This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TAFFC.2023.3286351 IEEE IEEE Transactions on Affective Computing

MGEED: A Multimodal Genuine Emotion and Expression Detection Database

Yiming Wang, Hui Yu, *Senior Member, IEEE,* Weihong Gao, Yifan Xia and Charles Nduka

***Abstract*—Multimodal emotion recognition has attracted increasing interest from academia and industry in recent years, since it enables emotion detection using various modalities, such as facial expression images, speech and physiological signals. Although research in this field has grown rapidly, it is still challenging to create a multimodal database containing facial electrical information due to the difficulty in capturing natural and subtle facial expression signals, such as optomyography (OMG) signals. To this end, we present a newly developed Multimodal Genuine Emotion and Expression Detection (MGEED) database in this paper, which is the first publicly available database containing the facial OMG signals. MGEED consists of 17 subjects with over 150K facial images, 140K depth maps and different modalities of physiological signals including OMG, electroencephalography (EEG) and electrocardiography (ECG) signals. The emotions of the participants are evoked by video stimuli and the data are collected by a multimodal sensing system. With the collected data, an emotion recognition method is developed based on multimodal signal synchronisation, feature extraction, fusion and emotion prediction. The results show that superior performance can be achieved by fusing the visual, EEG and OMG features. The database can be obtained from https://github.com/YMPort/MGEED.**

***Index Terms*—Emotion recognition, facial expression analysis, multimodal emotion database, affective sensing and analysis**

I. Introduction

Automatic emotion detection is a crucial part of affective computing and has been successfully applied to various applications, such as multimedia [1], [2], biopsychosocial healthcare [3] and human computer interaction (HCI) [4]. With the development of the wearable sensing and computing techniques, a lot of efforts have been made on the multi-sensory emotional data acquisition and analysis [5], [6]. Multi-sensory data are also referred to as multimodal data, which are captured with multiple different sensors and collected in multiple modalities, such as facial expression images, vocal & speech signals and physiological signals.

Facial expression is one of the most important means for humans to express emotions. In the past decade, image-based facial emotion and expression recognition methods have made great progress [7], [8], [9], [10]. The mainstream research in this field is based on six universal and recognizable categories: happy, sad, angry, fear, disgust and surprise [11]. Thus, most existing facial emotional recognition databases are created by capturing facial expression images or videos for six primary emotion recognition [12], [13]. There are also some other popular facial expression databases created for the facial Action Unit (AU) detection [12], [14] and dimensional Valence-Arousal (VA) estimation [13], [15], [16]. The main advantage of vision-based databases is that the image or video data are easy to obtain and annotate. Compared with the physiological data, the facial images can be simply obtained using a camera. There are also rich online resources for facial expression image data. Moreover, it is relatively easier for human experts to annotate facial expression images compared with physiological signals.

The above-mentioned emotional datasets are conducted on a single modality of visual sensor data. With the recent development of human-computer interfaces, novel emotion recognition databases have been created using multimodal sensors. Most existing multimodal datasets are developed by capturing the physiological signals, such as electroencephalography (EEG), facial electromyography (EMG), electrocardiography (ECG) and galvanic skin response (GSR). A multimodal database is often created in laboratory conditions where the emotion of a participant is evoked by watching affective stimuli videos. DEAP was one of the earliest multimodal datasets where physiological signals and facial images of participants in response to the video stimuli are recorded [17]. Soleymani et al. [6] presented a multimodal dataset including facial image data and EEG signal data, and further proved that EEG signals indeed carried expression-related information [18]. There are also some other emotion-related modalities, such as speech signals [19], eye gaze directions [20] and facial thermal images [21].

Although the modern wearable and easy-to-use physiological sensing devices have promoted the research of multimodal emotional analysis [22], [23], it is still a challenging task to collect multimodal data [24]. This is mainly because of the need for interdisciplinary knowledge to process different data from different sensing devices [6]. A specific sensing device requires a human expert with domain knowledge to solve the technical problems. There will be complex teamwork for constructing a multimodal data acquisition system regarding experimental setup, data recording, signal synchronisation and annotation. Therefore, it is still challenging to create a multimodal database.

In general, the study of multimodal databases faces two problems. Firstly, the facial electrical data might not reveal the natural expressions. The commonly used facial electrical signal is EMG. To capture EMG signals, the EMG sensors need to be directly attached to a human face, which may cause uncomfortable feelings to the participants and thus affects the participants to perform natural facial expressions. Secondly, the problem of low-level facial expression intensity is not addressed in the existing multimodal emotional databases. Empirically, annotating a facial expression relies heavily on the visual modality rather than other modalities, which intuitively requires the facial expressions performed in relatively high intensities [6], [18]. However, watching affective stimuli videos often stimulates subtle facial expressions with relatively low intensities, such that genuine emotions cannot be reflected by visual modality alone. Thus, there is an urgent demand to create a multimodal emotion database that addresses the acquisition of natural facial electrical signals and the problem of low-intensity facial expressions.

In this paper, we present a new multimodal dataset called Genuine Emotion and Expression Detection (MGEED) by using modalities of the optomyography (OMG [25], EEG, ECG signals and facial images, as well as depth maps. The "genuine expressions" can also be referred to as real, natural and spontaneous expressions. This term can be explained in two aspects. Firstly, the "genuine" emotion (expression)

Y. Wang is with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom

H. Yu, W. Gao are with the School of Creative Technologies, University of Portsmouth, PO1 2DJ, United Kingdom

Y. Xia is with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China

C. Nduka is with the Emteq Ltd., Sussex Innovation Center, Brighton, BN1 9SB, United Kingdom

Corresponding author: Hui Yu, hui.yu@port.ac.uk

should be mostly expressed in a relatively low intensity. It has also been shown that facial expressions in real life are often low intensity [26]. Compared with the existing databases consisting of many high-intensity facial expression images, the majority of MGEED facial images contain relatively low expression intensities. Secondly, the “genuine” emotions can be natural and spontaneous. Considering that the expressions may be unnatural if the participants wear the EMG sensors, MGEED replaces the EMG with OMG signals which uses the latest non-contact OMG sensors to capture facial electric signals. By using the OMG sensors that effectively avoid skin contact, the participants can perform more natural expressions when wearing OMG sensors.

The OMG sensing equipment is the Emteq smart glasses which benefit from a wide range of sensing channels and the latest non-contact sensing technique. The sensing channels guarantee that enriched expression-related signals can be captured and the non-contact sensing technique enables sensors to avoid physical contact with the human face in order to avoid the uncomfortable feelings of the participants. Besides OMG, the facial images & videos, EEG and ECG signals are also recorded in MGEED database. According to the video recordings, the participants rarely perform macro (high-intensity) facial expressions, which makes it difficult to identify the true emotion from visual modality. Therefore, the physiological signals, especially the OMG signal, are more important than the visual modality in this database. To this extent, the MGEED database is designed to contribute to the study and benchmark of genuine emotion detection.

In this experiment, a data acquisition system is designed to simultaneously capture the multimodal data, including the EEG, OMG, ECG signals, facial videos and depth maps. With this sensing system ready, 17 participants have been recruited to participate in the experiment. During the experiment, they watch a group of emotional videos, and then frame-by-frame self-report their emotional responses in terms of one of the six basic emotions and VA levels. In general, the collected MGEED database contains 17 different subjects with totally 150497 facial images, 147539 frames of depth map and three types of physiological signals including 70-channel EEG, 20-channel OMG and single-channel ECG signals.

With the collected dataset, a baseline multimodal emotion recognition method is developed including multimodal, data synchronisation, feature extraction, feature fusion and emotion recognition. We firstly propose an extended Convolutional Neural Network (CNN) method to extract and fuse the facial image feature and depth feature. Then the EEG, ECG and OMG signals are synchronised, segmented and normalized to their respective compact features. Finally, all the features are concatenated and fed to a prediction network.

The main contribution of this paper can be summarized as follow:

1) A new MGEED database is created for multimodal genuine emotion analysis. This database is characterised by a large number of facial expression images and depth maps, highly reliable expression-related OMG signals and emotion-related physiological signals. MGEED database contributes to the benchmarking of genuine emotion detection where the real emotions can be better reflected by the modalities of physiological signals rather than the facial expression images due to the relatively low intensities of the facial expressions.
2) As far as we are aware, this is the first study on creating a public emotional dataset that contains OMG signal recordings.
3) Based on the MGEED database, a multimodal signal synchronisation, feature extraction and fusion method is proposed for data analysis. The experimental results demonstrate that OMG features indeed improve emotion recognition accuracy, and the best performance is achieved by combining image, OMG and EEG features.

II. Related Work

A. Vision-based Facial Emotion Recognition

Although human emotions can be detected in multiple modalities, the majority of emotion recognition research is still conducted on the visual modality (facial expression images or videos). Facial expression recognition (FER) aims to use advanced computer vision and machine learning methods for image/video-based facial affective analysis [27], [28], [29]. Till now, there is a large number of vision-based FER databases that are publicly available. Some popular FER databases and their attributions are shown in Table I. In this section, these databases are introduced in the view of conditions (laboratory vs in-the-wild conditions), scales (small vs large) and benchmarks (categorical model vs continuous dimensional model).

Laboratory VS in-the-wild conditions: Early studies and benchmarking of FER are normally conducted based on laboratory-controlled conditions, where the facial expression images of the participants are captured in a laboratory. The emotions are either directly acted by the participants or evoked by video stimuli. Related databases include CK+ [12], MMI [30], Multi-PIE [31], DISFA [32] and SAMM [14]. The main problