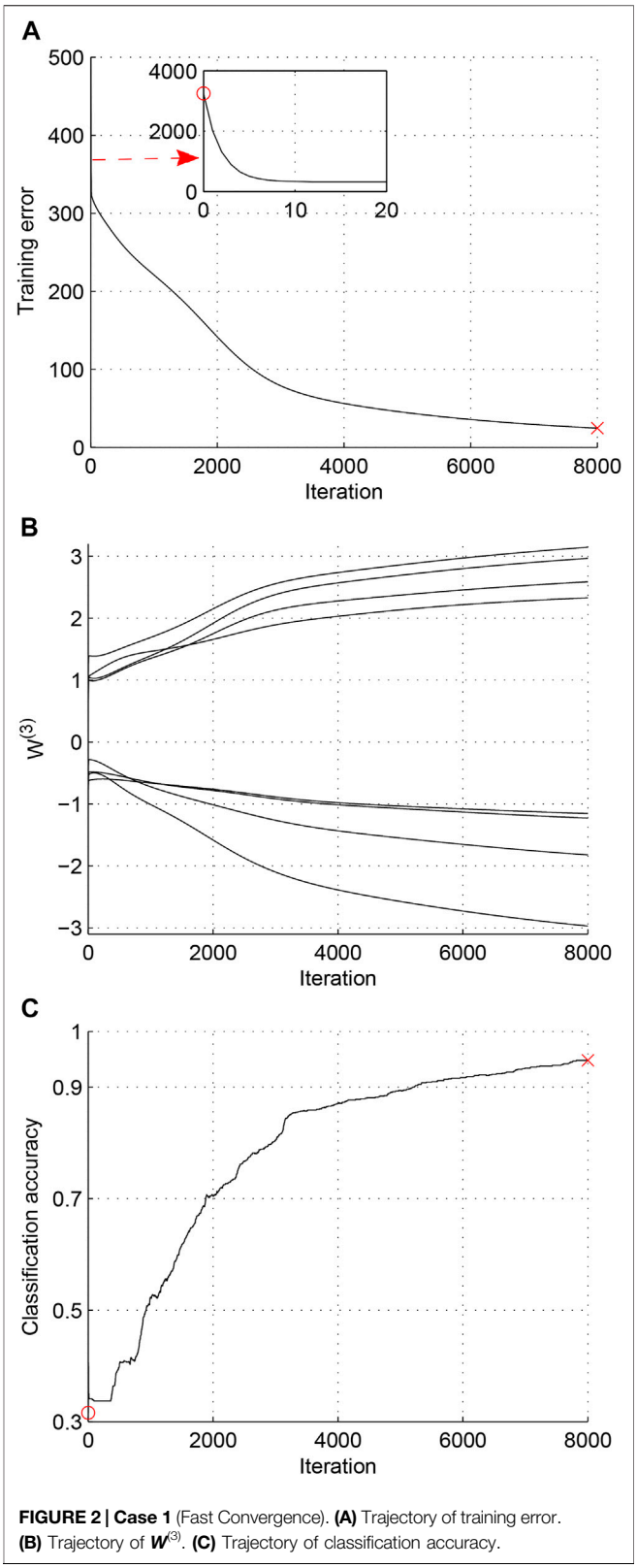


FIGURE 2 | Case 1 (Fast Convergence). **(A)** Trajectory of training error. **(B)** Trajectory of $\boldsymbol{W}^{(3)}$. **(C)** Trajectory of classification accuracy.

Next, we analyze the types of singularities. From **Eq. 4**, we can see that if one output weight equals zero, e.g., $W_j^{(3)} = 0$, whatever the values of $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}_j^{(2)}$ be, the output of unit j will be always 0

and the unit seems to be vanished. As the values of $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}_j^{(2)}$ have no effect on the output of the deep MLPs, the training process will encounter difficulties on the subspace $R_1 = \{\boldsymbol{\theta} | W_j^{(3)} = 0\}$. Besides the above singularity, if there are two elements of weight $\boldsymbol{W}^{(2)}$ overlap, e.g., $\boldsymbol{W}_i^{(2)} = \boldsymbol{W}_j^{(2)}$, then

$$\begin{aligned} & W_i^{(3)} \phi(\phi(\boldsymbol{x}, \boldsymbol{W}^{(1)}), \boldsymbol{W}_i^{(2)}) \\ & \quad + W_j^{(3)} \phi(\phi(\boldsymbol{x}, \boldsymbol{W}^{(1)}), \boldsymbol{W}_j^{(2)}) \\ & = (W_i^{(3)} + W_j^{(3)}) \phi(\phi(\boldsymbol{x}, \boldsymbol{W}^{(1)}), \boldsymbol{W}_i^{(2)}) \end{aligned}$$

remains the same value when $W_i^{(3)} + W_j^{(3)}$ takes a fixed value, regardless of particular values of $W_i^{(3)}$ and $W_j^{(3)}$. Therefore, we can identify their sum $W = W_i^{(3)} + W_j^{(3)}$; nevertheless, each of $W_i^{(3)}$ and $W_j^{(3)}$ remains unidentifiable. Thus, the training will also suffer difficulties on the subspace $R_2 = \{\boldsymbol{\theta} | \boldsymbol{W}_i^{(2)} = \boldsymbol{W}_j^{(2)}\}$.

To sum up the above analysis, it can be seen that there are at least two types of singularities:

(1) Zero weight singularity: $R_1 = \{\boldsymbol{\theta} | W_j^{(3)} = 0\}$,
(2) Overlap singularity: $R_2 = \{\boldsymbol{\theta} | \boldsymbol{W}_i^{(2)} = \boldsymbol{W}_j^{(2)}\}$.

Till now, we have theoretically analyzed the types of singularity that existed in the parameter space of deep MLPs; in the next section, we will numerically analyze the influence of singularities to solve EEG-based emotion recognition problem.

4 NUMERICAL ANALYSIS OF LEARNING DYNAMICS NEAR SINGULARITIES

In this section, we take the numerical analysis of singularities by taking experiments on the dataset of EEG signals. For the EEG datasets, the SEED dataset is a typical benchmark dataset that is developed by SJTU and has been widely used to evaluate the proposed methods on EEG-based emotion recognition. In this paper, the training process will be carried out using the SEED dataset.

4.1 Data Preprocessing

The SEED dataset (Zheng and Lu, 2015) is collected from 62-channel EEG device and contains EEG signals of three emotions (positive, neutral, and negative) from 15 subjects. Due to the low signal-to-noise ratio of raw EEG signals, it is rather necessary to take the preprocessing step to extract meaningful features. As is known, there are five frequency bands for each EEG channel: delta (1–3 Hz), theta (4–7 Hz), alpha (8–13 Hz), beta (14–30 Hz), and gamma (31–50 Hz). That means, for one subject, the data are the form 5×62, the dimension of raw EEG signal is very large, and then we use principal component analysis (PCA) (Abdi and Williams, 2010) to extract the features of the EEG signal. After the PCA step, the form of EEG signals becomes 5×5, and then by putting every element of the data to a vector, the dimension of the input can be finally reduced to be 25.

4.2 Learning Trajectories Near Singularities

Now, we take experiments on the SEED dataset, and the learning dynamics near singularities will be numerically analyzed. We choose the neuron numbers of two hidden layers as $L_1 = 8$ and $L_2 = 8$; thus, the architecture of the deep MLPs is 25–8–8–1. As there are three emotions in the SEED dataset, we set values 1, 2, and 3 corresponding to labels positive, neutral, and negative, respectively. We choose the training sample number and

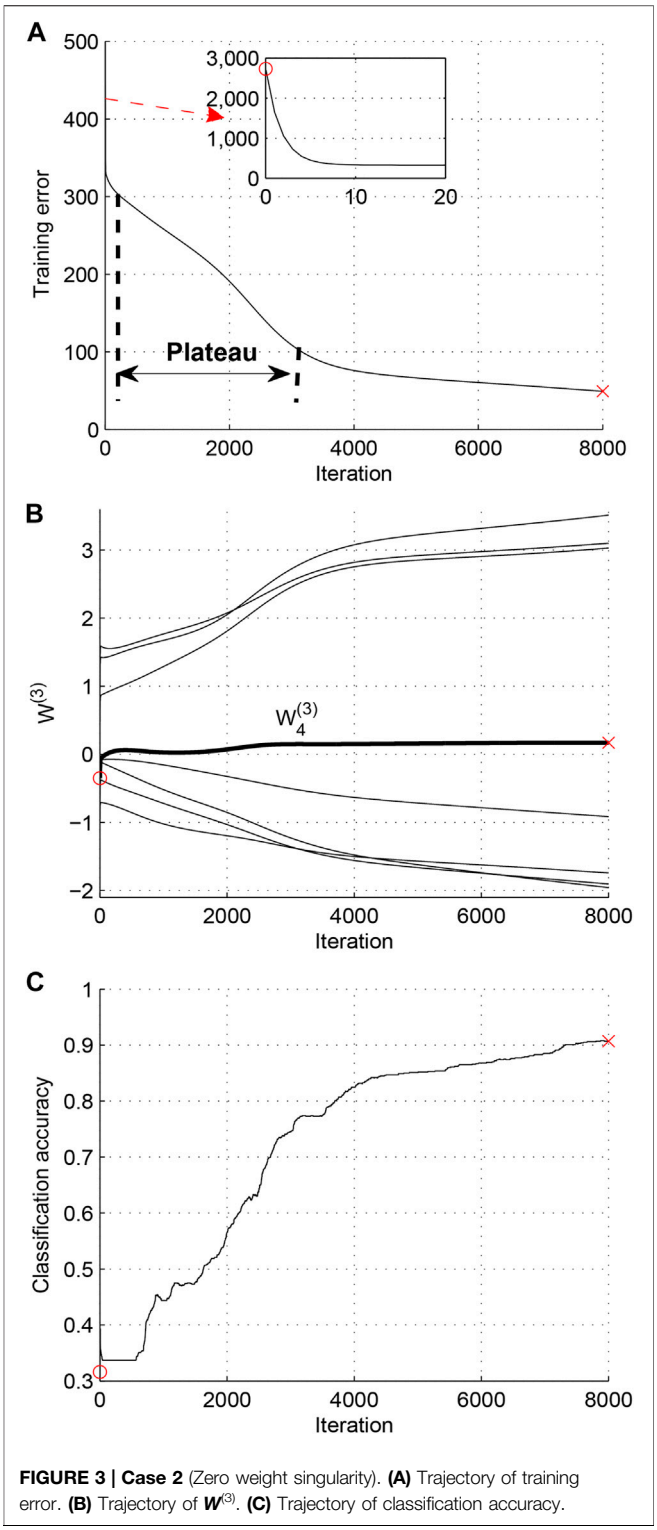


FIGURE 3 | Case 2 (Zero weight singularity). **(A)** Trajectory of training error. **(B)** Trajectory of $\boldsymbol{W}^{(3)}$. **(C)** Trajectory of classification accuracy.

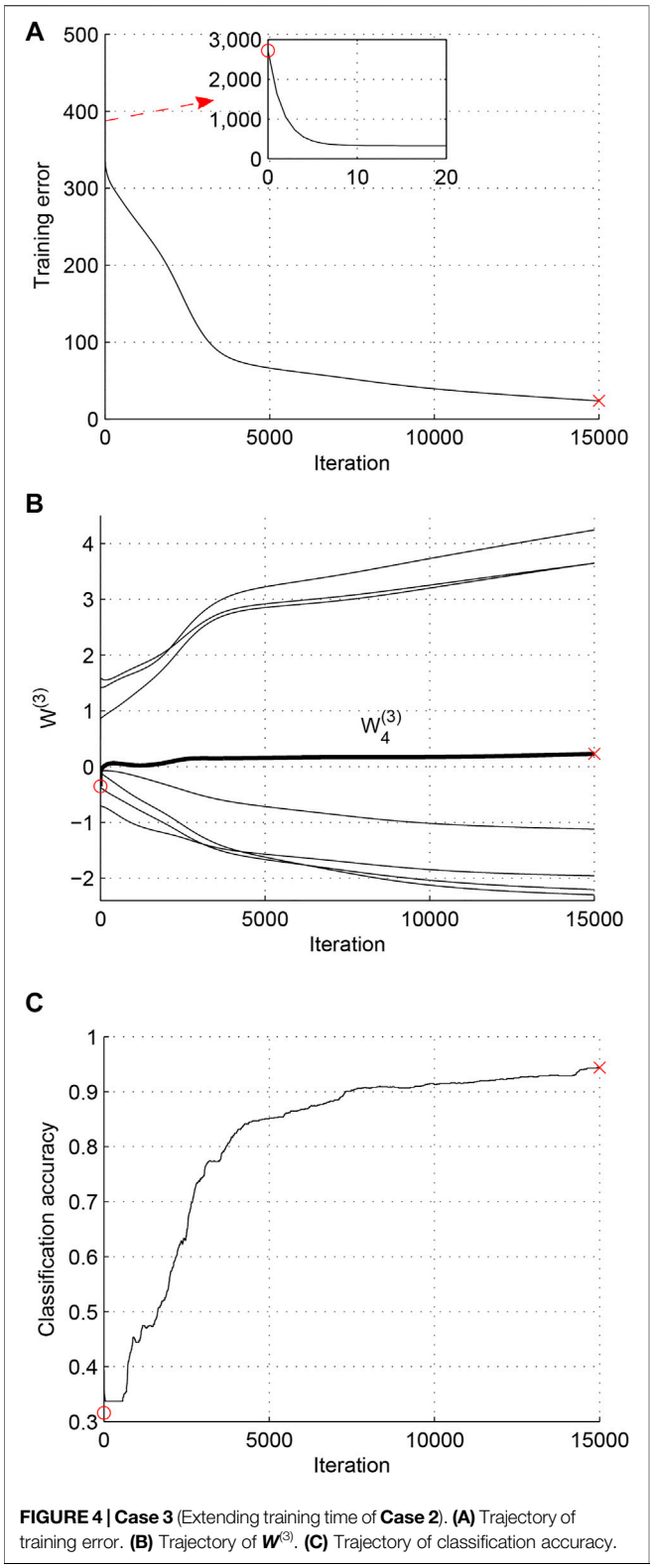


FIGURE 4 | Case 3 (Extending training time of **Case 2**). **(A)** Trajectory of training error. **(B)** Trajectory of $\boldsymbol{W}^{(3)}$. **(C)** Trajectory of classification accuracy.

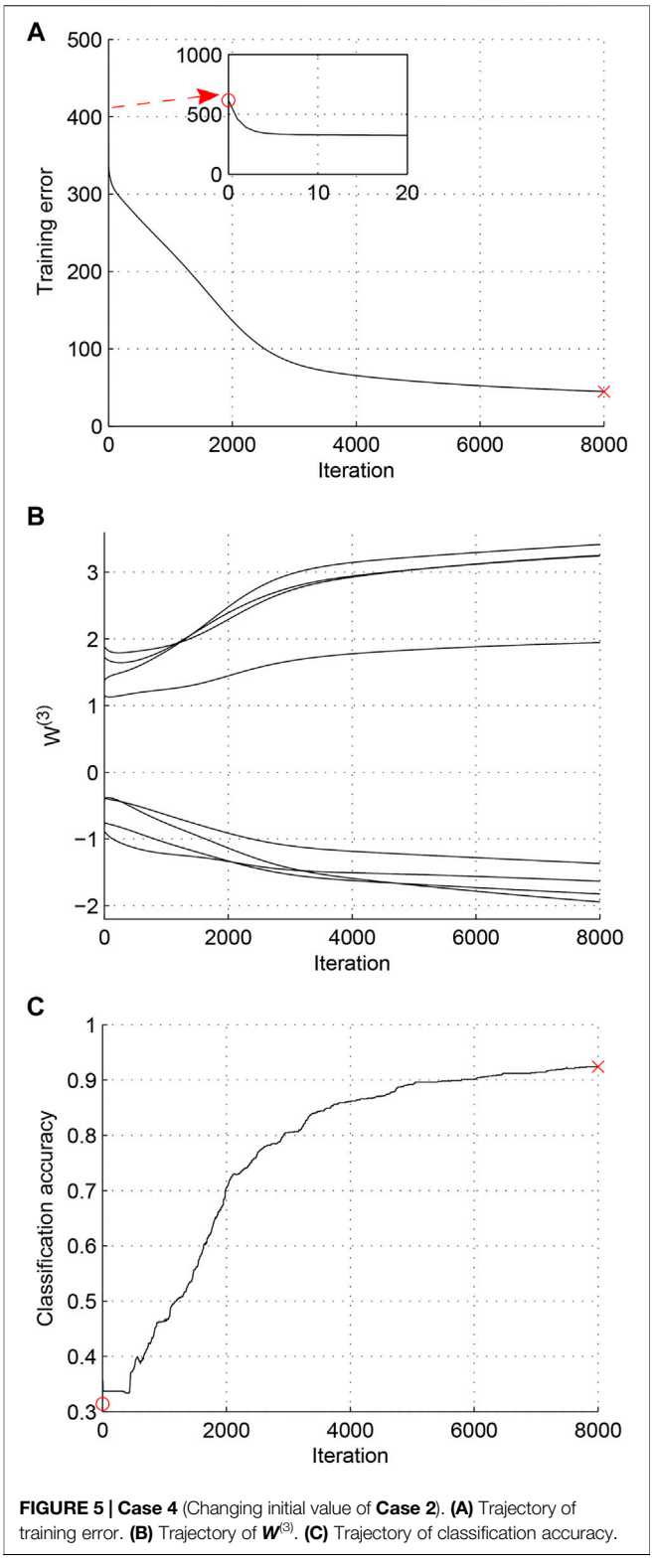


FIGURE 5 | Case 4 (Changing initial value of **Case 2**). **(A)** Trajectory of training error. **(B)** Trajectory of $\boldsymbol{W}^{(3)}$. **(C)** Trajectory of classification accuracy.

testing sample number to be 1,000 and 500, respectively. Then, by setting the learning rate to $\eta = 0.002$, the target error to 0.05, and the maximum epochs to 8,000, we use **Eq. 3** to accomplish the experiment. By analyzing the experiment results, two cases of learning dynamics will be shown. Besides training error, classification accuracy is also used to measure the performance. In the following figures of experiment results, "∘" and "×" represent the initial state and final state,

TABLE 1 | Training and testing classification accuracy.

	Iteration number	Training classification accuracy	Testing classification accuracy
Case 1	8,000	0.948	0.941
Case 2	8,000	0.901	0.894
Case 3	15,000	0.944	0.938
Case 4	8,000	0.924	0.920

respectively. The experiments were run by using Matlab 2013a on a PC with an Intel Core i7-9700K CPU @3.60 GHz, 32 GB RAM and NVIDIA GeForce RTX 2070 GPU.

Case 1. Fast convergence: The learning process fast converges to the global minimum.

For this case, the learning dynamics does not suffer from any influence of singularity and the parameters fast converge to the optimal value. The initial value of $\boldsymbol{W}^{(3)}$ is $\boldsymbol{W}^{(3)(0)}$ = [0.8874, 0.6993, 0.5367, −0.9415, −0.8464, −0.9280, 0.3335, −0.7339]^T and the final value of $\boldsymbol{W}^{(3)}$ is $\boldsymbol{W}^{(3)}$ = [3.1443, 2.5868, 2.3291, −1.1544, −1.2281, −2.9704, 2.9650, −1.8221]^T. The experiment results are shown in **Figure 2**, which represent the trajectories of training error, output weights $\boldsymbol{W}^{(3)}$, and classification accuracy, respectively.

As can be seen from **Figure 2A**, the learning dynamics quickly converge to the global minimum and have not been affected by any singularity.

Case 2. Zero weight singularity: the learning process is affected by the elimination singularity.

For this case, one output weight crosses 0 during the learning process and a plateau phenomenon can be obviously observed. The initial value of $\boldsymbol{W}^{(3)}$ is $\boldsymbol{W}^{(3)(0)}$ = [0.4825, 0.9885, −0.9522, −0.3505, −0.5004, 0.9749, −0.9111, −0.5056]^T, and the final student parameters are $\boldsymbol{W}^{(3)}$ = [3.0297, 3.1006, −1.7413, 0.1717, −1.9567, 3.5131, −1.9037, −0.9143]^T. The experiment results are shown in **Figure 3**, which represent the trajectories of training error, output weights $\boldsymbol{W}^{(3)}$ and classification accuracy, respectively.

From **Figure 3B**, we can see that $W_4^{(3)}$ crosses 0 in the learning process and the learning process is affected by elimination singularity. During the stage $W_4^{(3)}$ crosses 0, the plateau phenomenon can be obviously observed (**Figure 3A**). Then, the student parameters escape the influence of elimination singularity. After the training process, we can see that the training error is bigger than that in Case 1 and the classification accuracy is also lower than that in Case 1, which means that the parameters do not reach the optimum.

Case 3. Extending training time of Case 2.

In this experiment, we only increase the training epochs to 15,000, and the rest of the experiment setup remains the same with that in Case 2. The experiment results are shown in **Figure 4**. Compared to **Figure 3**, it can be seen that the learning process that is affected by the zero weight singularity can arrive at the optimum, but it costs much more time. This means that the zero weight singularities will greatly reduce the efficiency of deep MLPs.

Case 4. Changing initial value of Case 2.

In order to confirm that the plateau phenomenon corresponds to the zero weight singularity, a supplementary experiment is carried out here where only the initial value of $\boldsymbol{W}^{(3)(0)}$ has been changed and the rest of the experiment setup remains the same. The initial value of $\boldsymbol{W}^{(3)}$ is $\boldsymbol{W}^{(3)(0)}$ = [−0.5056, −0.9111, 1.7749, −0.5004, 1.6495, −0.9522, 0.9885, 1.2825]^T, and the final student parameters are $\boldsymbol{W}^{(3)}$ = [−1.3660, −1.8232, 3.2529, −1.9425, 3.2450, −1.6325, 1.9452, 3.4158]^T. The experiment results are shown in **Figure 5**, which represent the trajectories of training error, output weights $\boldsymbol{W}^{(3)}$, and classification accuracy, respectively. As can be seen in **Figure 5**, there is not any weight of $\boldsymbol{W}^{(3)}$ that becomes zero. Also, no plateau phenomenon can be observed, and the classification accuracy has reached a comparatively high value. By comparing the experiment results shown in **Figures 3**, **5**, we can conclude that the plateau phenomenon is indeed caused by zero weight singularity.

Remark 1. From the results shown in **Figures 2**–**5** and **Table 1**, we can see that the training and testing accuracy in Case 2 is the lowest. This means that when the training process is affected by the zero weight singularity, the parameters cannot achieve the optimum after the same training time with that in fast convergence case. When we extend the training time in Case 2, the parameters can escape the influence of zero weight singularity and finally arrive at the optimum, which is shown in Case 3. Thus, the points in zero weight singularity are saddle points, not local minimum. To sum up, the zero weight singularity will seriously delay the training process, and it is worthy to investigate algorithms to overcome the influence of zero weight singularities.

Remark 2. When taking the experiments, we do not observe the learning dynamics of deep MLPs that are affected by overlap singularities. The results are in accordance with the conclusion where we analyze the learning dynamics of shallow neural networks (Guo et al., 2018); i.e., the overlap singularities mainly influence the neural networks with low dimension and the large-scale networks predominantly suffer from zero weight singularities. Thus, we should pay more attention to how to overcome the influence of zero weight singularities.

In this section, we have numerically analyzed the learning dynamics near singularities of deep MLPs for EEG-based emotion recognition and showed the singular case. We can obtain that the

learning dynamics of deep MLPs are mainly influenced by zero weight singularities and rarely affected by overlap singularities.

5 CONCLUSION AND DISCUSSION

Deep learning technology has been widely used in EEG-based emotion recognition and has shown superior performance compared to traditional methods. However, for various DNNs, there exist singularities in the parameter space, which cause singular behaviors in the training process. In this paper, we investigate the singular learning dynamics of DNNs when applied to EEG-based emotion recognition. By choosing deep MLPs as the learning machine, we firstly take the theoretical analysis of singularities of deep MLPs, and obtained that there are at least two types of singularities: overlap singularity and zero weight singularity. Then, by doing several experiments, the numerical analysis is taken. The experiment results show that the learning dynamics of deep MLPs are seriously influenced by zero weight singularities and rarely affected by overlap singularities. Furthermore, the plateau phenomenon is caused by zero weight singularity. Thus, we should pay more attention to how to overcome the serious influence of zero weight singularity to improve the efficiency of DNNs in EEG-based emotion recognition in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

WG and GL: Methodology. WG and JY: Validation and investigation. WG: Writing—original draft preparation. GL, JL, and JY: Formal analysis, data curation. JL and JY: Writing—reviewing and editing, and supervision. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61906092, 61802059, 62006119, and 61876085, the Natural Science Foundation of Jiangsu Province of China under Grant Nos. BK20190441, BK20180365, and BK20190444, and the 973 Program No. 2014CB349303.

REFERENCES

Abdi, H., and Williams, L. J. (2010). Principal Component Analysis. *Wires Comp. Stat.* 2, 433–459. doi:10.1002/wics.101

Ainsworth, M., and Shin, Y. (2020). *Plateau Phenomenon in Gradient Descent Training of Relu Networks: Explanation, Quantification and Avoidance*.

Amari, S.-i., Park, H., and Ozeki, T. (2006). Singularities Affect Dynamics of Learning in Neuromanifolds. *Neural Comput.* 18, 1007–1065. doi:10.1162/neco.2006.18.5.1007

Atmaja, B. T., and Akagi, M. (2020). "Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition," in Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Auckland, New Zealand, December 7–10, 2020 (Auckland: IEEE), 325–331.

Basodi, S., Ji, C., Zhang, H., and Pan, Y. (2020). Gradient Amplification: An Efficient Way to Train Deep Neural Networks. *Big Data Min. Anal.* 3, 196–207. doi:10.26599/bdma.2020.9020004

Cao, Q., Zhang, W., and Zhu, Y. (2020). Deep Learning-Based Classification of the Polar Emotions of "moe"-Style Cartoon Pictures. *Tsinghua Sci. Technology* 26, 275–286.

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., and Chen, X. (2020). Eeg-based Emotion Recognition Using an End-To-End Regional-Asymmetric Convolutional Neural Network. *Knowledge-Based Syst.* 205, 106243. doi:10.1016/j.knosys.2020.106243

Fang, Y., Yang, H., Zhang, X., Liu, H., and Tao, B. (2020). Multi-feature Input Deep forest for Eeg-Based Emotion Recognition. *Front. Neurorobotics* 14, 617531.

Guo, W., Ong, Y.-S., Zhou, Y., Hervas, J. R., Song, A., and Wei, H. (2019). Fisher Information Matrix of Unipolar Activation Function-Based Multilayer Perceptrons. *IEEE Trans. Cybern.* 49, 3088–3098. doi:10.1109/tcyb.2018.2838680

Guo, W., Wei, H., Ong, Y., Hervas, J. R., Zhao, J., Wang, H., et al. (2018). Numerical Analysis Near Singularities in RBF Networks. *J. Mach. Learn. Res.* 19, 1–39.

Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., and Fortino, G. (2019). Human Emotion Recognition Using Deep Belief Network Architecture. *Inf. Fusion* 51, 10–18. doi:10.1016/j.inffus.2018.10.009

Li, J.-L., Huang, T.-Y., Chang, C.-M., and Lee, C.-C. (2020). A Waveform-Feature Dual branch Acoustic Embedding Network for Emotion Recognition. *Front. Comput. Sci.* 2, 13. doi:10.3389/fcomp.2020.00013

Li, J., Zhang, Z., and He, H. (2018). Hierarchical Convolutional Neural Networks for Eeg-Based Emotion Recognition. *Cogn. Comput.* 10, 368–380. doi:10.1007/s12559-017-9533-x

Li, P., Liu, H., Si, Y., Li, C., Li, F., Zhu, X., et al. (2019). EEG Based Emotion Recognition by Combining Functional Connectivity Network and Local Activations. *IEEE Trans. Biomed. Eng.* 66, 2869–2881. doi:10.1109/tbme.2019.2897651

Liao, Z., Drummond, T., Reid, I., and Carneiro, G. (2020). Approximate Fisher Information Matrix to Characterize the Training of Deep Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 15–26. doi:10.1109/tpami.2018.2876413

Ma, J., Tang, H., Zheng, W., and Lu, B. (2019). "Emotion Recognition Using Multimodal Residual LSTM Network," in Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, October 21–25, 2019, 176–183. doi:10.1145/3343031.3350871

Mauss, I. B., and Robinson, M. D. (2009). Measures of Emotion: A Review. *Cogn. Emot.* 23, 209–237. doi:10.1080/02699930802204677

Natarajan, Y., Srihari, K., Chandragandhi, S., Raja, R. A., Dhiman, G., and Kaur, A. (2021). Analysis of Protein-Ligand Interactions of Sars-Cov-2 against Selective Drug Using Deep Neural Networks. *Big Data Min. Anal.* 4, 76–83.

Nawaz, R., Cheah, K. H., Nisar, H., and Yap, V. V. (2020). Comparison of Different Feature Extraction Methods for Eeg-Based Emotion Recognition. *Biocybernetics Biomed. Eng.* 40, 910–926. doi:10.1016/j.bbe.2020.04.005

Ng, H., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, November 09–13, 2015. Editors Z. Zhang, P. Cohen, D. Bohus, R. Horaud, and H. Meng, 443–449. doi:10.1145/2818346.2830593

Nitta, T. (2016). "On the Singularity in Deep Neural Networks," in Proceedings of 23rd International Conference on Neural Information Processing, Kyoto, Japan, October 16–21, 2016 (Kyoto: Lecture Notes in Computer Science), 389–396. doi:10.1007/978-3-319-46681-1_47

Nitta, T. (2018). Resolution of Singularities via Deep Complex-Valued Neural Networks. *Math. Meth Appl. Sci.* 41, 4170–4178. doi:10.1002/mma.4434

Song, T., Liu, S., Zheng, W., Zong, Y., and Cui, Z. (2020). "Instance-adaptive Graph for EEG Emotion Recognition," in Proceedings of the Thirty-Fourth AAAI

Conference on Artificial Intelligence, New York, February 7–12, 2020, 2701–2708. doi:10.1609/aaai.v34i03.5656*Aaai*34

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., et al. (2020). Eeg-based Emotion Recognition via Channel-wise Attention and Self Attention. *IEEE Trans. Affective Comput.*, 1. doi:10.1109/TAFFC.2020.3025777

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal. Process.* 11, 1301–1309. doi:10.1109/jstsp.2017.2764438

Wei-Long Zheng, W., and Bao-Liang Lu, B. (2015). Investigating Critical Frequency Bands and Channels for Eeg-Based Emotion Recognition with Deep Neural Networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175. doi:10.1109/tamd.2015.2431497

Yang, Y., Fu, Z., Zhan, D., Liu, Z., and Jiang, Y. (2021a). Semi-supervised Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. *IEEE Trans. Knowl. Data Eng.* 33, 696–709.

Yang, Y., Wu, Q. M. J., Zheng, W.-L., and Lu, B.-L. (2018a). Eeg-based Emotion Recognition Using Hierarchical Network with Subnetwork Nodes. *IEEE Trans. Cogn. Dev. Syst.* 10, 408–419. doi:10.1109/tcds.2017.2685338

Yang, Y., Wu, Y., Zhan, D., Liu, Z., and Jiang, Y. (2018b). "Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, August 19–23, 2018, 2594–2603.

Yang, Y., Zhan, D., Wu, Y., Liu, Z., Xiong, H., and Jiang, Y. (2021b). Semi-supervised Multi-Modal Clustering and Classification with Incomplete Modalities. *IEEE Trans. Knowl. Data Eng.* 33, 682–695.

Yang, Y., Zhou, D., Zhan, D., Xiong, H., and Jiang, Y. (2019). "Adaptive Deep Models for Incremental Learning: Considering Capacity Scalability and Sustainability," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, August 4–8, 2019, 74–82.

Yin, Y., Zheng, X., Hu, B., Zhang, Y., and Cui, X. (2021). EEG Emotion Recognition Using Fusion Model of Graph Convolutional Neural Networks and LSTM. *Appl. Soft Comput.* 100, 106954. doi:10.1016/j.asoc.2020.106954

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2019). Spatial-temporal Recurrent Neural Network for Emotion Recognition. *IEEE Trans. Cybern.* 49, 839–847. doi:10.1109/tcyb.2017.2788081

Zhao, L., Yan, X., and Lu, B. (2021). "Plug-and-play Domain Adaptation for Cross-Subject Eeg-Based Emotion Recognition," in Proceedings of thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event, February 2–9, 2021 863–870.

Zheng, W. (2017). Multichannel Eeg-Based Emotion Recognition via Group Sparse Canonical Correlation Analysis. *IEEE Trans. Cogn. Dev. Syst.* 9, 281–290. doi:10.1109/tcds.2016.2587290

Zheng, W., Zhu, J., Peng, Y., and Lu, B. (2014). Eeg-based Emotion Classification Using Deep Belief Networks. *IEEE Int. Conf. Multimedia Expo*, 1–6. doi:10.1109/icme.2014.6890166

Zhong, P., Wang, D., and Miao, C. (2020). Eeg-based Emotion Recognition Using Regularized Graph Neural Networks. *IEEE Trans. Affective Comput.*, 1. doi:10.1109/TAFFC.2020.2994159

Zhu, K., and Zhang, T. (2021). Deep Reinforcement Learning Based mobile Robot Navigation: A Review. *Tsinghua Sci. Technol.* 26, 674–691. doi:10.26599/tst.2021.9010012

Zong, Y., Zheng, W., Cui, Z., and Li, Q. (2016). Double Sparse Learning Model for Speech Emotion Recognition. *Electron. Lett.* 52, 1410–1412. doi:10.1049/el.2016.1211

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 January 2022
doi: 10.3389/fpsyg.2021.784834

Spontaneous Facial Expressions and Micro-expressions Coding: From Brain to Face

Zizhao Dong[1], *Gang Wang*[2], *Shaoyuan Lu*[1,3], *Jingting Li*[1], *Wenjing Yan*[4] *and Su-Jing Wang*[1,3]*

[1] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [2] School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, China, [3] Department of Psychology, University of the Chinese Academy of Sciences, Beijing, China, [4] Department of Applied Psychology, College of Teacher Education, Wenzhou University, Zhejiang, China

Facial expressions are a vital way for humans to show their perceived emotions. It is convenient for detecting and recognizing expressions or micro-expressions by annotating a lot of data in deep learning. However, the study of video-based expressions or micro-expressions requires that coders have professional knowledge and be familiar with action unit (AU) coding, leading to considerable difficulties. This paper aims to alleviate this situation. We deconstruct facial muscle movements from the motor cortex and systematically sort out the relationship among facial muscles, AU, and emotion to make more people understand coding from the basic principles:

1. We derived the relationship between AU and emotion based on a data-driven analysis of 5,000 images from the RAF-AU database, along with the experience of professional coders.
2. We discussed the complex facial motor cortical network system that generates facial movement properties, detailing the facial nucleus and the motor system associated with facial expressions.
3. The supporting physiological theory for AU labeling of emotions is obtained by adding facial muscle movements patterns.
4. We present the detailed process of emotion labeling and the detection and recognition of AU.

Based on the above research, the video's coding of spontaneous expressions and micro-expressions is concluded and prospected.

Keywords: expressions, micro-expressions, action unit, coding, cerebral cortex, facial muscle

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Ming Yin,
Jiangsu Police Officer College, China
Xunbing Shen,
Jiangxi University of Chinese Medicine, China

***Correspondence:**
Su-Jing Wang
wangsujing@psych.ac.cn

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *28 September 2021*
Accepted: *17 November 2021*
Published: *04 January 2022*

Citation:
Dong Z, Wang G, Lu S, Li J, Yan W and Wang S-J (2022) Spontaneous Facial Expressions and Micro-expressions Coding: From Brain to Face. Front. Psychol. 12:784834. doi: 10.3389/fpsyg.2021.784834

1. INTRODUCTION

Emotions are the experience of a person's attitude toward the satisfaction of objective things and are critical to an individual's mental health and social behavior. Emotions consist of three components: subjective experience, external performance, and physiological arousal. The external performance of emotions is often reflected by facial expression, which is an important tool for expressing and recognizing emotions (Ekman, 1993). Expressing and recognizing facial expressions are crucial skills for human social interaction. It has been demonstrated by much research that inferences of

emotion fromfacial expressions are based on facial movement cues, i.e., muscle movements of the face (Wehrle et al., 2000).

Based on the knowledge of facial muscle movements, researchers usually described facial muscle movement objectively by creating facial coding systems, including Facial Action Coding System (FACS) (Friesen and Ekman, 1978), Face Animation Parameters (Pandzic and Forchheimer, 2003), Maximally Discriminative Facial Movement Coding System (Izard and Weiss, 1979), Monadic Phases Coding System (Izard et al., 1980), and The Facial Expression Coding System (Kring and Sloan, 1991). Depending upon the instantaneous changes in facial appearance produced by muscle activity, majority of these facial coding systems divide facial expressions into different action units (AUs), which can be used to perform quantitative analysis on facial expressions.

In addition to facial expression research based on psychology and physiology, artificial intelligence plays a vital role in affective computing. Notably, in recent years, with the rapid development of computer science and technology, the deep learning methods begin to be widely adopted to detect and recognize automatically by facial action units and makes automatic expression recognition possible in practical applications, including the field of security (Ji et al., 2006), clinical (Lucey et al., 2010), etc. The boom in expression recognition is attributed to many labeled expression datasets. For example, EmotioNet has a sample size of 950,000 (Fabian Benitez-Quiroz et al., 2016), which is large enough to fit the tens of millions of learned parameters in deep learning networks. The AU and emotion labels are the foundation for training the supervised deep learning networks and evaluating the algorithm performances. In addition, many algorithms are developed based on AU because of its importance (Niu et al., 2019; Wang et al., 2020).

However, the researchers found that ordinary facial expressions, i.e., macro-expressions, can not reflect a person's true emotions all the time. By contrast, the emergence of micro-expression has been considered as a significant clue to reveal the real emotion of humans. Studies have demonstrated that people would show micro-expressions in high-risk situations when they try to hide or suppress their genuine subjective feelings (Ekman and Rosenberg, 1997). Micro-expressions are brief, subtle, and involuntary facial expressions. Unlike macro-expression, micro-expression lasts only 1/25–1/5 s (Yan et al., 2013).

Micro-expression spotting and recognition have played a vital role in defense, suicide intervention, and criminal investigation. The AU-based study has also contributed to micro-expressions analysis. For instance, Davison et al. (2018) created an objective micro-expression classification system based on AU combinations; Xie et al. (2020) proposed an AU-assisted graph attention convolutional network for micro-expression recognition. Micro-expression has the characteristics of short duration and subtle movement amplitude, which causes that the manual annotation of ME videos requires the data processing personnel to view the video sample frame by frame slowly and attentively. Accordingly, long working hours increase the risk of errors. Furthermore, the current sample size of micro-expressions is still relatively small due to the difficulty of elicitation and annotation.

The prevailing annotation method is to annotate the AU according to the FACS proposed by Ekman et al. (Friesen and Ekman, 1978). FACS is the most widely used face coding system, and the manual is over 500 pages long. The manual covers Ekman's detailed explanation of each AU and its meaning, providing schematics and possible combinations of AUs. However, when AU is regarded as one of the criteria for classifying facial expressions (macro-expressions and micro-expressions), a FACS-certified expert is generally required to perform the annotation. The lengthy manual and the certification process have raised the barrier for AU coders.

Therefore, this paper focuses on macro-expression or micro-expression that responds to genuine emotions and analyzes the relationship between the cerebral cortex, which controls facial muscle movements, facial muscles, action units, and expressions. We theoretically deconstruct AU coding based on these analyses, systematically highlight the specific regions for each emotion. Finally, we provide an annotation framework for the annotator to facilitate the AU coding, expression labeling, and emotion classification.

This paper is an extended version of our ACM International Conference on Multimedia(ACM MM) paper (Zizhao et al., 2021), in which we make a brief guide to coding for spontaneous expressions and micro-expressions in video, and make the beginner to code get started as quickly as possible. In this paper, We discuss in further detail the principles of facial muscle movement from the brain to the face. Specifically, we show the cortical network system of facial muscle movement, introduce the neural pathways of the facial nucleus that control facial muscles, and the influence of other motor systems on the motor properties of the face. Secondly, we explain the relationship between AU and the six basic emotions with a physiological explanation. Finally, the coding of spontaneous expressions and micro-expressions is summarized in emotion label and AU detection and recognition research.

The following of this article is organized as follows: section 2 introduces the relationship between AU and emotions through the analysis of 5,000 images in RAF-AU database; section 3 demonstrates the nervous system of facial muscle movement; section 4 describes the muscles groups targeting the facial expression; section 5 exhibits the process of emotion labeling; section 6 shows detection and recognition research of AU; section 7 presents our conclusion and perspective on coding for spontaneous expressions and micro-expressions in videos.

2. ACTION UNITS AND EMOTIONS

Human muscle movements are innervated by nerves, and the majority of facial muscle movements are controlled by the seventh nerve in the brain, the facial nerve (Cranial Nerve VII, CN VII). The CN VII is divided into five branches, including the *temporal* branch, *zygomatic* branch, *buccal* branch, *marginal*

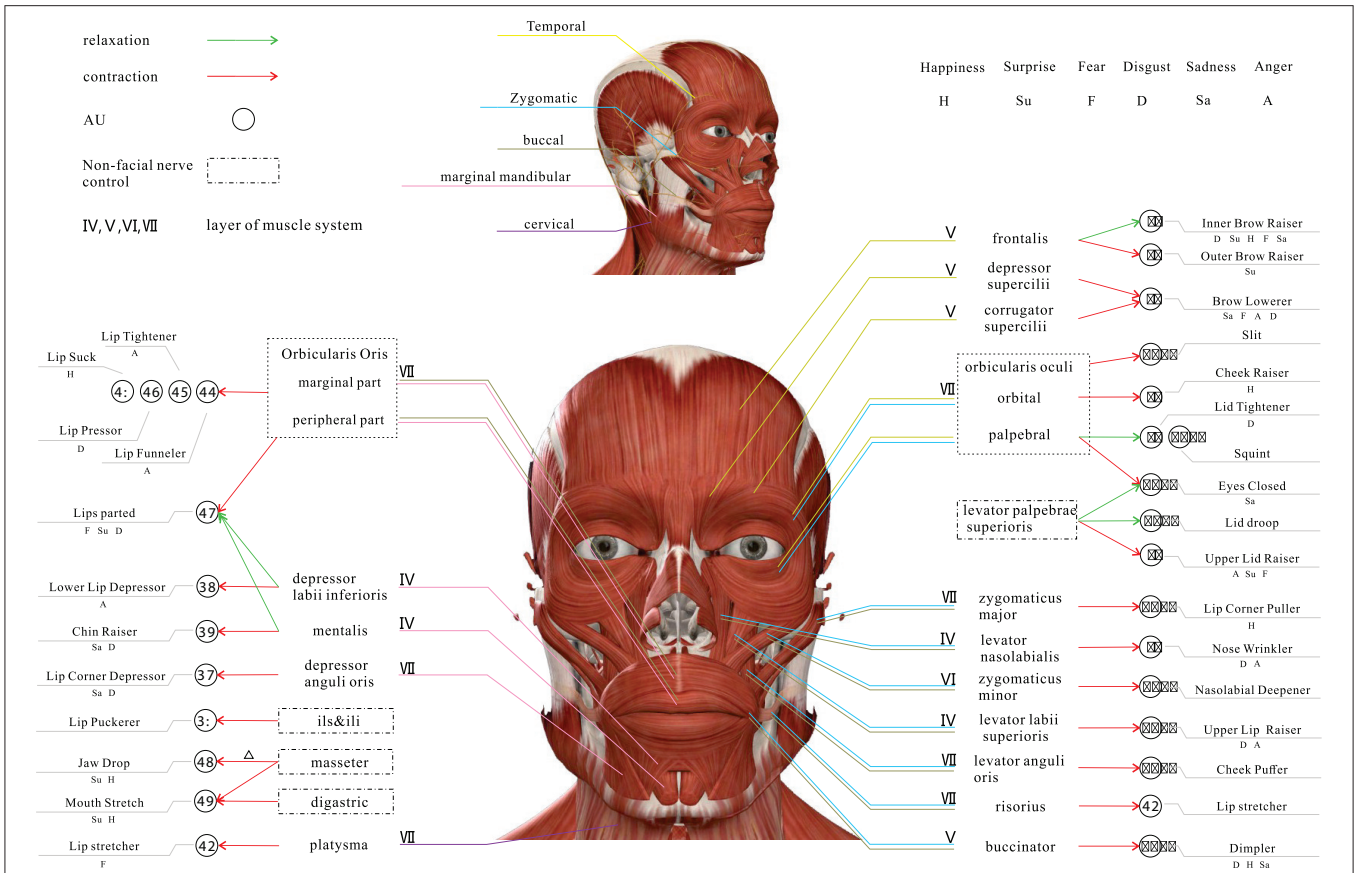


FIGURE 1 | An overview of relationships of facial muscle, AU and emotion based on facial nerve (Zizhao et al., 2021).

mandibular branch and *cervical* branch (Drake et al., 2009). These branches are illustrated in the upper part of **Figure 1**.

The *temporal* branch of the CN VII is located in the upper and anterior part of the auricle and innervates the *frontalis*, *corrugator supercilii*, *depressor supercilii*, *orbicularis oculi*. The *zygomatic* branch of the CN VII begins at the *zygomatic bone* and ends at the lateral orbital angle, innervates the *orbicularis oculi* and *zygomaticus*. The *buccal* branch of the CN VII is located in the inferior box area and around the mouth and innervates the *Buccinator*, *orbicularis oris* and other orbicularis muscles. The *marginal mandibular* branch of the CN VII is distributed along the lower edge of the mandible and ends in the descending *depressor anguli oris*, which innervates the lower lip and chin muscles. The *cervical* branch of the CN VII is distributed in the cervical region and innervates the *platysma*.

All facial muscles are controlled by one or two terminal motor branches of the CN VII, as shown in **Figure 1**. One or more muscle movements can constitute AUs, and different combinations of AUs show a variety of expressions, which ultimately reflect human emotions. Therefore, it is a complex process from muscle movements to emotions. We conclude the relationship between AU and emotion based on the images in the RAF-AU database (Yan et al., 2020) and the experience of professional coders.

2.1. The Data-Driven Relationship Between AU and Emotion

All the data, nearly 5,000 images used to analyze, are from RAF-AU (Yan et al., 2020). The database consists of face images collected from social networks with varying covering, brightness, resolution, and annotated through human crowdsourcing. Six basic emotions and one neutral emotion were used in the samples. Crowdsourced annotation is a method, which may help sag facial expressions in a natural setting by allowing many observers to tag a target heuristically. Finally, the probability score that the picture belongs to a specific emotion is calculated. The database contains about 200,000 facial expressions labeling because that about 40 independent observers tagged each image. It should be noted that although the source image materials are diverse, the judging group of raters is relatively narrow because the taggers are all students.

The corresponding annotation contains both the expert's AU labels and the emotion score obtained from the crowdsourcer's label statistics for each image. We analyzed only the contribution of AUs to the six basic emotions with two methods. One method is to take the highest score as the emotion of the image and then combine it with the labeled AU. In this method, repeated combinations must be removed to avoid the effect on the results due to the predominance of one sample type, i.e., to

mitigate the effect of sample imbalance. Another is to count the weighted sum of the contributions of all AUs to the six emotions without removing repetitions. The pseudocode details of these two methods are shown in Algorithms 1 and 2. **Tables 1**, **2** list the Top 10 AUs contributing to the six basic emotions, respectively. From **Table 1**, it can be seen that the contribution of AU25 is very high in the six basic emotions, which makes no sense because the movement of opening the corners of the mouth in AU25 is caused by the relaxation of the lower lip muscles, the relaxation of the genital muscles, and the orbicularis oris muscle. According to our subjective perception, AU25 rarely appears when we have three emotions: happiness, sadness, and anger. The abnormal top statistical data in **Table 1** may be caused by the shortcomings of crowdsourced annotations, i.e., the subjective tendency or random labeling of some individuals.

Algorithm 1

Initialization: AU's contribution array to emotions $C[6][M] = \{0\}$
M: Max AU number, N: Number of samples, $i = 0$.
repeat
  $i \leftarrow i + 1$
  Split the AU combination into a single AU set
  Take the maximum score of the six emotions as the emotion of the sample, defined as E
  if the combination of AU and emotion E first appears **then**
    Add the emotion score of the sample to the emotion AU
  end if
until $i > N$
Calculate the proportion of AU in each emotion
Sort C in descending order

Output: Contribution array C

Algorithm 2

Initialization: AU's contribution array to emotions $C[6][M] = \{0\}$
M: Max AU number, N: Number of samples, $i = 0$.
repeat
  $i \leftarrow i + 1$
  Split the AU combination into a single AU set
  Add the score of each emotion in the sample to C
until $i > N$
Sort C in descending order

Output: Contribution array C

However, there is room for improvement in the results obtained through the above data-driven approaches. The data-driven results can be affected by many aspects. Primarily, by the data source, such as the possible homogeneity of the RAF-AU database (number of subjects, gender, race, age, etc.), the uneven distribution of the samples, and the subjective labels based on human perception resulting from crowdsourcing annotation. Furthermore, the analysis method we

TABLE 1 | Top 10 of AU's contribution to the six basic emotions (Method 1).

Emotion	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score
Happiness	25	0.1577	12	0.1424	10	0.0725	1	0.0649	6	0.0616	2	0.0565	26	0.0506	9	0.0498	4	0.0472	27	0.0447
Surprise	25	0.1760	1	0.1093	5	0.1088	2	0.0962	26	0.0772	12	0.0683	10	0.0499	27	0.0499	16	0.0473	4	0.0452
Anger	25	0.1454	10	0.1118	4	0.1034	9	0.0997	16	0.0701	12	0.0506	27	0.0502	5	0.0480	26	0.0402	7	0.0365
Fear	25	0.1669	12	0.0997	1	0.0873	27	0.0866	5	0.0835	4	0.0742	10	0.0734	16	0.0703	2	0.0580	26	0.0479
Disgust	4	0.1303	25	0.1226	10	0.1199	9	0.0782	17	0.0611	1	0.0488	12	0.0488	7	0.0470	26	0.0448	6	0.0398
Sadness	4	0.1526	25	0.1249	10	0.0745	1	0.0689	17	0.0578	12	0.0566	26	0.0505	15	0.0455	9	0.0455	16	0.0375

TABLE 2 | Top 10 of AU's contribution to the six basic emotions (Method 2).

Emotion	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score	AU	Score
Happiness	12	0.2312	35	0.2059	19	0.2015	2	0.1746	6	0.1635	28	0.1598	30	0.1543	27	0.1513	26	0.1470	14	0.1470
Surprise	2	0.3742	5	0.3625	35	0.3186	1	0.3051	26	0.2859	27	0.2832	21	0.2783	34	0.2721	28	0.2588	25	0.2398
Anger	9	0.3387	33	0.3333	16	0.2726	23	0.2662	7	0.2604	10	0.2549	30	0.2521	24	0.2346	29	0.2332	32	0.2328
Fear	20	0.2500	27	0.2054	33	0.1944	5	0.1854	16	0.1813	2	0.1674	1	0.1631	12	0.1444	30	0.1398	21	0.1307
Disgust	24	0.3112	32	0.3026	17	0.2942	15	0.2789	19	0.2642	14	0.2622	7	0.2484	4	0.2387	18	0.2377	9	0.2325
Sadness	39	0.4294	15	0.2810	43	0.2543	17	0.2355	4	0.1957	14	0.1686	6	0.1621	28	0.1598	7	0.1490	24	0.1467

TABLE 3 | The relationship between AU and emotion.

Emotion	AU
Happiness	1, **6**, **12**, 14, 26, 27, **28**
Surprise	1, **2**, 5, 25, 26, 27
Anger	4, 5, 9, 10, **16**, **22**, **23**
Fear	1, 4, 5, **20**, 25
Disgust	1, 4, **7**, 9, 10, 14, 15, 17, **24**, 25
Sadness	1, 4, 14, 15, 17, **43**

The bolded AU in the table indicates that the AU is only associated with the corresponding specific emotion, and not with other emotions.

used is based on a maximum value and probability weighting. Although straightforward, such analytical approaches represent the contribution of AU to the six basic emotions, are less comprehensive. More analysing methods are also needed in dealing with unbalanced data. In response to the challenges posed by data and analytical methods to data-driven methods, we could combine data-driven and experience-driven research methods. In this way, we could draw on the objectivity of data-driven and the robustness of experience-driven to realize the construction of the AU coding system for macro-expressions/micro-expressions.

2.2. The Experience-Driven Relationship Between AU and Emotion

There usually exists difficulties for the data-driven methods to analyze with theoretic basis. For example, the typical "black box" characteristic brings the problem of poor interpretability. Meanwhile, the results by data-driven are highly dependent on the quality (noiseless) and quantity (wide and massive) of the database. By comparison, the experience-driven method, based on the knowledge of coding and the common sense, is a way to label emotion. Three advantages are listed below: (1) The experience-driven method can help reduce the noise by using coding and common sense knowledge. (2) Experience-driven method has a reliable theory as a support, making the results convincing. (3) Experience-driven can often solve most universal laws with just a few simple formulas. Therefore, we combine experience-driven and data-driven methods to get the final AU and emotional relationship summary table, as shown in **Table 3**, by using their respective advantages.

Specifically, firstly, based on the analysis results listed in **Tables 1**, **2** (data-driven), the preliminary selection is made by comparing the description and legend of each AU in FACS, and combining with the meaning of emotion. We obtained a preliminary AU system for emotion. Then, with large amounts of facial expression images on search engines such as Google and Baidu, the preliminary AU system for emotion was screened by eliminating non-compliant AU in these images. In this way, the ultimate relationship is shown in **Figure 1** and **Table 3**.

Based on **Table 3**, we assume that the sets of six basic emotions containing AU are S_1, S_2, S_3, S_4, S_5, and S_6. Let $S = \{S_1, S_2, S_3, S_4, S_5, S_6\}$, then

$$Q_i = S_i \backslash \bigcap_j S_j \tag{1}$$

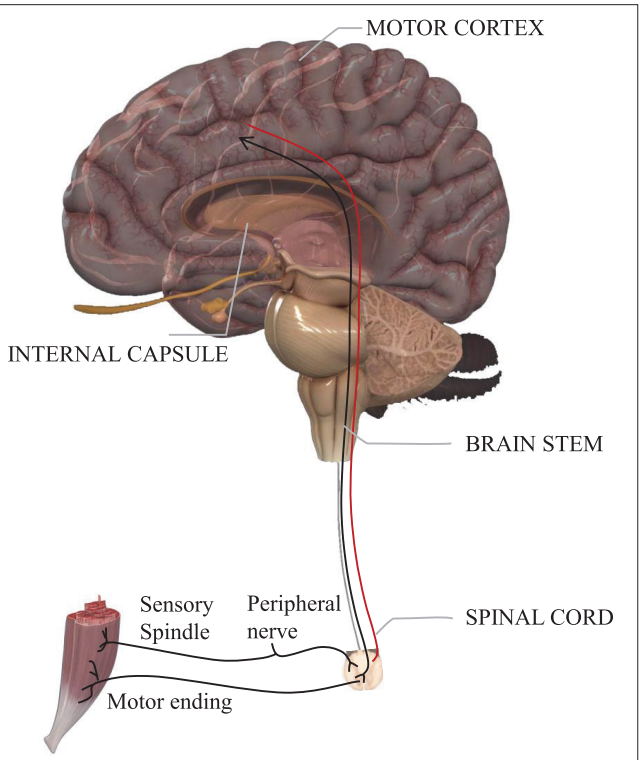


FIGURE 2 | Motor neurons from the brain to the muscle.

where $i = 1, ..., 6$, and $j = \{1, ..., i-1, i+1, ..., 6\}$. $\bigcap$ is the intersection operation of the set. \ represents the set of symmetric difference, for example, we assume that $A = \{3, 9, 14\}$, $B = \{1, 2, 3\}$, then $A \backslash B = \{9, 14\}$. Q_i denotes the AU set that is exclusive to the S_i emotion.

According to **Table 3**, we can infer that the bloded AU is only associated with the corresponding specific emotion, and not with other emotions. See **Table 3** in bold for details. Therefore, we can conclude that the appearance of certain AU represents related emotion. For instance, if AU20 appears, we assume that fearful emotion emerges.

3. COMPLEX CORTICAL NETWORKS OF FACIAL MOVEMENT

The facial motor system is a complex network of specialized cortical areas dependent on multiple parallel systems, voluntary/involuntary motor systems, emotional systems, visual systems, etc., all of which are anatomically and functionally distinct and all of which ultimately reach the facial nucleus to govern facial movements (Cattaneo and Pavesi, 2014). The nerve that emanates from the facial nucleus is the facial nerve. The facial nerve originates in the brainstem, and its pathway is commonly divided into three parts: intra-cranial, intra-temporal, and extra-cranial (see **Figure 2**).

3.1. Facial Nucleus Controls Facial Movements

The human facial motor nucleus is the largest of all motor nuclei in the brainstem. It is divided into two parts: upper and lower. The upper part is innervated by the motor areas of the cerebral cortex bilaterally and sends motor fibers to innervate the muscles of the ipsilateral upper face; the lower part of the nucleus is innervated by the contralateral cerebral cortex only and sends motor fibers to innervate the muscles of the ipsilateral lower face. It contains around 10,000 neurons and consists mainly of the cell bodies of motor neurons (Sherwood, 2005).

A large number of neurons in the facial nucleus provides the anatomical basis for the various reflex responses of the facial muscles to different sensory modalities. For example, in the classic study by Penfield and Boldrey, it was found that the sensation of facial movement and the urge/desire to move the face was elicited by electrical stimulation of the cerebral cortex, causing movement of different parts of the face, as well as occurring in the absence of movement. Movements of the eyebrows and forehead were less frequent than those of the eyelids, and movements of the lips were the most frequent (Penfield and Boldrey, 1937).

Another way to assess the mechanism of inhibition within the cerebral cortex is to study the cortical resting period of transcranial magnetic stimulation. The cortical resting period is a period of inactivity called the silent period, when spontaneous muscle contraction is followed by a pause in myoelectric activity after the generation of motor evoked potentials by transcranial magnetic stimulation in the corresponding functional areas of the cerebral cortex. Studies on facial muscle movements have found that the silent period occurs after motor-evoked potentials in the pre-activated lower facial muscles (Curra et al., 2000), (Paradiso et al., 2005).

3.2. Cortical Systems Controls Facial Movement

The earliest studies on facial expressions date back to the nineteenth century. For example, the French neurophysiologist Duchenne de Boulogne (1806–1875) used electrical stimulation to study facial muscle activity (Duchenne, 1876). He used this experimental method to define for the first time expressions in different emotional states, including attention, relaxation, aggression, pain, happiness, sadness, cry, surprise, and fear, showing that each emotional state is expressed with specific facial muscle activity. Also, Duensing observed that there might be different neural structures involved between involuntary and emotional facial movements. Duensing's theory also influenced Charles Darwin's book *The expression of the emotions in man and animals* (1872) (Darwin, 1872).

Meanwhile, facial movements depend on multiple parallel systems that ultimately all reach the facial nucleus to govern facial movements. We focus on facial movements of expressions or micro-expressions, and two systems related to them have been discussed here: the voluntary/involuntary motor system and the emotional system.

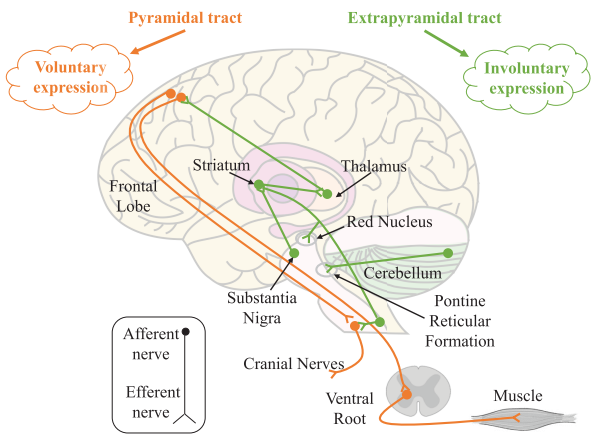


FIGURE 3 | The somatic motor system. Voluntary and involuntary expressions are controlled by the pyramidal tract (orange trajectory) and extrapyramidal tract (green trajectory), respectively.

3.2.1. The Somatic Motor System

According to the form of movement of skeletal muscles, body movements are divided into voluntary and involuntary movements. Voluntary movements are emitted from the cortical centers of the brain and are movements executed according to one's consciousness, characterized by sensation followed by movement; involuntary movements are spontaneous movements that are not controlled by consciousness, such as chills. Meanwhile, the neuroanatomical distinction between voluntary and involuntary expressions has been established in clinical neurology (Matsumoto and Lee, 1993). Voluntary expressions are thought to emanate from the cortical motor tract and enter the facial nucleus through the pyramidal tract; involuntary expressions originate from innervation along the external pyramidal tract. See **Figure 3**.

Most facial muscles are overlapping, rarely contracting individually, and usually being brought together in synergy. In particular, these synergistic movements always occur during voluntary movements. For example, the *orbicularis oculi* and *zygomaticus* have a synergistic effect during the voluntary closure of the eyelid. In contrast, asymmetric movements of the face are usually thought to be the result of facial nerve palsy or involuntary movements (Devoize, 2011), for example, simultaneous contraction of the ipsilateral *frontalis* and *orbicularis oculi*, i.e., raising the eyebrows and closing the eyes at the same time. Babinski, a professor of neurology, considers that combined movements such as these cannot be activated by central mechanisms and cannot be replicated by volition. Therefore, facial asymmetry is always considered to be one of the characteristics of micro-expressions.

3.2.2. The Emotional Motor Systems

Facial expressions are stereotyped physiological responses to specific emotional states, controlled by the voluntary and somatic systems controlled by the emotion-motor system (Holstege, 2002). Expression is only one of the somatic motor components of emotion, which also includes body posture and voice changes. However, in humans, facial expressions are external manifestations of emotions and are an essential part of human non-verbal communication (Müri, 2016), and a significant factor in the cognitive process of emotion. The emotion-motor pathway originates in the gray matter around the amygdaloid nucleus, lateral hypothalamus, and striatum. Most of these gray matter projects, in turn to the reticular formation to control facial premotor neurons, and a few project to facial motor neurons to control facial muscles directly.

In the study of traumatic facial palsy, a separation between the emotional motor system and the voluntary motor system at the brainstem level was found between facial movements (Bouras et al., 2007). It indicates that these two systems are entirely independent before the facial nucleus. This could be the reason why it is not possible to generate true emotional expression through volition. Therefore, emotion elicitation is required to produce behavioral (expression/micro-expression) responses through stimuli that induce emotion of the subject. It is relatively such expressions that have emotional significance. Moreover, there is also a strong correlation between the different activity patterns among facial muscles and the emotional valence of external stimuli (Dimberg, 1982). Similarly, the emotional motor system and the voluntary motor system interact and confront each other, and the results of this interaction are usually non-motor (e.g., motor dissonance) (Bentsianov and Blitzer, 2004).

Similar to the involuntary motor system, there is a small degree of asymmetry in the facial movements produced by the emotional motor system. However, the conclusions of this asymmetry are controversial. Many studies in brain-injured, emotionally disturbed, or normal subjects have shown that the majority of emotion expression, recognition, and related behavioral control is in the right hemisphere; that the right hemisphere dominates in the production of basic emotions, i.e., happiness and sadness, and the left hemisphere dominates in the production of socially conforming emotions, i.e., jealousy and complacency; and that the right hemisphere specializes in negative emotions while the left hemisphere specializes in positive emotions (Silberman and Weingartner, 1986).

4. THE SPECIFICITY OF THE RELATIONSHIP BETWEEN FACIAL MUSCLE AND EMOTIONS

According to **Figure 1** and **Table 3**, we make further analysis of facial muscle and emotions to guarantee that each emotion can be targeted at a specific AU.

4.1. The Muscle That Classifies Positive and Negative Emotions

The basic dimensions for emotions are the two main categories, positive and negative emotions. Positive emotions are associated with the satisfaction of demand and are usually accompanied by a pleasurable subjective experience, which can enhance motivation and activity. By comparison, negative emotions represent a negative or aversive emotion such as sadness, disgust, etc., by an

individual. The *zygomaticus* is controlled by the *zygomatic* branch of the CN VII. The *zygomatic* branch of the CN VII begins at the *zygomatic bone* and ends at the lateral orbital angle, innervates the *orbicularis oculi* and *zygomaticus*. The *zygomaticus* includes the *zygomaticus major* and the *zygomaticus minor*. The *zygomaticus major* begins in the *zygomatic bone*, and ends at the *angulus oris*. The responsibility of *zygomaticus* is to pull the corners of the mouth back or up to smile. The *zygomaticus minor* begins in the lateral profile of *zygomatic bone*, and ends at the *angulus oris*. The function is to raise the upper lip, such as grinning.

The *corrugator supercilii* begins in the medial end of the arch of the eyebrow and ends at the skin of the eyebrow, which is located at the *frontalis* and *orbicularis oculi* muscles back. It is innervated by the *temporal* branch of the CN VII. The contraction of *corrugator supercilii* depresses the brow and generates a vertical frown.

It has been found that the *corrugator supercilii* induced by unpleasant stimuli is more intense than that induced by pleasant stimuli, and the *zygomaticus* is more intense by pleasant stimuli (Brown and Schwartz, 1980). In a word, pleasant stimuli usually lead to greater electromyography(EMG) activity in the *zygomaticus*, whereas unpleasant stimuli lead to greater EMG activity in the frowning muscle (Larsen et al., 2003).

In the AU encoding process, *zygomaticus* activity and *corrugator supercilii* activity can reliably recognize positive emotion and negative emotion respectively. This conclusion also supports the discrete emotion theory (Cacioppo et al., 2000). For example, oblique lip-corner contraction (AU12), together with cheek raising (AU6) can reliably signals enjoyment (Ekman et al., 1990), while brow furrowing (AU4) tends to signal negative emotion (Brown and Schwartz, 1980). The correlation between emotion and facial muscle activity can be summarized as follows: (1) The main muscle area of the zygomatic is a reliable discriminating area for positive emotion; (2) The corrugator muscle area is a reliable identification area for negative emotion.

As shown in **Figure 1**, AU4, which is controlled by contraction of the depressor supercilii and corrugator supercilii, is present in all negative emotions. Most of the AU associated with happiness is controlled by the zygomatic branch, which mainly innervates the zygomatic muscle. Therefore, the coder should focus more on the cheekbones, i.e., the middle of the face and the mouth if they want to catch the expressions or micro-expressions elicited by positive stimuli. For those elicited by negative stimuli, the coder should focus more on the forehead, i.e., the eyebrows and the upper part of the face.

4.2. Further Specific Classification of the Muscles of Negative Emotions

In the six basic emotions, the negative emotions usually manifested as sadness, disgust, anger and fear, which are all highly associated with the *corrugator supercilii*, the brow and upper region. Therefore, in combination with the lower face, launching a further distinguishing of these four emotions from facial muscles is crucial for emotional classification.

4.2.1. Muscle Group Specific for Sadness

The *depressor anguli oris* begins at the genital tubercle and the oblique line of the mandible, ends at the *angulus oris*. It is innervated by the *buccal* branch of the CN VII and the *marginal mandibular* branch. It serves to depress the *angulus oris*. The study found that when the participants produced happy or sad emotions by recalling, the facial EMG of the frowning muscle in the sadness was significantly higher than that in the happiness (Schwartz et al., 1976). This suggests that the combination of *corrugator supercilii* and textitdepressor anguli oris may be effective in classifying sad emotions.

4.2.2. Muscle Group Specific for Fear

The *frontalis* begins in the *epicranial aponeurosis*, and extends to terminates in the skin of the brow and nasal root, and into the *orbicularis oculi* and *corrugator supercilii*. It is innervated by the *auricular posterior* nerve and the *temporal* branch of the CN VII. The *frontalis* is a vertical movement that serves to raise the eyebrows and increase the wrinkles at the level of the forehead, often seen in expressions of surprise. In expression coding, the action of raising the inner brow is coded as AU1. The *orbicularis oculi* begin in the pars nasalis ossis frontalis, the frontal eminence of the upper skeleton and the medial palpebral ligament, surrounds the orbit and ends at the adjacent muscles. Anatomically it is divided into the orbital and palpebral portions. It is innervated by the *temporal* and *zygomatic* branches of the CN VII. The function is to close the eyelid. In the study of the positive intersection of facial expressions and emotional stimuli, the researchers asked the subjects to maintain the fear feature of facial muscles, involving *corrugator supercilii*, *frontalis*, *orbicularis oculi*, and *depressor anguli oris* (Tourangeau and Ellsworth, 1979).

4.3. Distinguish the Special Muscle of Surprise

Surprise is an emotion that is independent of positive and negative emotions. For example, pleasant surprise, shock, etc., fall within the category of surprise. The study of people's surprise emotion has been started since Darwin (Darwin, 1872), and it is ubiquitous in social life and belongs to one of the basic emotions. Moreover, surprise can be easily induced in the laboratory.

Landis conducted the earliest study of surprising expressions (Landis, 1924). About 30% of people raised their eyebrows, and about 20% of people's eyes widened when a firecracker landed on the back of the subject's chair. Moreover, in discussing the evidence for a strong dissociation between emotion and facial expression, the research measured facial movements associated with surprise twice (see experiments 7 and 8). When subjects experienced surprise, the facial movements were described as frowning, eye-widening, and eyebrow raising (Reisenzein et al., 2006). Also, in exploring the distinction in dynamics between genuine and deliberate expressions of surprise, it was found that all expressions of surprise consisted mainly of raised eyebrows and eyelid movements (Namba et al., 2021). The facial muscles involved in these movements were: *corrugator supercilii*, *orbicularis oculi*, and *frontalis*. Details are described in section 4.2. The AUs

associated with these facial muscles and movements include AU2, AU4, and AU5. As shown in **Figure 1** and **Table 3**.

5. EMOTION LABEL

Expressions are generally divided into six basic emotions, happiness, disgust, sadness, fear, anger and surprise. Micro-expressions are usually useful when there is a small negative micro-expression in a positive expression, such as "nasty-nice." For micro-expressions, therefore, they are usually divided into four types, positive, negative, surprise and other. To be specific, positive expression includes happy expressions, which is relatively easy to be induced because of some obvious characteristics. Negative expressions like disgust, sadness, fear, anger, etc., are relatively difficult to distinguish, but they are significantly different from positive expressions. Meanwhile, surprise, which expresses unexpected emotions that can only be interpreted according to the context, has no direct relationship with positive or negative expressions. The additional category, "Others," indicates expressions or micro-expressions that have ambiguous emotional meanings can be classified into the six basic emotions.

Emotion labeling requires the consideration of the components of emotions. Generally speaking, we need to take three conditions into account for the emotional facial action: AU label, elicitation material, and the subject's self-report of this video. Meanwhile, the influence of some habitual behaviors should be eliminated, such as frown when blinking or sniffing.

5.1. AU Label

For AU annotation, the annotator needs to be skilled in the facial coding system and watches the videos containing facial expression frame by frame. The three crucial frames for AU are the start frame (onset), peak frame (apex), and end frame (offset). Then we can get the expression time period for labeling AU. The start frame represents the time where the face changes from neutral expression. The peak frame is the time with the greatest extent of that facial expression. The end frame is the time where the expression ends and returns to neutral expression.

5.2. Elicitation Material

Spontaneous expressions have high ecological validity compared to posed expressions and are usually elicited with elicitation material. In psychology, researchers usually use different emotional stimuli to induce emotions with different properties and intensities. A stimulus is an important tool for inducing experimental emotions. We use stimuli materials, usually from existing emotional materials databases, to elicit different types of emotions of the subject.

5.3. Subject's Self-Report of This Video

After watching the video, the subjects need to evaluate the video according to their subjective feelings. This self-report is an effective means of testing whether emotions have been successfully elicited.

5.4. Reliability of Label

In order to ensure the validity or reliability of data annotation, the process of emotion labeling usually requires the participation of two coders and the calculation of inter-coders confidence must exist in a proper range. The formula is as follows 2:

$$R = \frac{N \times \left| \bigcap_{i=1}^{N} C_i \right|}{\left| \bigcup_{i=1}^{N} C_i \right|} \tag{2}$$

where C_i represents the set of labeled emotions in the facial expression images by coder i($2 \leq i \leq N$), respectively, and $|\cdot|$ represents the number of labeled emotion in the set after the intersection or merge operations.

The reason is that in the process of annotation, the coders must make subjective judgments based on their expertise. In order to make these subjective judgments as similar as possible to the perceptions of the majority of people, inter-rater reliability is of paramount importance. Inter-rater reliability is a necessary step for the validity of content analysis (emotion labeling) research. The conclusions of data annotation are questionable or even meaningless without this step.

It is mentioned above show that emotion labeling is a complex process, which needs coders to have the expertise with both psychology and statistics, increasing the threshold for being a coder. So we tried to find a direct relationship between emotions and facial movements to identify specific regions of emotions, as shown in **Figure 1**.

6. DETECTION AND RECOGNITION OF AU

Facial muscles possess complex muscle patterns. Researchers have developed facial motion coding systems, video recordings, electromyography, and other methods to study and analyze facial muscle contractions.

The FACS coding system developed by Friesen and Ekman (1978) is based on the anatomical structure of facial muscles and is composed of all visible facial motion units AU under different intensities. So far, more than 7,000 AU combinations have been found in a large number of expressions. However, even for FACS coders, such labeling is time-consuming and labor-intensive.

Since then, the researchers have made some automatic coding attempts. For example, by analyzing the images in the video, it can automatically detect, track and classify the AU or AU combination that causes facial expressions (Lien et al., 2000). Nevertheless, unfortunately, image quality is especially susceptible to illumination, which, to some extent, limits such visible spectrum imaging technology.

To surmount such problem, researchers used facial electromyography, which is widely used in clinical research, to record AU muscle electrical activity (even visually imperceptible). This technique is susceptible to measuring the dynamics and strength of muscle contractions (Delplanque et al., 2009). However, there still exist some shortcomings: objective factors such as electrode size and position, epidermal cleanliness, and muscle movement methods, may interfere with the accuracy of the final experimental results and cause deviations

in experimental conclusions. What's more, the number of muscles related to AU should theoretically be as much as that of electrodes, which also makes EMG a severe limitation as a non-invasive method.

Additionally, thermal imaging technology has also been applied to the study of facial muscle contraction and AU. Research has demonstrated that muscle contraction can cause skin temperature to increase (González-Alonso et al., 2000). For this reason, Jarlier et al. took thermal imaging as a tool to investigate specific facial heat patterns associated with the production of facial AUs (Jarlier et al., 2011). Therefore, thermal images can be used to detect and evaluate specific facial muscle thermal patterns (the speed and intensity of muscle contraction). Furthermore, this method avoids the lighting problems encountered when using traditional cameras and the influence of electrodes when using EMG.

7. CONCLUSION

In this article, with the help of statistical analysis, a data-driven approach is used to obtain a quantifiable system between AU and emotion. And then, we further obtain a robust correspondence system between AU and emotion by combining with an empirically driven comparison to actual data (from the web). In the next part, we introduce the cortical system that controls facial movements. Moreover, the physiological theoretical support for AU labeling of emotions was obtained by adding facial muscle movements. Finally, we sort out the process of emotion label and the research of AU recognition and detection. The main manifestations are listed below:

Based on the **Figure 1** and **Table 3**, the theories of sections 3 and 4, we sum up the main points of coding in the article:

1. When corners of lips pulled up (AU12) appears, it can be coded as a positive emotion, i.e., happy; In addition, cheek rise (AU6), lip suck (AU28) are both happy specific action units and can also be coded as positive emotions;
2. When brow rise (AU2) is present, it can be coded as surprise;
3. When frown (AU4) is present, it can be coded as a negative emotion;
4. It can be coded as anger when gnashing (AU16, AU22 or AU23), which only occur in the specific action units, appear;
5. When movements of the eyebrows (AU1 and AU4), eyes (AU5) and mouth (AU25) are present simultaneously, they can be coded as fear;
6. It can be coded as disgusted when the specific action unit of disgust, lower eyelid rise (AU7), mouth tightly closed (AU24), is present;
7. It can be coded as sad when frown (AU4) and eyes wide open (AU5) are present at the same time; eyes closed (AU43) is the specific action unit for sadness and can also be coded as sadness.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://whdeng.cn/RAF/model3.html#:$~$:text=a%20Real-world%20Affective%20Faces%20Action%20Unit%20%28RAF-AU%29database%20that,to%20annotating%20blended%20facial%20expressions%20in%20the%20wild.

AUTHOR CONTRIBUTIONS

ZD has contributed the main body of text and the main ideas. GW was responsible for empirical data analysis. SL was responsible for constructing the partial research framework. JL has contributed to the construction of the text and refinement of ideas and provided extensive feedback and commentary. S-JW led the project and acquired the funding support. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from National Natural Science Foundation of China (61772511, U19B2032), National Key Research and Development Program (2017YFC0822502), and China Postdoctoral Science Foundation (2020M680738).

ACKNOWLEDGMENTS

The authors would like to thank Xudong Lei for linguistic assistance during preparation of this manuscript.

REFERENCES

Bentsianov, B., and Blitzer, A. (2004). Facial anatomy. *Clin. Dermatol.* 22, 3–13. doi: 10.1016/j.clindermatol.2003.11.011

Bouras, T., Stranjalis, G., and Sakas, D. E. (2007). Traumatic midbrain hematoma in a patient presenting with an isolated palsy of voluntary facial movements: case report. *J. Neurosurg.* 107, 158–160. doi: 10.3171/JNS-07/07/0158

Brown, S.-L., and Schwartz, G. E. (1980). Relationships between facial electromyography and subjective experience during affective imagery. *Biol. Psychol.* 11, 49–62. doi: 10.1016/0301-0511(80)90026-5

Cacioppo, J., Berntson, G., Larsen, J., Poehlmann, K., Ito, T., Lewis, M., et al. (2000). *Handbook of Emotions*. New York, NY: The Guilford Press.

Cattaneo, L., and Pavesi, G. (2014). The facial motor system. *Neurosci. Biobehav. Rev.* 38, 135–159. doi: 10.1016/j.neubiorev.2013.11.002

Curra, A., Romaniello, A., Berardelli, A., Cruccu, G., and Manfredi, M. (2000). Shortened cortical silent period in facial muscles of patients with cranial dystonia. *Neurology* 54, 130–130. doi: 10.1212/WNL.54.1.130

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals by Charles Darwin*. London: John Murray. doi: 10.1037/10001-000

Davison, A. K., Merghani, W., and Yap, M. H. (2018). Objective classes for micro-facial expression recognition. *J. Imaging* 4:119. doi: 10.3390/jimaging4100119

Delplanque, S., Grandjean, D., Chrea, C., Coppin, G., Aymard, L., Cayeux, I., et al. (2009). Sequential unfolding of novelty and pleasantness appraisals of odors: evidence from facial electromyography and autonomic reactions. *Emotion* 9:316. doi: 10.1037/a0015369

Devoize, J.-L. (2011). Hemifacial spasm in antique sculpture: interest in the "other babinski sign". *J. Neurol. Neurosurg. Psychiatry* 82, 26–26. doi: 10.1136/jnnp.2010.208363

Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology* 19, 643–647. doi: 10.1111/j.1469-8986.1982.tb02516.x

Drake, R., Vogl, A. W., and Mitchell, A. W. (2009). *Gray's Anatomy for Students E-book*. London: Churchill Livingstone; Elsevier Health Sciences.

Duchenne, G.-B. (1876). *Mécanisme de la physionomie humaine ou analyse électro-physiologique de l'expression des passions*. Paris: J.-B. Bailliére et fils.

Ekman, P. (1993). Facial expression and emotion. *Am. Psychol.* 48:384. doi: 10.1037/0003-066X.48.4.384

Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The duchenne smile: emotional expression and brain physiology: II. *J. Pers. Soc. Psychol.* 58:342. doi: 10.1037/0022-3514.58.2.342

Ekman, P., and Rosenberg, E. L. (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY: Oxford University Press.

Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). "Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 5562–5570. doi: 10.1109/CVPR.2016.600

Friesen, E., and Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3:5.

González-Alonso, J., Quistorff, B., Krustrup, P., Bangsbo, J., and Saltin, B. (2000). Heat production in human skeletal muscle at the onset of intense dynamic exercise. *J. Physiol.* 524, 603–615. doi: 10.1111/j.1469-7793.2000.00603.x

Holstege, G. (2002). Emotional innervation of facial musculature. *Movement Disord.* 17, S12–S16. doi: 10.1002/mds.10050

Izard, C. E., Huebner, R. R., Risser, D., and Dougherty, L. (1980). The young infant's ability to produce discrete emotion expressions. *Dev. Psychol.* 16:132. doi: 10.1037/0012-1649.16.2.132

Izard, C. E., and Weiss, M. (1979). *Maximally Discriminative Facial Movement Coding System*. University of Delaware, Instructional Resources Center.

Jarlier, S., Grandjean, D., Delplanque, S., N'diaye, K., Cayeux, I., Velazco, M. I., et al. (2011). Thermal analysis of facial muscles contractions. *IEEE Trans. Affect. Comput.* 2, 2–9. doi: 10.1109/T-AFFC.2011.3

Ji, Q., Lan, P., and Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Trans. Syst. Man Cybernet. A Syst. Hum.* 36, 862–875. doi: 10.1109/TSMCA.2005.855922

Kring, A. M., and Sloan, D. (1991). The facial expression coding system (FACES): a users guide. Unpublished manuscript. doi: 10.1037/t03675-000

Landis, C. (1924). Studies of emotional reactions. II. General behavior and facial expression. *J. Comp. Psychol.* 4, 447–510. doi: 10.1037/h0073039

Larsen, J. T., Norris, C. J., and Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and *corrugator supercilii*. *Psychophysiology* 40, 776–785. doi: 10.1111/1469-8986.00078

Lien, J. J.-J., Kanade, T., Cohn, J. F., and Li, C.-C. (2000). Detection, tracking, and classification of action units in facial expression. *Robot. Auton. Syst.* 31, 131–146. doi: 10.1016/S0921-8890(99)00103-7

Lucey, P., Cohn, J. F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., et al. (2010). Automatically detecting pain in video through facial action units. *IEEE Trans. Syst. Man Cybernet. B* 41, 664–674. doi: 10.1109/TSMCB.2010.2082525

Matsumoto, D., and Lee, M. (1993). Consciousness, volition, and the neuropsychology of facial expressions of emotion. *Conscious. Cogn.* 2, 237–254. doi: 10.1006/ccog.1993.1022

Müri, R. M. (2016). Cortical control of facial expression. *J. Comp. Neurol.* 524, 1578–1585. doi: 10.1002/cne.23908

Namba, S., Matsui, H., and Zloteanu, M. (2021). Distinct temporal features of genuine and deliberate facial expressions of surprise. *Sci. Rep.* 11:3362. doi: 10.1038/s41598-021-83077-4

Niu, X., Han, H., Shan, S., and Chen, X. (2019). "Multi-label co-regularization for semi-supervised facial action unit recognition," in *Advances in Neural Information Processing Systems 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC: Vancouver Convention Center; Curran Associates, Inc.).

Pandzic, I. S., and Forchheimer, R. (2003). *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. Chennai: John Wiley & Sons. doi: 10.1002/0470854626

Paradiso, G. O., Cunic, D. I., Gunraj, C. A., and Chen, R. (2005). Representation of facial muscles in human motor cortex. *J. Physiol.* 567, 323–336. doi: 10.1113/jphysiol.2005.088542

Penfield, W., and Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* 60, 389–443. doi: 10.1093/brain/60.4.389

Reisenzein, R., Bordgen, S., Holtbernd, T., and Matz, D. (2006). Evidence for strong dissociation between emotion and facial displays: the case of surprise. *J. Pers. Soc. Psychol.* 91, 295–315. doi: 10.1037/0022-3514.91.2.295

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., and Klerman, G. L. (1976). Facial expression and imagery in depression: an electromyographic study. *Psychosom. Med.* 38, 337–347. doi: 10.1097/00006842-197609000-00006

Sherwood, C. C. (2005). Comparative anatomy of the facial motor nucleus in mammals, with an analysis of neuron numbers in primates. *Anat. Rec. A* 287, 1067–1079. doi: 10.1002/ar.a.20259

Silberman, E. K., and Weingartner, H. (1986). Hemispheric lateralization of functions related to emotion. *Brain Cogn.* 5, 322–353. doi: 10.1016/0278-2626(86)90035-7

Tourangeau, R., and Ellsworth, P. C. (1979). The role of facial response in the experience of emotion. *J. Pers. Soc. Psychol.* 37:1519. doi: 10.1037/0022-3514.37.9.1519

Wang, S., Peng, G., and Ji, Q. (2020). Exploring domain knowledge for facial expression-assisted action unit activation recognition. *IEEE Trans. Affect. Comput.* 11, 640–652. doi: 10.1109/TAFFC.2018.2822303

Wehrle, T., Kaiser, S., Schmidt, S., and Scherer, K. R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *J. Pers. Soc. Psychol.* 78:105. doi: 10.1037/0022-3514.78.1.105

Xie, H.-X., Lo, L., Shuai, H.-H., and Cheng, W.-H. (2020). "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA), 2871–2880. doi: 10.1145/3394171.3414012

Yan, W.-J., Li, S., Que, C., Pei, J., and Deng, W. (2020). "RAF-AU database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations," in *Proceedings of the Asian Conference on Computer Vision* (Kyoto).

Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., and Fu, X. (2013). How fast are the leaked facial expressions: the duration of micro-expressions. *J. Nonverb. Behav.* 37, 217–230. doi: 10.1007/s10919-013-0159-8

Zizhao, D., Gang, W., Shaoyuan, L., Wen-Jing, Y., and Wang, S.-J. (2021). "A brief guide: code for spontaneous expressions and micro-expressions in videos," in *ACM International Conference on Multimedia* (Chengdu).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neurorobotics

METHODS
published: 12 January 2022
doi: 10.3389/fnbot.2021.824592

Progressive Multi-Scale Vision Transformer for Facial Action Unit Detection

Chongwen Wang and Zicheng Wang*

School of Computer Science, Beijing Institute of Technology, Beijing, China

Facial action unit (AU) detection is an important task in affective computing and has attracted extensive attention in the field of computer vision and artificial intelligence. Previous studies for AU detection usually encode complex regional feature representations with manually defined facial landmarks and learn to model the relationships among AUs *via* graph neural network. Albeit some progress has been achieved, it is still tedious for existing methods to capture the exclusive and concurrent relationships among different combinations of the facial AUs. To circumvent this issue, we proposed a new progressive multi-scale vision transformer (PMVT) to capture the complex relationships among different AUs for the wide range of expressions in a data-driven fashion. PMVT is based on the multi-scale self-attention mechanism that can flexibly attend to a sequence of image patches to encode the critical cues for AUs. Compared with previous AU detection methods, the benefits of PMVT are 2-fold: (i) PMVT does not rely on manually defined facial landmarks to extract the regional representations, and (ii) PMVT is capable of encoding facial regions with adaptive receptive fields, thus facilitating representation of different AU flexibly. Experimental results show that PMVT improves the AU detection accuracy on the popular BP4D and DISFA datasets. Compared with other state-of-the-art AU detection methods, PMVT obtains consistent improvements. Visualization results show PMVT automatically perceives the discriminative facial regions for robust AU detection.

Keywords: affective computing, facial action unit recognition, multi-scale transformer, self-attention, cross-attention

OPEN ACCESS

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Xiaoya Zhang,
Nanjing University of Science and Technology, China
Yaohui Zhu,
Beijing Normal University, China
Hao Su,
Beihang University, China

****Correspondence:***
Chongwen Wang
wcwzzw@bit.edu.cn

Received: *29 November 2021*
Accepted: *10 December 2021*
Published: *12 January 2022*

Citation:
Wang C and Wang Z (2022) Progressive Multi-Scale Vision Transformer for Facial Action Unit Detection. Front. Neurorobot. 15:824592. doi: 10.3389/fnbot.2021.824592

1. INTRODUCTION

Facial expression is a natural way for non-verbal communication in our daily life and can be considered as an intuitive illustration of human emotions and mental states. There are some popular facial expression topics categorized as discrete facial expression categories, facial micro-expression, and the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). Among them, FACS is the most comprehensive, anatomical system for encoding expression. FACS defines a detailed set of about 30 atomic non-overlapping facial muscle actions, i.e., action units (AUs). Almost any anatomical facial muscle activity can be introduced *via* a combination of facial AUs. Automatic AU detection has drawn significant interest from computer scientists and psychologists over recent decades, as it holds promise to several practical applications (Bartlett et al., 2003; Zafar and Khan, 2014), such as human affect analysis, human-computer interaction, and pain estimation.

Thus, a reliable AU detection system is of great importance for the analysis of fine-grained facial expressions.

In FACS, different AUs are tightly associated with different facial muscles. It actually means we can observe the active AUs from specific facial regions. For example, the raising of the inner corners of the eyebrows means activated AU1 (inner brow raiser). Lowering the inner corners of the brows corresponds to AU4 (brow lowerer). AU annotators are ofter unable to describe the precise location and the facial scope of the AUs due to the ambiguities of the AUs and individual differences. Actually, the manually defined local AU regions are ambiguous. Existing methods (Li et al., 2017a,b, 2018a,b; Corneanu et al., 2018; Shao et al., 2018; Jacob and Stenger, 2021) usually use artificially define rectangle local regions, or use adaptive attention masks to focus on the expected local facial representations. However, the rectangle local regions violate the actual appearance of the AUs. Moreover, several AUs are simultaneously correlated with multiple and fine-grained facial regions. The learned adaptive attention masks fail to perceive the correlations among different AUs. Therefore, it is critical to automatically learn the AU-adaptive local representations and perceive the dependencies of the facial AUs.

To mitigate this issue, we introduce a new progressive multi-scale vision transformer (PMVT) to capture the complex relationships among different AUs for the wide range of facial expressions in a data-driven fashion. PMVT is based on the multi-scale self-attention mechanism that can flexibly attend to a sequence of image patches to encode the critical cues for AU detection. Currently, vision transformers (Dosovitskiy et al., 2020; Li et al., 2021) have shown promising performance across several vision tasks. The vision transformer models contain MSA mechanisms that can flexibly attend to a sequence of image patches to encode the dependencies of the image patches. The self-attention in the transformers has been shown to effectively learn global interactions and relations between distant object parts. A series of works on various tasks such as image segmentation (Jin et al., 2021), object detection (Carion et al., 2020), video representation learning (Girdhar et al., 2019; Fang et al., 2020) have verified the superiority of the vision transformer models. Inspired by CrossViT (Chen et al., 2021) that processes the input image tokens with two separate transformer branches, our proposed PMVT firstly uses the convolutional neural network (CNN) to encode the convolutional AU feature maps. Then PMVT obtains the multi-scale AU tokens with the small-patch and large-patch branches. The two branches receive different scale AU tokens and exchange semantic AU information *via* a cross attention mechanism. The self-/cross-attention mechanisms facilitate PMVT the content-dependent long-range interaction perceiving capabilities. Thus, PMVT can flexibly focus on the region-specific AU representations and encode the correlations among different AUs to enhance the discriminability of the AU representations. **Figure 1** shows the attention maps of several faces. It is clear that PMVT is capable of focusing on the critical and AU-related facial regions for a wide range of identities and races. More facial examples and detailed explanations can be seen in section 4.2.1.

In summary, the contributions of this study are as follows:

1. We introduce a PMVT for facial AU detection. PMVT does not rely on manually defined facial landmarks to extract the regional AU representations.
2. To further enhance the discriminability of the facial expression representation, PMVT consists of separate transformer branches that receive the multi-scale AU tokens as input. PMVT is capable of encoding multi-scale facial AU representations and perceiving the correlations among different AUs to facilitate representing different AU flexibly.
3. Experimental results demonstrate the advantages of the proposed PMVT over other state-of-the-art AU detection methods on two popular AU datasets. Visualization results show that PMVT is superior in perceiving and capturing the AU-specific facial regions.

2. RELATED WORK

We focus on the previous studies considering two aspects that are tightly related to the proposed PMVT, i.e., the facial AU detection and vision transformer.

2.1. Methods for Facial AU Detection

Action units detection is a multi-label classification problem and has been studied for decades. Several AU detection methods have been proposed (Zhao et al., 2016; Li et al., 2017a,b; Shao et al., 2018; Li and Shan, 2021). To achieve higher AU detection accuracy, different hand-crafted features have been used to encode the characteristics of AUs, such as Histogram of Oriented Gradient (HOG), local binary pattern (LBP), Gabor (Benitez-Quiroz et al., 2016) etc. Recently, AU detection has achieved considerable improvements due to deep learning. Since AU corresponds to the movement of facial muscles, many methods detect the occurrence of AU based on location (Zhao et al., 2016; Li et al., 2017a,b; Shao et al., 2018). For example, Zhao et al. (2016) used a regionally connected convolutional layer and learned the region-specific convolutional filters from the sub-areas of the face. EAC-Net (Li et al., 2017b) and ROI (Li et al., 2017a) extracted AU features around the manually defined facial landmarks that are robust with respect to non-rigid shape changes. SEV-Net (Yang et al., 2021) utilized the AU semantic description as auxiliary information for AU detection. Jacob and Stenger (2021) used a transformer-based encoder to capture the relationships between AUs. However, these supervised methods rely on precisely annotated images and often overfit on a specific dataset as a result of insufficient training images.

Recently, weakly-supervised (Peng and Wang, 2018; Zhao et al., 2018) and self-supervised (Wiles et al., 2018; Li et al., 2019b, 2020; Lu et al., 2020) methods have attracted a lot of attention to mitigate the AU data scarcity issue. Weakly supervised methods typically use the incomplete AU annotations and learn AU classifiers from the prior knowledge between facial expression and facial AU (Peng and Wang, 2018). The self-supervised learning approaches usually adopt pseudo supervisory signals to learn facial AU representation without manual AU annotations (Li et al., 2019b, 2020; Lu et al., 2020). Among them, Lu et al. (2020) proposed a triplet ranking loss to learn AU representations *via* capturing the temporal AU consistency. Fab-Net (Wiles et al.,

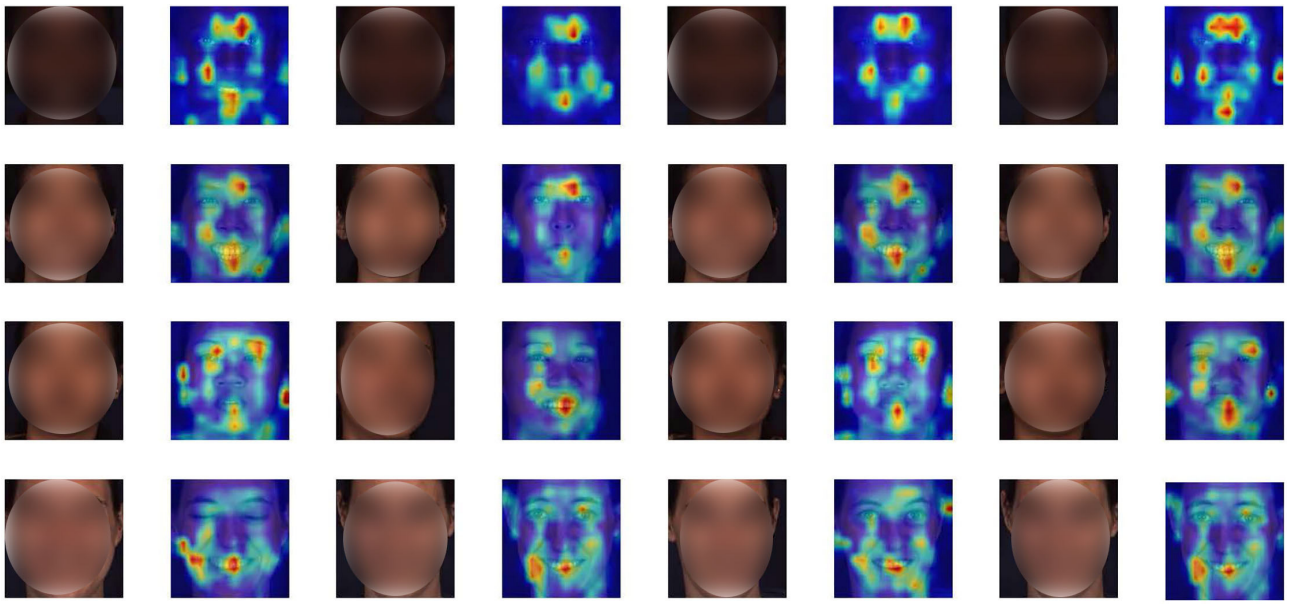


FIGURE 1 | Attention maps of some faces. Our proposed PMVT is capable of capturing the AU-specific facial regions for different identities with diverse facial expressions.

2018) was optimized to map a source facial frame to a target facial frame *via* estimating an optical flow field between the source and target frames. TCAE (Li et al., 2019b) was introduced to encode the pose-invariant facial AU representation *via* predicting separate displacements for pose and AU and using the cycle consistency in the feature and image domains simultaneously.

Our proposed PMVT differs from previous CNN-based or transformer-based (Jacob and Stenger, 2021) AU detection methods in two ways. One, PMVT does not rely on facial landmarks to crop the regional AU features. It is because the facial landmarks may suffer from considerable misalignments under severe facial poses. Under this condition, the encoded facial parts are not part-aligned and will lead to incorrect results. Two, PMVT is the multi-scale transformer-based and the self-attention and cross-attention mechanisms in PMVT can flexibly focus on a sequence of image fragments to encode the correlations among AUs. PMVT is potentially to obtain better facial AU detection performance than previous approaches. We will verify this in section 4.

2.2. Vision Transformer

Self-attention is capable of improving computer vision models due to its content-dependent interactions and parameter-independent scaling of the receptive fields, in contrast to previous parameter-dependent scaling and content-independent interactions of convolutions. Recently, self-attention-based transformer models have greatly facilitated research in machine translation and natural language processing tasks (Waswani et al., 2017). Transformer architecture has become the de-facto standard for a wide range of applications. The core intuition of the original transformer is to obtain self-attention by comparing a feature to all other features in the input sequence. In detail, features are first encoded to obtain a query (*Query*) and memory [(including key (*Key*) and value (*Value*)] embedding *via* linear projections. The product of *Query* with *Key* is used as the attention weight for *Value*. A position embedding is also introduced for each input token to remember the positional information which will be lost in the transformer, which is especially good at capturing long-range dependencies between tokens within an input sequence.

Inspired by this, many recent studies use transformers in various computer vision tasks (Dosovitskiy et al., 2020; Li et al., 2021). Among them, ViT (Dosovitskiy et al., 2020) introduces to view an image as a sequence of tokens and conduct image classification with a transformer encoder. To obtain the input patch features, ViT partition the input image into non-overlapping tokens with 16×16 spatial dimension and linearly project the tokens to match the encoder's input dimension. DeiT (Touvron et al., 2021) further proposes the data-efficient training and distillation for transformer-based image classification models. DETR (Carion et al., 2020) introduces an excellent object detection model based on the transformer, which considerably simplifies the traditional object detection pipeline and obtains comparable performances with prior CNN-based detectors. CrossViT (Chen et al., 2021) encodes small-patch and large-patch image tokens with two exclusive branches and these image tokens are then fused purely by a cross-attention mechanism. Subsequently, transformer models are further extended to other popular computer vision tasks such as segmentation (Jin et al., 2021), face recognition (Li et al., 2021), and 3D reconstruction (Lin et al., 2021). In this study, we extend CrossViT to facial AU detection and show its feasibility and superiority on two publicly available AU datasets.

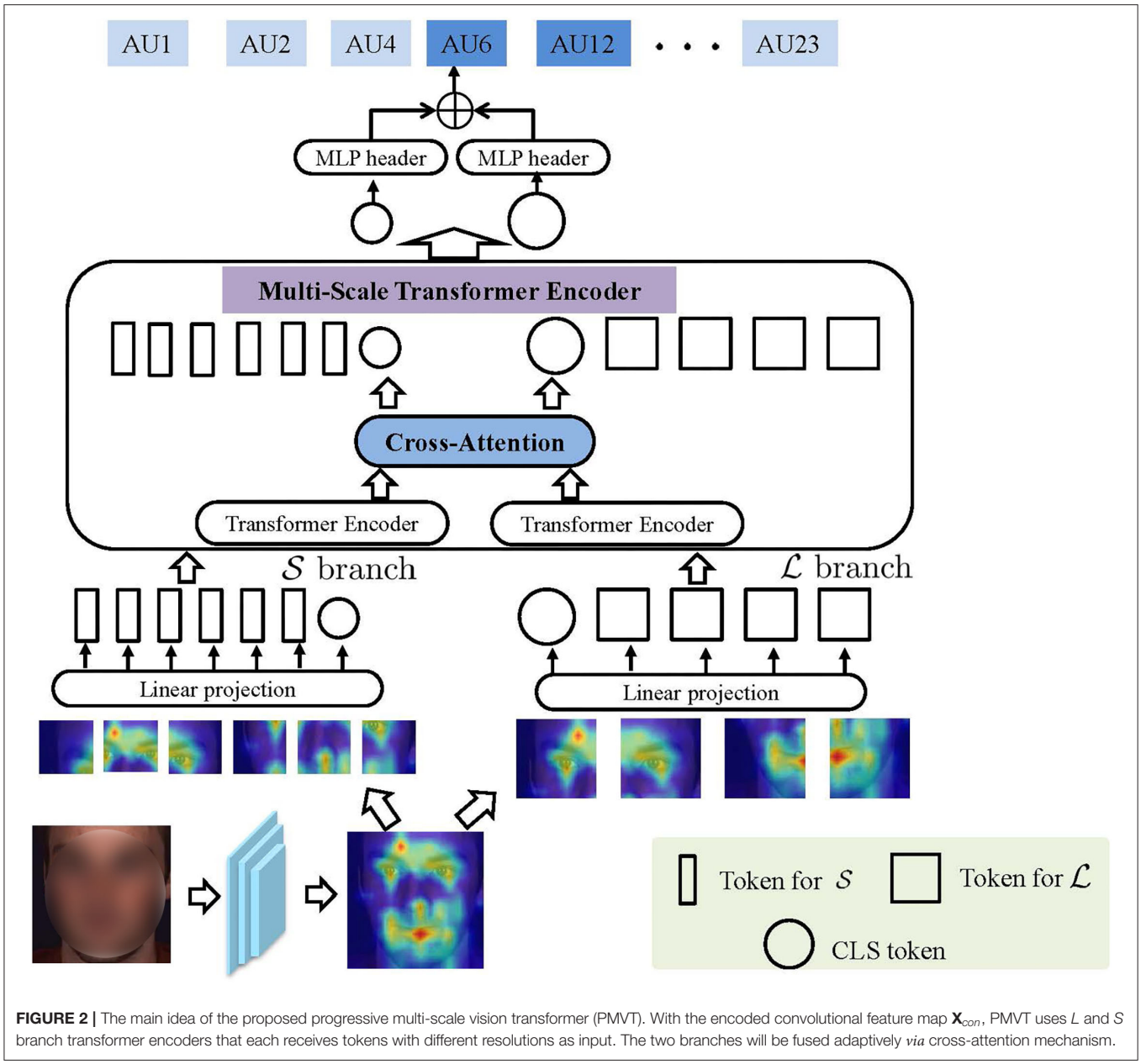


FIGURE 2 | The main idea of the proposed progressive multi-scale vision transformer (PMVT). With the encoded convolutional feature map $\mathbf{X}_{con}$, PMVT uses *L* and *S* branch transformer encoders that each receives tokens with different resolutions as input. The two branches will be fused adaptively *via* cross-attention mechanism.

3. METHOD

Figure 2 illustrates the main idea of the proposed PMVT. Given an input face, PMVT first extracts its convolutional feature maps *via* a commonly-used backbone network. Second, PMVT encodes the discriminative facial AU feature by the multi-scale transformer blocks. We will first review the traditional vision transformer and present our proposed PMVT afterward.

3.1. Revisiting Vision Transformer

We first revisit the critical components in ViT (Dosovitskiy et al., 2020) that mainly consist of image tokenization and several layers of the token encoder. Each encoder consists of two layers, i.e., multi-head self-attention (MSA) layer and feed-forward network (FFN) layer.

Traditional vision transformer typically receives a sequence of image patch embeddings as input. To obtain the token embeddings, ViT encodes the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ into a set of flattened two-dimensional image patches: $\mathbf{X}_p \in \mathbb{R}^{N\times P_2\times C}$. Among the mathematic symbols, H W, C denote the height, width, channel of the input image $\mathbf{X}$. P means the spatial resolution of each image patch $\mathbf{X}_p$. After the image tokenization, we can obtain $N = \frac{H\times W}{P^2}$ patches that will be treated as the sequential input for the transformer. These image patches are then flattened and projected to embeddings with a size of S. Typically, ViT adds an extra class token that will be concatenated

with the image embeddings, resulting in the input sequence with a size of $\mathbf{X}_t \in \mathbb{R}^{(N+1)\times S}$. Finally, the class token will serve as the image representation that will be used for image classification. ViT uses a residual connection for each encoder. The computation in each encoder can be formulated as:

$$\mathbf{X}_t{}' = \mathrm{LN}(\mathbf{X}_t + \mathrm{MSA}(\mathbf{X}_t)), \quad (1)$$

$$\mathbf{Y} = \mathrm{LN}(\mathbf{X}_t{}' + \mathrm{FFN}(\mathbf{X}_t{}')), \quad (2)$$

where $\mathbf{X}_t$ and $\mathbf{Y}$ denote the input and output of the encoder. $\mathbf{X}_t{}'$ is the output of the MSA layer. LN means layer normalization. MSA means multi-head self-attention which will be described next.

For the self-attention module in ViT, the sequential input tokens $\mathbf{X}_t \in \mathbb{R}^{(N+1)\times S}$ are linearly transformed into *Query*, *Key*, *Value* spaces. Typically, $Query \in \mathbb{R}^{(N+1)\times S}$, $Key \in \mathbb{R}^{(N+1)\times S}$, $Value \in \mathbb{R}^{(N+1)\times S}$. Afterward, a weighted sum over all values in the sequential tokens is computed as,

$$\mathrm{Attention}(Quey, Key, Value) = \mathrm{softmax}(\frac{Query \times Key^T}{\sqrt{S}})Value. \quad (3)$$

Then a linear projection is conducted to the weighted values Attention(*Quey*, *Key*, *Value*). MSA is a natural extension of the single-head self-attention described above. MSA splits *Query*, *Key*, *Value* for h times and performs the self-attention mechanism in parallel, then maps their concatenated outputs *via* linear transformation. In addition to the MSA module, ViT exploits the FFN module to conduct dimension adjustment and non-linear transformation on each image token to enhance the representation ability of the transformed tokens.

3.2. Progressive Multi-Scale Transformer

The direct tokenization of input images into large patches in ViT has been found to show its limitations (Yuan et al., 2021). On the one hand, it is difficult to perceive the important low-level characteristics (e.g., edges, colors, corners) in images; On the other hand, large CNN kernels for the image tokenization contain too many trainable parameters and are often difficult to optimize, and thus, ViT requires much more training samples. This is particularly impartial for facial AU detection. As AU annotation is time-consuming, cumbersome, and error-prone. Currently, the publicly available AU datasets merely contain limited facial images. To cope with this issue, we exploit the popular ResNet-based backbone to encode the input facial image $\mathbf{X}$ to obtain the convolutional feature map $\mathbf{X}_{con} = F(\mathbf{X})$, where F means the neural operation in the backbone network.

To obtain multi-scale tokens from $\mathbf{X}_{con}$, we use two separate branch transformer encoder that each receives tokens with different resolution as input. We illustrate the main idea of our proposed PMVT in **Figure 2**. Mathematically speaking, let us denote the two branches as $\mathcal{L}$ and $\mathcal{S}$, respectively. In PMVT, the $\mathcal{L}$ branch uses coarse-grained token as input while the $\mathcal{S}$ branch directly operates at a much more fine-grained token. Both branches are adaptively fused K times *via* a cross-attention mechanism. Finally, PMVT exploits the *CLS* token of the $\mathcal{L}$ and $\mathcal{S}$ branches for facial AU detection. For each token within each branch, PMVT introduces a trainable position embedding. Note that we can use multiple multi-scale transformer encoders (MST) or perform cross-attention times within each MST. We will analyze the performance variations in section 4.2.1.

Figure 3 illustrates the cross-attention mechanism in PMVT. To effectively fuse the multi-scale AU features, PMVT utilizes the *CLS* token at each branch (e.g., $\mathcal{L}$ branch) as an agent to exchange semantic AU information among the patch tokens from the other branch (e.g., $\mathcal{S}$ branch) and then project the *CLS* token back to its own branch (e.g., $\mathcal{L}$ branch). Such operation is reasonable because the *CLS* token in $\mathcal{L}$ or $\mathcal{S}$ branch already learns semantic features among all patch tokens in its own branch. Thus, interacting with the patch tokens at the other branch can absorb more semantic AU information at a different scale. We hypothesize such cross-attention mechanism will help learn discriminative AU features as different AU usually have different appearance scopes and there exist correlations among the facial AUs. The multi-scale features will help encode AUs more precisely and PMVT will encode the AU correlations with the self-/cross-attention mechanism.

Take $\mathcal{L}$ for example to show the cross-attention mechanism in PMVT. Specially, PMVT uses the *CLS* token $\mathbf{X}^l_{cls}$ from the $\mathcal{L}$ branch and patch tokens the $\mathbf{X}^s_i$ from $\mathcal{S}$ branch for feature fusing. PMVT uses $\mathbf{X}^l_{cls}$ to obtain a *query* and use $\mathbf{X}^s_i$ to obtain the *key* and *value*. The *query*, *key*, *value* will be transformed into a weighted sum overall values in the sequential tokens as that in Equation (3). Notably, such a cross-attention mechanism is similar to self-attention except that the *query* is obtained from the *CLS* token in another transformer branch. In **Figure 3**, $f(.)$ and $g(.)$ mean linear projections that aim the alignment of the feature dimension. We will evaluate the effectiveness of the proposed PMVT in the next section.

3.3. Training Objective

We utilize the multi-label sigmoid cross-entropy loss for training the facial AU detection model in PMVT, which can be formulated as:

$$\mathcal{L}^{AU} = -\sum_j^J z^j \log \hat{z}^j + (1 - z^j)\log(1 - \hat{z}^j), \quad (4)$$

where J denotes the number of facial AUs. z^j denotes the j-th ground truth AU annotation of the input AU sample. $\hat{z}^j$ means the predicted AU score. $z_i \in \{0, 1\}$ denotes the annotation with respect to the ith AU. 1 means the AU is active, 0 means inactive.

4. EXPERIMENT

4.1. Implementation Details

We adopted ResNet-34 (He et al., 2016) as the backbone network for PMVT due to its elegant network structure and excellent performance in image classification. We chose the output of the third stage as the convolutional feature maps: $\mathbf{X}_{con} \in \mathbb{R}^{14\times14\times512}$. For the L branch, the token size is set as $N = 5 \times 5$ *via* adaptative pooling operation. For the S branch, the token size is set as $N = 14 \times 14$. The pre-trained model based on the ImageNet dataset was used for initializing the

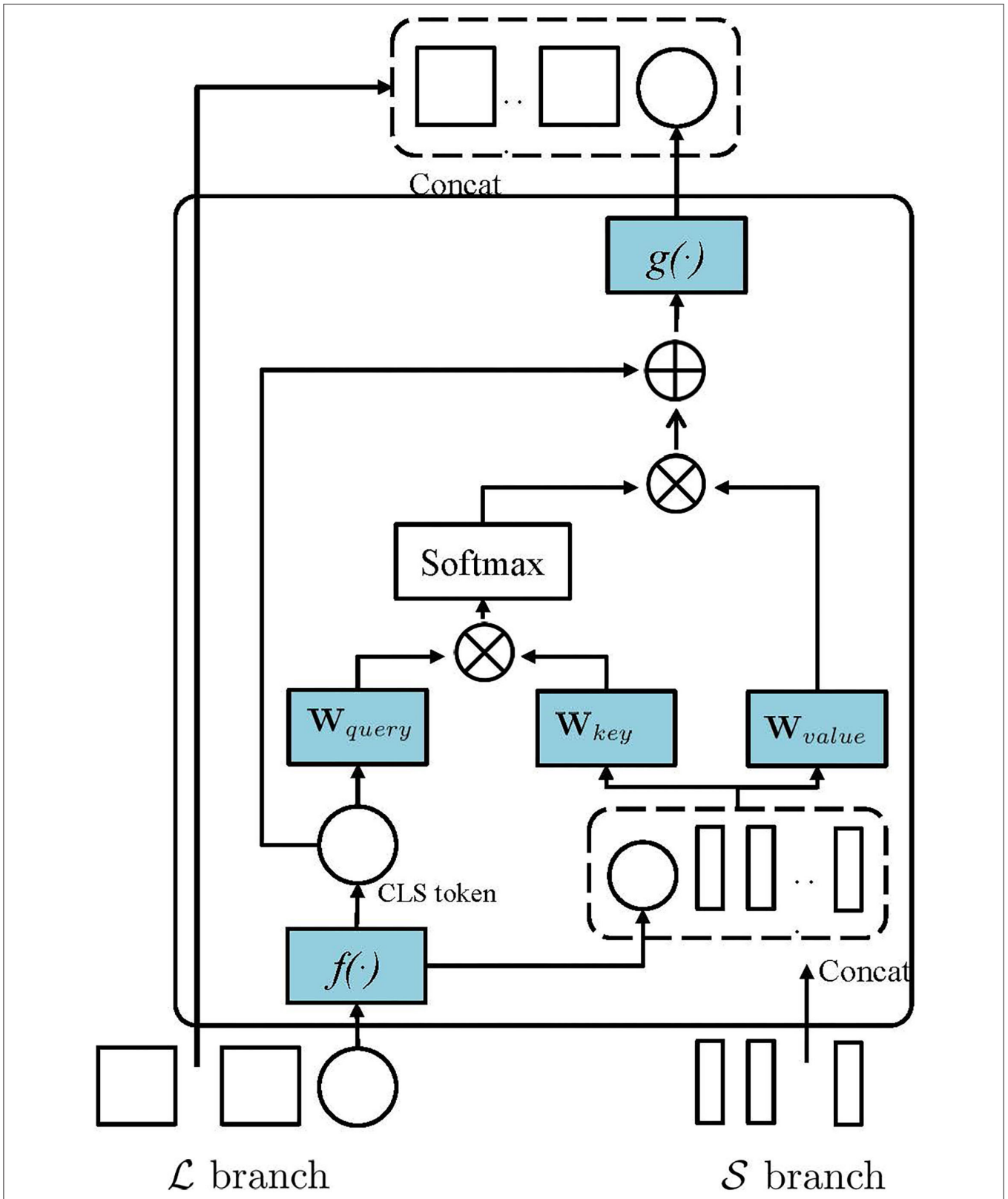


FIGURE 3 | The main idea of the cross-attention in PMVT. In this study, we show that PMVT utilizes the classification (*CLS*) token at the *L* branch as an agent to exchange semantic AU information among the patch tokens from the *S* branch. PMVT can also use the *CLS* token at *S* to absorb information among the tokens from the *L* branch.

backbone network. For the transformer part, we use one layer of transformer encoder that consists of two-layer cross-attention. We exploited a batch-based stochastic gradient descent method to optimize the proposed PMVT. During the training process, we set the batch size as 64 and the initial learning rate as 0.002. The momentum was set as 0.9 and the weight decay was set as 0.0005.

4.1.1. Datasets

For AU detection, we adopted BP4D (Zhang et al., 2013) and DISFA (Mavadati et al., 2013) datasets. Among them, BP4D is a spontaneous FACS dataset that consists of 328 videos for 41 subjects (18 men and 23 women). Eight different tasks are evaluated on a total of 41 participants, and their spontaneous facial expression variations were recorded in several videos. Each participant subject is involved in eight sessions, and their spontaneous facial expressions were captured in both 2D and 3D videos. A total of 12 AUs were annotated for the 328 videos, and there are approximately 1,40,000 frames with AU annotations. DISFA contains 27 participants that consists of 12 women and 15 men. Each subject is asked to watch a 4-min video to elicit their facial AUs. The facial AUs are annotated with intensities from 0 to 5. In our experiments, we obtained nearly 1,30,000 AU-annotated images in the DISFA dataset by considering the images with intensities greater than 1 as active. For BP4D and DISFA datasets, the images are split into 3-fold in a subject-independent manner. Based on the datasets, we conducted 3-fold cross-validation. We adopted 12 AUs in BP4D and 8 AUs in DISFA dataset for evaluation. For the DISFA dataset, we leveraged the model trained on BP4D to initialize the backbone network, following the same experimental setting of Li et al. (2017b).

4.1.2. Evaluation Metric

We adopted F1-score ($F1 = \frac{2RP}{R+P}$) to evaluate the performance of the proposed AU detection method, where R and P, respectively, denote recall and precision. We additionally calculated the average F1-score over all AUs (AVE) to quantitatively evaluate the overall facial AU detection performance. We show the AU detection results as $F1 \times 100$.

4.2. Experimental Results

We compare the proposed with the state-of-the-art facial AU detection approaches, including DRML (Zhao et al., 2016), EAC-Net (Li et al., 2017b), ROI (Li et al., 2017a), JAA-Net (Shao et al., 2018), OFS-CNN (Han et al., 2018), DSIN (Corneanu et al., 2018), TCAE (Li et al., 2019b), TAE (Li et al., 2020), SRERL (Li et al., 2019a), ARL (Shao et al., 2019), SEV-Net (Yang et al., 2021), and FAUT (Jacob and Stenger, 2021). Among them, most of the AU methods (Li et al., 2017a, 2019a; Corneanu et al., 2018; Shao et al., 2018) manually crop the local facial regions to learn the AU-specific representations with exclusive CNN branches. TAE (Li et al., 2020) utilize unlabeled videos that consist of approximately 7,000 subjects to encode the AU-discriminative representation without AU annotations. SEV-Net (Yang et al., 2021) introduce the auxiliary semantic word embedding and visual feature for AU detection. FAUT (Jacob and Stenger, 2021) introduce an AU correlation network based on a transformer architecture to perceive the relationships between different AU in an end-to-end manner.

Table 1 shows the AU detection results of our method and studies works on the BP4D dataset. Our PMVT achieves comparable AU detection accuracy with the best state-of-the-art AU detection methods in the average F1 score. Compared with other methods, PMVT obtains consistent improvements in the average accuracy (+14.6% over DRML, +7.0% over EAC-Net, +6.5% over ROI, +2.9% over JAA-Net, +4.0% over DSIN, +6.8% over TCAE, +2.6% over TAE). The benefits of our proposed PMVT over other methods can be explained in 2-fold. First, PMVT explicitly introduces transformer modules in the network structure. The self-attention mechanism in the transformer modules is capable of perceiving the local to global interactions between different facial AUs. Second, we use multi-scale features to better encode the regional features of the facial AUs, as different AUs have different appearance scopes. The cross-attention mechanism between the multi-scale features is beneficial for learning discriminative facial AU representations. **Table 2** shows the quantitative facial AU detection results of our PMVT and other methods on the DISFA dataset. PMVT achieves the second-best AU detection accuracy compared with all the state-of-the-art AU detection methods in the average F1 score. In detail, PMVT outperforms EAC-Net, JAA-Net, OFS-CNN, TCAE, TAE, SRERL, ARL, and SEV-Net with +12.4%, +4.9%, +9.5%, +7.3%, +15.9%, +9.4%, +5.0%, +2.2%, and +2.1% improvements in the average F1 scores. The consistent improvements over other methods on the two popular datasets verify the feasibility and superiority of our proposed PVMT. We will carry out an ablation study to investigate the contribution of the self-/cross-attention in PVMT and illustrate visualization results in the next section.

4.2.1. Ablation Study

We illustrate the ablation study experimental results in **Table 3**. In **Table 3**, we show the AU detection performance variations with different cross-attention layers ($CL = 1, 2, 3$) in the multi-scale transformer encoder and with different layers of multi-scale transformer encoders ($MS = 1, 2, 3$).

As shown in **Table 3**, PMVT shows its best AU detection performance with $CL = 2$ and $MS = 1$. It means PMVT merely contains one layer of the multi-scale transformer encoder, and the encoder contains two layers of cross-attention. With more MST encoders, PMVT will contain too many trainable parameters and will suffer from insufficient training images. With $CL = 1$ or $CL = 3$, PMVT shows degraded AU detection performance, and it suggests that information fusion should be performed twice to achieve the discriminative AU representations.

We additionally show the attention maps of PMVT on some randomly sampled faces in **Figure 4**. The visualization results show the benefits of the proposed PMVT for robust facial AU detection. It is obvious that PVMT shows consistent activation maps for each face under different races, expressions, lightings, and identities. For example, the third face in the second row

TABLE 1 | Action unit (AU) detection performance of our proposed progressive multi-scale vision transformer (PMVT) and state-of-the-art methods on the BP4D dataset.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	AVE
LSVM (Fan et al., 2008)	23.2	22.8	23.1	27.2	47.1	77.2	63.7	64.3	18.4	33.0	19.4	20.7	35.3
DRML (Zhao et al., 2016)	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-Net (Li et al., 2017b)	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
ROI (Li et al., 2017a)	36.2	31.6	43.4	77.1	73.7	85.0	87.0	62.6	45.7	58.0	38.3	37.4	56.4
JAA-Net (Shao et al., 2018)	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
DSIN (Corneanu et al., 2018)	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
TCAE (Li et al., 2019b)	43.1	32.2	44.4	75.1	70.5	80.8	85.5	61.8	34.7	58.5	37.2	48.7	56.1
TAE (Li et al., 2020)	47.0	45.9	50.9	74.7	72.0	82.4	85.6	62.3	48.1	62.3	45.9	46.3	60.3
SRERL (Li et al., 2019a)	46.9	45.3	55.6	77.1	78.4	83.5	**87.6**	63.9	52.2	**63.9**	47.1	53.3	62.9
ARL (Shao et al., 2019)	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	55.4	61.1
FAUT (Jacob and Stenger, 2021)	51.7	49.3	**61.0**	77.8	**79.5**	82.9	86.3	**67.6**	51.9	63.0	43.7	**56.3**	**64.2**
SEV-Net (Yang et al., 2021)	58.2	**50.4**	58.3	81.9	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9
PMVT (Ours)	**59.3**	43.0	59.3	**82.3**	73.6	82.6	86.1	57.6	**53.0**	60.2	**47.9**	50.6	62.9

The highest values are illustrated in Bold format.

TABLE 2 | Action unit detection performance of our proposed PMVT and state-of-the-art methods on the DISFA dataset.

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	ave
DRML (Zhao et al., 2016)	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-Net (Li et al., 2017b)	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
JAA-Net (Shao et al., 2018)	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
OFS-CNN (Han et al., 2018)	43.7	40.0	67.2	59.0	49.7	75.8	72.4	54.8	51.4
DSIN (Corneanu et al., 2018)	42.4	39.0	**68.4**	28.6	46.8	70.8	90.4	42.2	53.6
TCAE (Li et al., 2019b)	15.1	15.2	50.5	48.7	23.3	72.1	82.1	52.9	45.0
TAE (Li et al., 2020)	21.4	19.6	64.5	46.8	44.0	73.2	85.1	55.3	51.5
SRERL (Li et al., 2019a)	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
FAUT (Jacob and Stenger, 2021)	46.1	48.6	72.8	**56.7**	50.0	72.1	90.8	55.4	**61.5**
ARL (Shao et al., 2019)	43.9	42.1	63.6	41.8	40.0	76.2	95.2	**66.8**	58.7
SEV-Net (Yang et al., 2021)	**55.3**	53.1	61.5	53.6	38.2	71.6	95.7	41.5	58.8
PMVT (Ours)	50.0	**54.3**	63.2	55.6	40.0	72.2	**95.9**	56.3	60.9

The highest values are illustrated in Bold format.

TABLE 3 | Ablation studies on the BP4D and DISFA datasets.

Methods	BP4D	DISFA
CL=1	60.7	56.3
CL=2	62.9	60.9
CL=3	59.5	55.8
MS=1	62.9	60.9
MS=2	59.8	58.1
MS=3	55.0	51.1

is annotated with active AU1 (inner brow raiser), AU2 (outer brow raiser), AU6 (cheek raiser), AU7 (inner brow raiser), AU10 (inner brow raiser), and AU12 (inner brow raiser). The second face in the third row is annotated with active AU1 (inner brow raiser), AU10 (inner brow raiser), AU12 (inner brow raiser), and AU15 (lip corner depressor). The first face in the fourth row is annotated with active AU7 (inner brow raiser) and AU14 (dimpler). The attention maps of these faces are in line and consistent with the annotated AUs. The visualization maps in **Figure 4** show the generalization ability and feasibility of our proposed PVMT.

5. CONCLUSIONS

In this study, we propose a PMVT to perceive the complex relationships among different AUs in an end-to-end data-driven manner. PMVT is based on the multi-scale self-/cross-attention mechanism that can flexibly focus on sequential image patches to effectively encode the discriminative AU representation and perceive the correlations among different facial AUs. Compared with previous facial AU detection methods, PMVT obtains

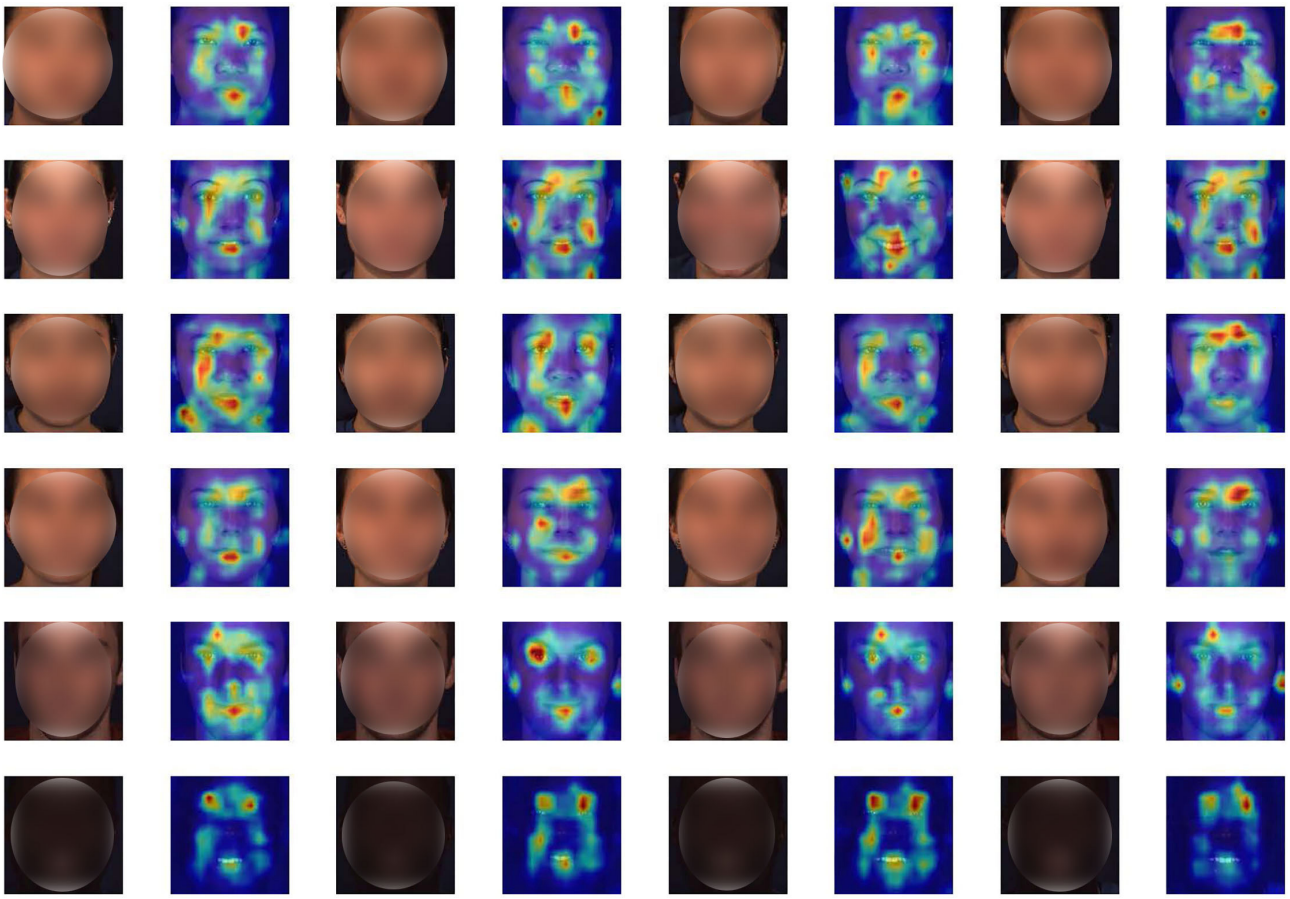


FIGURE 4 | Attention maps of some representative faces. We illustrate a subject with different facial expressions in each row. It is obvious that the proposed PMVT is capable of focusing on the most silent parts for facial AU detection. Deep red denotes high activation, better viewed in color and zoom in.

comparable AU detection performance. Visualization results show the superiority and feasibility of our proposed PMVT. For future study, we will explore utilizing PMVT for more affective computing tasks, such as facial expression recognition, AU density estimation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CW and ZW cooperatively completed the method design and experiment parts. CW wrote all the sections of the manuscript. ZW carried out the experiments and gave the detailed analysis. Both the two authors have carefully read, polished, and approved the final manuscript.

REFERENCES

Bartlett, M. S., Littlewort, G., Fasel, I., and Movellan, J. R. (2003). "Real time face detection and facial expression recognition: development and applications to human computer interaction," in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, Vol. 5 (Madison, WI: IEEE), 53–53.

Benitez-Quiroz, C. F., Srinivasan, R., and Martinez, A. M. (2016). "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 5562–5570.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *European Conference on Computer Vision* (Glasgow: Springer), 213–229.

Chen, C.-F., Fan, Q., and Panda, R. (2021). Crossvit: cross-attention multi-scale vision transformer for image classification. *arXiv preprint* arXiv:2103.14899.

Corneanu, C., Madadi, M., and Escalera, S. (2018). "Deep structure inference network for facial action unit recognition," in *ECCV* (Munich), 298–313.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929.

Ekman, P., and Friesen, W. V. (1978). *Manual for the Facial Action Coding System*. Consulting Psychologists Press.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874. Available online at: https://dl.acm.org/citation.cfm?id=1442794

Fang, Y., Gao, S., Li, J., Luo, W., He, L., and Hu, B. (2020). Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting. *Neurocomputing* 392, 98–107. doi: 10.1016/j.neucom.2020.01.087

Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. (2019). "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Long Beach, CA: IEEE), 244–253.

Han, S., Meng, Z., Li, Z., O'Reilly, J., Cai, J., Wang, X., et al. (2018). "Optimizing filter size in convolutional neural networks for facial action unit recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 5070–5078.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Jacob, G. M., and Stenger, B. (2021). "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7680–7689.

Jin, Y., Han, D., and Ko, H. (2021). Trseg: transformer for semantic segmentation. *Pattern Recognit. Lett.* 148, 29–35. doi: 10.1016/j.patrec.2021.04.024

Li, G., Zhu, X., Zeng, Y., Wang, Q., and Lin, L. (2019a). "Semantic relationships guided representation learning for facial action unit recognition," in *AAAI*, vol. 33, 8594–8601.

Li, W., Abtahi, F., and Zhu, Z. (2017a). "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *CVPR* (Honolulu, HI: IEEE).

Li, W., Abtahi, F., Zhu, Z., and Yin, L. (2017b). "Eac-net: a region-based deep enhancing and cropping approach for facial action unit detection," in *FG* (Washington, DC).

Li, Y., and Shan, S. (2021). Meta auxiliary learning for facial action unit detection. *arXiv preprint* arXiv:2105.06620.

Li, Y., Sun, Y., Cui, Z., Shan, S., and Yang, J. (2021). Learning fair face representation with progressive cross transformer. *arXiv preprint* arXiv:2108.04983.

Li, Y., Zeng, J., and Shan, S. (2020). Learning representations for facial actions from unlabeled videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 302–317. doi: 10.1109/TPAMI.2020.3011063

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018a). Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Process.* 28, 2439–2450. doi: 10.1109/TIP.2018.2886767

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018b). "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)* (Beijing: IEEE), 2209–2214.

Li, Y., Zeng, J., Shan, S., and Chen, X. (2019b). "Self-supervised representation learning from videos for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 10924–10933.

Lin, K., Wang, L., and Liu, Z. (2021). "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 1954–1963.

Lu, L., Tavabi, L., and Soleymani, M. (2020). "Self-supervised learning for facial action unit recognition through temporal consistency," in *Proceedings of the British Machine Vision Conference (BMVC). BMVA Press*.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: a spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* 4, 151–160. doi: 10.1109/T-AFFC.2013.4

Peng, G., and Wang, S. (2018). "Weakly supervised facial action unit recognition through adversarial training," in *CVPR* (Salt Lake City, UT), 2188–2196.

Shao, Z., Liu, Z., Cai, J., and Ma, L. (2018). "Deep adaptive attention for joint facial action unit detection and face alignment," in *ECCV* Munich.

Shao, Z., Liu, Z., Cai, J., Wu, Y., and Ma, L. (2019). Facial action unit detection using attention and relation learning. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2019.2948635

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021)."Training data-efficient image transformers and distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 10347–10357.

Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. et al. (2017)."Attention is all you need," in *Proceedings of the Conference on Neural Information Processing Systems* (Long Beach, CA), 1–11.

Wiles, O., Koepke, A., and Zisserman, A. (2018). Self-supervised learning of a facial attribute embedding from video. *arXiv preprint* arXiv:1808.06882.

Yang, H., Yin, L., Zhou, Y., and Gu, J. (2021). "Exploiting semantic embedding and visual feature for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN), 10482–10491.

Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., and Wu, W. (2021). Incorporating convolution designs into visual transformers. *arXiv preprint* arXiv:2103.11816.

Zafar, Z., and Khan, N. A. (2014). "Pain intensity evaluation through facial action units," in *2014 22nd International Conference on Pattern Recognition* (Stockholm:IEEE), 4696–4701.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., and Liu, P. (2013). "A high-resolution spontaneous 3d dynamic facial expression database," in *FG* (Shanghai: IEEE).

Zhao, K., Chu, W.-S., and Martinez, A. M. (2018). "Learning facial action units from web images with scalable weakly supervised clustering," in *CVPR* (Salt Lake City, UT), 2090–2099.

Zhao, K., Chu, W.-S., and Zhang, H. (2016). "Deep region and multi-label learning for facial action unit detection," in *CVPR* (Las Vegas, NV), 3391–3399.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 09 February 2022
doi: 10.3389/fcomp.2022.837269

Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children

Manuel Milling[1], Alice Baird[1], Katrin D. Bartl-Pokorny[1,2,3], Shuo Liu[1], Alyssa M. Alcorn[4], Jie Shen[5], Teresa Tavassoli[6], Eloise Ainger[4], Elizabeth Pellicano[7], Maja Pantic[5], Nicholas Cummins[8] and Björn W. Schuller[1,5]*

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [2] Research Unit iDN – Interdisciplinary Developmental Neuroscience, Division of Phoniatrics, Medical University of Graz, Graz, Austria, [3] Division of Physiology, Otto Loewi Research Center, Medical University of Graz, Graz, Austria, [4] Centre for Research in Autism and Education, UCL Institute of Education, London, United Kingdom, [5] Department of Computing, Imperial College London, London, United Kingdom, [6] School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom, [7] School of Education, Macquarie University, Sydney, NSW, Australia, [8] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London, London, United Kingdom

OPEN ACCESS

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Chuangao Tang,
Southeast University, China
Hongli Chang,
Southeast University, China

***Correspondence:**
Manuel Milling
manuel.milling@informatik.uni-augsburg.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *16 December 2021*
Accepted: *03 January 2022*
Published: *09 February 2022*

Citation:
Milling M, Baird A, Bartl-Pokorny KD, Liu S, Alcorn AM, Shen J, Tavassoli T, Ainger E, Pellicano E, Pantic M, Cummins N and Schuller BW (2022) Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children. Front. Comput. Sci. 4:837269. doi: 10.3389/fcomp.2022.837269

Individuals with autism are known to face challenges with emotion regulation, and express their affective states in a variety of ways. With this in mind, an increasing amount of research on automatic affect recognition from speech and other modalities has recently been presented to assist and provide support, as well as to improve understanding of autistic individuals' behaviours. As well as the emotion expressed from the voice, for autistic children the dynamics of verbal speech can be inconsistent and vary greatly amongst individuals. The current contribution outlines a voice activity detection (VAD) system specifically adapted to autistic children's vocalisations. The presented VAD system is a recurrent neural network (RNN) with long short-term memory (LSTM) cells. It is trained on 130 acoustic Low-Level Descriptors (LLDs) extracted from more than 17 h of audio recordings, which were richly annotated by experts in terms of perceived emotion as well as occurrence and type of vocalisations. The data consist of 25 English-speaking autistic children undertaking a structured, partly robot-assisted emotion-training activity and was collected as part of the DE-ENIGMA project. The VAD system is further utilised as a preprocessing step for a continuous speech emotion recognition (SER) task aiming to minimise the effects of potential confounding information, such as noise, silence, or non-child vocalisation. Its impact on the SER performance is compared to the impact of other VAD systems, including a general VAD system trained from the same data set, an out-of-the-box Web Real-Time Communication (WebRTC) VAD system, as well as the expert annotations. Our experiments show that the child VAD system achieves a lower performance than our general VAD system, trained under identical conditions, as we obtain receiver operating characteristic area under the curve (ROC-AUC) metrics of 0.662 and 0.850, respectively. The SER results show varying performances across valence and arousal depending on the utilised VAD system with a maximum concordance correlation coefficient (CCC) of 0.263 and a minimum

root mean square error (RMSE) of 0.107. Although the performance of the SER models is generally low, the child VAD system can lead to slightly improved results compared to other VAD systems and in particular the VAD-less baseline, supporting the hypothesised importance of child VAD systems in the discussed context.

Keywords: affective computing, voice activity detection, deep learning, speech emotion recognition, children with autism, robot human interaction

1. INTRODUCTION

Speech emotion recognition (SER) is a prominent subfield of Affective Computing as the complexity of the human speech apparatus together with the communicative importance of emotions in speech make a good understanding of the problem both difficult and desirable, which becomes apparent from the long history of emotion recognition challenges (Valstar et al., 2013; Ringeval et al., 2019; Stappen et al., 2021). The subjective nature of emotions leads to a variety of emotion recognition tasks, which make the possibility for a one-fits-all solution not the optimal approach to capture the subtle variation in emotion expression. As most models are only focused on a single corpus, which can range from acted emotions (Busso et al., 2008) via emotions induced by a trigger (Koelstra et al., 2012) to spontaneous emotions (Stappen et al., 2020), and is often recorded for adult individuals, the application of SER models needs to be chosen with care and in general adapted to the specific scenario.

Continuous SER tasks, especially in interactive scenarios, such as robot-assisted child-robot interactions, can be prone to auditory artefacts, and limited instances of speech, creating the need to discriminate between background noise and information-rich instances. Voice activity detection (VAD) systems are therefore commonly used in SER tasks to remove unvoiced segments of the audio signal, for instance displayed in Harár et al. (2017), Alghifari et al. (2019) and Akçay and Oğuz (2020). In a scenario with more than one speaker however, VAD alone might not be enough to filter out all non-relevant information about a specific speaker's affective state.

Autism is a neurodevelopmental condition that is associated with difficulties in social communication and restricted, repetitive patterns of behaviour, interests, or activities (American Psychiatric Association, 2013). The clinical picture of autism is heterogeneous, including diversity in autistic characteristics and spoken language skills, and frequently occurring comorbidities, such as anxiety disorder, attention-deficit hyperactivity disorder, developmental coordination disorder, or depressive disorders (Kopp et al., 2010; Lord et al., 2018; Zaboski and Storch, 2018; Hudson et al., 2019). Difficulties in socio-communicative skills and recognition and expression of emotion in autistic children can make interactions with their family, peers, and professionals challenging.

However, only few research projects have investigated how recent technology including Artificial Intelligence can help to better understand the needs and improve the conditions of children with autism: the ASC-inclusion project developed a platform aiming to playfully support children in understanding and expressing emotions through a comprehensive virtual world (Schuller, 2013), for instance through serious games (Marchi et al., 2018). The DE-ENIGMA project[1] focused on a better understanding of behaviour and needs of autistic children in a researcher-led robot-human-interaction (RCI) scenario, contributing to insights about robot predictability in RCI scenarios with children with autism (Schadenberg et al., 2021), as well as prediction of the severity of traits related to autism (Baird et al., 2017) and detection of echolalic vocalisations (Amiriparian et al., 2018), i.e., word or phrase repetitions of autistic children based on spoken utterances of their conversational partners. Schuller et al. introduced a task for the speech-based diagnosis of children with autism and other pervasive developmental disorders (Schuller et al., 2013). Particularly in the field of SER for individuals with autism, data appears quite sparse (Schuller, 2018), presumably caused in part due to the considerable time-expense needed to gather such data from autistic children. Rudovic et al. developed a personalised multi-modal approach based on deep learning for affect and engagement recognition in autistic children, achieving up to 60 % agreement with human annotators, aiming to enable affect-sensitive child-robot interaction in therapeutic scenarios (Rudovic et al., 2018). From this overview of related works, there have been limited works, which model emotions of autistic children with continuous labelling strategies. To the best of our knowledge, no research as of yet has explored how VAD can improve such modelling.

In this manuscript, we investigate a subset of data collected in the DE-ENIGMA project (Shen et al., 2018). The presented data consist of about 17 h of audio recordings and rich annotations including continuously perceived affective state, and manually performed speaker diarisation. The data poses numerous challenges commonly associated with in-the-wild data including noise (for instance from robot or furniture movements) or varying distances to microphones. Additionally, a particular challenge in the current dataset results from the sparsity of child vocalisations in the interaction between child, robot, and researcher, as several children who took part in the study had limited-to-no spoken communication. In contrast to common continuous emotion recognition tasks, we hypothesised that a model focusing on the child vocalisations alone would be able to outperform other models, as we expect the child vocalisations to contain the most information about the children's affective states. For this reason, in the current work, we implement a VAD system specifically trained for vocalisations of autistic

[1] https://de-enigma.eu/

children on the dataset and evaluate its performance against a trained general VAD system - trained on all vocalisations of our dataset - as well as an implementation of the Web Real-Time Communication (WebRTC) VAD (Google, 2021) and the manual speaker diarisation annotations, for the SER task at hand. The WebRTC VAD is based on Gaussian mixture models (GMMs) and log energies of six frequency bands.

The remainder of this manuscript is organised as follows. In section 2, we provide a detailed overview of the investigated dataset. Furthermore, we introduce the deep learning-based methodology for both the VAD and the SER task in section 3. Subsequently, we present experimental results for the isolated VAD experiments, as well as the SER task with a combined VAD-SER system in section 4. Finally, we discuss the results and the limitations of our approaches in section 5 before we conclude our work in section 6.

2. DATASET

The Experiments in this manuscript are based on a subset of data gathered in the DE-ENIGMA Horizon 2020 project, which were collected in a school-based setting in the United Kingdom and Serbia. In this work, we solely focus on audio data from the British study arm of the project, for which all relevant data streams and annotations are available. Here, autistic children undertook emotion-recognition training activities based on the Teaching Children with Autism to Mind-Read programme (Howlin et al., 1999), under guidance of a researcher. Ethical approval was granted for this study by the Research Ethics Committee at the UCL Institute of Education and the University College London (REC 796). Children were randomly assigned to researcher-only sessions, or to sessions, which were supported by the humanoid robot Zeno-R2. Zeno is capable of performing different emotion-related facial expressions, and which was controlled by the researcher via an external interface. The sessions were recorded with multiple cameras and microphones covering different angles of the room.

Each child attended between one and five daily sessions (3.4 on average), yielding a total of 84 sessions with an average length of 12.4 min from 25 children (19 males, 6 females), 13 participating in researcher-only sessions and 12 participating in robot-assisted sessions, with an average age of 8.2 yrs (standard deviation: 2.5 yrs), led by three different researchers (only one researcher per child). We divided the data in a speaker-independent manner with respect to the children. As there were overall three researchers in the data set, each child only interacting with one researcher, we group our data splits based on the researchers. We do so to avoid overfitting of our machine learning models on person-specific speech characteristics of the researchers, who largely contribute to the vocalisations. An overview of the partitions is given in **Table 1**; the partitioning is being used for both types of experiments.

The sessions were richly annotated in terms of both audio and video data, following a pre-defined annotation protocol, including instructions for speaker diarisation, vocalisation type, occurrences of echolalia, type of non-verbal vocalisations, as well as emotion in terms of valence and arousal. For our study, we exploit the speaker diarisation annotations, the origin of the labels for voice activity detection, as well as valence and arousal annotations as labels for the SER system.

2.1. Speaker Diarisation Annotation

The Speaker Diarisation (in the British study arm) was performed by fluent English speakers utilising the ELAN annotation tool[2]. The task was to highlight any vocalisation of any speaker present within the session, i.e., the child, the researcher, any additionally present person (generally a teacher), or the robot Zeno. The annotators were able to base their decisions on a combination of the available video streams together with one of the video cameras' native audio recordings, as well as the according depiction of the raw audio wave form. The annotation tool further allowed annotators to skip to arbitrary points of the recording. Overall, each session was assessed by one annotator.

2.2. Emotion Annotation

The emotion annotations in the database aim to capture the emotional dimensions valence and arousal, i.e., continuous representations of how positive or negative (valence) and how sleepy or aroused (arousal) an emotional state seems. Emotional dimensions are a commonly used alternative to categorical emotions, like happy, angry, etc., when assessing people's emotional states. Five expert raters, all either native or near native English speakers, annotated their perception of the valence and arousal values expressed by the children in each session under consideration of the same video and audio data as in the speaker diarisation task. For the annotation process, raters were given a joystick (model *Logitech Extreme 3D Pro*) in order to annotate valence and arousal separately. While annotators were watching the recordings of the sessions, they changed the position of the joystick, which was continuously sampled with a sampling rate of 50 Hz and indicated degree and sign of the estimated valence or arousal values (positive in an *up* position, negative in a *down* position). The annotations of the different annotators for each session are summarised in a single gold standard sequence utilising the evaluator weighted estimator (EWE) (Schuller, 2013) gold standard. The EWE gold standard is commonly used in emotion recognition tasks (Ringeval et al., 2017, 2019) and considers annotator-specific weights depending on the pairwise correlation of the annotations. For our experiments, we use only one emotion label per second by calculating a second-wise average over the gold standard annotations.

3. METHODOLOGY

To explore the task of VAD-based SER, we employ two separate models based on feature extraction and recurrent neural networks (RNNs) with long short-term memory (LSTM) cells. The first component is a VAD component and the second is a SER component. The VAD model is presented with 1 s long audio chunks, and aims to label segments of the audio signal with a vocalisation present. The SER model is then trained on

[2]https://archive.mpi.nl/tla/elan

TABLE 1 | Overview of the three partitions of the data set: train, development (dev.), and test.

Partition	# children	# sessions	# researchers	child vocalisations	total vocalisations	total duration
Train	12	41	1	1:26:39	6:42:15	9:43:34
Dev.	4	15	1	0:18:27	1:24:37	3:14:35
Test	9	28	1	0:32:42	2:35:21	4:22:03
Overall	25	84	3	2:17:49	10:42:14	17:20:13

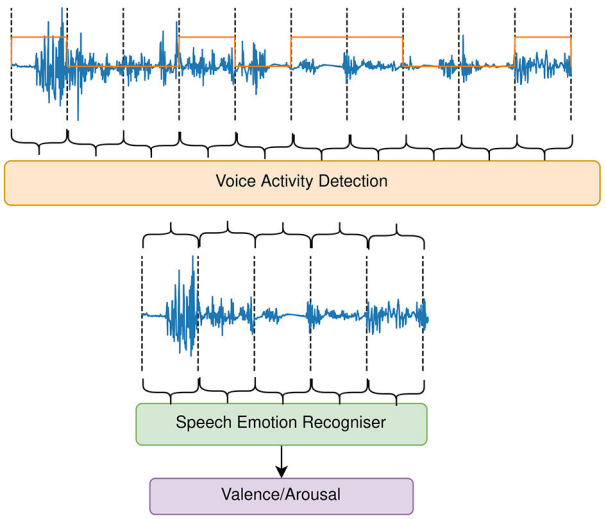


FIGURE 1 | Sequential employment of the (child) voice activity detection (VAD) and the speech emotion recognition (SER) system. The VAD system removes 1 s chunks of the audio signal, where no voice is detected. Any remaining 1 s chunks are fed into the SER system to predict continuous values for valence and arousal.

audio segments presumably containing speech, with the aim of predicting the affective dimensions valence and arousal in a continuous manner. An illustration of the combined system is depicted in **Figure 1**.

3.1. Voice Activity Detection

As the target of the VAD system is to remove as much information-shallow data from the audio data as possible, we compare several approaches here: at a first level, we try to filter for all vocalisations with general VAD systems, one specifically trained on our data set, the other one being an implementation of the WebRTC VAD system[3] (Google, 2021), commonly used as a comparison for other VAD systems, e.g., (Salishev et al., 2016; Nahar and Kai, 2020). The aggressiveness score of the WebRTC VAD is set equal to one. Additionally, we use the ground truth annotations for all vocalisations as a gold standard for a general VAD system. At a second level, we try to filter out only child vocalisations, which presumably contain the most information about the children's affective state. For this, we train a child VAD system on the data set mentioned above and use the ground truth annotations for child vocalisations for further comparison. Evaluations of the different impacts of general VADs and the child VAD are of further interest, as some information about the children's affective state could be retrieved from the interaction between the child and the researcher. Besides, a worse performance of the child VAD system compared to more robust general VAD systems could lead to detections of ambient noise and therefore potentially have a negative impact on the SER task.

Given the potentially short duration of vocalisations, we extract 130 ComParE2016 LLDs with a frame size of 10 ms and a hop size of 10 ms from the raw audio signal utilising the openSMILE toolkit (Eyben et al., 2010). The audio features are then fed into a two-layer bi-directional RNN with LSTM cells and a hidden layer size of 128 units, followed by a dense layer with a single output neuron indicating the confidence in the voice detection. The neural network architecture is similar to Hagerer et al. (2017), but has been adjusted based on preliminary experiments. We utilise a fixed sequence length of 100 samples during training time, i.e., the audio stream is cut into samples of 1 s length. During training of this regression problem, each frame is assigned the label 1 if speech is present or the label 0 if it is not.

The VAD models are trained for 8 epochs with a batch size of 256 utilising the Adam optimiser with a learning rate of 0.01 and mean square error (MSE) loss. We choose the rather small number of epochs based on the large amount of samples. Given that each second provides 100 sequence elements to the LSTM, the training includes around 2 000 optimisation steps. For the evaluation of the VAD system, we compute a receiver operating characteristic (ROC) curve, i.e., we vary the confidence threshold of the system, for which a frame is recognised as a detection in order to depict the relationship between true positive rate (TPR) and false positive rate (FPR).

For inference, we choose a confidence threshold, which corresponds to the equal-error-rate (EER), i.e., equal values of FPR and $1 - \text{TPR}$, visualised by the intersection of the ROC curve and the bisectional line $\text{TPR} + \text{FPR} = 1$. The VAD system is then used as a preprocessing step for the SER task, such that each second of audio is classified as containing voice activity if at least 25 % of the frames contained in 1 s are above the EER confidence threshold.

3.2. Speech Emotion Recognition

For the SER task we use 1 s chunks of audio extracted with the VAD system in order to predict a single continuous-valued valence (and arousal, respectively) value per audio chunk. The applied VAD system therefore impacts the SER task by the selection of audio chunks guided by the hypothesis that audio

[3]https://github.com/wiseman/py-webrtcvad

with child vocalisations contains the most information about the perceived affective states of the children and therefore leads to higher performance in the SER task.

Subsequently, we extract 88 functional features for each 1 s audio chunk according to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), a comprehensive expert-based audio feature selection (Eyben et al., 2015). The resulting sequence of features from one session is then used as an input to our deep learning model consisting of two RNN layers with LSTM cells and a hidden layer size of 128 units, followed by a dense layer with 128 neurons, a rectified linear unit (ReLU) activation and a dropout rate of 0.3. A final dense layer with a single neuron outputs the valence or arousal prediction for our task. The identical network architecture is trained independently for valence and arousal, respectively. With our methodology, we follow (Stappen et al., 2021), with an adjusted model architecture based on preliminary experiments.

The SER models are trained for 180 epochs with full batch optimisation – each session producing one sequence – utilising the Adam optimiser with a learning rate of 0.0001 and MSE loss. The much larger number of epochs compared to the VAD experiments is chosen based on the full batch optimisation, i.e., only one optimisation step is performed each epoch.

4. EXPERIMENTS

All experiments are implemented in Python 3 (Van Rossum and Drake, 2009), as well as TensorFlow 2 (Abadi et al., 2015) for deep learning models and training. The code is publicly available under[4].

4.1. Voice Activity Detection

For our VAD experiments, we train the architecture as described in section 3.1 with two different targets: (i) to recognise only child vocalisations, including overlap with others vocalisations and (ii) to recognise any vocalisation, including overlapping vocalisations. The two approaches are evaluated on the respective tasks. We thereby aim to evaluate the feasibility of training a general VAD system for the specifics and limitations of our dataset and to further investigate the presumably more challenging task of training a specialised VAD system for children with autism. Besides the evaluation of the VAD systems based on their raw performance, we further assess their impact on the SER task in the following section.

We report ROC-curves for both the child VAD system and the general VAD system on the respective tasks in **Figure 2**, as well as the EER and area-under-the-curve (AUC) in **Table 2**.

4.2. Speech Emotion Recognition

As described in section 3, we utilise our child VAD system and the general VAD system trained in the previous section in order to extract 1 s chunks from the session recordings if 25 out of the 100 frames within one second have a prediction confidence above the EER threshold. In a similar way 1 s chunks are extracted if the WebRTC VAD predicts a voice activity for at least 0.25 s of the audio. In the same manner, we use the ground truth annotations of child vocalisations, as well as ground truth annotations of all speakers to mimic a perfect child VAD and a perfect general VAD system. As a baseline, we use the audio without any VAD-based preprocessing (All Audio). **Figure 3** shows the distribution of valence and arousal values across partitions, as well as the test partition's adjusted distribution after filtering via the VAD systems and vocalisation annotations.

For evaluation, we use the root mean squared error (RMSE), as well as the concordance correlation coefficient (CCC) according to Lin (1989), which is defined between two distributions $\boldsymbol{x}$ and $\boldsymbol{y}$ as

$$\mathrm{CCC}(\boldsymbol{x},\boldsymbol{y}) = \frac{\rho(\boldsymbol{x},\boldsymbol{y})\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}}{\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{y}}^2 + (\mu_{\boldsymbol{x}} - \mu_{\boldsymbol{y}})^2}, \tag{1}$$

with the correlation coefficient ρ, as well as the mean μ and the standard deviation σ of the respective distribution. As the CCC is designed as a metric for sequences and has an inherent weakness for short sequences and sequences with little variation, we combine all predictions and labels from one data partition to two respective sequences when calculating the CCC. The results for valence and arousal are summarised in **Table 3**.

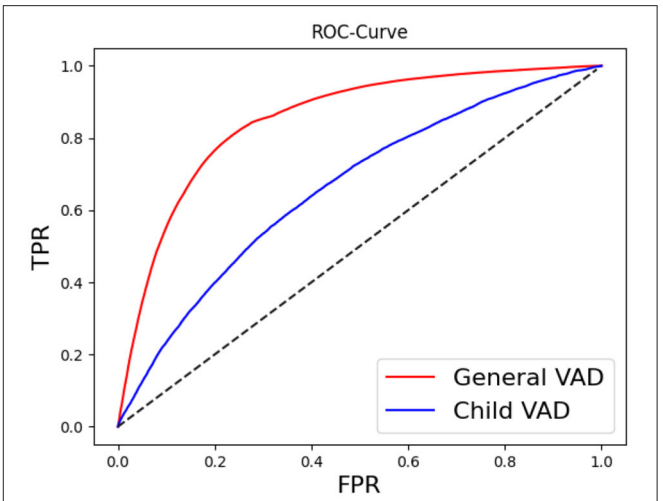


FIGURE 2 | Receiver operating characteristic (ROC)-curve of voice activity detection trained for child vocalisations specifically and for all vocalisations.

TABLE 2 | Equal-error-rates (EERs) and area-under-the-curve (AUC) for the child voice activity detection system and the general voice activity detection system evaluated on the respective task.

VAD System	EER	AUC
Child VAD	0.381	0.662
General VAD	0.215	0.850

[4]https://github.com/EIHW/VAD_SER_pipeline_ASC

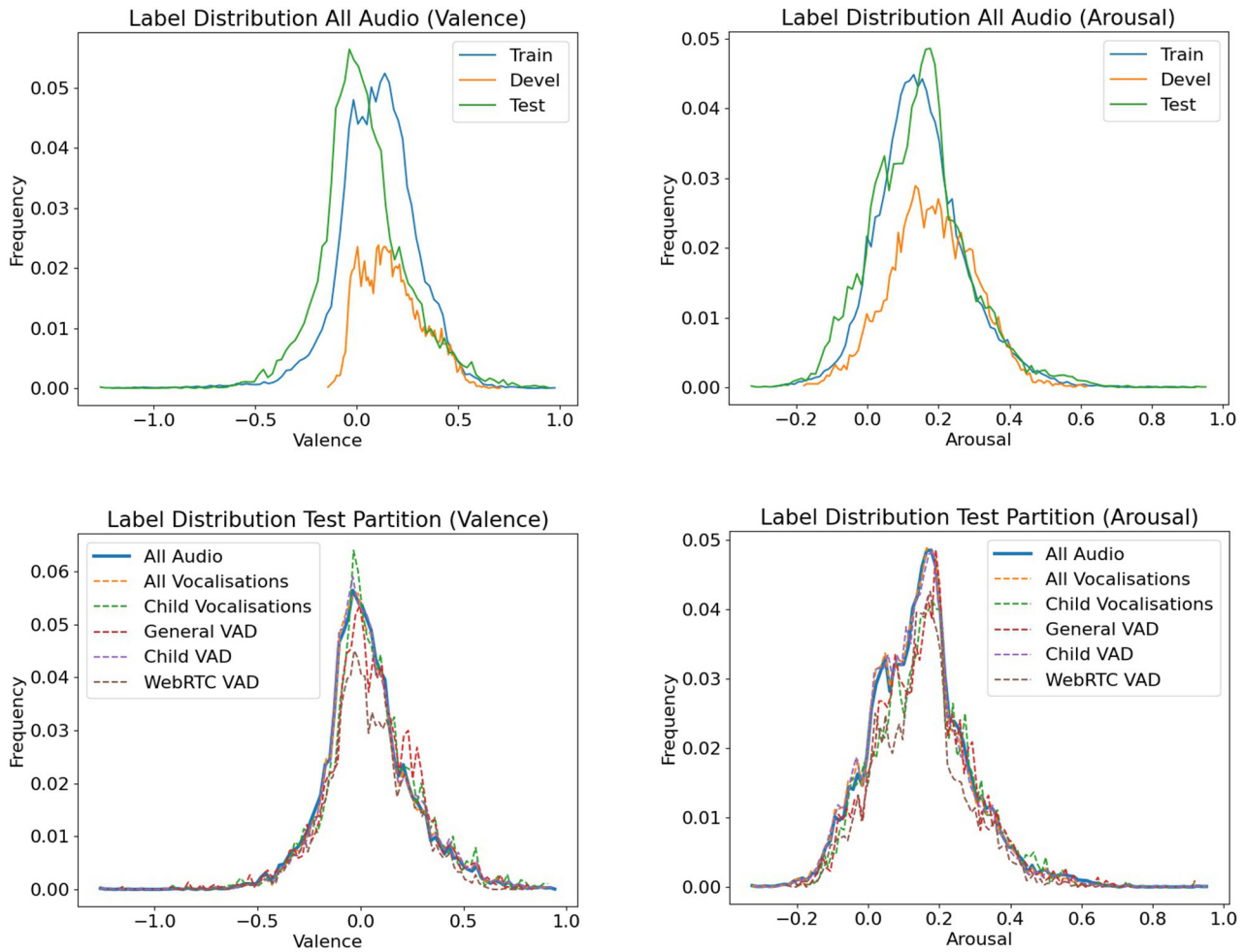


FIGURE 3 | Distributions of valence labels (left) and arousal labels (across) considering all audio data without VAD preprocessing across different partitions (top), as well the adjusted distributions of the test partition after preprocessing via the different VAD systems and vocalisation annotations (bottom).

TABLE 3 | Results of the speech emotion recognition (SER) task.

VAD System	# samples detected	Valence (CCC/RMSE)		Arousal (CCC/RMSE)	
		Dev	Test	Dev	Test
Child VAD	17,944	**0.200**/0.201	0.021/0.245	0.201/0.121	**0.168**/**0.138**
General VAD	40,013	0.012/**0.160**	0.117/0.260	0.100/0.120	0.154/0.142
WebRTC VAD	29,918	0.140/0.183	0.063/**0.224**	**0.263**/**0.107**	0.098/0.152
GT child vocalisations	10,961	0.153/0.169	0.085/0.277	0.182/0.115	0.145/0.143
GT all vocalisations	47,184	-0.032/**0.160**	**0.120**/0.231	0.166/0.114	0.105/ 0.156
All Audio	62,370	0.133/0.162	0.024/0.231	0.093/0.122	0.049/0.152

We report concordance correlation coefficient (CCC) and root mean squared error (RMSE) for valence and arousal with respect to the voice activity detection (VAD) system and ground truth (GT) annotations utilised for preprocessing of the data, as well as the baseline without a VAD preprocessing step (All Audio). Bold values indicate the best performance in each column.

5. DISCUSSION

Figure 2 and **Table 2** show that both a general voice activity system, as well as a child-specific voice activity system with a performance above-chance level can be trained from the data at hand. However, the general VAD system shows a clearly superior performance compared to the child-specific one. One apparent reason for this results from the dataset itself. **Table 1** highlights

that the dataset offers more than four times as many annotations for the general VAD compared to child VAD system, leading to a more unbalanced child VAD task. Moreover, the task of training a VAD system specialised and focused solely on autistic children appears to be generally more challenging, as the model not only needs to detect speech-typical characteristics, but also has to differentiate between speech characteristics of the speakers, i.e., the model has to find common patterns in vocalisations of children with different language levels and distinguish those from patterns in the researchers' voices. The different language levels of the children involved in the study, as well their unique ways of expression most likely made it difficult to uncover common characteristics.

Table 3 further shows that all considered VAD systems have a largely varying sensitivity. By the term sensitivity we mean in this context the total number of voice detection events independent of the correctness of the detections. The sensitivity of the child VAD system, aiming to detect the ground truth child vocalisations, can be considered too high with almost twice as many detected events compared to the number of human target annotations. The two remaining VAD systems naturally seem to be much more sensitive than the child VAD, as they do not aim at filtering out the child vocalisations only. However, both the out-of-the-box WebRTC VAD, as well as our trained general VAD both seem to show a lower sensitivity than the ground truth annotations of all speaker vocalisations with the WebRTC's deviation of detection events being considerably higher.

The top part of **Figure 3** shows that there are no large difference between the label distributions in the train and test partition for the SER task. The development partition however deviates substantially. The bottom part of **Figure 3** indicates the difference in emotion label distributions in the test set caused by the preprocessing via the different VAD approaches. Even though the choice of the VAD system has only little impact on the label distribution and therefore should not give any considerable, label-related advantage to any of the resulting SER experiments, it still affects the comparability of the results as it alters the test data.

According to **Table 3**, the best test results for arousal in our SER experiments are obtained with the child VAD preprocessing, even outperforming the preprocessing based on ground truth annotations. These results seem in-line with the hypothesis that considering only child vocalisations could improve the performance of SER systems for autistic children and they further suggest a reasonable system performance of the child VAD. However, this analysis only holds to a certain amount for the arousal development set and even less for the valence experiments, which tend to achieve lower performance in acoustic SER tasks compared to arousal experiments. Nevertheless, the VAD-based systems outperform the VAD-less system in most experiments, suggesting a clear advantage of VAD-based systems for the task at hand. Limitations to the expressiveness of the results discussed here have to be taken into account, as small improvements together with a low overall performance of the SER models are not always consistent across the investigated evaluation metrics.

Future work shall further investigate the impact of a child-specific VAD system in a multi-modal emotion recognition approach. Given the complex scenarios resulting from sessions with autistic children, it is inevitable that not all modalities are available at all times, as children for instance move out of the focus of the cameras or are silent for an extended period of time. The detection and consideration of those missing modalities, for instance in form of a VAD system contributing to a weighted feature fusion, might therefore have a substantial influence on model behaviour and even help with explaining the decisions of applied approaches.

6. CONCLUSION

With this contribution, we discussed the feasibility and utility of a VAD system, specifically trained on autistic child vocalisations, for SER tasks in robot-assisted intervention sessions for autistic children in order to improve programme success for children with autism. Given the size as well as the noise-heavy quality of the dataset, we showed that the voice activity component could be trained with reasonable performance, while being inferior to an identically trained general VAD system. Our results further suggest that the use of VAD systems, and in particular child VAD systems, could lead to slight improvements of continuous SER for autistic children, even though an overall low performance across SER models, most likely caused by the challenges of the task at hand, weaken the expressiveness of the results. Further research based on this work will examine the use of child VAD systems as a basis for missing data strategies in multi-modal SER tasks.

DATA AVAILABILITY STATEMENT

The data set presented in this article cannot be made publicly available in its current form due to ethical reasons.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee at UCL Institute of Education, University College London (REC 796). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MM: literature search, experimental design, experiment implementation, computational analysis, manuscript preparation, and drafting manuscript. AB: experimental design, feature extraction, literature search, annotation guideline development and implementation, and manuscript preparation. KB-P: literature search, manuscript preparation, consultancy on experimental task and study-related aspects of child development. SL: experimental design and advise on implementation. AA, JS, and EA: study design and implementation, annotation guideline development and implementation. TT: study design and implementation. EP:

study design and manuscript preparation. NC: study design, annotation guideline development and implementation, and manuscript preparation. MP: study design. BS: study design and manuscript editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the EU's Horizon 2020 Programme under grant agreement No. 688835 (RIA DE-ENIGMA), the European Commission's Erasmus+ project – 'EMBOA, Affective loop in Socially Assistive Robotics as an intervention tool for children with autism' under contract No. 2019-1-PL01-KA203-065096, and the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

ACKNOWLEDGMENTS

We are enormously grateful to all of the children, parents, teachers, researchers, and schools who so generously took part in our study. We would also like to thank the DE-ENIGMA student volunteers at the UCL Institute of Education for their contributions to school studies and data annotation.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems.* Available online at: https://www.tensorflow.org/ (accessed December 13, 2021).

Akçay, M. B., and Oğuz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* 116, 56–76. doi: 10.1016/j.specom.2019.12.001

Alghifari, M. F., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., Janin, Z., et al. (2019). On the use of voice activity detection in speech emotion recognition. *Bull. Elect. Eng. Inf.* 8, 1324–1332. doi: 10.11591/eei.v8i4.1646

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, Arlington, VA: APA.

Amiriparian, S., Baird, A., Julka, S., Alcorn, A., Ottl, S., Petrović, S., et al. (2018). "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Proceedings of the Interspeech 2018* (Hyderabad), 2334–2338.

Baird, A., Amiriparian, S., Cummins, N., Alcorn, A. M., Batliner, A., Pugachevskiy, S., et al. (2017). "Automatic Classification of autistic child vocalisations: a novel database and results," in *Proceedings of the Interspeech 2017* (Stockholm), 849–853.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi: 10.1007/s10579-008-9076-6

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 190–202. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *MM '10* (New York, NY: Association for Computing Machinery), 1459–1462.

Google (2021). *WebRTC.* Available online at: https://webrtc.org/. (accessed December 13, 2021).

Hagerer, G., Pandit, V., Eyben, F., and Schuller, B. (2017). "Enhancing lstm rnn-based speech overlap detection by artificially mixed data," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, Erlangen.

Harár, P., Burget, R., and Dutta, M. K. (2017). "Speech emotion recognition with deep learning," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)* (Noida: IEEE), 137–140.

Howlin, P., Baron-Cohen, S., and Hadwin, J. (1999). *Teaching Children With Autism to Mind-Read: A Practical Guide for Teachers and Parents*. Chichester: J. Wiley & Sons Chichester.

Hudson, C. C., Hall, L., and Harkness, K. L. (2019). Prevalence of depressive disorders in individuals with autism spectrum disorder: A meta-analysis. *J. Abnormal Child Psychol.* 47, 165–175. doi: 10.1007/s10802-018-0402-1

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). Deap: A database for emotion analysis ;using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Kopp, S., Beckung, E., and Gillberg, C. (2010). Developmental coordination disorder and other motor control problems in girls with autism spectrum disorder and/or attention-deficit/hyperactivity disorder. *Res. Develop. Disabil.* 31, 350–361. doi: 10.1016/j.ridd.2009.09.017

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268. doi: 10.2307/2532051

Lord, C., Elsabbagh, M., Baird, G., and Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *Lancet* 392, 508–520. doi: 10.1016/S0140-6736(18)31129-2

Marchi, E., Schuller, B., Baird, A., Baron-Cohen, S., Lassalle, A., O'Reilly, H., et al. (2018). The asc-inclusion perceptual serious gaming platform for autistic children. *IEEE Trans. Games* 11:328–339. doi: 10.1109/TG.2018.2864640

Nahar, R., and Kai, A. (2020). "Effect of data augmentation on dnn-based vad for automatic speech recognition in noisy environment," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)* Kobe, 368–372.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al. (2019). "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop AVEC '19* (New York, NY: Association for Computing Machinery), 3–12.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., et al. (2017). "Avec 2017: real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17* (New York, NY: Association for Computing Machinery), 3–9.

Rudovic, O., Lee, J., Dai, M., Schuller, B., and Picard, R. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science* 3:eaao6760. doi: 10.1126/scirobotics.aao6760

Salishev, S., Barabanov, A., Kocharov, D., Skrelin, P., and Moiseev, M. (2016). "Voice activity detector (vad) based on long-term mel frequency band features," in *Text, Speech, and Dialogue*, eds P. Sojka, A. Horák, I. Kopeček, and Pala, K. (Cham: Springer International Publishing), 352–358.

Schadenberg, B. R., Reidsma, D., Evers, V., Davison, D. P., Li, J. J., Heylen, D. K., et al. (2021). Predictable robots for autistic children-variance in robot behaviour, idiosyncrasies in autistic children's characteristics, and child–robot engagement. *ACM Trans. Comput. Human Interact.* 28, 1–42. doi: 10.1145/3468849

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association* (Lyon).

Schuller, B. W. (2013). *Intelligent Audio Analysis* (Berlin: Springer Publishing Company, Incorporated).

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 90–99. doi: 10.1145/3129340

Shen, J., Ainger, E., Alcorn, A., Dimitrijevic, S. B., Baird, A., Chevalier, P., et al. (2018). Autism data goes big: A publicly-accessible multi-modal database of child interactions for behavioural and machine learning research. In *International Society for Autism Research Annual Meeting* (Kansas City, MO).

Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., et al. (2020). "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop MuSe'20* (New York, NY), 35–44.

Stappen, L., Meßner, E.-M., Cambria, E., Zhao, G., and Schuller, B. W. (2021). "Muse 2021 challenge: Multimodal emotion, sentiment, physiological-emotion, and stress detection," in *Proceedings of the 29th ACM International Conference on Multimedia MM '21* (New York, NY), 5706–5707.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "Avec 2013 - the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge AVEC '13* (New York, NY), 3–10.

Van Rossum, G., and Drake, F. L. (2009). *Python 3 Reference Manual*. (Scotts Valley, CA: CreateSpace).

Zaboski, B. A., and Storch, E. A. (2018). Comorbid autism spectrum disorder and anxiety disorders: a brief review. *Future Neurol.* 13, 31–37. doi: 10.2217/fnl-2017-0030

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neurorobotics

METHODS
published: 24 February 2022
doi: 10.3389/fnbot.2022.834952

Linking Multi-Layer Dynamical GCN With Style-Based Recalibration CNN for EEG-Based Emotion Recognition

*Guangcheng Bao[1], Kai Yang[1], Li Tong[1], Jun Shu[1], Rongkai Zhang[1], Linyuan Wang[1], Bin Yan[1] and Ying Zeng[1,2]**

[1] Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategic Support Force Information Engineering University, Zhengzhou, China, [2] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

Electroencephalography (EEG)-based emotion computing has become one of the research hotspots of human-computer interaction (HCI). However, it is difficult to effectively learn the interactions between brain regions in emotional states by using traditional convolutional neural networks because there is information transmission between neurons, which constitutes the brain network structure. In this paper, we proposed a novel model combining graph convolutional network and convolutional neural network, namely MDGCN-SRCNN, aiming to fully extract features of channel connectivity in different receptive fields and deep layer abstract features to distinguish different emotions. Particularly, we add style-based recalibration module to CNN to extract deep layer features, which can better select features that are highly related to emotion. We conducted two individual experiments on SEED data set and SEED-IV data set, respectively, and the experiments proved the effectiveness of MDGCN-SRCNN model. The recognition accuracy on SEED and SEED-IV is 95.08 and 85.52%, respectively. Our model has better performance than other state-of-art methods. In addition, by visualizing the distribution of different layers features, we prove that the combination of shallow layer and deep layer features can effectively improve the recognition performance. Finally, we verified the important brain regions and the connection relationships between channels for emotion generation by analyzing the connection weights between channels after model learning.

Keywords: electroencephalography (EEG), emotion recognition, graph convolutional neural networks (GCNN), convolutional neural networks (CNN), style-based recalibration module (SRM)

OPEN ACCESS

Edited by:
Yuan Zong,
Southeast University, China

Reviewed by:
Sheng Ge,
Southeast University, China
Yu Zhang,
Lehigh University, United States

***Correspondence:**
Ying Zeng
yingzeng@uestc.edu.cn

Received: *14 December 2021*
Accepted: *24 January 2022*
Published: *24 February 2022*

Citation:
Bao G, Yang K, Tong L, Shu J, Zhang R, Wang L, Yan B and Zeng Y (2022) Linking Multi-Layer Dynamical GCN With Style-Based Recalibration CNN for EEG-Based Emotion Recognition. Front. Neurorobot. 16:834952. doi: 10.3389/fnbot.2022.834952

INTRODUCTION

Human emotion is a state that reflects the complex mental activities of human beings. In recent years, new modes of human-computer interaction, such as voice, gesture, and force feedback, have sprung up. Although significant progress has been made in the field of human-computer interaction, it still lacks one of the indispensable functions of human-computer interaction, emotional interaction (Sebe et al., 2005). However, the prerequisite for realizing human-computer emotional interaction is to recognize human emotional state in real time. Human emotions

come in many forms, which can be recognized by human facial expressions (Harit et al., 2018), body movements (Ajili et al., 2019), and physiological signals (Goshvarpour and Goshvarpour, 2019; Valderas et al., 2019). But humans can control their facial expressions, body movements to hide or disguise their emotions, and physiological signals such as electroencephalogram, electrocardiogram, and electromyography have the advantage of being difficult to hide or disguise. With the rapid development of non-invasive, portable, and inexpensive EEG acquisition equipment, EEG-based emotion recognition has attracted the attention of researchers.

EEG signals are collected through electrodes distributed in various brain regions on the cerebral cortex, which has the advantages of non-invasiveness, convenience, and fast. In addition, EEG have the advantages of high time resolution, and are considered to be one of the most reliable signals. However, EEG also has some shortcomings, such as low spatial resolution and low signal-to-noise ratio. Moreover, the EEG is non-stationary, and there are great differences among subjects. Studies have shown that some cortical and subcortical brain systems may play a key role in the evaluation or reaction phase of emotion generation (Clore and Ortony, 2008; Kober et al., 2008). However, it is difficult to use EEG to model brain activity and interpret the activity state of brain regions. Therefore, high-precision recognition of emotions based on EEG is still a challenge.

In these decades of development, researchers have proposed many machine learning and signal processing methods for EEG emotion recognition. Traditional EEG emotion recognition methods usually include two aspects: EEG feature extraction and emotion classification used to distinguish emotion categories. The EEG features used for emotion recognition are mainly divided into three parts: time-domain features, frequency-domain features, and time-frequency features. Time domain features mainly include statistics (Jenke et al., 2017), Hjorth features (Hjorth, 1970), non-stationary index (NSI) (Kroupi et al., 2011), fractal dimension (Sourina and Liu, 2011; Liu and Sourina, 2013), sample entropy (Jie et al., 2014), and higher order crossings (HOC) (Petrantonakis and Hadjileontiadis, 2011). These features mainly describe the temporal characteristics and complexity of EEG signals. Frequency domain feature refers to the use of Fourier Transform (TF) and other information analysis methods to transform EEG signals from time domain to frequency domain, and then extract emotion related information from frequency domain as features. At present, one of the most commonly used frequency domain feature extraction methods is to divide EEG signals into five bands: Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Beta (12–30 Hz), Gamma (30–64 Hz). Emotion Feature Extraction in frequency domain mainly includes power spectral density (PSD) (Alsolamy and Fattouh, 2016), differential entropy (DE) (Duan et al., 2013), differential asymmetry (DASM) (Liu and Sourina, 2013), rational asymmetry (RASM) (Lin et al., 2010), and differential causality (DCAU) (Zheng and Lu, 2015). Time frequency feature refers to the use of time-frequency analysis methods, such as short-time Fourier transform (STFT) (Lin et al., 2010), wavelet transform (WT) (Jatupaiboon et al., 2013) and Hilbert Huang transform (HHT) (Hadjidimitriou and Hadjileontiadis, 2012). Due to the typical non-stationary signal of EEG, the traditional frequency domain analysis method such as Fourier transform is not suitable for analyzing the signal whose frequency changes with time, while the time-frequency analysis method provides the joint distribution information of time domain and frequency domain.

The classifiers based on EEG emotion recognition are mainly divided into traditional machine learning method and deep network method. Among the traditional machine learning methods, support vector machine (SVM) (Koelstra et al., 2010; Hatamikia et al., 2014), k-nearest neighbor (KNN) (Mehmood and Lee, 2015), linear discriminant analysis (LDA) (Zong et al., 2016) and other methods are used for emotion classification based on EEG. Among them, SVM has better performance and is usually used as baseline classifier. However, due to the complexity of EEG-based emotion features, the current method is to extract the artificial features, and then use machine learning method to classify the extracted features, which leads to the traditional machine learning method cannot get better classification performance. Therefore, researchers turn their attention to deep learning methods. Zhang X. et al. (2019) summarized the work of using deep learning technology to study brain signals in recent years. In EEG-based emotion recognition based on neural network, the input is usually artificial features, and then the neural network is used to learn deeper features to improve the performance of emotion recognition. Zheng et al. (2014) used deep belief networks (DBNs) to learn and classify the frequency bands and channels of EEG-based emotion, which is a great improvement compared to SVM. In recent years, many deep networks have emerged in this field to extract spatiotemporal features of EEG-based emotions. Jia et al. (2020) proposed a spatial-spectral-temporal based Attention 3D Dense Network (SST-EmotionNet) for EEG emotion recognition. Li Y. et al. (2018) and Li et al. (2020) proposed BiDANN and BiHDM networks for EEG emotion recognition, considering the asymmetry of emotion response between left and right hemispheres of human brain. Li et al. (2021) proposed a Transferable Attention Neural Network (TANN), which considers local and global attention mechanism information for emotion recognition. In addition, some researchers considered the spatial information of EEG features, and arrange and distribute the features of each channel through the physical location before inputting them into the neural network. Li J. et al. (2018) arranged the DE features of different leads into a two-dimensisonal feature matrix according to their physical locations before entering the network. Bao et al. (2021) mapped the DE feature to a two-dimensional feature matrix through an interpolation algorithm according to the physical location.

Although researchers currently use neural network to consider the temporal and spatial information, the EEG signals of each channel are distributed in different regions of the brain, which can be regarded as a non-Euclidean data. However, convolution neural network processing EEG will ignore the spatial distribution information. In order to solve this problem, graph convolution neural network (GCNN) (Defferrard et al.,

2016) is introduced to process non-Euclidean data. Zhao et al. (2022) proposed a new dynamic graph convolutional network (dGCN) to learn the potentially important topological information. Song et al. (2020) used dynamic graph convolution network (DGCNN) for the first time in the EEG-based emotion recognition task. The network constructed graph data more in line with the brain activity state by learning the connections between different channels, and achieved better performance. Zhong et al. (2020) proposed a regularized graph neural network (RGNN), which considers the global and local relationships of different EEG channels. Zhang T. et al. (2019) proposed GCB-Net, which combines GCN and CNN to extract deep-level features and introduces a generalized learning system (BLS) to further improve performance.

However, the brain activity in emotional state is more complex, and multiple brain regions participate in interaction. The traditional convolutional neural network cannot effectively learn the interaction between brain regions.

However, the networks proposed in the above studies all use one layer of GCN, and (Kipf and Welling, 2017) concluded that using 2-3 layers is the best. In addition, the receptive field of single-layer GCN is limited and cannot extract spatial information well. The brain activity in emotional state is more complex, and multiple brain regions participate in interaction. Therefore, the characteristics of single network learning are relatively single, and cannot well reflect the complex emotional state. For this reason, in this paper, we proposed a multi-layer dynamic graph convolutional network-style-based recalibration convolutional neural network (MDGCN-SRCNN) to extract shallow layer and deep layer features. The shallow layer features include the features of different levels of GCN learning, which contain different levels of spatial information. Deep layer features are mainly learned by SRCNN, because CNN has a strong ability to learn abstract features. In addition, by adding the style-based recalibration module, when CNN extracts features, it emphasizes the information related to emotion and ignores other information, which greatly enhances the representation ability of CNN. The shallow layer and deep layer features are connected to form a multi-level rich feature, and finally the fully connected layer search is used to classify the features that are distinguishable from various emotions.

The main contributions of this paper are as follows:

1) MDGCN-SRCNN framework composed of multi-layer GCN and multi-layer style-based recalibration CNN is used to learn features at different levels. In the shallow layer network, GCN learns different levels of spatial features. In the deep layer network, CNN learns abstract features, using a fully connected layer to fuse the shallow layer spatial features with deep layer abstract features and search for highly distinguishable features for emotion classification.
2) SEED and SEED-IV data sets are used to verify the performance of the emotion recognition framework MDGCN-SRCNN proposed in this paper. Compared with the existing models, the framework proposed in this paper obtains the best results, which proves that the network proposed in this paper has a strong classification ability in EEG emotion recognition.

METHODS

In this section, we introduce in detail the framework MDGCN-SRCNN proposed in this paper.

Model Framework

As shown in **Figure 1**, we propose the MDGCN-SRCNN framework for EEG-based emotion recognition tasks. The MDGCN-SRCNN model consists of four blocks: graph construction block, graph convolutional block, SRM-based convolutional block and classification block. We will give the specific model architecture below.

Graph Construction Block

We considered that EEG is non-Euclidean data. EEG data is collected by many electrodes, which are distributed in different parts of the brain. The construction of a graph requires three parts: nodes, features, and edge sets. For EEG signals, the nodes of the graph are the EEG signal channels. Different acquisition devices have different channel numbers. Currently, 16 channels, 32 channels, 64 channels, and 128 channels are commonly used.

The feature is the data collected by each channel, which can be the original collected data or manually extracted features. Most of the current researches use artificial features for EEG-based emotion recognition. Therefore, in this paper, the DE features of five bands are extracted as the features of the graph. Short Time Fourier Transform (STFT) is used to transform each segment of data. The formula of DE features is as follows:

$$h(X) = -\int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx$$

$$= \frac{1}{2}\log\left(2\pi e\sigma^2\right) \tag{1}$$

where $X \sim N(\mu, \sigma^2)$ is the input raw signal, x is a variable, and e and π are constants.

The edge set of the graph describes the connected relationship between nodes. Currently, Pearson correlation coefficient (PCC) (Faskowitz et al., 2020), coherence value (Wagh and Varatharajah, 2020), phase locked value (PLV) (Wang et al., 2019), and physical distance (Song et al., 2020) are mainly used to describe the connection between channels. In this paper, PCC is used as the weighted adjacency matrix of each channel, and its calculation formula is as follows:

$$A(i,j) = abs(PCC(x_i, x_j)) = abs(\frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}) \tag{2}$$

where $i, j = 1, 2, ..., n$, n are the number of channels of EEG signals. $x_{i/j}$ represents the EEG signal of the i/j-th channel. $cov(\cdot)$ refers to covariance.

Graph Convolutional Block

In the graph convolutional block, we use graph convolution network as a shallow layer network to learn the spatial information of EEG signals.

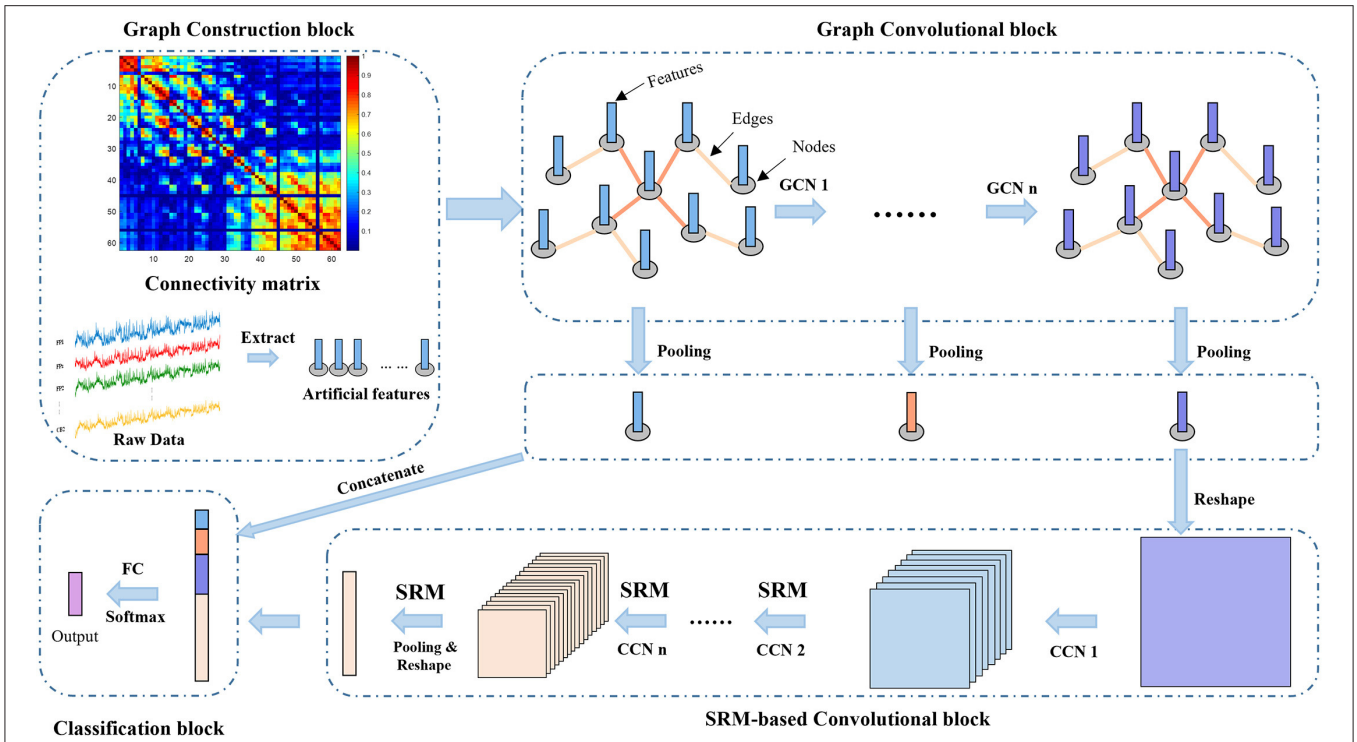


FIGURE 1 | The overall architecture of the MDGCN-SRCNN model consists of four blocks: graph construction block, graph convolutional block, SRM-based convolutional block, and classification block. The output of the model is a predicted label with probability.

The graph convolutional neural network is the network using convolution operations on the graph. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ refers to the vertex set with $|\mathcal{V}| = n$ nodes, and $\mathcal{E}$ is a set of edges between nodes. Data on vertex $\mathcal{V}$ can be represented by a set of feature matrix $\mathbf{X} \in \mathbb{R}^{n \times f}$, where n represents the number of nodes and f represents the feature dimension. The edge set $\mathcal{E}$ can be represented by a set of weighted adjacency matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$ describing the connections between nodes. Kipf and Welling (2017) proposed the propagation rules of Graph Convolutional Networks (GCN):

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{3}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix of the undirected graph $\mathcal{G}$ with additional self-connections, and $\mathbf{I}$ is the identity matrix. $\tilde{\mathbf{D}}$ is the diagonal matrix of $\tilde{\mathbf{A}}$, that is, $\tilde{\mathbf{D}}_{\mathbf{ii}} = \sum_j \tilde{\mathbf{A}}_{\mathbf{ij}}$, $\mathbf{W}^{(l)}$ is the training parameter matrix of the l-th layer. $\mathbf{H}^{(l)}$ is the transformation matrix of the l-th layer. σ refers to the activation function.

Next, GCN is analyzed by spectral convolution. The Laplacian operator matrix of the graph $\mathcal{G}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, the normalized Laplacian operator can be expressed as $\hat{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, and the characteristic decomposition of $\hat{\mathbf{L}}$ is $\hat{\mathbf{L}} = \mathbf{U}\lambda\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}$ is the orthonormal eigenvector matrix, and $\Lambda = diag(\lambda_1, ..., \lambda_n)$ is the diagonal matrix of the corresponding characteristic.

For the input signal X, the graph Fourier Transform is:

$$\hat{X} = \mathbf{U}^T X \tag{4}$$

The inverse Fourier transform is as follows:

$$X = \mathbf{U}\hat{X} \tag{5}$$

The generalized convolution on the graph can be defined as the product of signal X and filter g_θ in Fourier domain:

$$g_\theta * \mathbf{X} = \mathbf{U}((\mathbf{U}^{\mathrm{T}} g_\theta) \odot (\mathbf{U}^T \mathbf{X})) = \mathbf{U} g_\theta(\Lambda)\mathbf{U}^T\mathbf{X} \tag{6}$$

where $\odot$ refers to the element-wise multiplication, and $g_\theta(\Lambda) = diag(g_{\theta_1}, ..., g_{\theta_n})$represents the diagonal matrix with$n$spectral filtering coefficients.

If formula 6 is calculated directly, the amount of calculation is very large. For a large graph, it costs a lot to calculate all the features of Laplacian matrix, and it needs$\mathcal{O}(n^2)$times to multiply with Fourier basis$\mathbf{U}$. Therefore, Defferrard et al. (2016) proposed that the diagonal matrix $g_\theta(\Lambda)$ of spectral filtering coefficients can be approximated to K^{th} by the truncated expansion of Chebyshev polynomials:

$$g_\theta(\Lambda) \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\Lambda}) \tag{7}$$

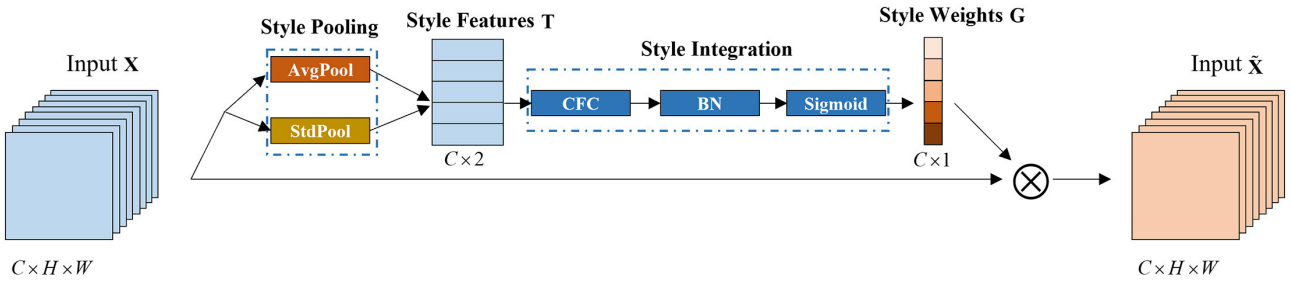


FIGURE 2 | SRM module. This module is mainly composed of two parts: style pooling and style integration. AvgPool refers to global average pooling, StdPool refers to global standard deviation pooling; CFC refers to the channel fully connected layer; BN refers to batch standardization.

where, $\tilde{\Lambda} = \frac{2}{\lambda_{\max}}\Lambda - \mathbf{I}$, $\lambda_{\max}$ refer to the largest eigenvalues of L. θ is a vector of Chebyshev coefficients. Chebyshev polynomials $T_k(\cdot)$ can be recursively computed as $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, where $T_0(x) = 1$ and $T_1(x) = x$. Then the graph filtering operation can be written as:

$$g_\theta * \mathbf{X} \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\mathbf{L}})\mathbf{X} \tag{8}$$

where $\tilde{\mathbf{L}} = \frac{2}{\lambda_{\max}}\tilde{\mathbf{L}} - \mathbf{I}$ is the normalized Laplacian. Then equation 8 is the Laplacian polynomial. In this case, the computational complexity is reduced to $\mathcal{O}(|\mathcal{E}|)$.

The GCN proposed by Thomas et al., based on Equation 8, sets $K = 1$, $\lambda_{\max} = 2$, $\theta_0 = -\theta_1$, then Equation 8 becomes Equation 3.

The EEG signal is converted into graph structure data by graph construction block and input into graph convolution network. Assuming that the initial data of the input graph rolled into the network is $\mathbf{H}^{(0)}$, the output of the l-th graph convolutional layer is shown in formula 3.

SRM-Based Convolutional Block

In the SRM-based convolutional block, we use a convolutional neural network combined with a style-based recalibration module as a deep layer network to learn abstract features related to emotions. The style-based recalibration module can be regarded as an attention module. But different from the traditional attention mechanism, the style-based recalibration module dynamically learns the recalibration weight of each channel based on the importance of the task style, and then merges these styles into the feature map, which can effectively enhance the representation ability of convolutional neural network.

Given an input $\mathbf{X} \in \mathbb{R}^{N\times C\times H\times W}$, SMR generates a channel-based recalibration weight $\mathbf{G} \in \mathbb{R}^{N\times C}$ through the style of X, where$N$refers to the number of samples in the minimum batch training, C represents the number of channels, H and W represent the spatial dimensions. This module is mainly composed of style pooling and style integration, as shown in **Figure 2**.

In the style pooling module, using the mean and standard deviation of the channel as style features, the extracted style features are $\mathbf{T} \in \mathbb{R}^{N\times C\times 2}$. Compared with other types of style features, using the mean and standard deviation of the channel can better describe the overall style information of each sample and channel (Lee et al., 2019). In the style integration module, the style features are converted into channel-related style weights through the channel fully connected layer, batch standard layer, and sigmoid activation function, which can simulate the importance of styles related to a single channel, thereby emphasizing, or suppressing them accordingly.

The output $\mathbf{H}^{(l)}$ of the convolutional network from the l-th graph is globally superposed and pooled as the input of the convolutional neural network, and then SRM is used in the middle of the convolutional layer to extract information related to the task style. Then each convolutional layer can be written as:

$$C_k = SRM(conv(C_{k-1}, h)) \tag{9}$$

where $h = 1, 2, 3$ represents the size of convolution kernel dimension, which is related to the input data type. In this paper, $h = 2$, $C_0 = Pool(\mathbf{H}^{(l)})$. $conv(\cdot, h)$ refers to h-dimensional convolution operation. k is the number of convolution layers. $SRM(\cdot)$ refers to SRM operation.

Classification Block

In the classification block, the learned features are input to the multi-layer fully connected layer for feature aggregation, and then the softmax layer is used for classification. After the shallow layer features and deep layer features are extracted, the multi-level features are spliced together, then the connected features can be written as:

$$\mathbf{F} = [Pool(\mathbf{H}^{(1)}), Pool(\mathbf{H}^{(2)}), ..., Pool(\mathbf{H}^{(l)}), Pool(C_k)] \tag{10}$$

where $Pool(\cdot)$ refers to the global pooling operation, in which the global sum pooling operation is used in the graph convolutional network. Compared with the maximum pooling and the average pooling, the sum pooling shows a stronger expressive ability (Xu et al., 2019). In convolutional neural networks, maximum pooling is used.

The classification prediction of the input EEG signal is:

$$\hat{y} = softmax(FC(\mathbf{F})) \tag{11}$$

where $FC(\cdot)$ refers to the fully connected layer operation, and $\hat{y} \in \mathbb{R}^C$ is the predicted label of class C.

We use DGCNN to learn the adjacency matrix of the graph by optimizing the loss function. Then use the optimizer to optimize the cross entropy loss:

$$\mathcal{L} = cross\_entropy(y, \hat{y}) + \alpha \|\Theta\|_2 \tag{12}$$

where y refers to the true label of the sample. θ is the matrix of all the parameters learned in the MDGCN-SRCNN model, α is the regularization coefficient, $cross\_entropy(\cdot)$ refers to the calculation of cross entropy, and $\|\cdot\|_2$ refers to the calculation of the second norm.

We use the Adam optimizer to learn the adjacency matrix$\mathbf{A}$:

$$\mathbf{A}^* = \mathbf{A} - lr\frac{\hat{m}^*}{\sqrt{\hat{v}^*} + \varepsilon} \tag{13}$$

$$\hat{m}^* = \frac{m^*}{1-\beta_1} = \frac{\beta_1 m + (1-\beta_1)\nabla_\theta \theta}{1-\beta_1} \tag{14}$$

$$\hat{v}^* = \frac{v^*}{1-\beta_2} = \frac{\beta_2 v + (1-\beta_2)(\nabla_\theta \theta)^2}{1-\beta_2} \tag{15}$$

where $\mathbf{A}^*$ is the adjacency matrix after learning and $\mathbf{A}$ is initialization value. lr is learning rate. $m = 0$, $v = 0$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. θ is all parameters of the network.

Algorithm 1 summarizes the specific implementation steps of the MDGCN-SRCNN model.

Details of the MDGCN-SRCNN Model

We consider that the amount of EEG data is too small, so the network cannot be designed too deep to prevent overfitting. In addition, the graph convolutional network cannot be superimposed too much, which will affect the performance, generally within 5 layers. After a small amount of trial and error experiments, we have observed that MDGCN-SRCNN achieves a higher accuracy rate under the two-layer graph convolutional layer and the two-layer convolutional layer plus the three-layer fully connected layer. The detailed description of the MDGCN-SRCNN model is shown in **Table 1**.

EXPERIMENTAL SETTINGS

In this section, we introduce the data sets and model settings used in the experiment.

Datasets

We used two datasets SEED (Zheng and Lu, 2015) and SEED-IV (Zheng et al., 2018) to evaluate our proposed model.

Algorithm 1: The training process of MDGCN-SRCNN.

Input : A labeled training data set$\{\mathbf{X}, \mathbf{Y}\} = \{x_i, y_i\}_{i=1}^N$, the maximum number of training epochs T; the initialize adjacency matrix$\mathbf{A}$,regularization coefficientα.
Output: The learned adjacency matrix$\hat{\mathbf{A}}$, the model parameterΘ for MDGCN-SRCNN and the predicted label$\hat{y}$.
Step 1 : Initialize the model parametersΘ in MDGCN-SRCNN model. Set iteration unit iter = 1;
Step 2 : **while**iter < T**do**
Step 3 : **for**$k = 1, ..., l$**do**
Step 4 : Calculate the k-th graph convolutional layer$\mathbf{H}^{(k)}$via Eq. (1) and calculate the k-th sum pooling layer$Pool(\mathbf{H}^{(k)})$;
Step 5 : **for**$k = 1, ..., l$**do**
Step 6 : Calculate the k-th SMR-based convolution layer$C_k$via Eq. (9);
Step 7 : Concatenate the different layers of features$\mathbf{F}$via Eq. (10);
Step 8 : Calculate the prediction label$\hat{y}$ via Eq. (11);
Step 9 : Update the adjacency matrix$\mathbf{A}$and the model parametersΘ via optimizer according to the cross-entropy loss.
Step 10: iter =iter+1;
Step 11: **end while**

SEED

The SEED data set contains EEG data of 15 subjects (7 males and 8 females), which were collected through 62 channels of ESI neuroscan system when they watched movie clips. All participants watched 15 movie clips, which contained five positive emotions, five neutral emotions, and five negative emotions. Each movie clip lasted about 4 min. There were three periods of data collection, and each subject collected a total of 45 experiments. The original EEG data were de sampled and the artifacts such as EOG and EMG were removed. The EEG data of each channel is divided into 1s segments without overlapping, and then the differential entropy characteristics of the five bands (Delta, Theta, Alpha, Beta, and Gamma) of the linear dynamic system smoothing (LDS) (Duan et al., 2013) are calculated for the segmented data segments.

SEED-IV

The EEG data of 15 healthy subjects (7 males and 8 females) were collected in the SEED-IV dataset using the same equipment as the SEED dataset. The data set selected 72 video clips to induce four different emotions (happy, neutral, sad, and fear). Each video clip lasted about 2 min. Each experiment conducted 24 experiments (6 experiments for each emotion). Each subject participated in three experiments at different times, and a total of 72 experiments were collected. Each experiment was divided into non-overlapping data segments of 4 s, each segment of data as a sample. Same as SEED, the differential entropy characteristics of five frequency bands are calculated.

TABLE 1 | MDGCN-SRCNN architecture.

Block	Layer	Kernel size	Stride	Input	Output	Activation
Graph convolution	Input				(*n*, *f*)	
	GCN1			(*n*, *f*)	(*n*, 16)	Leaky_ReLU
	Global_add_pool			(*n*, 16)	16	
	GCN2			(*n*, 16)	(*n*, 64)	Leaky_ReLU
	Global_add_pool			(*n*, 64)	64	
SMR-based convolution	Reshape			64	(8,8,1)	
	Conv1	(2,2)	2	(8,8,1)	(7,7,16)	Leaky_ReLU
	SMR1			(7,7,16)	(7,7,16)	Sigmoid
	Conv2	(2,2)	2	(7,7,16)	(6,6,32)	Leaky_ReLU
	SMR1			(6,6,32)	(6,6,32)	Sigmoid
	Max_pool	(2,2)		(6,6,32)	(3,3,32)	
Classifier	Reshape			(3,3,32)	3*3*32	
	FC1			16+64+3*3*32	256	Leaky_ReLU
	FC2			256	128	Leaky_ReLU
	FC3			128	C	Softmax

TABLE 2 | Compare the accuracy rate (mean/std) with different existing methods on the SEED data set.

Model	Delta band	Theta band	Alpha band	Beta band	Gamma band	All bands
SVM (Zheng and Lu, 2015)	60.50/14.14	60.95/10.20	66.64/14.41	80.76/115.6	79.56/11.38	83.99/9.72
GSCCA (Zheng, 2017)	63.92/11.16	64.64/10.33	70.10/14.76	76.93/11.00	77.98/10.72	82.96/9.95
DBN (Zheng and Lu, 2015)	64.32/12.45	60.77/10.42	64.01/15.97	78.92/12.48	79.19/14.58	86.08/8.34
STRNN (Zhang et al., 2017)	**80.90/12.27**	**83.35/9.15**	**82.69/12.99**	83.41/10.16	69.61/15.65	89.50/7.63
GCNN (Song et al., 2020)	72.75/10.85	74.40/8.23	73.46/12.17	83.24/9.93	83.36/9.43	87.40/9.20
DGCNN (Song et al., 2020)	74.25/11.42	71.52/5.99	74.43/12.16	83.65/10.17	85.73/10.64	90.40/8.49
BiDANN (Li Y. et al., 2018)	76.97/10.95	75.56/7.88	81.03/11.74	**89.65/9.59**	88.64/9.46	92.38/7.04
GCB-net (Zhang T. et al., 2019)	80.38/10.04	76.09/7.54	81.36/11.44	88.05/9.84	88.45/9.67	92.30/7.40
GCB-net+BLS (Zhang T. et al., 2019)	79.98/8.93	76.51/9.56	81.97/11.05	89.06/8.69	89.10/9.55	94.24/6.70
RGNN (Zhong et al., 2020)	76.17/7.91	72.26/7.25	75.33/8.85	84.25/12.54	**89.23/8.9**	94.24/**5.95**
MDGCN-SRCNN	77.73/10.23	77.27/9.38	80.47/13.22	87.59/12.13	89.02/9.13	**95.08/**6.12

Bold represents the best result.

Model Settings

The parameter selection of the MDGCN-SRCNN model is based on previous experience and a small number of experiments. The Adam optimizer is used to optimize the loss function, and the learning rate is selected in the range of [0.001, 0.01]. L2 regular term coefficient $\alpha = 0.01$. The fully connected layer in the SMR-based convolution block uses a dropout rate of 0.7. In the SEED data set, the batch size used is 16, and in SEED-IV, the batch size used is 9.

RESULTS AND ANALYSIS

In this section, we will evaluate the effectiveness and advancement of the propos ed model on the two data sets described in section Experimental Settings.

Overall Performance

Performance on SEED

In the SEED data set, we refer to the settings of Zheng and Lu (2015), Song et al. (2020), and Li Y. et al. (2018). Each subject contains 15 trials per experiment. Therefore, the first 9 trials are used as the training set and the remaining 6 trials are used as the test set. The final accuracy and variance are the average results of 15 subjects.

The MDGCN-SRCNN model proposed in this paper is compared with the latest methods such as Support Vector Machine (SVM), Deep Belief Network (DBN), DGCNN, RGNN, GCB-net, STRNN, and BiHDM. In addition, we evaluated the performance of the related model on the 5 frequency bands of the DE feature. The comparison results of these models are shown in **Table 2**.

It can be seen in **Table 2** that the model MDGCN-SRCNN proposed in this paper has achieved the best performance in the full-band features, with an average recognition accuracy rate of 95.08% (standard deviation of 6.12%). The performance in each frequency band is also very good. Compared with the low-frequency band (Delta band, Theta band and Alpha band) features, the high-frequency band (Beta band and Gamma band) features are more related to human brain activity. Compared with DGCNN and CGB-net, the accuracy rate of the whole

frequency band is improved by 4.68 and 2.78%, respectively, and the stability of our proposed model is better.

Performance on SEED-IV

On the SEED-IV data set, in order to better compare other methods, we have the same settings as Zheng et al. (2018) and Li et al. (2020). Each subject has a total of 24 trials in an experiment. The first 16 trials are selected as the training set, and the remaining 8 trials are used as the test set. The 8 trials in the test set include 2 trials of happy, neutral, sad, and fear.

In order to evaluate the performance of the MDGCN-SRCNN model proposed in this paper on the SEED-IV dataset, we compared the baseline methods SVM, DBN, DGCNN, etc., and also compared the current latest methods RGNN, BiHDM, SST-EmotionNet, etc. We conduct experiments and comparisons on the DE features of the whole frequency band (Delta, Theta, Alpha, Beta, and Gamma). The results are shown in **Table 3**.

In **Table 3**, it can be seen that the MDGCN-SRCNN model proposed in this paper achieves the most advanced performance at present, with an average accuracy of 85.2%, which is 15.64 and 6.15% higher than the similar graph networks DGCNN and RGNN, respectively. It shows that MDGCN-SRCNN model has a good advantage in emotion recognition task.

TABLE 3 | The accuracy of the proposed method is compared with the existing methods on the SEED-IV dataset.

Model	ACC (%)	STD (%)
SVM (Zhong et al., 2020)	56.61	20.05
DBN (Zhong et al., 2020)	66.77	7.38
GSCCA (Zheng, 2017)	69.08	16.66
DGCNN(Zhong et al., 2020)	69.88	16.29
BiDANN (Li Y. et al., 2018)	70.29	12.63
EmotionMeter (Zheng et al., 2018)	70.58	17.01
BiHDM (Li et al., 2020)	74.35	14.09
RGNN (Zhong et al., 2020)	79.37	10.54
SST-EmotionNet (Jia et al., 2020)	84.92	**6.66**
MDGCN-SRCNN	**85.52**	11.58

Bold represents the best result.

Visualization of Results

In order to intuitively distinguish between different emotions, we draw the confusion matrix of SEED data set and SEED-IV data set. As shown in **Figure 3**, the positive and neutral emotions of SEED dataset are better distinguished than negative emotions, and the neutral emotions will have certain negative emotions. Fear emotions in the SEED-IV data set are relatively difficult to distinguish. On the contrary, sad emotions are the best to distinguish among the four types of emotions, followed by neutral and happy emotions.

In addition, we performed a visual analysis of feature distribution to evaluate the influence of the corresponding modules in the MDGCN-SRCNN model. We use t-SNE to reduce the dimensionality of the features output in different layers, and draw a two-dimensional feature distribution map. **Figure 4** shows the original artificial feature distribution of the SEED data set and the SEED-IV data set and the output feature distribution of different layers. It can be seen from **Figure 4** that the output features of a single layer will be confused with some samples to varying degrees, resulting in a decrease in classification accuracy. In addition, the features learned by two-layer GCN are more representative than those learned by single-layer GCN. Moreover, the deep features learned by SRCNN can better express each type of emotion. Therefore, by combining the shallow GCN features and the deep SRCNN features, the features that express various emotions can be fully learned, and the robustness of the model is improved.

Study of Brain Connection

We analyzed the connections between the brain regions in human emotion. We standardize the initial adjacency matrix and the adjacency matrix learned by network, and the range of their values is [0, 1]. We select the top 10 strongest connection weights in the SEED dataset and the SEED-IV dataset, respectively, and draw their connection diagram, as shown in **Figure 5**. **Figures 5A,B** show the initial connection

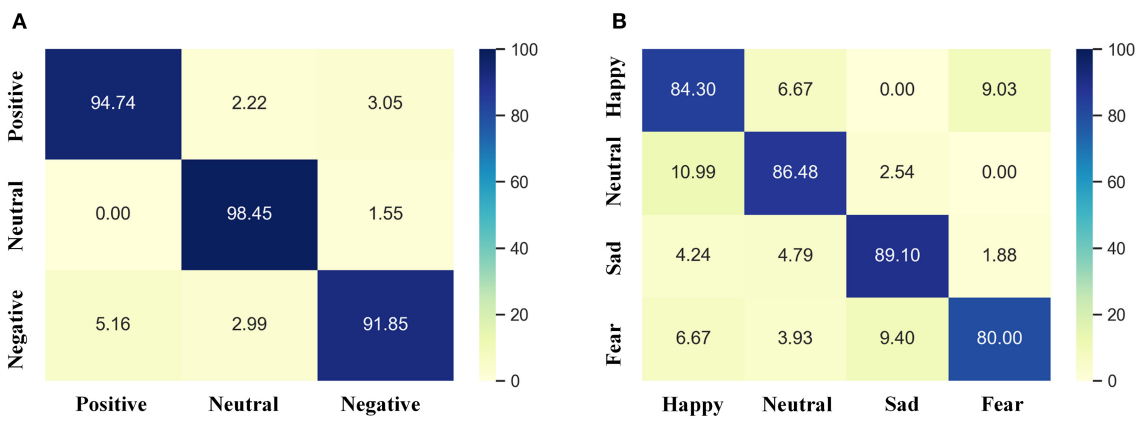


FIGURE 3 | Confusion matrix of different data sets. **(A)** is the confusion matrix of the SEED data set; **(B)** is the confusion matrix of the SEED-IV data set.

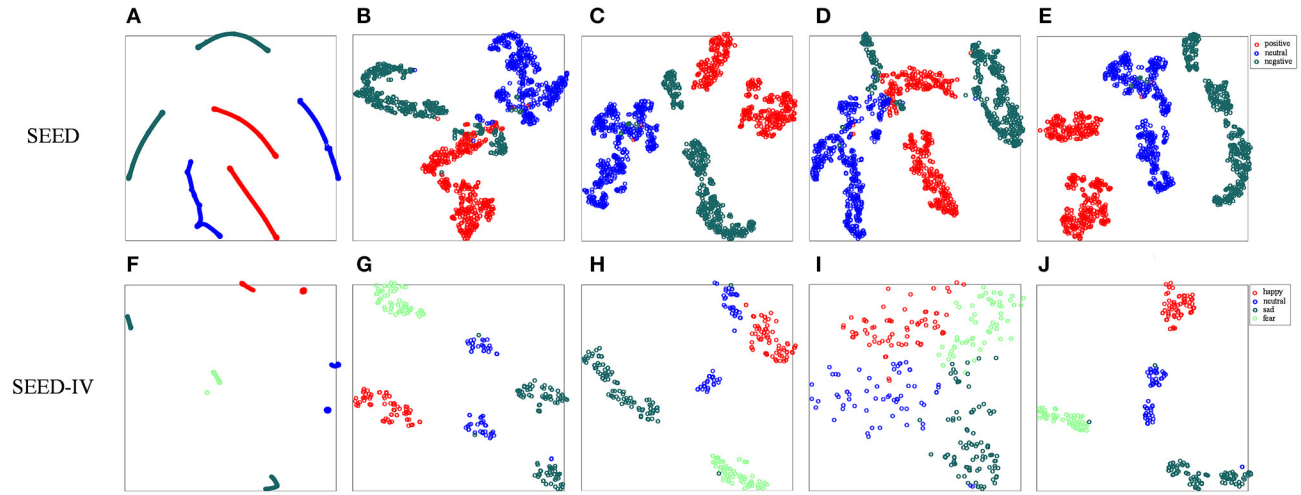


FIGURE 4 | Visualization of t-SNE output from different layers. **(A,F)** are the original data; **(B,G)** are the feature distributions output by the first layer of GCN; **(C,H)** are the feature distributions output by the second layer of GCN; **(D,I)** is the feature distribution of the output of the convolutional neural network; **(E,J)** are the feature distributions after connecting the two layers of GCN and SRCNN. Different colors represent different emotions.

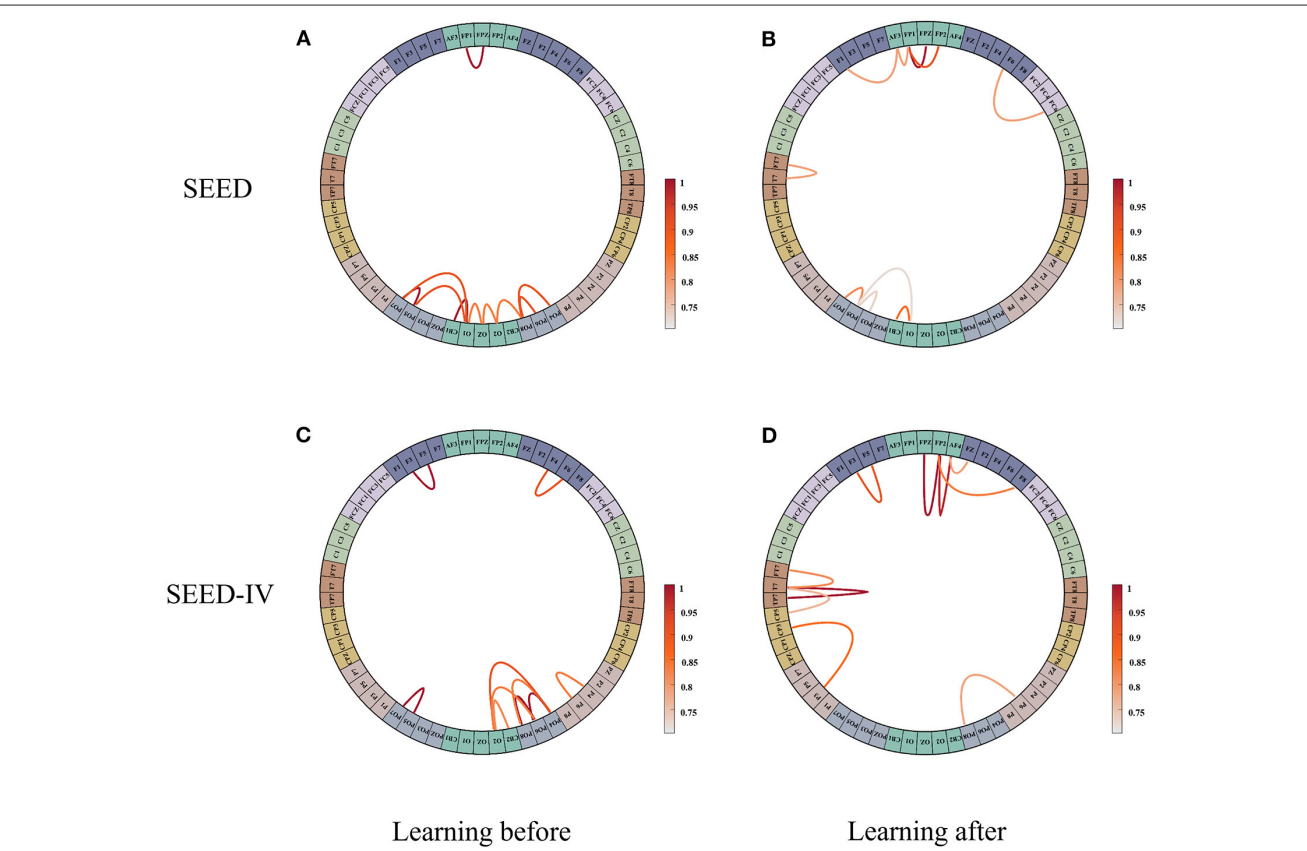


FIGURE 5 | The connection weights between the first 10 channels are selected from the initial adjacency matrix and the learned adjacency matrix. **(A)** is the initial adjacency matrix of the SEED data set, **(B)** is the adjacency matrix learned from SEED dataset. **(C)** is the initial adjacency matrix of the SEED-IV data set, **(D)** is the adjacency matrix learned from SEED-IV dataset.

TABLE 4 | The SEED data set and SEED-IV data set are compared by using different adjacency matrix A initialization methods.

Method	SEED		SEED-IV	
	ACC(%)	STD(%)	ACC(%)	STD(%)
PCC	**95.08**	**6.12**	**85.52**	11.58
RGNN	91.98	7.21	84.16	**10.93**
PLV	92.04	7.56	80.92	13.48
Random	91.83	8.47	82.39	11.74

Bold represents the best result.

TABLE 5 | The results of ablation experiments on SEED and SEED-IV (mean/std), "∼" represents the module is removed.

Model	SEED	SEED-IV
MDGCN-SRCNN	**95.08/**6.12	**85.52/**11.58
∼SRM	93.36/6.49	83.63/10.20
∼SRCNN	91.38/7.74	81.15/10.89
One-layer GCN	89.72/6.52	79.73/9.61

Bold represents the best result.

and the learned connection selected on the SEED data set, respectively. **Figures 5C,D** show the initial connection and the learned connection selected on the SEED -IV data set, respectively. It can be seen from **Figures 5A,C** that the initial connection between the left and right hemispheres of the brain is symmetrical and concentrated in the occipital lobe, while the subjects' movie clips are mainly visual stimulation, and the visual information is mainly processed in the occipital lobe, which is in line with the common sense. After learning, the connection between the left and right hemispheres of the brain becomes asymmetric, as shown in **Figures 5B,C**, especially in the temporal lobe, frontal lobe, and parietal lobe, where the asymmetry is the strongest, indicating that these regions are crucial to emotional activity. Among the local connections, (FT7-T7), (FP2-FPZ), (FP2-AF4), and (T7-TP7) are the strongest connections, and in the global connection (FP1-FP2), is the strongest connection. It shows that emotional activities in the brain are mainly local connections, and global connections are complementary connections. In addition, the more complex emotions are, the more brain areas need to be used. The more complex the connections between brain areas, the greater the strength of local connections.

In order to explore the impact of the initial method of the adjacency matrix**A**on the performance of the model, we chose the common initial methods, such as phase locking (PLV), Pearson correlation coefficient (PCC), local, and global connections used in RGNN and in [0,1] and random values. We extracted DE features in the SEED data set and SEED-IV data set for comparison. **Table 4** shows the effect of using different initial methods of adjacency matrix on the performance of the MDGCN-SRCNN model on the SEED dataset and the SEED-IV dataset. The results show that using PCC as the initialization method of the adjacency matrix achieves the best performance. In RGNN, a global connection is added on the basis of relative physical distance, and a great improvement has been made on the SEED-IV data set. The performance of PLV as an initialization method of the adjacency matrix is equivalent to that of random value selection.

Ablation Results

In order to verify the contribution of each module of our proposed model, we conducted a series of ablation experiments. The results are shown in **Table 5**. After removing the SRCNN module, the performance is significantly reduced. The accuracy on SEED and SEED-IV decreased by 3.7 and 4.37%, respectively, indicating the importance of CNN in extracting deep abstract features related to emotion. In addition, the accuracy on SEED and SEED-IV decreased by 1.72 and 1.89% respectively after removing the SRM module, which proved that the attention mechanism such as SRM module can effectively emphasize emotion related features and abandon useless features, so as to improve the recognition performance of the model. Compared with the one-layer GCN, the recognition performance of two-layer GCN on SEED and SEED-IV is improved by 1.66 and1.42%, respectively, indicating that there is a certain complementarity between global features and local features.

CONCLUSIONS

In this paper, we propose a multi-layer dynamic graph convolutional network-style-based recalibration convolutional neural network (MDGCN-SRCNN) model for EEG-based emotion recognition. In our model, EEG data is considered to be non-Euclidean structure, and dynamic graph neural network is used to learn the connection relationship between each channel of EEG signal as a shallow layer feature. Because analyzing emotions through EEG signals is very complicated. We use a style-based recalibration convolutional neural network to further extract abstract deep layer features. Finally, the fully connected layer is used to search for the features most relevant to emotions in the shallow layer and deep layer features for recognition. We conducted systematic experimental verification on the SEED data set and the SEED-IV data set. MDGCN-SRCNN model has achieved better performance on the two public data sets, surpassing the state-of-the-art RGNN. The recognition accuracy on the SEED data set and SEED-IV data set is 95.08 and 85.52%, respectively, and the standard deviation is 6.12 and 11.58%, respectively. Based on using PCC as the initialization method of the adjacency matrix, the MDGCN-SRCNN model is used to learn the local connections and global connections that are most relevant to emotions, such as (FT7-T7), (FP2-FPZ), (FP2-AF4), (T7-TP7), and (FP1-FP2), these connections are mainly distributed in the temporal lobe, frontal lobe, and parietal lobe, proving that these brain regions play a vital role in inducing emotions. In addition, we also found that the more complex emotions are processed, the more brain regions are involved, the more complex the connections, and the greater the strength of local connections.

It is worth noting that using different initial methods of adjacency matrix has a great influence on the connection

relationship between graph neural network learning and task. Therefore, it is very important to build the initial connection relationship related to the task. In the future, our main work direction is to build more complex network based on GCN to solve the differences between subjects. And further explore the differences of adjacency matrix under different emotional states, and then analyze the differences of brain activity under different emotional states.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

GB was mainly responsible for research design, data analysis, and manuscript writing of this study. KY was mainly responsible for data analysis. LT and BY was mainly responsible for research design. JS was mainly responsible for data collection and production of charts. RZ was mainly responsible for production of charts. LW was mainly responsible for data analysis and document retrieval. YZ was mainly responsible for data collection and manuscript modification. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Key Research and Development Plan of China under Grant 2017YFB1002502, in part by the National Natural Science Foundation of China under Grant 61701089, and in part by the Natural Science Foundation of Henan Province of China under Grant 162300410333.

REFERENCES

Ajili, I., Mallem, M., and Didier, J. Y. (2019). Human motions and emotions recognition inspired by LMA qualities. *Vis. Comput.* 35, 1411–1426. doi: 10.1007/s00371-018-01619-w

Alsolamy, M., and Fattouh, A. (2016). "Emotion estimation from EEG signals during listening to Quran using PSD features," in *7th International Conference on Computer Science and Information Technology (CSIT)*, 1–5. doi: 10.1109/CSIT.2016.7549457

Bao, G., Zhuang, N., Tong, L., Yan, B., Shu, J., Wang, L., et al. (2021). Two-level domain adaptation neural network for EEG-based emotion recognition. *Front. Hum. Neurosci.* 14a, 605246. doi: 10.3389/fnhum.2020.605246

Clore, G., and Ortony, A. (2008). "Appraisal theories: how cognition shapes affect into emotion," in *Handbook of Emotions*, eds M. Lewis, J. M. Haviland-Jones, and L. F. Barrett (The Guilford Press), 628–642.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems* NIPS'16 (Red Hook, NY, USA: Curran Associates Inc.), 3844–3852.

Duan, R.-N., Zhu, J.-Y., and Lu, B.-L. (2013). "Differential entropy feature for EEG-based emotion classification," in *6th International IEEE/EMBS Conference on Neural Engineering (NER)* (IEEE), 81–84. doi: 10.1109/NER.2013.6695876

Faskowitz, J., Esfahlani, F., Jo, Y., Sporns, O., and Betzel, R. (2020). Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. *Nat. Neurosci.* 23, 1–11. doi: 10.1038/s41593-020-00719-y

Goshvarpour, A., and Goshvarpour, A. (2019). The potential of photoplethysmogram and galvanic skin response in emotion recognition using nonlinear features. *Austral. Phys. Eng. Sci. Med.* doi: 10.1007/s13246-019-00825-7. [Epub ahead of print].

Hadjidimitriou, S. K., and Hadjileontiadis, L. J. (2012). Toward an EEG-based recognition of music liking using time-frequency analysis. *IEEE Trans. Biomed. Eng.* 59, 3498–3510. doi: 10.1109/TBME.2012.2217495

Harit, A., Joshi, J., and Gupta, K. (2018). Facial emotions recognition using gabor transform and facial animation parameters with neural networks. *IOP Conf. Series* 331, 012013. doi: 10.1088/1757-899X/331/1/012013

Hatamikia, S., Maghooli, K., and Motie Nasrabadi, A. (2014). The emotion recognition system based on autoregressive model and sequential forward feature selection of electroencephalogram signals. *J. Med. Signals Sens.* 4, 194–201. doi: 10.4103/2228-7477.137777

Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* 29, 306–310. doi: 10.1016/0013-4694(70)90143-4

Jatupaiboon, N., Pan-ngum, S., and Israsena, P. (2013). "Emotion classification using minimal EEG channels and frequency bands," in *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 21–24. doi: 10.1109/JCSSE.2013.6567313

Jenke, R., Peer, A., and Buss, M. (2017). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 327–339. doi: 10.1109/TAFFC.2014.2339834

Jia, Z., Lin, Y., Cai, X., Chen, H., Gou, H., and Wang, J. (2020). SST-EmotionNet: Spatial-Spectral-Temporal based Attention 3D Dense Network for EEG Emotion Recognition. *in* 2909–2917. doi: 10.1145/3394171.3413724

Jie, X., Rui, C., and Li, L. (2014). Emotion recognition based on the sample entropy of EEG. *Biomed. Mater. Eng.* 24, 1185. doi: 10.3233/BME-130919

Kipf, T. N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*.

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., and Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage* 42, 998–1031. doi: 10.1016/j.neuroimage.2008.03.059

Koelstra, S., Yazdani, A., Soleymani, M., Mühl, C., Lee, J.-S., Nijholt, A., et al. (2010). "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos, in *International Conference on Brain Informatics*, eds Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang, Vol 6334 (Berlin; Heidelberg: Springer) doi: 10.1007/978-3-642-15314-3_9

Kroupi, E., Yazdani, A., and Ebrahimi, T. (2011). "EEG correlates of different emotional states elicited during watching music videos," in *Affective Computing and Intelligent Interaction*, 457–466. doi: 10.1007/978-3-642-24571-8_58

Lee, H., Kim, H.-E., and Nam, H. (2019). "SRM: a style-based recalibration module for convolutional neural networks," in 2019 *IEEE/CVF International Conference on Computer Vision* (IEEE), 1854–1862. doi: 10.1109/ICCV.2019.00194

Li, J., Zhang, Z., and He, H. (2018). Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cognit. Comput.* 10, 368–380. doi: 10.1007/s12559-017-9533-x

Li, Y., Fu, B., Li, F., Shi, G., and Zheng, W. (2021). A novel transferability attention neural network model for EEG emotion recognition. *Neurocomputing* 447, 92–101. doi: 10.1016/j.neucom.2021.02.048

Li, Y., Wang, L., Zheng, W., Zong, Y., and Song, T. (2020). "A novel bi-hemispheric discrepancy model for EEG emotion recognition", *IEEE Transactions on Cognitive and Developmental Systems*, 1–1.

Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2018). "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition", in *IEEE Transactions on Affective Computing*, 1–1.

Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., et al. (2010). EEG-based emotion recognition in music listening. *IEEE Trans. Biomed. Eng.* 57, 1798–1806. doi: 10.1109/TBME.2010.2048568

Liu, Y., and Sourina, O. (2013). "Real-time fractal-based valence level recognition from EEG," in *Transactions on Computational Science XVIII*, eds, M. L. Gavrilova, C. J. K. Tan and A. Kuijper, vol 7848 (Berlin, Heidelberg: Springer) 101–120. doi: 10.1007/978-3-642-38803-3_6

Mehmood, R. M., and Lee, H. J. (2015). "Emotion classification of EEG brain signal using SVM and KNN," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. doi: 10.1109/ICMEW.2015.7169786

Petrantonakis, P. C., and Hadjileontiadis, L. J. (2011). Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Trans. Affect. Comput.* 1, 81–97. doi: 10.1109/T-AFFC.2010.7

Sebe, N., Cohen, I., and Huang, T. S. (2005). "Multimodal emotion recognition," in *Handbook of Pattern Recognition and Computer Vision,* 387–409. doi: 10.1142/9789812775320_0021

Song, T., Zheng, W., Song, P., and Cui, Z. (2020). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affective Comput.* 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Sourina, O., and Liu, Y. (2011). "A fractal-based algorithm of emotion recognition from eeg using arousal-valence model," in *BIOSIGNALS 2011 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, 209–214.

Valderas, M. T., Bolea, J., Laguna, P., Bailón, R., and Vallverdú, M. (2019). Mutual information between heart rate variability and respiration for emotion characterization. *Physiol. Meas.* 40:84001. doi: 10.1088/1361-6579/ab310a

Wagh, N., and Varatharajah, Y. (2020). *EEG-GCNN: Augmenting Electroencephalogram-based Neurological Disease Diagnosis using a Domain-guided Graph Convolutional Neural Network*.

Wang, Z., Tong, Y., and Heng, X. (2019). Phase-locking value based graph convolutional neural networks for emotion recognition. *IEEE Access* 7, 93711–93722. doi: 10.1109/ACCESS.2019.2927768

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). *How Powerful are Graph Neural Networks?*.

Zhang, T., Wang, X., Xu, X., and Chen, C. (2019). GCB-Net: graph convolutional broad network and its application in emotion recognition. *IEEE Trans. Affect. Comput.* 1:1. doi: 10.1109/TAFFC.2019.2937768

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2017). Spatial-temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* 49, 839–847. doi: 10.1109/TCYB.2017.2788081

Zhang, X., Yao, L., Wang, X., Monaghan, J., and McAlpine, D. (2019). *A Survey on Deep Learning based Brain Computer Interface: Recent Advances and New Frontiers*. *CoRR* abs/1905.04149. Available online at: http://arxiv.org/abs/1905.04149

Zhao, K., Duka, B., Xie, H., Oathes, D. J., Calhoun, V., and Zhang, Y. (2022). A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in ADHD. *Neuroimage* 246, 118774. doi: 10.1016/j.neuroimage.2021.118774

Zheng, W. (2017). Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Trans. Cogn. Dev. Syst.* 9, 281–290. doi: 10.1109/TCDS.2016.2587290

Zheng, W.-L., Zhu, J.-Y., Peng, Y., and Lu, B.-L. (2014). "EEG-based emotion classification using deep belief networks," in *Proceedings - IEEE International Conference on Multimedia and Expo*. doi: 10.1109/ICME.2014.6890166

Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2018). EmotionMeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1–13. doi: 10.1109/TCYB.2018.2797176

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 1–1. doi: 10.1109/TAMD.2015.2431497

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* 1, 1–1. doi: 10.1109/TAFFC.2020.2994159

Zong, Y., Zheng, W., Huang, X., Yan, K., Yan, J., and Zhang, T. (2016). Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis. *J. Multimodal User Interfaces* 10, 163–172. doi: 10.1007/s12193-015-0210-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neurorobotics

METHODS
published: 16 March 2022
doi: 10.3389/fnbot.2022.851847

Unsupervised Facial Action Representation Learning by Temporal Prediction

Chongwen Wang and Zicheng Wang*

School of Computer Science, Beijing Institute of Technology, Beijing, China

Due to the cumbersome and expensive data collection process, facial action unit (AU) datasets are generally much smaller in scale than those in other computer vision fields, resulting in overfitting AU detection models trained on insufficient AU images. Despite the recent progress in AU detection, deployment of these models has been impeded due to their limited generalization to unseen subjects and facial poses. In this paper, we propose to learn the discriminative facial AU representation in a self-supervised manner. Considering that facial AUs show temporal consistency and evolution in consecutive facial frames, we develop a self-supervised pseudo signal based on temporally predictive coding (TPC) to capture the temporal characteristics. To further learn the per-frame discriminativeness between the sibling facial frames, we incorporate the frame-wisely temporal contrastive learning into the self-supervised paradigm naturally. The proposed TPC can be trained without AU annotations, which facilitates us using a large number of unlabeled facial videos to learn the AU representations that are robust to undesired nuisances such as facial identities, poses. Contrary to previous AU detection works, our method does not require manually selecting key facial regions or explicitly modeling the AU relations manually. Experimental results show that TPC improves the AU detection precision on several popular AU benchmark datasets compared with other self-supervised AU detection methods.

Keywords: facial action unit recognition, self-supervised learning, contrastive learning, temporal predictive coding, representation learning

OPEN ACCESS

Edited by:
Yong Li,
Nanjing University of Science and Technology, China

Reviewed by:
Zhenguo Yang,
Guangdong University of Technology, China
Hao Su,
Beihang University, China
Xiaoya Zhang,
Nanjing University of Science and Technology, China

****Correspondence:***
Chongwen Wang
wcwzzw@bit.edu.cn

Received: *10 January 2022*
Accepted: *31 January 2022*
Published: *16 March 2022*

Citation:
Wang C and Wang Z (2022) Unsupervised Facial Action Representation Learning by Temporal Prediction. Front. Neurorobot. 16:851847. doi: 10.3389/fnbot.2022.851847

1. INTRODUCTION

Facial expression recognition technology offers the opportunity to seamlessly capture the expressed emotional experience of humans and facilitates unique human-computer interaction experiences. Over the past decades, facial expression recognition and analysis have been a hot research topic in the field of computer vision and human-computer interaction. To precisely characterize facial expressions, Ekman *et al.* developed the facial action coding system (FACS) (Ekman and Friesen, 1978). FACS has been widely used for describing and measuring facial behavior and has been the most comprehensive, anatomical system for describing facial expressions. FACS defines a detailed set of about 30 atomic non-overlapping facial muscle actions, i.e., action units (AUs). Almost any anatomical facial muscle activity can be characterized *via* a combination of facial AUs. Automatic AU detection has been a vital task for facial expression analysis, with a variety of applications in psychological and behavioral research, mental health assessment, and human-computer interaction (Bartlett et al., 2003; Zafar and Khan, 2014). Therefore, a reliable AU detection system is of vital importance for precise human emotion analysis.

Benefiting from the promising advancement in deep learning research, the performance and accuracy of AU detection has been improved by virtue of the convolutional neural network (CNN) based approaches in recent years (Li et al., 2017a,b, 2018a,b, 2020a; Corneanu et al., 2018; Jacob and Stenger, 2021). However, the CNN-model-based AU detection approaches are quite data starved. What is worse is that AU annotation is time-consuming, labor-intensive, cumbersome, and error-prone. Thus, many existing works propose to exploit the auxiliary information for precise AU detection, e.g., Yang et al. (2021) proposed to use the semantic embedding and visual feature (SEV-Net) for AU detection. SEV-Net obtains AU semantic embeddings through both intra-AU and inter-AU attention components to capture the relationships among words within each sentence that describes individual AU. Li and Shan (2021) use the categorical facial expression images as auxiliary training data to boost the AU detection performance in a meta-learning manner. These pioneering works have inspired us to use a large amount of unlabeled facial videos to learn the AU representation unsupervised, as the unlabeled facial videos are easy to obtain and they consist of a large amount of subjects with diverse facial expressions.

Recently, self-supervised learning (SSL) has shown promising potential in learning discriminative features from the unlabeled data *via* various different manually defined pretext tasks (Wang et al., 2020; Cai et al., 2021; Hu et al., 2021; Kotar et al., 2021; Luo et al., 2021; Sun et al., 2021). For the task of AU detection, Li et al. (2019b) proposed to predict the optical flow caused by AUs and poses between two randomly sampled facial frames in a video sequence. The optical flow of the AUs and poses are then linearly combined to obtain the overall displacements between the two sampled faces. Lu et al. (2020) leveraged the temporal consistency to learn the AU feature *via* a self-supervised temporal ranking constraint. To capture the AU correlations in an input facial image, Yan et al. (2021) disentangled the global feature into multiple AU-specific features *via* a contrastive loss and then compute the feature for each AU by aggregating the features from the other AU-specific features with a transformer component. To bridge the performance gap between the fully supervised and self-supervised AU detection methods, we propose a self-supervised pseudo signal based on the temporally predictive coding (TPC) to capture the temporal characteristics of the AUs. Specially, we construct a model that combines an AU feature extraction network with a convolutional gated recurrent unit (GRU) unit (Zonoozi et al., 2018), and a prediction head on top of the GRU that can make temporal predictions. We train the constructed model *via* TPC loss, which will be detailed in Section 3.1.

To further learn the per-frame discriminativeness between the sibling facial frames within a video clip, we propose a frame-wisely temporal contrastive learning mechanism. The AU detection model is tasked to perceive the temporal consistency and frame-wisely discriminativeness self-supervised. The AU detection backbone is trained end-to-end with the linear combination of the two contrastive losses on the unlabeled facial videos. Afterward, we additionally train a linear classifier with the pre-trained AU detection backbone with the scarce AU annotations.

In summary, the core contributions of this work can be summarized as follows:

1. We introduce self-supervised TPC for facial AU representation learning. TPC does not rely on AU annotations to learn the discriminative AU representations.
2. To further enhance the discriminability of the AU representation, TPC consists of a frame-wisely temporal contrastive learning constraint. TPC is capable of perceiving the temporal consistency and frame-wisely discriminativeness self-supervised.
3. Experimental results demonstrate the advantages of the proposed TPC over other state-of-the-art self-supervised AU detection methods on two popular AU datasets. Image retrieval results show that the learned AU representation in TPC is superior in spotting and capturing the AU similarities between different faces.

2. RELATED WORK

A number of AU detection approaches have been proposed recently (Zhao et al., 2016; Li et al., 2017a,b; Li and Shan, 2021). AU detection approaches are deep learning-based mostly. Since AU actually means the movement of the facial muscles, many approaches detect the active/inactive states of AUs locally (Zhao et al., 2016; Li et al., 2017a,b). Among them, Zhao et al. (2016) used a locally connected convolutional layer to learn the AU-specific convolutional filters. SEV-Net (Yang et al., 2021) exploited the AU semantic word embedding as the auxiliary labels. FAUT was (Jacob and Stenger, 2021) proposed to capture the relationships between AUs *via* a transformer. These supervised AU detection methods need manually labeled training facial data. As training images are scarce, these methods often overfit on a specific dataset and cannot generalize well.

Recently, self-supervised (Wiles et al., 2018; Li et al., 2019b, 2020b; Lu et al., 2020) and weakly-supervised (Peng and Wang, 2018; Zhao et al., 2018) methods have been proposed to learn the deep learning-based models from unlabeled or partially labeled images. The former usually adopts the manually defined pseudo supervisory signals to learn the facial AU representation (Li et al., 2019b, 2020b; Lu et al., 2020). Among them, Fab-Net (Wiles et al., 2018) was trained to map a source facial frame to a target facial frame *via* estimating an optical flow field between the source and the target faces. Twin-cycle autoencoder (TCAE and TAE) (Li et al., 2019b, 2020b) were proposed to learn the pose-invariant facial action features by estimating the respective optical flows for the poses and AUs *via* the cycle-consistency in the image and representations. Lu et al. (2020) proposed a temporally sensitive triplet-based metric learning to learn the facial AU representations *via* capturing the temporal AU consistency. It actually learns to rank the neighboring faces from the sequential frames in the correct order. Our proposed TPC differs from previous methods in three aspects. First, TPC is self-supervised in the pre-training stage. Second, TPC does not crop the regional AU features to learn the region-specific AU feature. Instead, it uses an abundant number of unlabeled videos to enhance the AU detection performance. Finally, TPC

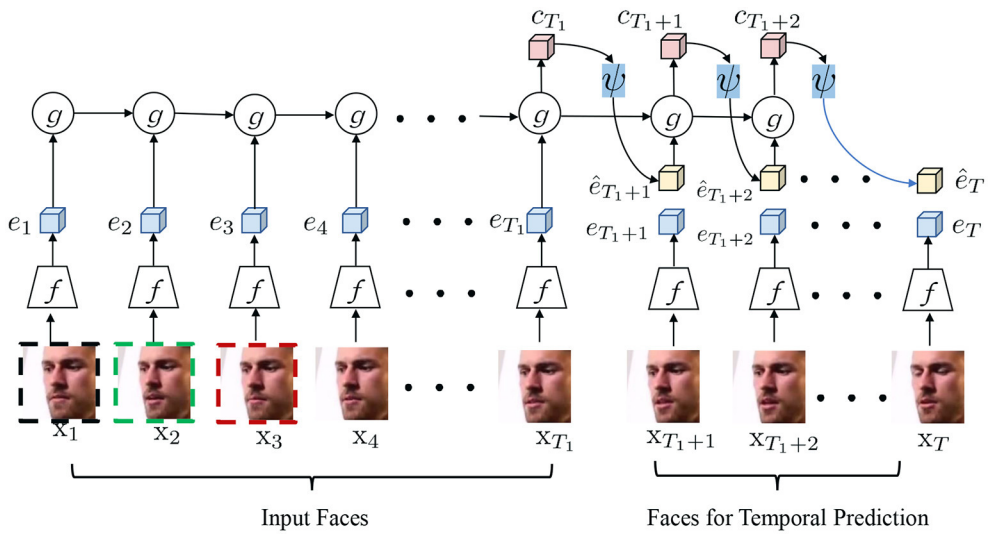


FIGURE 1 | Main idea of the proposed self-supervised temporally predictive coding (TPC) for facial AU representation learning. Given a facial sequence with T faces, we use the preceding T_1 faces as input and exploit the left faces for temporal prediction. Besides, we randomly sampled some triplets in each facial sequence to perceive the temporal consistency and frame-wisely discriminativeness self-supervised. ψ takes the context representation c_t as input and estimates the features for the future frame recursively. Better viewed in color and zoom in.

is proposed to encode the temporal dynamics and consistencies to encode the characteristics of the facial AUs.

3. METHOD

Figure 1 illustrates the main framework of the proposed TPC for AU representation learning. Given an input facial sequence sampled from an unlabeled facial video, TPC first extracts the convolutional feature maps of each face *via* a commonly-used backbone network such as ResNet-50. Second, TPC learns the discriminativeness between different facial frames *via* temporal contrastive learning. We will introduce the proposed TPC and present the temporal contrastive learning paradigm in our proposed TPC as below.

3.1. Temporal Predictive Coding

Videos are very appealing as a data source for self-supervision as there are many forms of pseudo signal. In detail, the self-supervision in the video sequence generally originates from three types: spatial, spatio-temporal, and sequential. Among the three kinds of self-supervised signal, spatial supervision can be derived from the structures in the static frame, spatio-temporal supervision naturally reflects the correlation across the different frames, and sequential supervision signifies the temporal coherence. Therefore, we exploit the sequential self-supervision to learn a robust model for facial AU detection that is capable of capturing the temporal dynamics as well as temporal consistency of the facial AUs.

Let $\mathrm{X} = \{x_t\}_{t=1}^{T}$ denotes a consecutive sequence of T facial frames within an unlabeled video, where $\mathrm{x}_t \in \mathbb{R}^{H\times W\times C}$ means the input t-th facial image of size $H \times W \times C$. Our goal here is to learn a model that predicts a slowly varying semantic representation based on the recent past. As illustrated in **Figure 1**, we partition a facial video clip into two parts: input part I and output part O:

$$\mathrm{I} = \{\mathrm{x}_t\}_{t=1}^{T_1}, \tag{1}$$

$$\mathrm{O} = \{\mathrm{x}_t\}_{t=T_1+1}^{T}, \tag{2}$$

where T_1 is the length of the input facial sequence. First, a backbone network $f(.)$ maps each facial frame x_t to its latent convolutional map representation $e_t \in \mathbb{R}^{H'\times W'\times C'}$, organized as height $\times$ width $\times$ channels. Then, we use a convolutional GRU to aggregate the sequential latent representations into a context representation c_t. Mathematically, GRU uses the same gated principal of LSTM but with a simpler architecture. The below equations describe the mathematical model for the GRU:

$$z_t = \sigma(W_{hz}h_{t-1} + W_{xz}e_t + b_z), \tag{3}$$

$$r_t = \sigma(W_{hr}h_{t-1} + W_{xr}e_t + b_r), \tag{4}$$

$$\hat{h}_t = \Phi(W_h(r_t \odot h_{t-1}) + W_x e_t + b), \tag{5}$$

$$c_t = h_t = (1 - z_t) \odot h_{t-1} + z \odot \hat{h}_t, \tag{6}$$

where h_t is the hidden state, r_t and z_t are the reset gate value and update gate value at frame t. The functions $\sigma(.)$ and $\Phi(.)$ denote the sigmoid and tangent activation functions, respectively. The reset gate r_t can decide whether or not to forget the previous activation. $\odot$ means the element-wise multiplication. **Figure 2** shows the main idea of the convolutional GRU.

With the encoded context representation c_t, we exploit a prediction head ψ to predict the convolutional latent representation of the feature. In detail, ψ takes the context

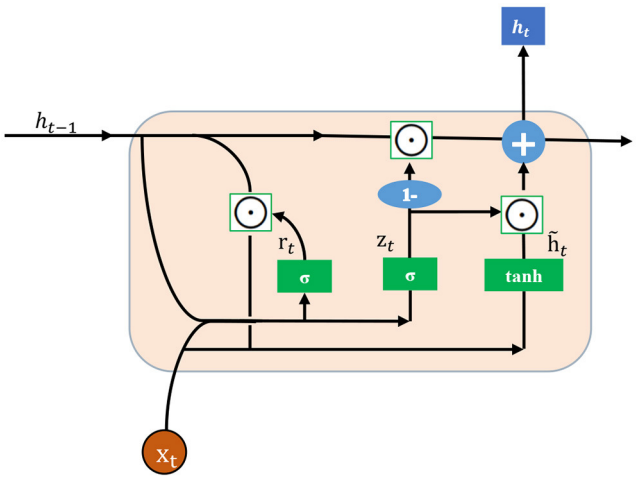


FIGURE 2 | Illustration of the convolutional gated recurrent unit (GRU).

representation c_t as input and estimates the features for the future frame recursively:

$$e_{t+1} = \psi(c_t), \tag{7}$$
$$e_{t+2} = \psi(c_{t+1}), \tag{8}$$

where c_t means the context feature from time step 1 to t, and e_{t+1} means the estimated latent convolutional feature of the time step $t+1$. Similarly, we can predict the latent convolutional feature maps for the $t+2$ facial frame, in a recursive manner. Such a recursive TPC manner enforces the prediction to be conditioned on all previous predictions and observations. The intuition behind the TPC is that the model is tasked to infer future AU semantics from the context representations c_t and thus c_t has to encode temporal consistency and dynamics of the facial AUs.

The learning of the TPC is accomplished *via* a noise contrastive estimation, where our goal is the classify the real from the noisy samples. We denote the feature vector in each spatial location of the encoded and the predicted convolutional feature maps as $e_{i,k}$ and $\hat{e}_{i,k}$, where i denotes the temporal index and k means the spatial index in the convolutional features, $k \in \{(1,1),(1,2),\cdots,(H',W')\}$. Finally, we can formulate the learning objective as follows:

$$\mathcal{L}_{pred} = -\sum_{i,k} \log \frac{\exp(\hat{e}_{i,k} \cdot e_{i,k})}{\sum_{j,m} \exp(\hat{e}_{i,k} \cdot e_{i,m})}. \tag{9}$$

The goal of $\mathcal{L}_{pred}$ is to classify the positive pair $(\hat{e}_{i,k}, e_{i,k})$ among a set of constructed pairs. A positive pair consists of two elements that are located in the same spatial location and at the same time step. All the other pairs $(\hat{e}_{i,k}, e_{j,m})$ that satisfy $(i,k) \neq (j,m)$ are negative pairs. $\mathcal{L}_{pred}$ is optimized such that the similarities of the positive pairs are higher than the similarities of the negative pairs. While the proposed TPC can spot the temporal consistency and dynamics of the input facial sequences, the discriminativeness of the nearby facial frames can be further enhanced so that the encoded AU representation can be more discriminative. We will explain how we use the temporal contrastive learning paradigm to achieve this goal in the next section.

3.2. Temporal Contrastive Learning

To learn the frame wisely discriminativeness of the input facial images, we introduce a temporal contrastive learning goal by adding multiple triplet losses (Schroff et al., 2015), each measuring the pairwise distance between the adjacent frames to the anchor frame. Learning to rank through triplet loss actually trains an AU detection backbone that learns to make the distance between the anchor and the positive face smaller than the distance between the anchor and the negative face.

Let us denote a triplet that consists of three facial frames as (x_a, x_p, x_n), where x_a, x_p, and x_n mean the anchor face, positive sample, negative sample, respectively. Note that x_a, x_p, and x_n are consecutive facial frames randomly sampled from the input facial sequence $\mathrm{X} = \{x_t\}_{t=1}^{T}$. Intuitively, (x_a, x_p) should have more similar facial expressions than (x_a, x_n) because the time interval is smaller between x_a and x_p. Inspired by intuition, we randomly sampled M triplets from the input facial sequence X and expect that the sum of M triplet losses would enable the AU detection backbone to learn to perceive the facial expression difference in the nearby facial frames. The learning target of the proposed temporal contrastive learning paradigm can be formulated as:

$$\mathcal{L}_{tcl} = \left[D(f(x_a^{i,1}), f(x_p^{i,j})) - D(f(x_a^{i,1}), f(x_n^{i,j+1})) + m \right]_+, \tag{10}$$

where D is the cosine similarity of the input frame pairs. i is the sequence index, j is the frame index within the i-th input facial sequence. m is the margin that ensures $\mathcal{L}_{tcl}$ will not be zero until the difference between the distances of the negative and positive frame from the anchor is greater than m. For each training facial sequence with T faces, we randomly sampled P triplets.

3.3. Overall Training Objective of TPC

For pre-train, we use the linear combination of $\mathcal{L}_{pred}$ and $\mathcal{L}_{tcl}$ as below:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{tcl}, \tag{11}$$

where λ means the importance of the temporal triplet loss, which will be discussed in the experimental section.

For AU detection, we finetune the pre-trained model with the annotated AU labels. Mathematically, we exploit the multi-label sigmoid cross-entropy loss for optimizing the AU classification head and the pre-trained backbone model, which can be formulated as:

$$\mathcal{L}^{AU} = -\sum_{m}^{M} z^m \log \hat{z}^m + (1 - z^m) \log(1 - \hat{z}^m), \tag{12}$$

where M denotes the number of facial AUs. z^m denotes the m-th ground truth AU annotation of the input AU sample. $\hat{z}^m$ means the predicted AU score. $z_i \in \{0,1\}$ means the labels w.r.t the ith AU. 0 means the AU is inactive, and 1 means the AU is active.

4. EXPERIMENT

4.1. Implementation Details

We adopted ResNet-18 (He et al., 2016) as the backbone network for pretrain. We optimized the proposed backbone model *via* a batch-based stochastic gradient descent method. During training, we set the batch size as 64 on 4 GPU units and the initial learning rate as 0.001. For each video, we randomly sampled $T = 10$ consecutive faces for training, we used the first 8 eight faces as the input and the left 2 faces for prediction. Additionally, we randomly sampled $P = 4$ triplets from each facial sequence for temporal contrastive learning. During finetuning, we dropped the convolutional GRU and added a linear classifier layer for AU prediction. We set the momentum as 0.9 and the weight decay as 0.0005. We use the popular Voxceleb dataset (Nagrani et al., 2020) for pre-training. The dataset consists of about 6,000 subjects and hundreds of thousands of videos. All the videos only contain a subject with varying expressions and no AU or facial expression annotations.

4.1.1. Datasets and Evaluation Metric

For AU detection, we adopted the denver intensity of spontaneous facial action (DISFA) (Mavadati et al., 2013) and binghamton-pittsburgh 3D dynamic spontaneous facial expression database (BP4D) (Zhang et al., 2013) datasets. BP4D consists of a total of 328 videos recorded for 41 subjects (18 men and 23 women). A total of 8 different experimental tasks are evaluated on the 41 subjects, and their spontaneous facial AUs variations were recorded in the videos. There are nearly 14,0000 frames with 12 facial AUs labeled. DISFA contains 27 participants. Each participant is asked to watch a video to elicit his/her facial expressions. The facial AUs are annotated with intensities ranging from 0 to 5. There are about 130,000 AU-annotated images in the DISFA dataset by setting the images with intensities greater than 1 as active. For the two datasets, the facial images are split into 3-fold in a subject-independent manner. We used the 3-fold cross-validation and adopted 12 AUs in BP4D and 8 AUs in DISFA dataset for evaluation.

We adopted F1-score to evaluate the performance of the proposed AU detection method. The F1-score can be calculated as $F1 = \frac{2RP}{R+P}$, where R and P, respectively, denote the recall and precision. We also use the average F1-score over all the evolved AUs (Ave) to evaluate the overall facial AU detection precision.

4.2. Experimental Results

For the supervised methods, we compare the proposed TPC with deep region and multi-label (DRML) (Zhao et al., 2016), enhancing and cropping net (EAC-Net) (Li et al., 2017b), deep structure inference network (DSIN) (Corneanu et al., 2018), local relationship learning with person-specific shape regularization (LP-Net) (Niu et al., 2019), semantic relationship embedded representation learning (SRERL) (Li et al., 2019a), uncertain graph neural networks (UGN) (Song et al., 2021), semantic embedding and visual feature net (SEV-Net) (Yang et al., 2021) and facial action unit detection with transformers (FAUT) (Jacob and Stenger, 2021), meta auxiliary learning (MAL) (Li and Shan, 2021). It is worth noting that some of the AU detection approaches (Li et al., 2017b, 2019a; Corneanu et al., 2018; Jacob and Stenger, 2021) learn the AU-specific representations with exclusive CNN branches *via* cropping the local facial regions. SEV-Net (Yang et al., 2021) proposes to learn robust visual features for AU detection via introducing the auxiliary AU descriptions. UGN (Song et al., 2021) learn to model the uncertainty of the AU annotations.

For the self-supervised methods, we compare the proposed TPC with TCAE (Li et al., 2019b), TAE (Li et al., 2020b), triplet ranking loss (TRL) (Lu et al., 2020). Among the compared methods, in TRL (Lu et al., 2020) proposed an aggregate ranking loss by taking the sum of multiple triplet losses to allow pairwise comparisons between the adjacent facial frames. In TRL, they learn to rank the faces through triplet loss involves training an encoder that learns to force the distance between the anchor face and the positive face smaller than the distance between the anchor face and the negative face.

Table 1 shows the AU detection accuracy comparison of our TPC and previous methods on BP4D dataset. TPC obtains comparable AU detection accuracy in the average accuracy. In detail, TPC shows its superiority over DRML, EAC-Net, DSIN, LP-Net, with +12.8%, +5.2%, +2.2%, +0.1% improvements, respectively. Notably, TPC does not rely on facial landmarks to extract specified local facial regions, which will bring out a heavy computation burden in the training and inference phase. Besides, TPC does not need to use auxiliary AU description word embeddings or a large amount of annotated facial expression data for auxiliary learning. As different AUs are associated with specific facial muscles and corresponds to fine-grained local facial regions, learning region-specific AU representations is beneficial. The success of the region-based AU detection approaches (Li et al., 2017b, 2019a, 2020b; Corneanu et al., 2018; Jacob and Stenger, 2021) have verified the benefits of the region-based AU detection approaches. We will explore this in future work.

Table 2 shows the AU detection accuracy comparison of our TPC and previous methods on the DISFA dataset. TPC achieves slightly superior AU detection accuracy with the best state-of-the-art self-supervised AU detection methods in the average F1 score, with 0.8% improvements over TAE, 7.3% improvements over TCAE, and 12.9% improvements over TRL. Notably, TPC shows its superiority in AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU6 (Cheek Raiser), AU12 (Lip Corner Puller), and obtains comparable AU detection performance in AU9 (Nose Wrinkler) and AU25 (Lips part). In summary, the benefits of the proposed TPC over other self-supervised AU detection methods can be summarized in 2-fold. First, TPC explicitly learns to encode the temporal evolution and consistency of the facial Aus in the temporal sequences. The self-attention mechanism in the transformer modules is capable of perceiving the local to global interactions between different facial AUs. Second, TPC incorporates the frame-wisely temporal contrastive learning into the self-supervised paradigm to further learn the per-frame discriminative-ness between the nearby facial frames. Thus, TPC is capable of perceiving the temporal consistency and the frame-wisely discriminativeness of the facial AUs self-supervised. The consistent improvements over other self-supervised AU detection methods have verified the feasibility of TPC. We will

TABLE 1 | Action unit (AU) detection accuracy of the proposed temporally predictive coding (TPC) and state-of-the-art approaches on BP4D dataset.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Ave
DRML Zhao et al. (2016)	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-Net Li et al. (2017b)	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
DSIN Corneanu et al. (2018)	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
LP-Net Niu et al. (2019)	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
UGN Song et al. (2021)	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	**48.5**	53.6	63.3
SRERL Li et al. (2019a)	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	**63.9**	47.1	53.3	62.9
FAUT Jacob and Stenger (2021)	51.7	49.3	**61.0**	77.8	**79.5**	82.9	86.3	**67.6**	51.9	63.0	43.7	**56.3**	**64.2**
SEV-Net Yang et al. (2021)	**58.2**	**50.4**	58.3	**81.9**	73.9	**87.8**	87.5	61.6	**52.6**	62.2	44.6	47.6	63.9
MAL Li and Shan (2021)	47.9	49.5	52.1	77.6	77.8	82.8	**88.3**	66.4	49.7	59.7	45.2	48.5	62.2
TCAE Li et al. (2019b)	43.1	32.2	44.4	**75.1**	70.5	80.8	85.5	61.8	34.7	58.5	37.2	**48.7**	56.1
TAE Li et al. (2020b)	**47.0**	**45.9**	50.9	74.7	**72.0**	**82.4**	85.6	62.3	48.1	**62.3**	45.9	46.3	60.3
TRL Lu et al. (2020)	42.3	24.3	44.1	71.8	67.8	77.6	83.3	61.2	31.6	51.6	29.8	38.6	52.0
TPC (Ours)	43.2	44.6	**52.8**	72.6	71.9	84.9	**86.9**	**64.8**	**50.3**	61.5	**55.6**	43.7	**61.1**

The best results in the supervised and self-supervised methods are illustrated in Bold.

TABLE 2 | Action unit detection accuracy of the proposed TPC and state-of-the-art approaches on the DISFA dataset.

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Ave
DRML Zhao et al. (2016)	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-Net Li et al. (2017b)	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
OFS-CNN Han et al. (2018)	43.7	40.0	67.2	59.0	49.7	75.8	72.4	54.8	51.4
DSIN Corneanu et al. (2018)	42.4	39.0	**68.4**	28.6	46.8	70.8	90.4	42.2	53.6
SRERL Li et al. (2019a)	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net Niu et al. (2019)	29.9	24.7	72.7	46.8	49.6	72.9	93.8	65.0	56.9
FAUT Jacob and Stenger (2021)	46.1	48.6	72.8	**56.7**	50.0	72.1	90.8	55.4	**61.5**
SEV-Net Yang et al. (2021)	**55.3**	53.1	61.5	53.6	38.2	71.6	95.7	41.5	58.8
UGN Song et al. (2021)	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
MAL Li and Shan (2021)	**43.8**	**39.3**	**68.9**	47.4	**48.6**	**72.7**	**90.6**	52.6	**58.0**
TCAE Li et al. (2019b)	15.1	15.2	50.5	48.7	23.3	72.1	82.1	52.9	45.0
TAE Li et al. (2020b)	21.4	19.6	**64.5**	46.8	**44.0**	73.2	**85.1**	**55.3**	**51.5**
TRL Lu et al. (2020)	18.7	27.4	35.1	33.6	20.7	67.5	68.0	43.8	39.4
TPC (Ours)	**22.8**	**30.8**	59.6	**53.9**	42.7	**75.3**	82.1	51.6	52.3

The best results in the supervised and self-supervised methods are illustrated in Bold.

TABLE 3 | Ablation studies on the BP4D and DISFA datasets.

Methods	BP4D	DISFA
$\mathcal{L}_{pred}$	58.7	49.8
$\mathcal{L}_{tcl}$	57.9	50.8
$\lambda = 10.0$	55.2	47.1
$\lambda = 1.0$	59.3	48.6
$\lambda = 0.1$	61.1	52.3

carry out an ablation study to investigate the contribution of the two components in TPC in the next section.

4.2.1. Ablation Study

Table 3 shows the ablation experimental results. In **Table 3**, we show the accuracy variations with a different self-supervised components, and show the influence with different λ. As shown in **Table 3**, TPC shows the best AU detection performance with the linear combination of $\mathcal{L}_{pred}$ and $\mathcal{L}_{tcl}$ with $\lambda = 0.1$. It means both components in TPC contribute to its success in learing discriminative AU representations. Without either of the two self-supervised targets, TPC will show degraded AU detection accuracies. Besides, TPC also suffers from low accuracy with $\lambda = 1.0$ and $\lambda = 10.0$, which suggests the two self-supervised learning targets should be appropriately balanced to achieve the discriminative AU representations.

5. CONCLUSION

Within this paper, we aim to propose a self-supervised pseudo signal based on TPC to capture the temporal characteristics of the facial AUs in the sequential facial frames. To further learn the per-frame discriminativeness between the nearby faces, TPC incorporates the frame-wisely temporal contrastive learning into the self-supervised paradigm. The proposed TPC can be pre-trained without AU annotations, which facilitates making use of a large amount of unlabeled facial videos to learn the AU features that are robust to other undesired nuisances. Compared with supervised facial AU detection methods, TPC obtains comparable AU detection performance. Besides, TPC is superior to other self-supervised AU detection approaches. For future work, we will explore learning to perceive the regional and structural AU features in the temporal contrastive learning paradigm.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CW completed the algorithm design and wrote all parts of the manuscript. CW and ZW cooperatively conducted the experimental evaluation and cooperatively gave a detailed experimental analysis. ZW carefully checked the manuscript and polished the paper. Both authors have carefully read, polished, and approved the final manuscript.

REFERENCES

Bartlett, M. S., Littlewort, G., Fasel, I., and Movellan, J. R. (2003). "Real time face detection and facial expression recognition: development and applications to human computer interaction," in *2003 Conference on Computer Vision and Pattern Recognition Workshop, Vol. 5* (Madison, WI: IEEE), 53–53.

Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C., Tu, Z., and Soatto, S. (2021). "Exponential moving average normalization for self-supervised and semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN), 194–203.

Corneanu, C., Madadi, M., and Escalera, S. (2018). "Deep structure inference network for facial action unit recognition," in *ECCV* (Munich), 298–313.

Ekman, P., and Friesen, W. V. (1978). *Manual for the Facial Action Coding System*. Consulting Psychologists Press.

Han, S., Meng, Z., Li, Z., O'Reilly, J., Cai, J., Wang, X., et al. (2018). "Optimizing filter size in convolutional neural networks for facial action unit recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 5070–5078.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Hu, K., Shao, J., Liu, Y., Raj, B., Savvides, M., and Shen, Z. (2021). "Contrast and order representations for video self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal), 7939–7949.

Jacob, G. M., and Stenger, B. (2021). "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN), 7680–7689.

Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., and Mottaghi, R. (2021). "Contrasting contrastive self-supervised representation learning pipelines," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal), 9949–9959.

Li, G., Zhu, X., Zeng, Y., Wang, Q., and Lin, L. (2019a). Semantic relationships guided representation learning for facial action unit recognition. *AAAI* 33, 8594–8601. doi: 10.1609/aaai.v33i01.33018594

Li, W., Abtahi, F., and Zhu, Z. (2017a). "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE).

Li, W., Abtahi, F., Zhu, Z., and Yin, L. (2017b). "Eac-net: a region-based deep enhancing and cropping approach for facial action unit detection," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (Washington, DC: IEEE).

Li, Y., and Shan, S. (2021). Meta auxiliary learning for facial action unit detection. *IEEE Trans. Affect. Comput*. doi: 10.1109/TAFFC.2021.3135516

Li, Y., Zeng, J., Liu, X., and Shan, S. (2020a). Progress and challenges in facial action unit detection. *J. Image Graphics (in Chinese)* 25, 2293–2305. doi: 10.11834/jig.200343

Li, Y., Zeng, J., and Shan, S. (2020b). Learning representations for facial actions from unlabeled videos. *IEEE Trans. Pattern Anal. Mach. Intell*. 44, 302–317. doi: 10.1109/TPAMI.2020.3011063

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018a). Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Process*. 28, 2439–2450. doi: 10.1109/TIP.2018.2886767

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018b). "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)* (Beijing: IEEE), 2209–2214.

Li, Y., Zeng, J., Shan, S., and Chen, X. (2019b). "Self-supervised representation learning from videos for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 10924–10933.

Lu, L., Tavabi, L., and Soleymani, M. (2020). "Self-supervised learning for facial action unit recognition through temporal consistency," in *Proceedings of the British Machine Vision Conference (BMVC)* (BMVA Press).

Luo, C., Yang, X., and Yuille, A. (2021). "Self-supervised pillar motion learning for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN), 3183–3192.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE TAC* 4, 151–160. doi: 10.1109/T-AFFC.2013.4

Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang*. 60, 101027. doi: 10.1016/j.csl.2019.101027

Niu, X., Han, H., Yang, S., Huang, Y., and Shan, S. (2019). "Local relationship learning with person-specific shape regularization for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 11917–11926.

Peng, G., and Wang, S. (2018). "Weakly supervised facial action unit recognition through adversarial training," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 2188–2196.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). "Facenet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 815–823.

Song, T., Chen, L., Zheng, W., and Ji, Q. (2021). "Uncertain graph neural networks for facial action unit detection," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol.* 1.

Sun, Y., Zeng, J., Shan, S., and Chen, X. (2021). "Cross-encoder for unsupervised gaze representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Nashville, TN), 3702–3711.

Wang, G., Han, H., Shan, S., and Chen, X. (2020). Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Trans. Inf. Forensics Security* 16, 56–69. doi: 10.1109/TIFS.2020.3002390

Wiles, O., Koepke, A., and Zisserman, A. (2018). Self-supervised learning of a facial attribute embedding from video. *arXiv preprint* arXiv:1808.06882. doi: 10.1109/ICCVW.2019.00364

Yan, J., Wang, J., Li, Q., Wang, C., and Pu, S. (2021). "Self-supervised regional and temporal auxiliary tasks for facial action unit recognition," in *Proceedings of the 29th ACM International Conference on Multimedia* (Chengdu), 1038–1046.

Yang, H., Yin, L., Zhou, Y., and Gu, J. (2021). "Exploiting semantic embedding and visual feature for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN), 10482–10491.

Zafar, Z., and Khan, N. A. (2014). "Pain intensity evaluation through facial action units," in *2014 22nd International Conference on Pattern Recognition* (Stockholm: IEEE), 4696–4701.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., et al. (2013). "A high-resolution spontaneous 3d dynamic facial expression database," in *FG* (Shanghai: IEEE).

Zhao, K., Chu, W.-S., and Martinez, A. M. (2018). "Learning facial action units from web images with scalable weakly supervised clustering," in *CVPR* (Salt Lake City, UT), 2090–2099.

Zhao, K., Chu, W.-S., and Zhang, H. (2016). "Deep region and multi-label learning for facial action unit detection," in *CVPR* (Nevada), 3391–3399.

Zonoozi, A., Kim, J.-J., Li, X.-L., and Cong, G. (2018). "Periodic-crn: a convolutional recurrent model for crowd density prediction with recurring periodic patterns," in *IJCAI*, 3732–3738.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 23 March 2022
doi: 10.3389/fcomp.2022.773154

An Estimation of Online Video User Engagement From Features of Time- and Value-Continuous, Dimensional Emotions

Lukas Stappen[1]*, *Alice Baird*[1], *Michelle Lienhart*[1], *Annalena Bätz*[1] and *Björn Schuller*[1,2,3]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [2] audEERING GmbH, Gilching, Germany, [3] GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom

OPEN ACCESS

Edited by:
Zhen Cui,
Nanjing University of Science and Technology, China

Reviewed by:
Robertas Damasevicius,
Silesian University of Technology, Poland
Tong Zhang,
Nanjing University of Science and Technology, China
Ming Yin,
Jiangsu Police Officer College, China
Xiaoya Zhang,
Nanjing University of Science and Technology, China

***Correspondence:**
Lukas Stappen
stappen@ieee.org

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: 09 September 2021
Accepted: 24 February 2022
Published: 23 March 2022

Citation:
Stappen L, Baird A, Lienhart M, Bätz A and Schuller B (2022) An Estimation of Online Video User Engagement From Features of Time- and Value-Continuous, Dimensional Emotions. Front. Comput. Sci. 4:773154. doi: 10.3389/fcomp.2022.773154

Portraying emotion and trustworthiness is known to increase the appeal of video content. However, the causal relationship between these signals and online user engagement is not well understood. This limited understanding is partly due to a scarcity in emotionally annotated data and the varied modalities which express user engagement online. In this contribution, we utilize a large dataset of YouTube review videos which includes ca. 600 h of dimensional arousal, valence and trustworthiness annotations. We investigate features extracted from these signals against various user engagement indicators including views, like/dislike ratio, as well as the sentiment of comments. In doing so, we identify the positive and negative influences which single features have, as well as interpretable patterns in each dimension which relate to user engagement. Our results demonstrate that smaller boundary ranges and fluctuations for arousal lead to an increase in user engagement. Furthermore, the extracted time-series features reveal significant ($p < 0.05$) correlations for each dimension, such as, count below signal mean (arousal), number of peaks (valence), and absolute energy (trustworthiness). From this, an effective combination of features is outlined for approaches aiming to automatically predict several user engagement indicators. In a user engagement prediction paradigm we compare all features against semi-automatic (cross-task), and automatic (task-specific) feature selection methods. These selected feature sets appear to outperform the usage of all features, e.g., using all features achieves 1.55 likes per day (Lp/d) mean absolute error from valence; this improves through semi-automatic and automatic selection to 1.33 and 1.23 Lp/d, respectively (data mean 9.72 Lp/d with a std. 28.75 Lp/d).

Keywords: user engagement, explainable machine learning, popularity of videos, affective computing, YouTube, continuous emotion annotation

1. INTRODUCTION

Online video content hosted by platforms such as YouTube is now gaining more daily views than traditional television networks (Battaglio, 2016). There are more than 2 billion registered users on YouTube, and a single visitor will remain on the site for at least 10 min (Cooper, 2019). Viewers rate of retention for a single video is between 70–80%, and such retention times may be due to (cross-) social network effects (Roy et al., 2013; Yan et al., 2015; Tan and Zhang, 2019) and

the overall improvement in content and connection quality in recent years (Dobrian et al., 2011; Lebreton and Yamagishi, 2020), but arguably caused by intelligent mechanisms (Cheng et al., 2013), e.g., 70% of videos watched on YouTube are recommended from the previous video (Cooper, 2019). To this end, gaining a better understanding of what aspects of a video a user engages with has numerous real-life applications (Dobrian et al., 2011). For example, videos such as misinformation, fake messages, and hate speech are strongly emotionally charged (Knuutila et al., 2020) and detection using conventional methods such as natural language processing is to date a tremendous challenge (Stappen et al., 2020b). Another application is the use by creators who adapt their content to have a greater prospect of the video becoming *viral* (Trzciński and Rokita, 2017) and thus improve advertising opportunities.

Positive emotion (Berger and Milkman, 2012) and trust of the individuals in videos (Nikolinakou and King, 2018) have shown to affect user (i.e., content) engagement (Shehu et al., 2016; Kujur and Singh, 2018). In traditional forms of entertainment (i.e., film) portraying emotion captivates the audiences improving their ability to remember details (Subramanian et al., 2014) and similar *persuasion appeals* are applied within shorter-form YouTube videos (English et al., 2011). When emotion is recognized computationally, research has shown that the emotion (arousal and valence) of a video can be an indicator of popularity, particularly prominent when observing audio features (Sagha et al., 2017).

The frequency of comments by users is also a strong indicator of how engaged or not users are with a video (Yang et al., 2016). Furthermore, understanding the sentiment of comments (i.e., positive, neutral, or negative) can offer further insights on the type of view engagement, e.g., more positive sentiment correlates to longer audience retention (Yang et al., 2016). In a similar way to the use of emotions, developing trust between the viewer (trustor) and the presenter (trustee) has also shown to improve user engagement. It is a common strategy by content creators to facilitate what is known as a *parasocial relationship*. A parasocial relationship develops when the viewer begins to consider the presenter as a friend without having ever met them (Chapple and Cownie, 2017).

With this in mind, we unite multiple emotional signals for an explicit engagement analysis and prediction in this current contribution. Thereby, we focus on the utilization of the emotional dimensions of arousal and valence and extend the typical Russel circumplex model for emotion, by adding trustworthiness as a continuous signal. Hereby, we follow a two-step approach: First, we aim to understand better continuous factors which improve metadata-related (i.e., views, likes, etc.) and comment-related (i.e., sentiment of comments, positive-negative ratios, likes of comments etc.) user engagement across modes (i.e., emotional signals to text-based indicators). To do this, we collect the metadata as well as more than 75 k comments from the videos. We annotate a portion of these comments to be used in combination with other data sets for training a YouTube comment sentiment predictor for the automatic assessment of the unlabeled comments. Furthermore, we utilize a richly annotated data set of ca. 600 h of continuous annotations (Stappen et al., 2021), and derive cross-task features from this initial correlation analysis. Second, we compare these engineered, lean features, to a computationally intensive feature selection approach and to all features when predicting selected engagement indicators (i.e., views, likes, number of comments, likes of the comments). We predict these indicators as a regression task, and train *interpretable* (linear kernel) Support Vector Regressors (SVR). The main contributions to the research community are two-fold:

1. To the best of the authors' knowledge, there has been no research which analyses YouTube video user engagement against trustability time-series features.
2. Furthermore, we are the first to predict cross-modal user popularity indicators as a regression task—purely based on emotional signal features without using typical text, audio, or images/video features as input.

This article is organized as follows; firstly, in Section 2, we provide a brief background on the core concepts which relate to emotions and user-generated content. We then introduce the data that is used within the experiments in cf. Section 3. This is followed by the experimental methodology, in Section 4, including feature extraction from signals and sentiment extraction from text, and the machine learning pipeline overall. The results are then extensively analyzed and discussed in Section 5, with a mention of study limitations in Section 6. Finally, we offer concluding remarks and future outlook in Section 7. The newly designed and extended datasets, code, and the best models will be made publicly available on in our project repository.

2. BACKGROUND

Within our contribution, the concept of emotions for user-generated content is extended from the conventional Russel concept of emotion dimensions, valence, and arousal (Russell, 1980), to include a continuous measure for trustworthiness. In the following, we introduce these core concepts and related studies.

2.1. Concepts of Emotion and Trustworthiness

There are two predominant views in the field of affective science: the first assumes that emotions are discrete constructs, each acting as an independent emotional system of the human brain, and hence, can be expressed by discrete categories (Ekman, 1992). The second assumes an underlying interconnected dimensional signal system represented by continuous affective states.

For emotion recognition using continuous audio-video signals, the circumplex model of emotion developed by Russel is the most prominent (Russell, 1980) and applied (Busso et al., 2008; Kossaifi et al., 2019; Stappen et al., 2021) approach of the latter idea. This representation of affect typically consists of continuous valence (the positiveness/ negativity of the emotion) and arousal dimensions (the strength of the activation of the emotion), as well as an optional third focus dimension (Posner et al., 2005).

In the past, both approaches to classify emotions in user-generated content (Chen et al., 2017) rely on Ekmann's model to predict six emotional classes in YouTube videos. Similarly, Zadeh et al. (2018) annotated YouTube videos with labels for subjectivity and sentiment intensity (Wöllmer et al., 2013) was the first to transfer the dimensional concept to YouTube videos. Recently, Kollias et al. (2019) annotated 300 videos (ca. 15 h) of "in-the-wild" data, predominantly YouTube videos under the creative commons license.

However, none of the mentioned datasets allows the bridging of annotated or predicted emotional signals with user engagement data from videos. We fill this research gap utilizing continuous emotional signals and corresponding data, as well as providing insights into the novel dimension of trustworthiness, entirely without relying on word-based, audio, or video feature extraction.

Although general literature lacks in providing a consistent concept of trustworthiness (Horsburgh, 1961; Moturu and Liu, 2011; Cox et al., 2016), in this work, we define trust as the ability, benevolence, and integrity of a trustee analogous to Colquitt et al. (2007). In the context of user-generated reviews, the viewers assess from their perspective if and to what extent the reviewer communicates unbiased information. In other words, how truthful and knowledgeable does the viewer feel a review is at every moment? As we mentioned, building this trust is part of developing a parasocial relationship with the audience, and in doing so, likely increases repeated viewing (Lim et al., 2020).

2.2. Sentiment Analysis of YouTube Comments

Sentiment Analysis studies the extraction of opinions, sentiments, and emotions (e.g., "positive," "negative," or "neutral") of user-generated content. The analyzed content usually consists of text (Boiy et al., 2007; Gilbert and Hutto, 2014), such as in movie and product reviews, as well as comments (Singh et al., 2013; Siersdorfer et al., 2014). In recent years, the methods for text classification have developed rapidly. Earlier work using rule-based and classical word embedding approaches is now being replaced by what is known as *transformer networks*, predicting *context-based* word embeddings (Devlin et al., 2019). State-of-the-art accuracy results on sentiment benchmark datasets using these methods (Cui et al., 2019) range from 77.3 for the 3-classes twitter (Nakov et al., 2013) and between 72.4 and 75.0 on a 2-classes YouTube comment data sets (Uryupina et al., 2014).

In contrast to the literature, our approach utilizes the predicted sentiment of a fine-tuned Word Embedding Transformer ALBERT (Lan et al., 2020) to automatically classify comments on a large scale to investigate the cross-modal relationship to the continuous emotion and trustworthiness signals.

2.3. Analysis of YouTube Engagement Data and Cross-Modal Studies

YouTube meta and engagement data are well researched (Yan et al., 2015) with contributions exploring across domains (Roy et al., 2013; Tan and Zhang, 2019), and focusing on both long (Biel and Gatica-Perez, 2013) and short form video sharing (Cheng et al., 2013; Garroppo et al., 2018).

Most previous work analyse view patterns, users' opinions (comments) and users' perceptions (likes/dislikes), and their mutual influence (Bhuiyan et al., 2017). Khan and Vong (2014) correlated these reaction data, while (Rangaswamy et al., 2016) connects them to the popularity of a video.

An extended comment analysis has been conducted by Severyn et al. (2016) predicting the type and popularity toward the product and video. The comment ratings, thus the community acceptance, was predicted by Siersdorfer et al. (2010) using the comment language and discrete emotions. Moreover, in Wu and Ito (2014) the authors correlated popularity measures and the sentiment of the comments. Data of other social networking platforms combine sentiment analysis and social media reactions (Ceron et al., 2014; Gilbert and Hutto, 2014), and (Preoţiuc-Pietro et al., 2016) attempted to map Facebook posts to the circumplex model to predict the sentiment of new messages.

To the best of our knowledge, no work has so far attempt to investigate the relationship to sophisticated continuous emotional and trustworthiness signals and based on these, predict user engagement as regression tasks.

3. DATA

The base for our experimental work is the MUSE-CAR data set[1] (Stappen et al., 2021). MUSE-CAR is a multi-media dataset originally crafted to improve machine understanding of multi-modal sentiment analysis in real-life media. For the first time, it was used for the MUSE 2020 Challenge, which aimed to improve emotion recognition systems, focusing on the prediction of arousal and valence emotion signals (Stappen et al., 2020c). For a detailed description of typical audio-visual feature sets and baseline systems that are not directly related to this work, we point the reader to Stappen et al. (2020a).

3.1. Video, Meta- and Engagement Data

The dataset contains over 300 user-generated vehicle review videos, equal to almost 40 h of material that cover a broad spectrum of topics within the domain. The videos were collected from YouTube[2] and have an average duration of 8 min. The reviews are primarily produced by semi—("influencers") or professional reviewers with an estimated age range of the mid-20 s until the late-50 s. The speech of the videos is English. We refer the reader to Stappen et al. (2021) for further in-depth explanation about the collection, the annotator training, and the context of the experiments. Utilizing the YouTube ID, we extend the data set by user engagement data. The explicit user engagement indicators are calculated on a per-day basis (p/d) as the videos were uploaded on different days resulting in

[1] The raw videos and YouTube IDs are available for download: https://zenodo.org/record/4651164.

[2] All owners of the data collected for use within the MUSE-CAR data set were contacted in advance for the consent of use for research purposes.

FIGURE 1 | Example video (video ID 5): Four annotation signals for arousal and valence in addition to the fused EWE "gold-standard" signal (bold purple). The annotator 1 (green) has a negative correlation to the others on valence. The weight (r_1) is set to 0, not considered for the fused gold-standard by EWE.

views **(Vp/d)**, likes **(Lp/d)**, dislikes **(Dp/d)**, comments **(Cp/d)**, and likes of comments **(LCp/d)**. Per video the user engagement criteria is distributed (μ mean, σ standard deviation) as; Vp/d: $\mu = 863.88, \sigma = 2048.43$; Lp/d: $\mu = 9.73, \sigma = 28.75$, Dp/d: $\mu = 0.4125, \sigma = 1.11$; Cp/d: $\mu = 0.91, \sigma = 3.00$; and LCp/d: $\mu = 5.28, \sigma = 16.84$.

3.2. Emotion and Trustworthiness Signals

As with emotions in general, a certain level of disagreement due to subjectivity can be expected (Russell, 1980). For this reason, nine annotators were trained (Stappen et al., 2021) to have a common understanding of the arousal, valence, and trustworthiness concepts as discussed in Section 2.1. As well established (Busso et al., 2008; Kossaifi et al., 2019), the annotator moves the hand up and down using a *Logitech Extreme 3D Pro Joystick* to annotate one of three dimensions, while watching the videos. The movements are recorded over the entire duration of the video sequence and sampled with a bin size of 0.25 Hz on an axis magnitude between -1 000 and 1 000. Every annotation was checked by an auditor using quantitative and qualitative measures to ensure a high quality (Baird and Schuller, 2020). The time required for annotation alone stands for more than 600 working hours (40 h video * 3 dimensions * 5 annotators per dimension).

The annotation of five independent annotators for each video and signal type are fused to obtain a more objective gold-standard signal as depicted in **Figure 1**. For the fusion of the individual continuous signals, the widely established Evaluator Weighted Estimator (EWE) was computed (Schuller, 2013; Ringeval et al., 2017). It is an estimator of inter-rater agreement, hence, the personal reliability, in which the weighted mean corresponds to the calculated weights for each rater based on the cross-dependency of all other annotators. The EWE can be formulated as

$$y_n^{EWE} = \frac{1}{\sum_{a=1}^{A} r_a} \sum_{a=1}^{A} r_a y_{n,a}, \tag{1}$$

where y is a discrete point of the signal n and r_a is the reliability of the ath rater, consequently, A represents the whole population of raters. To use the data for later stages, we z-standardize them.

3.3. Video Comments

Based on the video IDs of the corpus, we collected more than 79 k YouTube comments and comment-related like counts excluding any other user information, such as the username. We focus exclusively on the parent comments, ignoring reaction from the child comments. The count of comment likes reflect the number of people sharing the same opinion and those who "liked" the comment. We randomly select 1 100 comments for

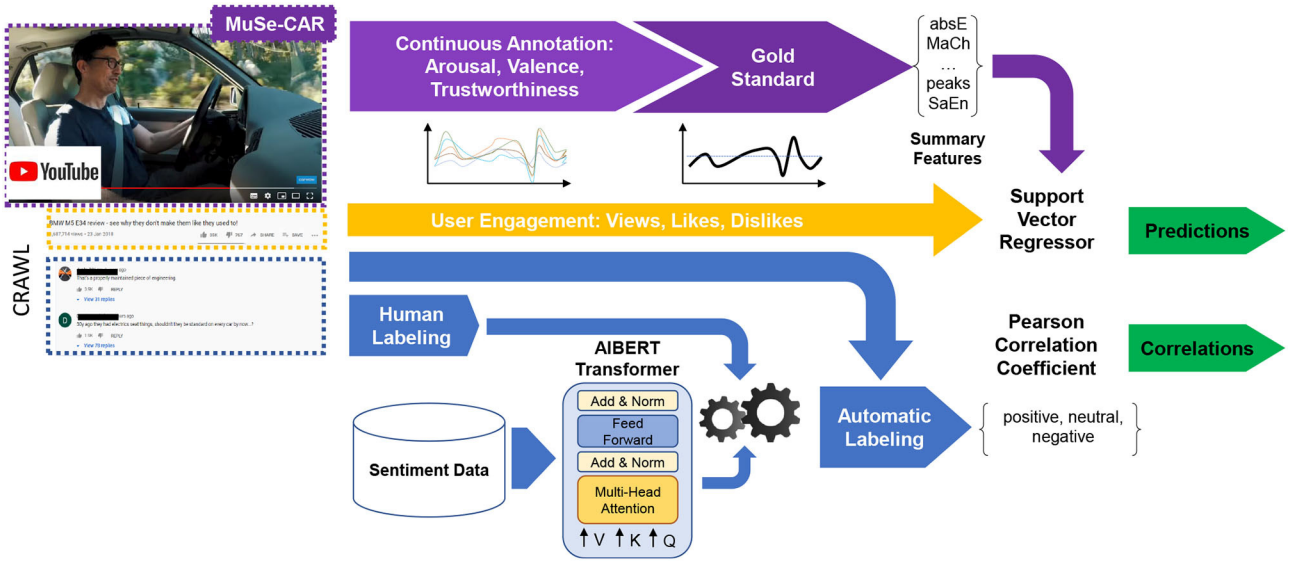


FIGURE 2 | A comprehensive overview of the approach, as explained in Section 4. Statistics derived from continuous emotion and trustworthiness (purple), automatic sentiment labeling of YouTube comments (blue), and user engagement data (orange) are preprocessed to investigate correlations and predict engagement from features. The development of our signal feature extraction as a cornerstone of our analysis is described in detail in Section 4.1. With its help, we aim to uncover relationships to engagement data (cf. Section 4.3) and the sentiment of the comments predicted by our trained network (cf. Section 4.2). They also serve as input to our regression experiments to predict user engagement (cf. Section 4.5).

labeling, which is used as a quantitative estimator of how accurate our prediction of the other unlabeled comments are. Three annotators labeled each of them as positive, neutral, negative, and not applicable. The average inter-rater joint probability is 0.47. We use a majority fusion to create a single ground truth, excluding texts where no majority is reached.

4. EXPERIMENTAL METHODOLOGY

Figure 2 gives an overview of our approach. As a cornerstone of our analysis (cf. Section 4.1), annotation of arousal, valence, and trustworthiness are annotated by five independent annotators. These signals are then fused (cf. Section 3.2) to a gold standard label, and meaningful features are extracted (**purple**). In addition, YouTube user engagement-related data (**yellow**) and the comments are scraped (**blue**) from each video. Several sentiment data sets are collected and merged in order to train a robust sentiment classifier using a Transformer network ALBERT to predict unlabeled YouTube comments after fine-tuning on several datasets and our labeled comments. Then, we first investigate correlations between the predicted sentiment of the YouTube comments, the YouTube metadata, and the statistics derived from the continuous signals (arousal, valence, and trustworthiness). Additionally, we use derived features to predict user engagement (Vp/d, Lp/d, Cp/d, and CLp/d) directly.

TABLE 1 | List of simple statistics and more complex time-series statistic features extracted by our framework.

Distribution statistics	
	Standard deviation (std)
	5%-, 25%-, 50%-, 75%-, and 95%-quantiles ($q_5, q_{25}, q_{50}, q_{75}, q_{95}$)
Time-series statistics	
Asymmetry	Dynamic sample skewness (skew)
	Kurtosis (kurt)
Energy-related	Absolute energy (absE)
	Sample entropy (SaEn)
Change-related	Absolute sum of changes (ASOC)
	Mean absolute change (MACh)
	Mean change (MCh)
	Mean value of a central approximation of the second derivatives (MSDC)
	Strike above the mean (LSAMe)
	Strike below the mean (LSBMe)
Relative points	Normalized percentage of reoccurring datapoints (PreDa)
	First and last location of the minimum and maximum (FLMi, LLMi, FLMa, and FLMa)
	Number of crossings of a point m (CrM)
	Peaks of the least support (peaks)

4.1. Feature Extraction From Signals

A signal is usually sampled to fine-grained, discrete points of regular intervals, which can be interpreted as a sequential set of successive data points over time (Adhikari and Agrawal, 2013). Audio, video, and psychological signals are widely used for computational analysis (Schuller, 2013; Schuller et al., 2020). Simple statistics and advanced feature extraction can be applied in order to condense these signals to meaningful summary

representations and make them more workable (Christ et al., 2018). In this work, we use common statistical measures such as the standard deviation (*std*), and 5-, 25-, 50-, 75-, and 95%-*quantiles* ($q_5, q_{25}, q_{50}, q_{75}, and q_{95}$) as they are less complex to interpret, and have been applied in related works (Sagha et al., 2017). Furthermore, to make better use of the changes over time, we manually select and calculate a wide range of time-series statistics following previous work in similar fields (Geurts, 2001; Schuller et al., 2002). For example, in computational audition (e.g., speech emotion recognition), energy-related features of the audio signals are used to predict emotions (Schuller et al., 2002).

We calculate the dynamic sample skewness (*skew*) of a signal using the adjusted Fisher-Pearson standardized moment coefficient, to have a descriptor for the asymmetry of the series (Ekman, 1992; Doane and Seward, 2011). Similarly, the kurtosis (*kurt*) measures the "flatness" of the distribution by utilizing the fourth moment (Westfall, 2014). Of the energy-related ones, the absolute energy (*absE*) of a signal can be determined by the sum over the squared values (Christ et al., 2018).

$$absE = \sum_{i=1,\ldots,n} x_i^2, \tag{2}$$

where x is the signal at point i. Also well known for physiological time-series signals is the sample entropy (*SaEn*), a variation of the approximate entropy, to measure the complexity independently of the series length (Richman and Moorman, 2000; Yentes et al., 2013). Several change-related features might be valuable to reflect the compressed signal (Christ et al., 2018): First, the sum over the absolute value of consecutive changes expresses the absolute sum of changes (*ASOC*):

$$ASOC = \sum_{i=1,\ldots,n-1} |x_{i+1} - x_i|. \tag{3}$$

Second, the mean absolute change (*MACh*) over the absolute difference between subsequent data points is defined as:

$$MACh = \frac{1}{n} \sum_{i=1,\ldots,n-1} |x_{i+1} - x_i|, \tag{4}$$

where n is the number of time-series points. Third, the general difference between consecutive points over time is called the mean change (*MCh*):

$$MCh = \frac{1}{n-1} \sum_{i=1,\ldots,n-1} x_{i+1} - x_i. \tag{5}$$

Fourth, the mean value of a central approximation of the second derivatives (*MSDC*) is defined as:

$$MSDC = \frac{1}{2 * (n-1)} \sum_{i=1,\ldots,n-1} 0.5 \cdot (x_{i+2} - 2 \cdot i + 1 + x_i). \tag{6}$$

Finally, the length of the normalized consecutive sub-sequence is named strike above (*LSAMe*) and below (*LSBMe*) the mean. To summarize the distribution similarity, the normalized percentage of reoccurring datapoints (*PreDa*) of non-unique single points can be calculated by taking the number of data points occurring more than once divided by the number of total points. Also early or late high and low points of the signal are of descriptive value. Four single points describe these: the first and last location of the minimum and maximum (*FLMi*, *LLMi*, *FLMa*, and *FLMa*) relatively to the length of the series. The last two count a) the number of crossings of a point m (here: m=0) (*CrM*), where for two successive time series steps are first lower (or higher) than m followed by two higher (or lower) ones (Christ et al., 2018) and b) the *peaks* of the least support n. A peak of support n is described as a subsequence of a series where a value occurs, bigger than its n neighbors to the left and the right (Palshikar, 2009; Christ et al., 2018). In total, we extract 24 features from one signal (cf. **Table 1**).

4.2. Sentiment Extraction From Comments

Given the vast amount of comments, we decided to carry out the labeling of the sentiment automatically and label only a small share of them by hand to quantify the prediction quality (cf. Section 3.3). For this reason, we built a robust classifier for automatic YouTube sentiment prediction using PyTorch. We opted to use ALBERT as our competitive Transformer architecture (Lan et al., 2020). Compared to other architectures, ALBERT introduces two novel parameter reduction methods: First, the embedding matrix is separated into two more compact matrices, and second, layers are grouped and used repeatedly. Furthermore, it applies a new self-supervised loss function that improves training for downstream fine-tuning tasks. These changes have several advantages, such as reducing the memory footprint, accelerating the converge of the network, and leading to state-of-the-art results in several benchmarks (Devlin et al., 2019).

Before training, we remove all words starting with a "#," "@," or "http" from all text sources and replace emotions' unicode by the name. We train ALBERT in a two-step procedure. First, we fine-tune the model for the down-stream task of general sentiment analysis. No extensive YouTube comment data set is available, which would span the wide range of writing styles and expressed opinions. Therefore, we aggregate several datasets which aim to classify whether a text is positive, negative, or neutral as our initial training data: all data sets from SemEval (the Semantic Evaluation challenge), a series of challenges for computer-based text classification systems with changing domains (Nakov et al., 2013) e.g., Twitter, SMS, sarcasm, from 2013 to 2017 consisting of more than 76 k data points; the popular US Airline Sentiment data set (Air, 2015) (14.5 k tweets), and finally, 35 k positive and 35 k negative text snippets are selected from Sentiment140 (Go et al., 2009). The 60 k positive, 32 k neutral, and 56 k negative text snippets are equally stratified and partitioned into 80-10-10 splits for training. We provide this selection for reproducibility in our code.

Following the authors' recommendation, ALBERT is trained using a learning rate of *1e-5*, a warmup ratio of *0.06*, ϵ set to *1e-8*, and gradient clipping set at *1.0*. In addition, we use half-precision training and a batch size of 12 to fit the GPU memory restrictions

TABLE 2 | Example comments and sentiment distribution within the YouTube comments predicted by our developed sentiment model.

Sentiment	# Comments	Predicted [%]	Example
Positive	26 032	33	"The metaphors are just flying like the raindrops in this video." #47620
Neutral	28 518	36	"Are engines for F30 made in Germany?" #4
Negative	24 494	31	"Poor review unfortunately, the microphone quality was very muffled..." #31

(32 GBs). Counteracting adverse effects of class imbalance, we further inject the class weight to each data point. The model converges after three epochs. Next, we use our own YouTube comment data set to validate the results and further fine-tune the model. This version is then further trained in a second fine-tuning step using the 60% of the YouTube comments and a reduced learning rate of *1e-6* for one epoch.

The relative distribution of the classified sentiment of the YouTube comments is given in **Table 2**. The model achieves an f1 score on the development of 81.13 and 78.09% on the test partitions, as well as 75.41% on the sample of our crawled and manually labeled YouTube test set.

4.3. Correlation Measure and Significance

The Pearson correlation (r) explores the relationship between two continuous variables (Ahlgren et al., 2003). Thereby, the relationship has to be linear, meaning that when one variable changes, the other also changes proportionally. r is defined by

$$r_{x,z} = \frac{cov(x,z)}{\sigma_x \cdot \sigma_z} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(z_i - \bar{z})^2}}, \quad (7)$$

where $\text{cov}(x, z)$ is the co-variance, a measure of the joint variability, of the variables X, Z, and σ_x, σ_z – the standard deviations of both variables (Surhone et al., 2010). The resulting correlation coefficient lies between -1 and $+1$. If the value is positive, the two variables are positively correlated. A value of ± 1 signifies a perfect positive or negative correlation. A coefficient equal to zero implies that there is no linear dependency between the variables.

For significance testing, we first compute the t-statistic, and then twice the survival function for the resulting t-value to receive a two-tailed p-value, in which the null hypothesis (two variables are uncorrelated) is rejected at levels of $\alpha = \{0.01, 0.05, 0.1\}$ (Sham and Purcell, 2014). Since, we intend to give the reader as much transparency as possible with regard to the robustness of the results obtained given the size of the data set, we report the results on three common significance levels (see **Appendix**). Therefore, results significant at an alpha level of 0.01 are also significant at 0.05 and 0.1.

4.4. Feature Selection

To the best of our knowledge, we are the first extracting advanced features directly from emotional signals. Usually, not all engineered features are equally relevant. Since no previous research can guide us to a reliable selection, we propose two ways for feature selection for our task of predicting user engagement. The first is a correlation-based, *cross-task semi-automatic selection* that uses the correlation between the feature and the target variables. Only those features whose mean value over all prediction tasks is between $-0.2 > r_{mean} > +0.2$ (minimum low positive/negative correlation) are selected.

The other concept is a regression-based, *task-specific automatic selection* with three steps. First, univariate linear regression (f) tests act as a scoring function and run successively to measure the individual effect of many regressors:

$$score(f, y) = \frac{X_{k_i} - \bar{X_{k_i}} \cdot (y - \bar{y})}{\sigma_{X_{k_i}} \cdot \sigma_y}, \quad (8)$$

where k_i is the feature index. The score is converted to an F-test estimate and then to a p-value. Second, the highest k number of features are selected based on the p-value. Finally, this procedure runs brute-force for all number of feature combinations, where $5 < k < k_{max-1}$. Brute-force implies an exhaustive search, which systematically checks all possible combinations until the best one is found based on the provided estimate.

4.5. SVR Training Procedure

For our regression experiments, we use a Support Vector Regression (SVR) with a linear kernel as implemented by the open-source machine learning toolkit Scikit-Learn (Pedregosa et al., 2011). The linear kernel allows us to interpret the weights from our various feature selections and has, among other applications, found wide acceptance in the selection of relevant genes from micro-array data (Guyon et al., 2002). Since the coefficients are orthogonal to the hyper-plane, a feature is useful to separate the data when the hyper-plane is orthogonal to the feature axis. The absolute size of the coefficient concerning the other features indicates the importance.

The training is executed on the 60-20-20 training/development/test partition split partitions, pre-defined in the MUSE-CAR emotion recognition sub-task (Stappen et al., 2020c) (cf. Section 3.1). During the training phase, we train a series of models on the training set with different parameters $C \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ up to 10 000 iterations and validate the performance on the development set. The best performing C value is then used to re-train the model on an enlarged, concatenated training and development set, to estimate the generalization performance on the hold-out test set. This method is repeated for each input signal (combination) on each target (%). Due to the various scales of the input features, we apply standardization to the data but leave the targets, as they allow interpretability of the results. The prediction results are evaluated using the Mean Absolute Error (MAE).

5. RESULTS AND DISCUSSION

Figure 3 depicts the Pearson correlations for the user engagement indicators, and we see that the number of Vp/d, Lp/d, Dp/d, and Cp/d are highly correlated. The

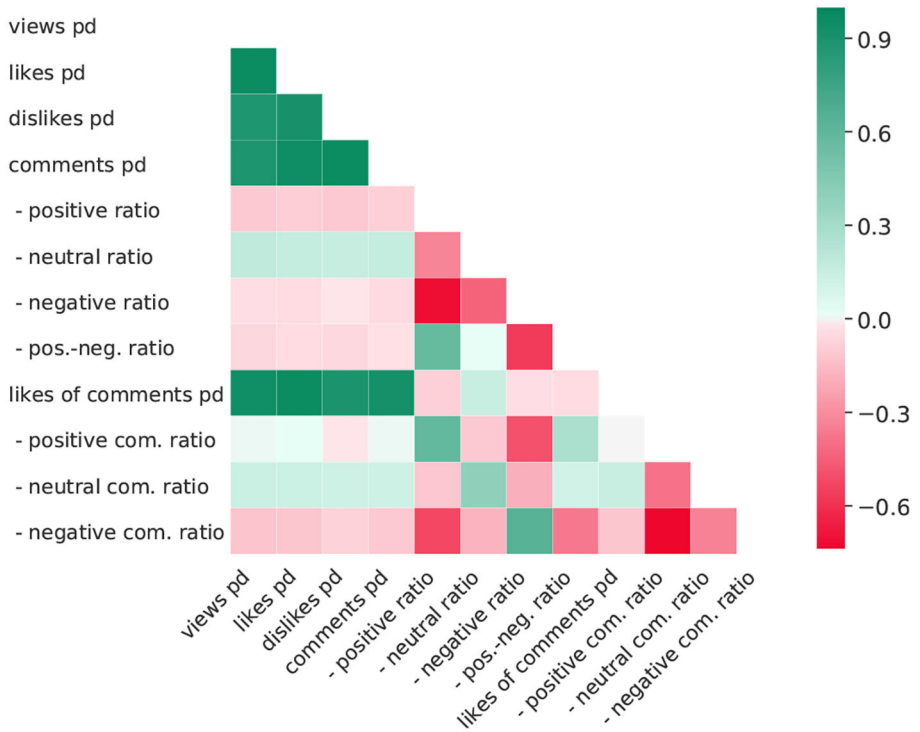


FIGURE 3 | Pearson correlation matrix of metadata. All results are considered to be significant at a 0.01 level. Com, comment; pos, positive; neg, negative; pd, per day.

correlations are based on both, absolute values and ratios. When a correlation to one of the variables occurs, it is likely to be accompanied by correlations to others. We would like to note that all absolute values correlate positively with each other, as all metrics have a positive correlation to the absolute popularity of the video. Therefore, a stronger distribution of the video also increases the absolute number of likes, dislikes, comments, etc., albeit to different magnitudes. For example, the average relationship between likes and dislikes in our crawled videos is not as antagonistic as one might expect, which means that as the number of Lp/d increases; so too does the Dp/d. Another example are the number of likes of the comments which increases with the number of dislikes of the video since the number of comments and of dislikes are interdependent. This may relate to the topic of the dataset, being that it is review videos, and the like or dislike may be more objective than other video themes. The correlation in terms of ratios gives a more definite picture in this context.

5.1. Relationship Between Features and User Engagement

Within this section, we discuss the correlation results for each emotional dimension separately. We report Pearson correlation coefficients, as depicted in **Figure 4**. Detailed results (r and significance level) can be found in **Figure 4**.

Arousal: The statistics extracted from the arousal signal indicate several correlations to the engagement data. When the *standard deviation* or the level of the $quantile_{0.95}$ increases, the number of Vp/d, Lp/d, Cp/d, and CLp/d slightly decreases (e.g., $r_{(views,std)} = -0.293, r_{(views,q_{95})} = -0.212$) with direct effect on the comment-like ratio (clr), e.g., $r_{std} = -0.271$. In contrast, the level of the $quantile_{.05}$ has the opposite effect on all these metrics (e.g., $r_{(views,q_{0.05})} = 0.231, r_{(clr,q_{0.05})} = -0.248$. Of the more complex time-series statistics, the *peaks* as well as the *CBM* have the strongest correlations across most indicators. These indicates a moderate positive linear relationship, for instance, to Vp/d and Lp/d $r_{(views,peaks)} = 0.440, r_{(likes,CBMe)} = 0.456$as well as Cp/d $r_{peaks} = 0.409$. Further, when these features increase, the share of neutral comments increases much less than the share of positive and negative comments. The next strongest correlated features, *CrM*, *aSoc*, *abE*, and *PreDa*, also represent upward correlation slopes to the user-engagement criteria. Although these features reflect the general change in engagement, no conclusions can be drawn regarding sentiment of the engagement, as there is no significant correlation of any feature to the ratios (e.g., like-dislike, and positive-negative comments).

Valence: Most statistics of the signal distribution are below $r = 0.2$, suggesting that there are only very weak linear dependencies with the engagement indicators. The only exceptions is the positive-negative ratio for the comments ($r = -0.276$) – a lower *standard deviation* leads to an increase in the

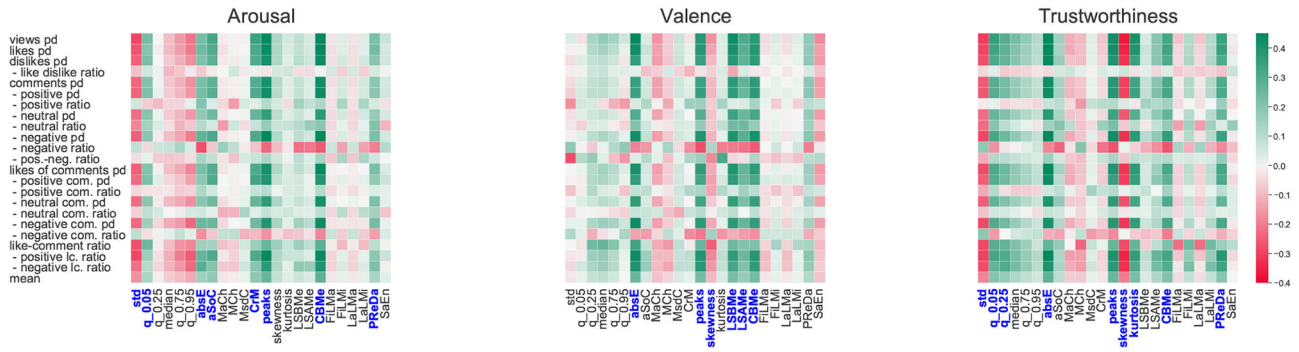


FIGURE 4 | Pearson correlation matrix of user engagement indicators and the statistics/ features extracted from each dimension. The latter are standard deviation (*std*), quantile (q_x), absolute energy (*absE*), mean absolute change (*MACh*), mean change (*MCh*), mean central approximation of the second derivatives (*MSDC*), crossings of a point *m* (*CrM*), peaks, skewness, kurtosis, strike above the mean (*LSAMe*), strike below the mean (*LSBMe*), count below mean (*CBMe*), absolute sum of changes (*ASOC*), first and last location of the minimum and maximum (*FLMi*, *LLMi*, *FLMa*, *FLMa*), perc. of reoccurring datapoints (*PreDa*), and sample entropy (*SaEn*). Features in blue are utilized as cross-task, semi-automatic features for user engagement prediction.

proportion of positive comments. Furthermore, higher values around the centre of the distribution (kurtosis – $r = -0.313$) to more likes per comment. The strongest positively correlated feature is *absE* e.g., $r_{views} = 0.467, r_{likes} = 0.422, r_{dislikes} = 0.355, r_{comments} = 0.350$, followed by the *peaks*, *CBME* and *LSBMe*, which suggest the greater the value of these features, the greater the user engagement In contrast, the *MaCh* and the *SaEn* have significant slight negative correlations, which implies that when the valence signal of a video has a high complexity, the video has a higher tendency to receive fewer user engagement.

Trustworthiness: The higher the level of $quantile_{0.05}$, $quantile_{0.25}$, *median*, and $quantile_{0.75}$ (all slightly positively correlated, with decreasing relevance e.g., $r_{(views,q_{0.05})} = 0.356, r_{(likes,q_{0.75})} = 0.175$), the higher the Vp/d, Lp/d, Dp/d, Cp/d, and CLp/d. Similar to the valence dimension, we see that there is a negative effect on these engagement indicators when the standard deviation in the trustworthiness signal is higher e.g., $r_{(views,std)} = -0.304, r_{(likes,std)} = -0.287$. As for the other features, the *absE* and the number of *peaks* have a moderate positive correlation. The *skewness* shows a significant negative correlation above $r < -0.3$ for most indicators. In other words, a negative *skew* of the trustworthiness signal, when the mass of the distribution is concentrated to the right (left-skewed), has a positive influence on user engagement. Regarding the positiveness/negativeness sentiment ratios (like-dislike, comments positive-negative ratio), none of the features show significant associations.

Result Discussion: When observing the results from the above sections, we see several patterns between the emotion (including trust) signal statistics and user engagement. While the standard statistics of arousal show that bounded arousal (higher lower quantiles and lower high quantiles) and higher trustworthiness scores (all quantiles are positively correlated, with lower quantiles at a higher level) leads to more user engagement, the sentiment of a video seems less influential contrary to the findings of (Sagha et al., 2017). Regarding the time-series features, the number of peaks with support $n = 10$ seems a stable indicator across all signals. The energy-related features of valence and trustworthiness (*valence* $= r_{(views,absE)} = 0.467$, *trustworthiness* $= r_{(views,absE)} = 0.497$) seem to have a medium-strong relationship and most likely a valuable predictive feature.

Regarding the comments, independently of the type of signal and statistic, the negative comments seem to be higher correlated consistently, followed by the number of likes and positive comments. Overall, mostly slight to modest correlations are found. However, significant correlations, especially to the more complex time-series features, between valence, arousal, and trustworthiness levels in a video to the user engagement (number of users who watch it, like it, dislike it, or leave a comment) is evident.

5.2. Predicting User Engagement From Features of Emotion and Trustworthiness Signals

Table 3 shows the results of the prediction tasks Vp/d, Lp/d, Cp/d, and CLp/D. It is worth noting that the scores vary according to the underlying scale of the target variables (cf. Section 3).

The features utilized from the cross-task semi-automatic feature selection method are highlighted (in blue) in **Figure 4** for each feature type. Across the seven experiments the automatic selection process selected on average the following number of features per each criteria; 7.6 Vp/d, 23.3 Cp/d, 29.3 Lp/d, and 20.1 LCp/d. For each dimension, an average of 9.3 for arousal, 9.5 for valence, and 6.0 for trustworthiness was selected. **Figure 5** illustrates an example of both selection methods for predicting CLp/d from a fusion of all three feature types. The *p*-values of the automatic (univariate) selection and the corresponding weights of all resulting SVMs are shown, indicating the relevance of each feature for the prediction. The interested reader is pointed to Chang and Lin (2008) for an in-depth methodical explanation. The most informative features (largest *p*-values) also receive

TABLE 3 | Prediction of views, likes, comments, and likes of comments aggregated per day utilizing features extracted and crafted from Arousal (A), Valence (V), and Trustworthiness (T).

Type	Views						Likes						Comments						Likes of Comments					
	dev			test			dev			test			dev			test			dev			test		
	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%	all MAE	sel. rel.%	auto. rel.%
A	231.8	+6.8	+5.0	220.3	+9.9	+3.1	2.30	-0.3	+2.6	1.55	+5.9	+3.0	0.288	-0.1	+3.7	0.154	+2.5	+0.6	1.19	+5.7	+5.9	0.50	-19.1	-22.7
V	253.1	+8.7	+7.2	223.8	+17.4	+24.3	2.29	+0.6	+1.0	1.61	+17.6	+24.0	0.288	+3.1	+3.9	0.154	+5.1	+2.4	1.17	-1.4	+3.6	0.51	-2.8	-18.4
T	237.4	+11.8	+16.3	207.9	-5.2	-9.7	2.21	+5.3	+14.4	1.92	+13.3	+3.6	0.262	+5.8	+6.4	0.225	+2.1	-5.3	1.11	-0.1	+9.5	0.75	+8.8	+6.7
A+V	237.6	-1.0	+2.1	210.7	+4.1	+18.3	2.27	-11.4	+0.3	1.79	+24.2	-0.7	0.277	-4.3	+3.4	0.161	+9.9	+0.1	1.16	+0.1	+2.0	0.54	+16.8	-27.3
A+T	240.3	+9.2	+15.7	207.9	-6.7	-3.9	2.26	+4.8	+10.6	2.02	+11.8	+10.3	0.268	+1.6	+7.2	0.182	-34.9	-1.1	1.11	-0.2	+3.7	0.59	-17.9	-14.7
V+T	249.1	+15.5	+20.0	205.8	-3.1	-2.6	2.07	-11.8	-0.2	1.99	+17.2	+0.1	0.262	-2.7	+5.5	0.188	+10.9	-24.7	1.04	-6.2	+0.3	0.78	+11.4	-0.1
A+V+T	228.9	-1.2	+8.7	205.9	-8.4	+0.2	2.06	-12.6	+0.6	2.08	-22.9	+0.3	0.264	-0.0	+2.7	0.192	-7.5	+0.5	1.10	+0.9	+4.3	0.60	-16.6	+8.4

We report C: parameter of the SVR, optimized for from 0.00001 to 1, using the best M: mean absolute error on the dev(elopement) set to define C for test set prediction. (%) Indicates the relative change of the automatic (auto.) and semi-automatically selected (sel.) in % to the unchanged features, "+" indicates an improvement, thus a decrease of the MAE compared to the original feature sets.

most weight from the corresponding SVM, indicating that the automatic selection is sensible. In this particular case, the hand selected features have almost identical weights as the automatic ones, whereby the missing features are enough to make the results worse than in the case of the other two (cf. **Table 3**), indicating a high sensitivity if certain features are left out.

Views Per day: When observing the Vp/d prediction from all features, we obtain the best result when performing an early fusion of the valence and trustworthiness signals, and with the addition of arousal, there is a minor decrease (205.8 and 205.8 MAE respectively); this demonstrates the predictive potential of all signals. However, when applying our semi-automatic cross-task feature selection, there is a more substantial improvement particularly for arousal and valence as mono signals, obtaining 198.5, and 184.8 MAE, respectively. This improvement is increased further for valence through automatic feature selection, with our best results for Vp/d of 169.5 MAE. Feature selection appears in all cases to not be beneficial for fused features, with arousal and valence improving slightly but no more than if the signal was alone. Without any feature selection trustworthiness is our strongest signal, for further investigation exploring why trustworthiness does not improve at all with either of the feature selection methods (218.7 and 228.0, for sel. and auto., respectively) would be of interest.

Likes per day: As with Vp/d, we see that arousal and valence are strong as singular signals when utilizing all summary features; however, in this case, there is no improvement found through the fusion of multiple feature types. Further to this, the cross-task selection method appears to improve results across all types, aside from the fusion of arousal, valence, and trustworthiness. As with Vp/d, valence again obtains our best result, improved even further by the automatic selection, up to 1.23 MAE Vp/d. Although the automatic selection appears valid for valence, this was not consistent across all the feature type variations. Trustworthiness appears much weaker than all other features types in this case, although when observing scores on the development set; we see that trustworthiness is our strongest singular signal (2.21), even showing promise when fused with the other feature types and from the automatic feature selection.

Comments per day: Results obtained from Cp/d continue to show the trend of valence being a meaningful signal. Again for all features, as singular signal both arousal and valence show the best score (0.154 MAE for both). Valence improves by the auto-selection process, and performs better with the cross-task method. Fusion in this case generally does not show much benefit assigned from the combination of arousal and valence, in which our best Cp/d score is obtained from cross-task selection of 0.145 MAE. As previously, trustworthiness is again not the strongest signal on test, however, we see a similar strength on development set.

Likes of comments per day: Arousal achieves the strongest result from all features for CLp/d. Unlike the other user engagement criteria, we see a large decrease across most results from the both selection methods. The best improvement comes from the fusion of arousal and valence with the task specific selection method. However, from automatic

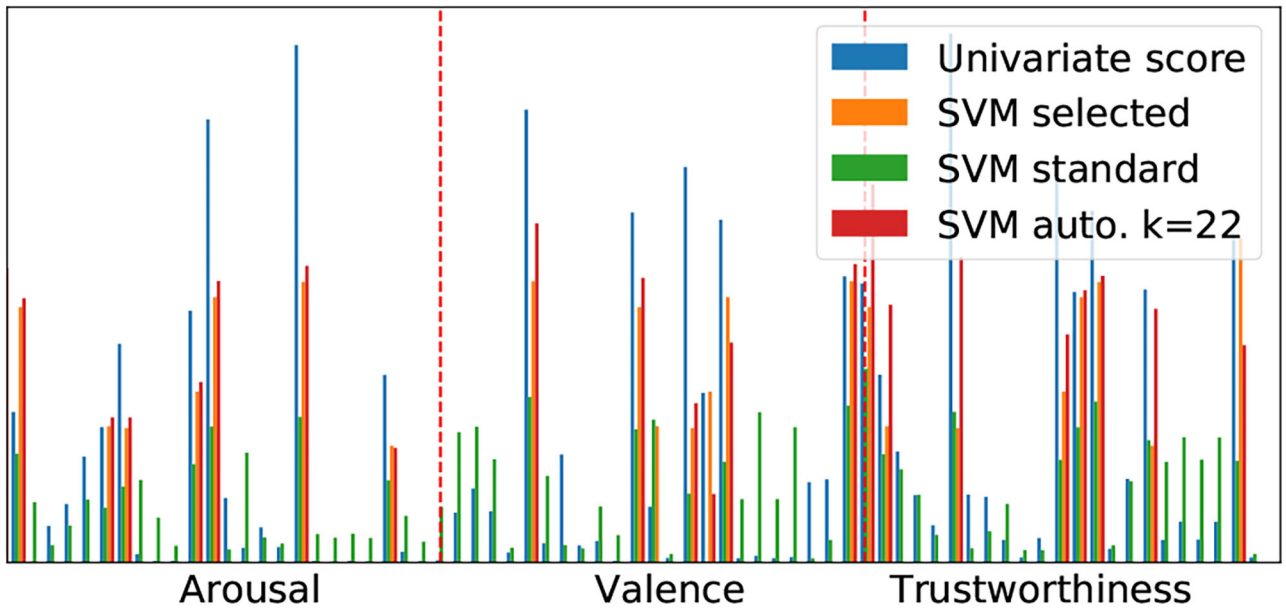


FIGURE 5 | SVM weights of arousal, valence, and trustworthiness features predicting the likes of comments comparing the use of all features, manually selected (24), and automatically selected $k = 22$ features. To receive a real output and fit the p-values of the automatic selection in scale, we apply a base 10 logarithm and divide the result by 10 ($-Log(p_{value})/10$).

selection, there is a large decrease. As in other criteria, trustworthiness again performs better than other signals on development, and poorly on test, although the cross-task selection does show improvement for trustworthiness on test, but the absolute value still does not beat that of the arousal and valence.

Result Discussion: When evaluating all results across each user-engagement criteria, it appears that our cross-task feature selection approach obtains the best results more consistently than either automatic selection or all features indicating that a more general selection stabilizes generalization. Through these feature selection approaches valence appears to be a more meaningful signal for most criteria, which can be expected given the positive:negative relationship that is inherent to all the criteria. Furthermore, without any selection, arousal is clearly a strong signal for prediction: with fusion of arousal and valence for Vp/d there is also an improvement. To this end, fusion in general does in no case obtain sustainable better results. With this in mind, further fusion strategies incorporating multiple modes at various stages in the network may be beneficial for further study.

Trustworthiness is consistently behind arousal and valence for all criteria. A somewhat unexpected result, although this may be caused by generalizability issues on the testing set, further shown by the strong results during development. Interestingly, as a single signal trustworthiness performs better than arousal and valence without feature selection for Vp/d. This result is promising, as it shows a tendency that trust is generally valuable for viewership, a finding which is supported by the literature in regard to building a parasocial relationship (Lim et al., 2020).

5.3. General Discussion

When observing the literature concerning user engagement and the potential advantage of performing this automatically—we see that one essential aspect is the ability for a content creator to develop the parasocial relationship with their viewers (Chapple and Cownie, 2017). In this regard, we see that the features from each emotional dimension (arousal, valence and trustworthiness) can predict core user engagement criteria. Most notably, as we mention previously short-term **fluctuations in arousal** appear to increase user engagement, and therefore it could be assumed such emotional understanding of video content will lead to higher user-engagement (i.e., an improved **parasocial relationship**).

Furthermore, the YouTube algorithm itself is known to bias content which has higher user engagement criterion, e.g., comments and likes per day. With this in mind, integration of the emotional features identified herein (which could be utilized for predicting forthcoming user engagement, cf. Section 6) may result in higher user engagement in other areas, e.g., views per days, resulting in better financial outcomes for the creator. The correlations between these aspects, i.e., the increase of comments per day, vs views per day should be further researched concerning these emotional dimensions.

We had expected trustworthiness to be useful for predicting user-engagement, given the aforementioned parasocial relationship theory. The results are promising for the prediction of trustworthiness. However, this does not appear to be as successful as the more conventional arousal valence emotional dimensions. The current study implements an arguably conventional method for prediction task and is limited by the data domain. Applying the trustworthiness dimension to other datasets of different domains (perhaps more popular topics, such

as comedy or infotainment) where similar metadata is available may show to be more fruitful for exploring the link of trust and improved user engagement.

6. LIMITATIONS AND FUTURE WORK

In this section, we would like to point out some aspects of our work that need further exploration, given the novelty of the proposed idea to use continuous emotion signals for modeling explicit user engagement.

As with MuSe-CaR, some previously collected datasets harvested YouTube as their primary source (Wöllmer et al., 2013; Zadeh et al., 2018). However, they either do not provide continuous emotion signals or the video **metadata** (e.g., unique video identifiers) of these datasets. Therefore, MuSe-CaR is currently the only dataset that allows studies similar to this, limits extensive exploration in other domains. We want to encourage future dataset creators using social media to provide such identifiers.

When choosing the **prediction method**, we had to make the difficult choice between interpretability and accuracy. For this study, we opted to use SVMs because we believe that initially, conceivable interactions matter more than a highly optimized outcome. This way, we can reason about relationships between influencing variables and the output predictions and compare them to ones, extracted from potential other datasets in the future. We are fully aware that state-of-the-art black-box methods, e.g., deep learning, may achieve better results but lack in clarity around inner workings and may rely on spurious and non-causal correlations that are less generalisable. However, this does not mean there are no other high non-linearity interactions between inputs, which we want to explore in future work.

Another point for future exploration is the **emotional spectrum**. Although MuSe-CaR provides arousal and valence, which are the most consistently used dimensions in previous research, also other third focus dimensions, for example, dominance (Grimm et al., 2008) and likeability (Kossaifi et al., 2019) have previously been annotated. Another interesting aspect might be categorical ratings which summarize an entire video. However, we expect much lower predictive value because of the highly compressed representation of such categories summarizing the emotional content (one value instead of several dynamically extracted features based on a video-length signal).

So far, no link existed between the use of emotional signals and user engagement. That is why, the aim of our paper was to provide a proof of concept that it is valuable to leverage such signals. However, utilizing human annotations can only be the first step since they are very limited in **scalability**. The annotations are usually the prediction target for developing robust emotion recognition models. Our final process is intended to be twofold: (i) using audio-visual features to learn to predict the human emotional signals (ii) using the predicted emotional signals on unseen, unlabeled videos to extract our feature set and predict user-engagement. (i) is very well researched in the field achieving CCCs of more than 0.7 (high correlation between predicted and human emotional annotations) on similar data sets (Huang et al., 2020). Recent advances aim at understanding contextual factors affecting multi-modal emotion recognition, such as the gender of the speaker and the duration of the emotional episode (Bhattacharya et al., 2021) and the use of non-intrusive on-body electromyography (EMG) sensors as additional input signals (Tamulis et al., 2021). For a broad overview of various (mutlimodal) emotion recognition research, we refer the interested reader to the surveys by Soleymani et al. (2017) and Tian et al. (2022). By using human annotations, we aimed to demonstrate the relationship in a vanilla way (using the targets) to avoid wrong conclusions based on any introduced prediction error bias. We also plan to explore (ii) in-depth in the near future. Another exciting research direction is to incorporate the uncertainty of multi-modal emotion recognition systems (Han et al., 2017), hence, how sure is the system in its prediction based on the availability of (or missing) audio, video, and text data, into the prediction of popularity. Thus, in parallel to the emotion, a measure of uncertainty could be given, which is then factored in the popularity prediction.

Through a bridge of emotion recognition and user engagement, we see novel **applications**. The link between emotional and user engagement provides information about what and when (e.g., a part of a video with many arousal peaks) exactly causes a user to feel e.g., aversion, interest or frustration (Picard, 1999). Two parties may particularly benefit from these findings: (a) Social media network providers: the relationships discovered are directly related to the user retention (e.g., user churn rate) (Lebreton and Yamagishi, 2020) and activity (e.g., recommender systems) (Zhou et al., 2016). These are the most common and important tasks of these platforms and are still extremely difficult to model to this day (Lin et al., 2018; Yang et al., 2018; Liu et al., 2019). Maybe more importantly, critical, emotionally charged videos (e.g., misinformation, fake messages, hate speech) can be recognized and recommendation systems adapted accordingly. (b) Content creators (marketing, advertising): companies act as (video) creators to interact with customers. In our work, we focused to show a connection between generalizable emotional characteristics and user engagement. However, we believe that there are various weaker/stronger influenced subgroups. A company can identify and target such groups or even explicitly fine-tune their content.

7. CONCLUSION

For the first time, we have empirically (and on a large-scale) presented in this contribution that there are both, intuitive and complex relationships between user engagement indicators and continuously annotated emotion, as well as trustworthiness signals in user-generated data. Of prominence, our contribution finds that emotion increases engagement when arousal is consistently bounded. In other words, the more consistent the portrayed arousal throughout a video, the better the engagement with it. This finding contradicted previous emotion literature (Sagha et al., 2017). Arousal shows consistently more robust prediction results, although valence innately (given the link of positive and negative) appears to be more valuable for prediction of video likes.

Further to this, we introduce trustworthiness as a continuous "emotion" dimension for engagement, and find when utilizing this for prediction, there is an overall value for monitoring user-engagement in social-media content. However, when fusing the signals, their appears to be little benefit from the current recognition paradigm. Furthermore, we assume that too strict feature selection causes generalization issues since often promising results on the development set seem non-transferable to the test set.

From the strong correlation of the results for trustworthiness, we consider that the addition of this dimension is of use for user engagement; however, further investigation in other domains would be valuable. When applying these metrics in a cross-modal sentiment paradigm, there may also be benefits for the prediction of audio-visual hate speech likelihood, as well as fake news.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://doi.org/10.5281/zenodo.4651164.

AUTHOR CONTRIBUTIONS

LS: literature analysis, data acquisition, data preparation, experimental design, computational analysis, and manuscript drafting and preparation. ABa: data acquisition, experimental design, and manuscript drafting and preparation. ML and ABä: data acquisition, data preparation, and computational analysis. BS: technical guidance and manuscript editing. All authors revised, developed, read, and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2022.773154/full#supplementary-material

REFERENCES

(2015). *Twitter Us Airline Sentiment.* Available online at: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Adhikari, R., and Agrawal, R. K. (2013). *An Introductory Study on Time Series Modeling and Forecasting*. LAP LAMBERT Academic Publishing.

Ahlgren, P., Jarneving, B., and Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *J. Am. Soc. Inf. Sci. Technol.* 54, 550–560. doi: 10.1002/asi.10242

Baird, A., and Schuller, B. (2020). Considerations for a more ethical approach to data in ai: on data representation and infrastructure. *Front. Big Data* 3, 25. doi: 10.3389/fdata.2020.00025

Battaglio, S. (2016). *Youtube Now Bigger Than TV Among Advertisers' Target Audience.* Available online at: https://www.latimes.com/entertainment/envelope/cotown/la-et-ct-you-tube-ad-spending-20160506-snap-story.html (October 15, 2020).

Berger, J., and Milkman, K. L. (2012). What makes online content viral? *J. Market. Res.* 49, 192–205. doi: 10.1509/jmr.10.0353

Bhattacharya, P., Gupta, R. K., and Yang, Y. (2021). Exploring the contextual factors affecting multimodal emotion recognition in videos. *IEEE Trans. Affect. Comput.*.

Bhuiyan, H., Ara, J., Bardhan, R., and Islam, M. R. (2017). "Retrieving youtube video by sentiment analysis on user comment," in *2017 IEEE International Conference on Signal and Image Processing Applications* (Kuching: IEEE), 474–478.

Biel, J., and Gatica-Perez, D. (2013). The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Trans. Multimedia* 15, 41–55. doi: 10.1109/TMM.2012.2225032

Boiy, E., Hens, P., Deschacht, K., and Moens, M.-F. (2007). "Automatic sentiment analysis in on-line text," in *Proceedings of the 11th International Conference on Electronic Publishing* (Vienna), 349–360.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335. doi: 10.1007/s10579-008-9076-6

Ceron, A., Curini, L., Iacus, S. M., and Porro, G. (2014). Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New Media Soc.* 16, 340–358.

Chang, Y.-W., and Lin, C.-J. (2008). "Feature ranking using linear svm," in *Causation and Prediction Challenge* (PMLR), 53–64.

Chapple, C., and Cownie, F. (2017). An investigation into viewers' trust in and response towards disclosed paid-for-endorsements by youtube lifestyle vloggers. *J. Promotional Commun.* 5, 19–28.

Chen, Y.-L., Chang, C.-L., and Yeh, C.-S. (2017). Emotion classification of youtube videos. *Decis. Support Syst.* 101, 40–50. doi: 10.1016/j.dss.2017.05.014

Cheng, X., Liu, J., and Dale, C. (2013). Understanding the characteristics of internet short video sharing: a youtube-based measurement study. *IEEE Trans. Multimedia* 15, 1184–1194. doi: 10.1109/TMM.2013.2265531

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing* 307, 72–77. doi: 10.1016/j.neucom.2018.03.067

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psychol.* 92, 909. doi: 10.1037/0021-9010.92.4.909

Cooper, P. (2019). *23 YouTube Statistics That Matter To Marketers in 2020.*

Cox, J. C., Kerschbamer, R., and Neururer, D. (2016). What is trustworthiness and what drives it? *Games Econ. Behav.* 98, 197–218. doi: 10.1016/j.geb.2016.05.008

Cui, B., Li, Y., Chen, M., and Zhang, Z. (2019). "Fine-tune BERT with sparse self-attention mechanism," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (Hong Kong: ACL), 3548–3553.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (ACL), 4171–4186.

Doane, D. P., and Seward, L. E. (2011). Measuring skewness: a forgotten statistic? *J. Stat. Educ.* 19, 1–18 doi: 10.1080/10691898.2011.11889611

Dobrian, F., Sekar, V., Awan, A., Stoica, I., Joseph, D., Ganjam, A., et al. (2011). Understanding the impact of video quality on user engagement. *ACM SIGCOMM Comput. Commun. Rev.* 41, 362–373. doi: 10.1145/2043164.2018478

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emotion* 6, 169–200.

English, K., Sweetser, K. D., and Ancu, M. (2011). Youtube-ification of political talk: an examination of persuasion appeals in viral video. *Am. Behav. Sci.* 55, 733–748. doi: 10.1177/0002764211398090

Garroppo, R. G., Ahmed, M., Niccolini, S., and Dusi, M. (2018). A vocabulary for growth: topic modeling of content popularity evolution. *IEEE Trans. Multimedia* 20, 2683–2692. doi: 10.1109/TMM.2018.2811625

Geurts, P. (2001). "Pattern extraction for time series classification," in *European Conference on Principles of Data Mining and Knowledge Discovery* (Freiburg im Breisgau: Springer), 115–127.

Gilbert, C., and Hutto, E. (2014). "Vader: a parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media*, vol. 81 (Ann Arbor, MI), 82.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Rep. Stanford* 1, 2009.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). "The vera am mittag german audio-visual emotional speech database," in *2008 IEEE International Conference on Multimedia & Expo (ICME)* (Hannover: IEEE), 865–868.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *J. Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Han, J., Zhang, Z., Schmitt, M., Pantic, M., and Schuller, B. (2017). "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM International Conference on Multimedia* (New York, NY), 890–897.

Horsburgh, H. (1961). Trust and social objectives. *Ethics* 72, 28–40.

Huang, J., Tao, J., Liu, B., Lian, Z., and Niu, M. (2020). "Multimodal transformer fusion for continuous emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 3507–3511.

Khan, G. F., and Vong, S. (2014). Virality over youtube: an empirical analysis. *Internet Res.* 24, 19. doi: 10.1108/INTR-05-2013-0085

Knuutila, A., Herasimenka, A., Au, H., Bright, J., and Howard, P. N. (2020). Covid-related misinformation on youtube. *Oxford Memos: The Spread of Misinformation Videos on Social Media and the Effectiveness of Platform Policies*. (Oxford).

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., et al. (2019). Deep affect prediction in-the-wild: aff-wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vis.* 127, 1–23. doi: 10.1007/s11263-019-01158-4

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B. W., et al. (2019). Sewa db: a rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1022–1040. doi: 10.1109/TPAMI.2019.2944808

Kujur, F., and Singh, S. (2018). Emotions as predictor for consumer engagement in youtube advertisement. *J. Adv. Manag. Res.* 15, 184–197. doi: 10.1108/JAMR-05-2017-0065

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). "Albert: a lite bert for self-supervised learning of language representations," in *2020 International Conference on Learning Representations* (Addis Ababa).

Lebreton, P., and Yamagishi, K. (2020). Predicting user quitting ratio in adaptive bitrate video streaming. *IEEE Trans. Multimedia* 23, 4526–4540. doi: 10.1109/TMM.2020.3044452

Lim, J. S., Choe, M.-J., Zhang, J., and Noh, G.-Y. (2020). The role of wishful identification, emotional engagement, and parasocial relationships in repeated viewing of live-streaming games: a social cognitive theory perspective. *Comput. Hum. Behav.* 108, 106327. doi: 10.1016/j.chb.2020.106327

Lin, Z., Althoff, T., and Leskovec, J. (2018). "I'll be back: on the multiple lives of users of a mobile activity tracking application," in *Proceedings of the 2018 World Wide Web Conference (WWW)* (Geneva), 1501–1511.

Liu, Y., Shi, X., Pierce, L., and Ren, X. (2019). "Characterizing and forecasting user engagement with in-app action graph: a case study of snapchat," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)* (Anchorage, AK), 2023–2031.

Moturu, S. T., and Liu, H. (2011). Quantifying the trustworthiness of social media content. *Distrib. Parallel Databases* 29, 239–260. doi: 10.1007/s10619-010-7077-0

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). "SemEval-2013 task 2: sentiment analysis in Twitter," in *Second Joint Conference on Lexical and Computational Semantics, Proceedings of the Seventh International Workshop on Semantic Evaluation* (Atlanta, GA, ACL), 312–320.

Nikolinakou, A., and King, K. W. (2018). Viral video ads: emotional triggers and social media virality. *Psychol. Market.* 35, 715–726. doi: 10.1002/mar.21129

Palshikar, G. (2009). "Simple algorithms for peak detection in time-series," in *Proceedings of the 1st International Conference on Advanced Data Analysis, Business Analytics and Intelligence*. (Ahmedabad), vol. 122.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Picard, R. W. (1999). "Affective computing for hci," in *Human Computer Interaction* (Citeseer), 829–833.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Develop. Psychopathol.* 17, 715. doi: 10.1017/S0954579405050340

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., et al. (2016). "Modelling valence and arousal in facebook posts," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis co-located to Association for Computer Lingustics* (San Diego, CA: ACM), 9–15.

Rangaswamy, S., Ghosh, S., Jha, S., and Ramalingam, S. (2016). "Metadata extraction and classification of youtube videos using sentiment analysis," in *2016 IEEE International Carnahan Conference on Security Technology* (Orlando, FL: IEEE), 1–2.

Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circul. Physiol.* 278, 2039–2049. doi: 10.1152/ajpheart.2000.278.6.H2039

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., et al. (2017). "Avec 2017: real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (Mountain View, CA), 3–9.

Roy, S. D., Mei, T., Zeng, W., and Li, S. (2013). Towards cross-domain learning for social video popularity prediction. *IEEE Trans. Multimedia* 15, 1255–1267. doi: 10.1109/TMM.2013.2265079

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178.

Sagha, H., Schmitt, M., Povolny, F., Giefer, A., and Schuller, B. (2017). "Predicting the popularity of a talk-show based on its emotional speech content before publication," in *Proceedings 3rd International Workshop on Affective Social Multimedia Computing, Conference of the International Speech Communication Association (INTERSPEECH) Satellite Workshop* (Stockholm, ISCA).

Schuller, B., Lang, M., and Rigoll, G. (2002). "Automatic emotion recognition by the speech signal," in *Proccedings of SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics*. (Orlando).

Schuller, B. W. (2013). *Intelligent Audio Analysis*. (Lyon, Springer).

Schuller, B. W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., et al. (2020). "The interspeech 2020 computational paralinguistics challenge: elderly emotion, breathing & masks," in *Proceedings Conference of the International Speech Communication Association (INTERSPEECH)* (Shanghai).

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., and Filippova, K. (2016). Multi-lingual opinion mining on youtube. *Inf. Process. Manag.* 52, 46–60. doi: 10.1016/j.ipm.2015.03.002

Sham, P. C., and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* 15, 335–346. doi: 10.1038/nrg3706

Shehu, E., Bijmolt, T. H., and Clement, M. (2016). Effects of likeability dynamics on consumers' intention to share online video advertisements. *J. Interact. Market.* 35, 27–43. doi: 10.1016/j.intmar.2016.01.001

Siersdorfer, S., Chelaru, S., Nejdl, W., and San Pedro, J. (2010). "How useful are your comments? analyzing and predicting youtube comments and comment ratings," in *Proceedings of the 19th International Conference on World Wide Web (WWW)* (Raleigh, NC), 891–900.

Siersdorfer, S., Chelaru, S., Pedro, J. S., Altingovde, I. S., and Nejdl, W. (2014). Analyzing and mining comments and comment ratings on the social web. *ACM Trans. Web* 8, 1–39. doi: 10.1145/2628441

Singh, V. K., Piryani, R., Uddin, A., and Waila, P. (2013). "Sentiment analysis of movie reviews: a new feature-based heuristic for aspect-level sentiment classification," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing* (Kottayam: IEEE), 712–717.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image Vis. Comput.* 65, 3–14. doi: 10.1016/j.imavis.2017.08.003

Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., et al. (2020a). "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, co-located to ACM International Conference on Multimedia*. (Seattle, WA: ACM).

Stappen, L., Baird, A., Schumann, L., and Schuller, B. (2021). The multimodal sentiment analysis in car reviews (muse-car) dataset: collection, insights and improvements. *arXiv preprint* arXiv:2101.06053.

Stappen, L., Brunn, F., and Schuller, B. (2020b). Cross-lingual zero-and few-shot hate speech detection utilizing frozen transformer language models and axel. *arXiv preprint* arXiv:2004.13850.

Stappen, L., Schuller, B. W., Lefter, I., Cambria, E., and Kompatsiaris, I. (2020c). "Summary of muse 2020: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *28th ACM International Conference on Multimedia*. (Seattle, WA: ACM).

Subramanian, R., Shankar, D., Sebe, N., and Melcher, D. (2014). Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *J. Vis.* 14, 31–31. doi: 10.1167/14.3.31

Surhone, L., Timpledon, M., and Marseken, S. (2010). *Spearman's Rank Correlation Coefficient: Statistics, Non-Parametric Statistics, Raw Score, Null Hypothesis, Fisher Transformation, Statistical Hypothesis Testing, Confidence Interval, Correspondence Analysis*. Betascript Publishing.

Tamulis, Ž., Vasiljevas, M., Damaševicius, R., Maskeliunas, R., and Misra, S. (2021). "Affective computing for ehealth using low-cost remote internet of things-based emg platform," in *Intelligent Internet of Things for Healthcare and Industry*, vol. 67. (Springer).

Tan, Z., and Zhang, Y. (2019). Predicting the top-n popular videos via a cross-domain hybrid model. *IEEE Trans. Multimedia* 21, 147–156. doi: 10.1109/TMM.2018.2845688

Tian, L., Oviatt, S., Muszynski, M., Chamberlain, B., Healey, J., and Sano, A. (2022). *Applied Affective Computing*. (Morgan & Claypool).

Trzciński, T., and Rokita, P. (2017). Predicting popularity of online videos using support vector regression. *IEEE Trans. Multimedia* 19, 2561–2570. doi: 10.1109/TMM.2017.2695439

Uryupina, O., Plank, B., Severyn, A., Rotondi, A., and Moschitti, A. (2014). "SenTube: a corpus for1 sentiment analysis on YouTube social media," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (Reykjavik, ELRA), 4244–4249.

Westfall, P. H. (2014). Kurtosis as peakedness. *Am. Stat.* 68, 191–195. doi: 10.1080/00031305.2014.917055

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., et al. (2013). Youtube movie reviews: sentiment analysis in an audio-visual context. *IEEE Intell. Syst.* 28, 46–53. doi: 10.1109/MIS.2013.34

Wu, Z., and Ito, E. (2014). "Correlation analysis between user's emotional comments and popularity measures," in *2014 3rd International Conference on Advanced Applied Informatics* (Kokura: IEEE), 280–283.

Yan, M., Sang, J., Xu, C., and Hossain, M. S. (2015). Youtube video promotion by cross-network association:@ britney to advertise gangnam style. *IEEE Trans. Multimedia* 17, 1248–1261. doi: 10.1109/TMM.2015.2446949

Yang, C., Shi, X., Jie, L., and Han, J. (2018). "I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)* (London), 914–922.

Yang, R., Singh, S., Cao, P., Chi, E., and Fu, B. (2016). "Video watch time and comment sentiment: experiences from youtube," in *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)* (Washington, DC: IEEE), 26–28.

Yentes, J. M., Hunt, N., Schmid, K. K., Kaipust, J. P., McGrath, D., and Stergiou, N. (2013). The appropriate use of approximate entropy and sample entropy with short data sets. *Ann. Biomed. Eng.* 41, 349–365. doi: 10.1007/s10439-012-0668-3

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Melbourne, VIC), 2236–2246.

Zhou, P., Zhou, Y., Wu, D., and Jin, H. (2016). Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks. *IEEE Trans. Multimedia* 18, 1217–1229. doi: 10.1109/TMM.2016.2537216

Conflict of Interest: BS was employed by audEERING GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 March 2022
doi: 10.3389/fpsyg.2022.781448

A Study of Subliminal Emotion Classification Based on Entropy Features

Yanjing Shi[1,2], Xiangwei Zheng[1], Min Zhang[1], Xiaoyan Yan[3], Tiantian Li[4] and Xiaomei Yu[1]**

[1] School of Information Science and Engineering, Shandong Normal University, Jinan, China, [2] Network Information Center, Shandong University of Political Science and Law, Jinan, China, [3] Geriatrics Center, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China, [4] Faculty of Education, Shandong Normal University, Jinan, China

Electroencephalogram (EEG) has been widely utilized in emotion recognition. Psychologists have found that emotions can be divided into conscious emotion and unconscious emotion. In this article, we explore to classify subliminal emotions (happiness and anger) with EEG signals elicited by subliminal face stimulation, that is to select appropriate features to classify subliminal emotions. First, multi-scale sample entropy (MSpEn), wavelet packet energy (E_i), and wavelet packet entropy (WpEn) of EEG signals are extracted. Then, these features are fed into the decision tree and improved random forest, respectively. The classification accuracy with E_i and WpEn is higher than MSpEn, which shows that E_i and WpEn can be used as effective features to classify subliminal emotions. We compared the classification results of different features combined with the decision tree algorithm and the improved random forest algorithm. The experimental results indicate that the improved random forest algorithm attains the best classification accuracy for subliminal emotions. Finally, subliminal emotions and physiological proof of subliminal affective priming effect are discussed.

Keywords: EEG, subliminal emotion, feature extraction, subliminal emotion classification, improved random forest

OPEN ACCESS

Edited by:
Zhen Cui,
Nanjing University of Science and Technology, China

Reviewed by:
Serap Aydin,
Hacettepe University, Turkey
Tong Zhang,
Nanjing University of Science and Technology, China
Chuangao Tang,
Southeast University, China

****Correspondence:***
Xiaoyan Yan
sharon.yan@163.com
Xiaomei Yu
yxm0708@126.com

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *22 September 2021*
Accepted: *24 February 2022*
Published: *25 March 2022*

Citation:
Shi Y, Zheng X, Zhang M, Yan X, Li T and Yu X (2022) A Study of Subliminal Emotion Classification Based on Entropy Features. Front. Psychol. 13:781448. doi: 10.3389/fpsyg.2022.781448

1. INTRODUCTION

Affective computing is a multidisciplinary field involving computer science, psychology, and cognitive science and its potential applications include disease diagnosis, human-computer interaction (HCI), entertainment, autonomous driving assistance, marketing, teaching, etc., (Bota et al., 2019). The intelligent brain-computer interface (BCI) systems based on electroencephalogram (EEG) can promote the continuous monitoring of fluctuations in the human brain area under the emotional stimulation, which is of great significance for the development of brain emotional mechanisms and artificial intelligence for medical diagnosis (Gu et al., 2021).

Emotion research also has important findings for neurological-marketing that local neuronal complexity is mostly sensitive to the affective valance rating, while regional neuro-cortical connectivity levels are mostly sensitive to the affective arousal ratings (Aydın, 2019). There is an attention bias processing mechanism for emotions. Some studies have shown that angry faces can automatically stimulate attention, that is, there is an anger dominance effect. On the contrary, some studies have shown the existence of a happiness dominance effect (Xu et al., 2019). Most psychologists regard subjective experiences as the central component of emotion, emphasizing the role of consciousness in emotional production and emotional state. However, the discussion of emotional issues from the perspective of unconscious emotion also has a profound tradition in

psychological research. Unconscious emotion, also named subliminal emotion, refers to the change of thoughts and emotions caused by certain emotional states (Li and Lv, 2014). This emotional state is independent of his conscious awareness, and the induction of this emotional state is unconscious (Jiang and Zhou, 2004; Wataru et al., 2014; Zheng et al., 2021a). The presentation of stimuli subliminally is an important research topic in the field of unconscious perception.

Researchers use subliminal stimulus to trigger unconscious states to analyze changes in mood, cognition, social information processing, and physiological signals. Emotional faces are an important and unique visual stimulus and humans are very sensitive to emotional faces and have complex and efficient recognition ability of them. Subliminal emotional face experiments use emotional faces as stimulus materials for subliminal presentation, and they will trigger unconscious emotion (Yin et al., 2021).

We have con'ducted a study on multi-scale sample entropy (MSpEn) (Shi et al., 2018) and in this article, we study new features which are also suitable for the subliminal emotion classification based on EEG signals. These features including MSpEn, WpEn, and E_i extracted from EEG signals have been employed as input feature vectors for classification of subliminal emotions. We combine them with decision tree algorithm and improved random forest to classify subliminal emotions.

This article is structured as follows: Section 2 introduces related work. Section 3 presents the experimental process, subjects, the feature extraction method, and classification algorithms. Section 4 describes the experiments and results. Finally, section 5 concludes this article.

2. RELATED WORK

The methods based on physiological signals are more effective and reliable because humans can not control them intentionally, such as electroencephalogram, electromyogram (EMG), electrocardiogram (ECG), skin resistance (SC) (Kim et al., 2004; Kim and Andr, 2008), pulse rate, and respiration signals. Among these methods, EEG-based emotion recognition has become quite common in recent years. There are many research projects focusing on EEG-based emotion recognition (Hosseini and Naghibi-Sistani, 2011; Colic et al., 2015; Bhatti et al., 2016). Jatupaiboon et al. (2013) indicated that the power spectrum from each frequency band is used as features and the accuracy rate of the SVM classifier is about 85.41%. Bajaj and Pachori (2014) proposed new features based on multiwavelet transform for the classification of human emotions from EEG signals. Duan et al. (2013) proposed a new effective EEG feature named differential entropy to represent the characteristics associated with emotional states.

Extracting effective features is the key to the subliminal emotion recognition of EEG signals. Four different features (time domain, frequency domain, time-frequency based, and non-linear) are commonly identified in the feature extraction phase. Compared to traditional time domain and frequency domain analysis, time-frequency based, and non-linear are more widely used (Vijith et al., 2017). Wavelet packet transform is a typical linear time-frequency analysis method. Wavelet packet decomposition is a wavelet transform that provides a time-frequency decomposition of multi-level signals. Murugappan et al. (2008) used video stimuli to trigger emotional responses and extract wavelet coefficients to obtain the energy of the frequency band as input features. Verma and Tiwary (2014) used discrete wavelet transform for feature extraction and classified emotions with support vector machine (SVM), multilayer perceptron (MLP), K nearest neighbor, and metamulticlass (MMC). In recent years, many scholars have tried to analyze EEG signals by non-linear dynamics methods. Commonly used methods are correlation dimension, Lyapunov exponent, Hurst exponent, and other entropy-based analysis methods (Sen et al., 2014). Hosseini and Naghibi-Sistani (2011) proposed an emotion recognition system using EEG signals, and a new approach to emotion state analysis by approximate entropy (ApEn) and wavelet entropy (WE) is integrated. Xin et al. (2015) proposed an improved multi-scale entropy algorithm for emotion EEG features extraction. Michalopoulos and Bourbakis (2017) applied multi-scale entropy (MSE) to EEG recordings of subjects who were watching musical videos selected to elicit specific emotions and found that MSE is able to discover significant differences in the temporal organization of the EEG during events that elicit emotions with low/high valence and arousal.

The upsurge in the study of emotion research attracts scholars to explore and discover subliminal emotions. The analysis and processing of EEG signals have become an indispensable research focus in emotion recognition.

3. EEG DATA ACQUISITION AND ANALYSIS METHODS

The process of subliminal emotion classification consists of several steps as shown in **Figure 1**. First, a stimulus such as picture, audio, or movie is needed. During the experiments, the participant is exposed to the stimuli to elicit emotion, and EEG signals are recorded accordingly. In order to trigger subliminal emotion, we set the presentation time as 33 ms. Then, artifacts that contaminate EEG signals are removed. EEG signals are analyzed and relevant features are extracted. Some data are used to train the classification model, and the remainder is used for the test which is classified using this model to compute accuracy (Zheng et al., 2021b). Age and gender specifications of the subjects would be given in the **Supplementary Material**.

3.1. Method

3.1.1. Feature Extraction

This article mainly adopts three features, including MSpEn, wavelet packet energy (Ei) and wavelet packet entropy (WpEn). MSpEn is a combination of sample entropy and multi-scale analysis (Klauer and Musch, 2003; Bai et al., 2007). The calculation steps of MSpEn are described in detail in Shi et al. (2018).

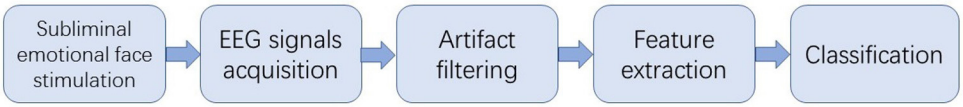


FIGURE 1 | The process of subliminal emotion classification.

3.1.1.1. Wavelet Packet Energy

Wavelet packet decomposition is a generalization of the wavelet transform, which is with multi-resolution characteristics. It can finely analyze signals more than wavelet analysis, so it is very suitable for processing non-stationary signals such as EEG signals and has been widely used in the field of EEG signal processing. Wavelet transform is a multi-scale signal analysis method. It can characterize local features of signals in both time and scale (Deng et al., 2011). The continuous wavelet transform of the signal $f(t)$ is defined as

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int f(t)\psi(\frac{t-b}{a})dt \quad (1)$$

where a is the scaling parameter, b is the translation parameter, $\psi(t)$ is the wavelet function, and t is the time.

The discrete wavelet transform is defined as

$$C_{j,k} = \int_{-\infty}^{+\infty} f(t)\psi_{j,k}(t)dt \quad (2)$$

where $\psi_{j,k}(t) = 2^{-\frac{J}{2}}\psi(2^{-j}t - k)$.

Wavelet analysis has been widely used in various fields as the main tool for time-frequency analysis. Compared to Fourier transform and short-time Fourier transform, wavelet analysis has the advantage of multi-resolution analysis, which can reflect the local details of signals on multiple scales. However, the traditional wavelet transform only further decomposes the low-frequency components of each decomposed signal. The high frequency components are ignored and the signal details are not adequately reflected. Wavelet packet decomposition is a generalization of the wavelet transform, which is with multi-resolution characteristics. In order to extract the EEG features, the original signal is decomposed by the Mallat algorithm, and the wavelet coefficients of the corresponding nodes are reconstructed to obtain the final coefficients.

To reduce noise and other factors, high frequency components are filtered out, leaving a frequency range below 256 Hz. After four layers wavelet decomposition, the original signal can be decomposed into 16 bands.

The wavelet packet energy of band i (E_i) is defined as

$$E_i = \sum_{i=1}^{N} |\left|n_j\right||^2 \quad (3)$$

where N is the number of corresponding band coefficients, n_i is the wavelet packet coefficient. The total wavelet packet energy is defined as

$$E_{total} = \sum_{i=1}^{2^i} E_i \quad (4)$$

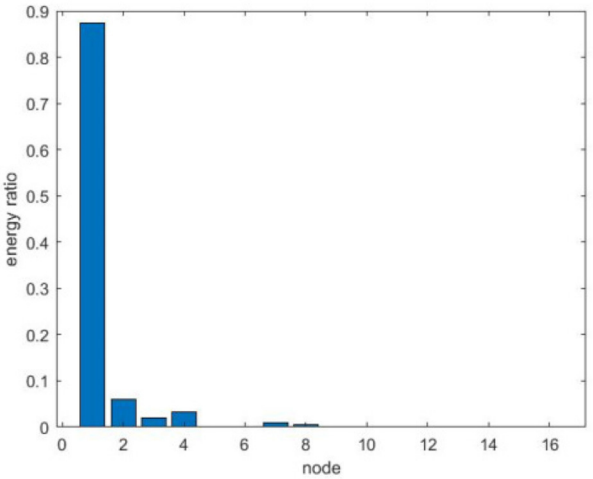


FIGURE 2 | Energy ratio of 4-layer wavelet packet decomposition.

The wavelet packet energy distribution is expressed as

$$P_i = \frac{E_i}{E_{total}} \quad (5)$$

The energy ratio of each wavelet packet node is calculated after the wavelet packet decomposition of the original EEG signals. The result is shown in **Figure 2**. The energy is concentrated in the frequency band corresponding to the first four wavelet packet nodes after the wavelet packet is decomposed. The energy ratio and corresponding frequency range of the first four wavelet packet nodes are shown in **Table 1**. More than 98% of the energy is concentrated in the wavelet packet nodes (4,0) ∼ (4,3), this is because the human EEG sub-band intervals are as follows: delta, 0.5–4 Hz; theta, 4.5–8 Hz; alpha, 8.5–16 Hz; beta, 16.5–32 Hz; gamma, 32.5–60 Hz. According to the EEG rhythm theory, it means that we only need to extract some features that can cover the human brain frequency range.

According to the above analysis, EEG activities are mainly concentrated in the (4,0) ∼ (4,3). Therefore, it is not necessary to use all frequency bands in actual analysis. In order to cover the EEG rhythm as much as possible and avoid the effects of noise and artifacts in EEG records, this article only deals with wavelet packet nodes. The wavelet packet energy of the packet node (4,0) ∼ (4,3) is extracted and analyzed whether it is a contribution to subliminal emotion face recognition.

TABLE 1 | The 4-layer wavelet packet decomposition frequency intervals and energy ratio.

Wavelet packet node	Wavelet packet energy distribution	Frequency interval (Hz)
(4,0)	87.3802%	0 ∼ 16
(4,1)	5.9086%	16 ∼ 32
(4,2)	3.1553%	32 ∼ 48
(4,3)	2.004%	48 ∼ 64

3.1.1.2. Wavelet Packet Entropy

Information entropy can provide a quantitative measure of information contained in various probability distributions. It is a measure of the degree of uncertainty and can be used to estimate the complexity of random signals. The energy distribution of wavelet packet decomposition coefficient and information entropy are combined to define WpEn as:

$$\mathrm{WpEn} = -\sum P_i \ln P_i \tag{6}$$

3.1.2. Classification Algorithm

This study employed and evaluated two classifiers, the decision tree algorithm (Yang and XU, 2017) and the improved random forest algorithm (Paul et al., 2018), for subliminal emotion classification. This study systematically compared the effects of all the feature types (MSpEn, WpEn, and E_i) on the classification performance.

3.1.2.1. Decision Tree Algorithm

Classification is one of the most widely studied and applied methods in the field of data mining. The decision tree algorithm is widely used because of its fast classification, high precision, and easy-to-understand classification rules. The popular algorithms in the decision tree algorithm are ID3, C4.5, CART, and CHAID. ID3 algorithm based on information entropy is a classic algorithm of decision tree algorithm. The possibility of attribute splitting will increase as the information gain increases. However, ID3 can only deal with discrete properties, while the C4.5 algorithm can handle both discrete and continuous properties. C4.5 algorithm is one of the most widely studied algorithms in decision tree algorithms and is also one of the representative algorithms of decision trees.

The C4.5 algorithm is an improved algorithm of the ID3 algorithm. It uses the information gain rate to select attributes and prunes during tree construction. It can process both discrete and continuous attributes, as well as default data.

The core idea of the C4.5 decision tree algorithm is to use the principle of information entropy to select the attribute with the largest information gain rate as the classification attribute, recursively construct the branches of the decision tree, and complete the construction of the decision tree.

C4.5 algorithm can be described in the following steps:

Step 1: The training data set is preprocessed. If there are continuous attributes in the data set, it needs to be discretized first.

Step 2: The data is classified according to the respective attributes of the data set, and the information gain rate is calculated for each classification result.

Set the training set as D and $|D|$ indicates the number of records of D. The label set of class D is C, $C = \{C_1, C_2, ..., C_m\}$, where $|C_i|$ represents the number of records of C. The training set can be divided into m different subsets D_i, $(1 \leq i \leq m)$ according to labels. Set the attributes set of D as A_n, where $A_n = \{A_1, A_2, ..., A_n\}$, the ith attribute of A_i with w different values is defined as $\{a_{1i}, a_{2i}, ..., a_{wi}\}$. The data set is divided into w different subsets D_i^A, $(1 \leq i \leq w)$ according to the attribute, where $|D_i^A|$ represents the number of samples in the subset D_i^A, $|C_i^A|$ represents the number of C_i in the subset D_i^A. So, the information entropy is defined as

$$Entropy(D) = -\sum_{i=1}^{m} p_i lb(p_i) \tag{7}$$

where $p_i = |C_i|/|D|$. The information entropy of subset divided according to attribute A_i is defined as

$$EntropyA(D) = -\sum_{i=1}^{w} \frac{|D_i^A|}{|D|} Entropy(D_i^A) \tag{8}$$

where $Entropy(D_i^A)$ is the information entropy of subset D_i^A. The formula $Gain(A_i)$ represents the information gain of the training set divided by the attribute A_i, which is defined as

$$Gain(A_i) = Entropy(D) - EntropyA(D) \tag{9}$$

The split information $SplitInfoA(D)$ is defined as

$$SplitInfoA(D) = -\sum_{i=1}^{w} \frac{D_i^A}{D} lb \frac{D_i^A}{D} \tag{10}$$

The information gain ratio of the dataset divided by the attribute A_i is defined as

$$GainRatio(A_i) = \frac{Gain(A_i)}{SplitInfoA(D)} \tag{11}$$

Step 3: According to the attribute which is corresponding to the maximum information gain ratio, the current data set is divided into different subsets, the decision tree branches are established, and the new child nodes are formed.

Step 4: Steps 2 and 3 are recursively called for the new node until the class labels are the same in all nodes.

3.1.2.2. Improved Random Forest Algorithm

Random forest algorithm is a typical multi-classifier algorithm. The basic classifier that constitutes the random forest algorithm is the decision tree. The basic principle of the random forest algorithm is to use the resampling technique to form a new training set by randomly extracting samples. Then, decision tree is modeled and composed of random forests, and the classification results are used for voting decisions (Bo, 2017).

The random forest algorithm is similar to the Bagging algorithm in that it is resampled based on the bootstrap method to generate multiple training sets. The difference is that the random forest algorithm uses the method of randomly selecting the split attribute set when constructing a decision tree.

Random forest algorithm can be divided into the following steps:

Step 1: Use the bootstrap method to resample and randomly generate T training sets, $S_1, S_2, , S_T$.
Step 2: Generate a corresponding decision tree using each training set; before selecting attributes on each internal node, m attributes are randomly extracted from M attributes as the split attribute set of the current node, and the node is split by the best classification among the m attributes.
Step 3: Every tree grows intact without pruning.
Step 4: For the test set sample X, use each decision tree to test and get the corresponding category.
Step 5: Using the voting method, the category with the most output in the T decision trees is taken as the category to which the test set sample X belongs.

However, the random forest algorithm also has deficiencies. This article uses a random forest algorithm based on the C4.5 tree algorithm. The attribute division strategy is based on the level of information gain rate to select the characteristics of the division. The principle of attribute division has the disadvantage of biasing features with many values. The voting on the classification result of the decision tree adopts the "majority voting principle," which means the number of votes is the final classification result, and the strength of the decision tree classifier cannot be distinguished. This article will improve the random forest algorithm and apply it to the classification of subliminal emotional faces.

Random forest algorithm is improved from the following mechanisms:

(1) In the choice of test attributes, attribute division by information gain rate will have the characteristic of biasing the features with more values, the Pearson coefficient is introduced to compensate.

The C4.5 decision tree algorithm uses the information gain ratio to select the test attribute. The larger the information gain ratio, the stronger the correlation between the attribute and the class attribute, the possibility of the attribute being selected as the test attribute the larger.

The C4.5 decision tree algorithm takes into account the influence of attributes on class, but it does not involve the influence between attributes. If an attribute has a strong correlation with other attributes, there will exist redundancy between them. Therefore, the Pearson coefficient is used to express the temporal correlation of attributes in this article, and the influence of attributes with high correlation is eliminated. The quotient of the covariance and the standard deviation is defined as the Pearson correlation coefficient, which can reflect the degree of correlation between two variables. The Pearson coefficient is defined as

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \tag{12}$$

When the Pearson coefficient is 0, it means there is no correlation between two variables. When the Pearson coefficient is positive, it indicates that there is a positive correlation between the two variables. The larger the value, the greater the positive correlation between the two variables. When the Pearson coefficient is negative, it means that there is a negative correlation between two variables. The greater the value, the greater the negative correlation between two variables. The range of Pearson's coefficient is $(-1, 1)$.

The improved attribute division algorithm is defined as

$$GainRatio(A_i) = \frac{Gain(A_i)}{SplitInfoA(D) + ar} \tag{13}$$

According to the improved attribute selection method, the attribute which has the high information gain ratio and low correlation with other attributes, has the greater probability of being selected as the test attribute.

(2) The voting decision process is an important mechanism of the random forest algorithm. Random forest algorithm adopts the principle of majority voting, assigning each decision tree the same weight, and ignoring the difference between the strong classifier and the weak classifier, which affects the overall classification performance of random forest. This article uses the weighted voting principle to improve random forest algorithm. During the formation of the random forest, according to the classification performance of each decision tree, each decision tree is assigned a corresponding weight. Then, the final classification effect is obtained by weighted voting.

In the process of generating a decision tree, the bagging method is used to extract samples from original training set with replacement to form a sample set, and the decision tree classification accuracy rate Ac_tree corresponding to the sample set can be obtained. The larger Ac_tree, the better the classification effect of the decision tree, which belongs to the strong classifier. Ac_tree of each decision tree is used as the weight of the corresponding decision tree and add the weights corresponding to decision trees with the same output class target. Finally, classification result with higher weight is the final category. **Figure 3** shows the schematic diagram of the improved random forest algorithm.

4. RESULTS

First, low pass filtering is applied to each EEG signal segment. According to the sampling theorem, the maximum frequency of the signal is about 500 Hz. To reduce noise and other factors, high frequency components are filtered out, leaving a frequency range below 256 Hz. The sample entropy has a strong ability to characterize nonlinear sequences on a macroscopic scale and

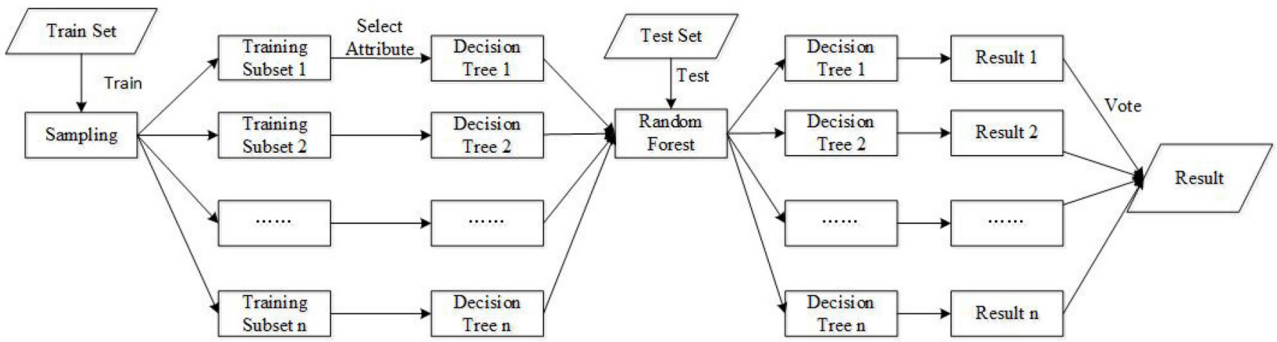


FIGURE 3 | The principle of improved random forest algorithm.

cannot describe the details. The wavelet packet decomposition has excellent description ability in detail. Therefore, MSpEn, WpEn, and $E_i$are extracted as feature vectors for the classifier, and the classification results of different features are compared. Bai et al. (2007) pointed out that when the sample entropy parameter $m = 2$ and $r = 0.2STD$ are selected through experiments, the classification effect is better. In addition, Duan et al. (2013) pointed out that the scale factor $t = 2$ is preferentially chosen. The wavelet basis function selects the db-4 wavelet. We can get 16 bands after the four-layer wavelet packet decomposition of the signals and we calculate the energy ratio of each node after wavelet packet decomposition according to the formula (5).

In our experiment, there were 80 sets of sample data for each subject, 40 groups were selected as training samples for training the proposed model, and the remaining 40 groups were used as test samples for verifying the performance of the model. WpEn and E_i extracted by wavelet packet transform and MspEn calculated by multi-scale sample analysis is put into decision tree algorithm, respectively. The averaged classification performance of the decision tree algorithm with MSpEn, WpEn, and E_i on 10 subjects is shown in **Table 2**.

Table 2 shows classification accuracy when MSpEn, E_i, and WpEn are input to the decision tree classifier. The experimental results show that decision tree algorithm can effectively classify subliminal emotional faces combined with the three feature vectors, and different feature vectors have different classification capabilities for subliminal emotional faces. The classification accuracy with E_i as a feature vector is significantly higher than other features, and its average classification accuracy is up to 94.33%. The classification accuracy using WpEn as the feature vector is slightly lower, and its average classification accuracy is 93.32%. The classification accuracy with MSpEn as the feature vector is the lowest, and its average classification accuracy is 75.52%.

In order to compare the classification performance of different features more intuitively, **Figure 4** shows the comparison of the classification accuracy of different features input to the decision tree. It can be seen from **Figure 4** that the classification accuracy using MSpEn as the feature vector is the lowest and the accuracy obtained by MSpEn fluctuates greatly in different subjects. When E_i and WpEn are adopted as the input feature vector of the decision tree classifier, the average classification accuracy is significantly higher than MSpEn and the classification accuracy is more stable.

In summary, a decision tree classifier can effectively classify subliminal emotional faces. In the perspective of feature vectors, classification effect of E_i and WpEn is better compared with MSpEn, which shows that wavelet packet decomposition features are more powerful for subliminal emotional face recognition.

This article further studies the classification effect of improved random forest algorithm on subliminal emotional faces. WpEn and E_i extracted by wavelet packet transform and MspEn calculated by multi-scale sample analysis are input into improved random forest, respectively. The averaged classification performance of the improved random forest algorithm with MSpEn, WpEn, and E_i on 10 subjects is shown in **Table 3**.

As we can see from **Table 3**, three feature extraction algorithms can identify unconscious emotions triggered by subliminal faces stimulus. The classification performance with E_i was evidently better than those based on other feature types under the same conditions, while the accuracy can reach up 96.75%. While WpEn and MSpEn are used as input feature vectors, the unconscious emotions can be classified combined with improved random forest. The highest classification accuracy can be achieved at 93.38 and 88.89%. For the classification results with E_i and the decision tree algorithms and the improved random forest algorithms, it was noted that the random forest algorithms outperformed the decision tree algorithm by $1 \sim 6\%$ for most subjects.

The comparison of classification accuracy of improved random forest classifier with different inputs feature vectors is shown in **Figure 5**. It can be seen from the figure that the average classification accuracy of a single subject and all subjects are significantly higher when using E_i and WpEn as the input feature vectors of the classifier compared to MSpEn. Due to the individual differences of the subjects, the classification accuracy of different subjects shows a small range of fluctuations, and the overall performance shows that

TABLE 2 | Average results of decision tree algorithm with multi-scale sample entropy (MSpEn), wavelet packet entropy (WpEn), and wavelet packet energy (E_i).

Method	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
MSpEn	80.25%	73.57%	72.15%	53.65%	83.52%	79.80%
E_i	**97.75%**	**85.02%**	96.77%	**96.05%**	**96.75%**	**91.87%**
WpEn	96.97%	83.27%	**97.60%**	95.30%	91.40%	88.68%
Method	**Subject 6**	**Subject 7**	**Subject 8**	**Subject 9**	**Subject 10**	**Average**
MSpEn	79.80%	75.80%	78.60%	84.50%	73.40%	75.52%
E_i	**91.87%**	89.07%	**94.82%**	**97.62%**	**97.57%**	**94.33%**
WpEn	88.68%	**91%**	94.08%	97.55%	97.39%	93.32%

TABLE 3 | Average results of random forest algorithm with MSpEn, WpEn, and E_i.

Method	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
MSpEn	87.65%	85%	86.25%	85%	91.25%	96.25%
E_i	**98.75%**	**95%**	**97.5%**	**93.75%**	**98.75%**	**97.5%**
WpEn	96.25%	86.25%	**93.75%**	90%	95%	95%
Method	**Subject 6**	**Subject 7**	**Subject 8**	**Subject 9**	**Subject 10**	**Mean**
MSpEn	96.25%	88.75%	86.25%	93.75%	88.75%	88.89%
E_i	**97.5%**	96.25%	**95%**	**97.5%**	**97.5%**	**96.75%**
sWpEn	95%	**91%**	91.25%	93.75%	96.25%	93.38%

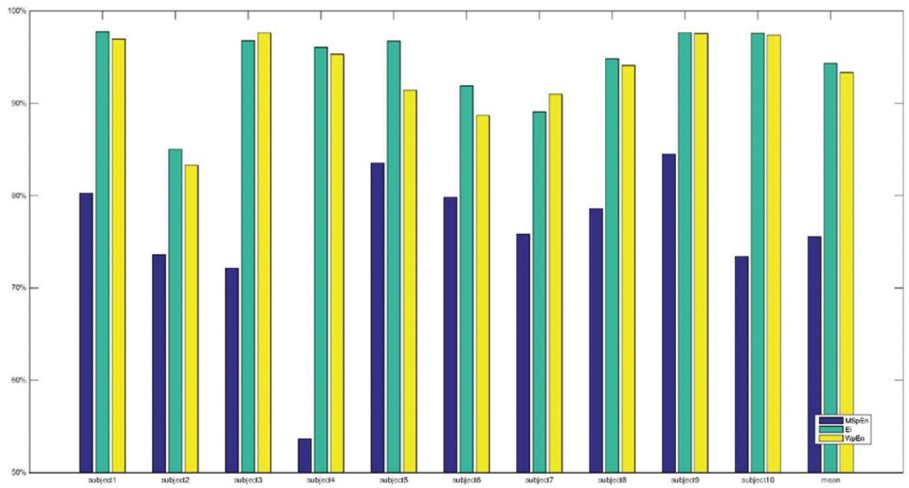


FIGURE 4 | Comparison of classification accuracy with three features and decision tree classifier.

E_i and WpEn have a stronger classification ability than the MSpEn.

This article further analyzes and compares the classification accuracy of the two classifiers under the same feature extraction method. The classification results are shown in **Figures 6–8**. **Figure 6** shows the classification results when using MSpEn as an input feature vector. The experimental results show that the improved random forest algorithm shows a stronger classification ability of 10 subjects, and the classification accuracy is significantly higher compared to the decision tree algorithm. **Figure 7** shows the classification results when using E_i as input feature vectors. It can be seen that the classification accuracy of the decision tree algorithm of only one subject is higher than that of the improved random forest algorithm. The classification accuracy of the improved random forest algorithm of the remaining 9 subjects is higher than that of the decision tree algorithm. Overall, the classification accuracy of the improved random forest algorithm is higher than that of the decision tree algorithm, and the average classification accuracy is improved by 2.42%. **Figure 8** shows the classification results when using WpEn as input feature vectors. When the WpEn is used as the classification feature, the classification accuracy fluctuation between different subjects is more obvious, and the two classification algorithms

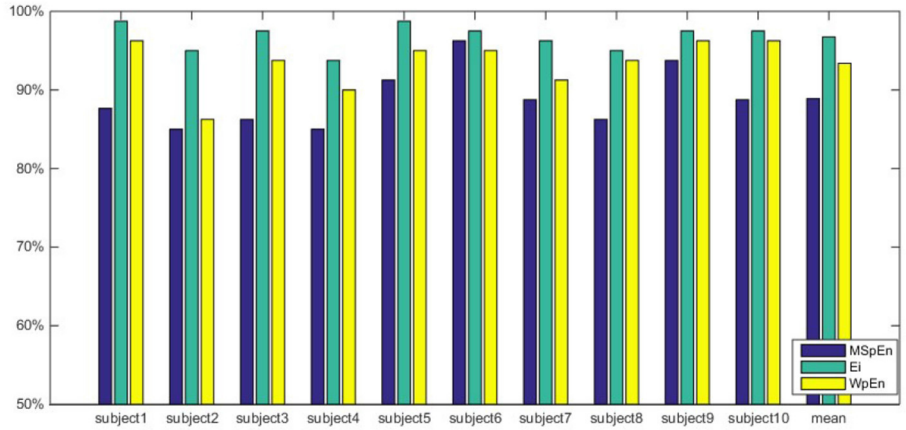


FIGURE 5 | Comparison of classification accuracy with three features and improved fandom forest.

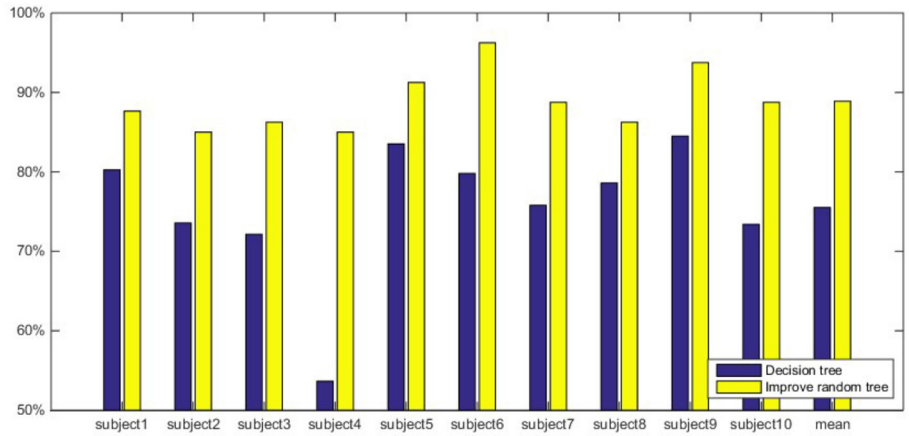


FIGURE 6 | Comparison of classification accuracy of two classifiers based on MSpEn.

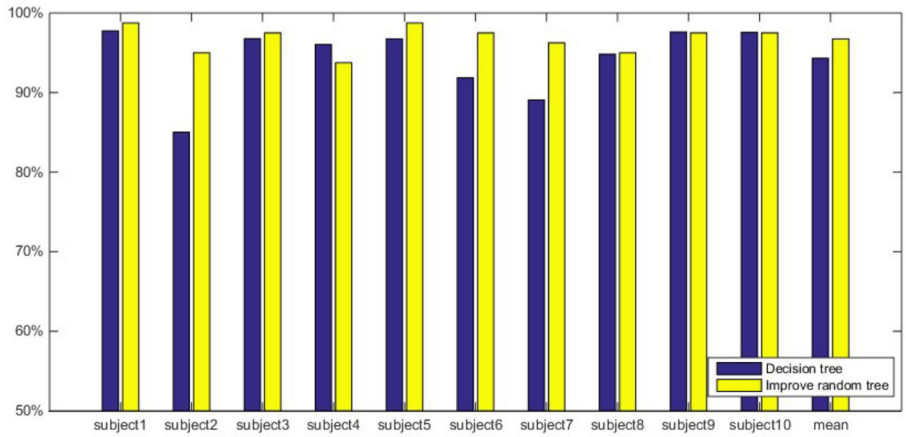


FIGURE 7 | Comparison of classification accuracy of two classifiers based on E_i.

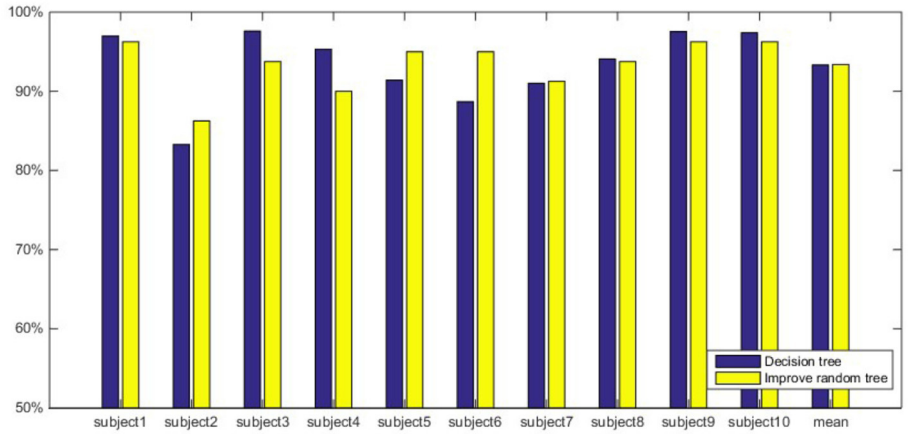


FIGURE 8 | Comparison of classification accuracy of two classifiers based on WpEn.

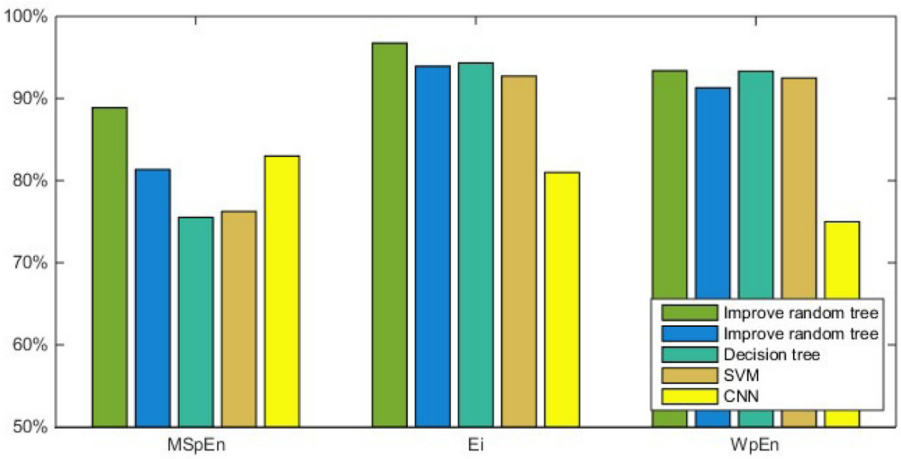


FIGURE 9 | Comparison of the average classification results with other classifiers.

show different advantages in different subjects. However, from the perspective of average classification accuracy, the classification accuracy of the improved random forest algorithm is slightly higher.

At present, there are few studies and references on subliminal unconscious emotions. In order to confirm the effectiveness of the proposed method, this article compares the classification results of several different classifier algorithms. The experimental results are shown in **Figure 9**.

In summary, combining three features with decision tree classifier and an improved random forest classifier can realize the classification of subliminal emotional faces. From the perspective of feature extraction, E_i and WpEn obtained by wavelet packet decomposition have obvious advantages for subliminal emotion face classification, and their ability to classify emotional faces is significantly stronger than MSpEn. From the perspective of the classifier, improved random forest is superior to decision tree.

5. CONCLUSION

This article studies features and classification of subliminal unconscious emotions based on EEG signals. We use the subliminal emotional faces as a starting stimulus, in fact, the subjects cannot recognize the emotional content of the face pictures. In the absence of clear emotional information, the human brain can still perform rapid, unconscious processing. We select three effective features first, and then they are combined with a decision tree algorithm and improved random forest algorithms to classify the unconscious emotions triggered by a subliminal stimulus. The experimental results show that classification accuracy of wavelet packet decomposition features

(E_i and WpEn) with two classifiers is significantly higher than MSpEn, which shows that wavelet packet decomposition can better characterize the EEG signals triggered by subliminal emotional face stimuli.From the perspective of psychology, we explore the neural mechanisms of brain activity under subliminal face stimulation (Zheng et al., 2021c). Psychological researches show that the presentation of face stimuli at a subliminal time can trigger an emotional priming effect, that is, the initiation of unconscious emotions. Researchers have conducted a lot of experimental investigations on this issue and explored the physiological proof of subliminal emotional priming effects. Some works have shown that the thalamus, hippocampus, amygdala, and their functional connections play an important role in the processing of subliminal emotional faces (Eickhoff et al., 2009). In unconscious situation, humans may have a faster way for processing of emotional faces (especially fearful faces). This way bypasses the primary visual cortex involved in conscious processing, along with the upper mound, the thalamus, and conveys to the amygdala, and then projects to other advanced cortical areas associated with emotional processing (Zhu et al., 2013). Dolan (2002) found that human emotional processing can occur when the subject is unconscious of the process. Smith (2011) work shows that multiple types of emotional faces as fearful faces can be processed at an unconscious level in an early stage.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The data were collected from college students in our university, while they are not be standardized and opened at present. Requests to access these datasets should be directed to Xiaomei Yu, yxm0708@126.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Shandong Normal University Ethics Committee. The participants provided written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XZ designed the project. YS wrote the code. MZ drafted this article. XYa analyzed the data. TL helped analyze the data. XYu revised this article. All authors read and approved this article.

FUNDING

This work was supported by the Shandong Provincial Project of Graduate Education Quality Improvement (Nos. SDYJG21104, SDYJG19171, and SDYY18058), the OMO Course Group Advanced Computer Networks of Shandong Normal University, the Teaching Team Project of Shandong Normal University, Teaching Research Project of Shandong Normal University (2018Z29), Provincial Research Project of Education and Teaching (No. 2020JXY012), and the Natural Science Foundation of Shandong Province (Nos. ZR2020LZH008, ZR2021MF118, and ZR2019MF071). The content of this manuscript has been presented in part at the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (Shi et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.781448/full#supplementary-material

REFERENCES

Aydın, S. (2019). Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips. *IEEE J. Biomed. Health Inf.* 24, 1695–1702. doi: 10.1109/JBHI.2019.2959843

Bai, D., Qiu, T., and Li, X. (2007). The sample entropy and its application in eeg based epilepsy detection. *J. Biomed. Eng.* 24, 200–205. doi: 10.3321/j.issn:1001-5515.2007.01.043

Bajaj, V., and Pachori, R. B. (2014). "Human emotion classification from eeg signals using multiwavelet transform," in *International Conference on Medical Biometrics* (Shenzhen).

Bhatti, A. M., Majid, M., Anwar, S. M., and Khan, B. (2016). Human emotion recognition and analysis in response to audio music using brain signals. *Comput. Hum. Behav.* 65, 267–275. doi: 10.1016/j.chb.2016.08.029

Bo, S. (2017). "Research on the classification of high dimensional imbalanced data based on the optimizational random forest algorithm," in *International Conference on Measuring Technology & Mechatronics Automation* (New York, NY).

Bota, P. J., Wang, C., Fred, A., and Silva, H. P. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access* 7, 140990–141020. doi: 10.1109/ACCESS.2019.2944001

Colic, S., Wither, R. G., Lang, M., Liang, Z., and Bardakjian, B. L. (2015). "Support vector machines using eeg features of cross-frequency coupling can predict treatment outcome in mecp2-deficient mice," in *Engineering in Medicine & Biology Society* (Milan).

Deng, W., Miao, D., and Xie, C. (2011). Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection. *Exp. Syst. Appl.* 38, 14314–14320. doi: 10.1016/j.eswa.2011.05.096

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358

Duan, R. N., Zhu, J. Y., and Lu, B. L. (2013). "Differential entropy feature for eeg-based emotion classification," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on* (San Diego, CA).

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., and Fox, P. T. (2009). Coordinate-based ale meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Map.* 30, 2907–2926. doi: 10.1002/hbm.20718

Gu, X., Cao, Z., Jolfaei, A., Xu, P., Wu, D., Jung, T.-P., et al. (2021). Eeg-based brain-computer interfaces (bcis): a survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM Trans. Comput. Biol. Bioinformat.* 18, 1645–1666. doi: 10.1109/TCBB.2021.3052811

Hosseini, S. A., and Naghibi-Sistani, M. B. (2011). Emotion recognition method using entropy analysis of eeg signals. *Int. J. Image Graph. Signal Process.* 3, 30–36. doi: 10.5815/ijigsp.2011.05.05

Jatupaiboon, N., Panngum, S., and Israsena, P. (2013). "Emotion classification using minimal EEG channels and frequency bands," in *The 2013 10th*

International Joint Conference on Computer Science and Software Engineering (JCSSE) (Khon Kaen).

Jiang, C., and Zhou, X. (2004). Emotional automatic processing and control processing. *Adv. Psychol. Sci.* 12, 688–692. doi: 10.3969/j.issn.1671-3710.2004.05.007

Kim, J., and Andr, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 2067–2083. doi: 10.1109/TPAMI.2008.26

Kim, K. H., Bang, S. W., and Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Med. Biol. Eng. Comput.* 42, 419–427. doi: 10.1007/BF02344719

Klauer, K. C., and Musch, J. (2003). "Affective priming: Findings and theories," in *The Psychology of Evaluation: Affective Processes in Cognition and Emotion* (New Jersey, NJ: Lawrence Erlbaum), 7–50.

Li, T., and Lv, Y. (2014). The subliminal affective priming effects of faces displaying various levels of arousal: an erp study. *Neurosci. Lett.* 583, 148–153. doi: 10.1016/j.neulet.2014.09.027

Michalopoulos, K., and Bourbakis, N. (2017). "Application of multiscale entropy on eeg signals for emotion detection," in *IEEE Embs International Conference on Biomedical & Health Informatics* (Orlando, FL), 341–344.

Murugappan, M., Rizon, M., Nagarajan, R., Yaacob, S., and Zunaidi, I. (2008). "Time-frequency analysis of EEG signals for human emotion detection," in *4th Kuala Lumpur International Conference on Biomedical Engineering* (Kuala Lumpur).

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., and Kundu, S. (2018). "Improved random forest for classification," in *IEEE Transactions on Image Processing (IEEE)*, 4012–4024.

Sen, B., Peker, M., Cavusoglu, A., and Celebi, F. V. (2014). A comparative study on classification of sleep stage based on eeg signals using feature selection and classification algorithms. *J. Med. Syst.* 38, 18. doi: 10.1007/s10916-014-0018-0

Shi, Y., Zheng, X., and Li, T. (2018). "Unconscious emotion recognition based on multi-scale sample entropy," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Madrid: IEEE), 1221–1226.

Smith, M. L. (2011). Rapid processing of emotional expressions without conscious awareness. *Cereb. Cortex* 22, 1748–1760. doi: 10.1093/cercor/bhr250

Verma, G. K., and Tiwary, U. S. (2014). Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signals. *Neuroimage* 102, 162–172. doi: 10.1016/j.neuroimage.2013.11.007

Vijith, V. S., Jacob, J. E., Iype, T., Gopakumar, K., and Yohannan, D. G. (2017). "Epileptic seizure detection using non linear analysis of eeg," in *International Conference on Inventive Computation Technologies* (Coimbatore).

Wataru, S., Yasutaka, K., and Motomi, T. (2014). Enhanced subliminal emotional responses to dynamic facial expressions. *Front. Psychol.* 5, 994. doi: 10.3389/fpsyg.2014.00994

Xin, L., Xie, J., Hou, Y., and Wang, J. (2015). An improved multiscale entropy algorithm and its performance analysis in extraction of emotion eeg features. *Chin. High Technol. Lett.* 7, 436–439. doi: 10.1166/jmihi.2017.2031

Xu, Q., He, W., Ye, C., and Luo, W. (2019). Attentional bias processing mechanism of emotional faces: anger and happiness superiority effects. *Acta Physiologica Sinica* 71, 86–94. doi: 10.13294/j.aps.2018.0098

Yang, H., and XU, J. (2017). Android malware detection based on improved random forest. *J. Commun.* 38, 8–16. doi: 10.11959/j.issn.1000-436x.2017073

Yin, Y., Zheng, X., Hu, B., Zhang, Y., and Cui, X. (2021). Eeg emotion recognition using fusion model of graph convolutional neural networks and lstm. *Appl. Soft Comput.* 100, 106954. doi: 10.1016/j.asoc.2020.106954

Zheng, X., Liu, X., Zhang, Y., Cui, L., and Yu, X. (2021a). A portable hci system-oriented eeg feature extraction and channel selection for emotion recognition. *Int. J. Intell. Syst.* 36, 152176. doi: 10.1002/int.22295

Zheng, X., Yu, X., Yin, Y., Li, T., and Yan, X. (2021b). Three-dimensional feature maps and convolutional neural network-basedemotion recognition. *Int. J. Intell. Syst.* 36, 6312–6336. doi: 10.1002/int.22551

Zheng, X., Zhang, M., Li, T., Ji, C., and Hu, B. (2021c). A novel consciousness emotion recognition method using erp components and mmse. *J. Neural Eng.* 18, 046001. doi: 10.1088/1741-2552/abea62

Zhu, X. L., Xiao, L., and Wen, P. (2013). Subliminal emotional face and its brain mechanism. *Nat. Defense Sci. Technol.* 34, 16–20. doi: 10.3969/j.issn.1671-4547.2013.04.004

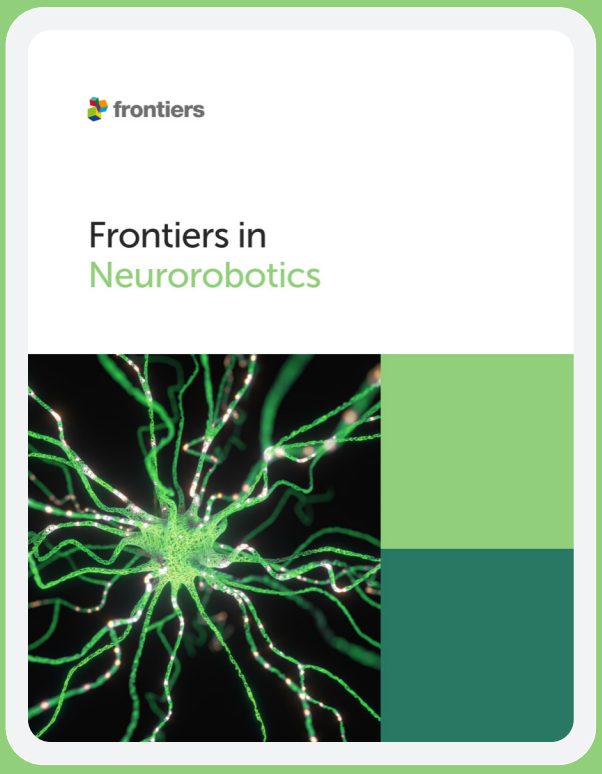
Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Frontiers in Neurorobotics

Investigates embodied autonomous neural systems and their impact on our lives Part of the most cited neuroscience series, this journal advances understanding of neurorobotics - from prosthetic devices to brain machine interfaces, and wearable systems to home appliances.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact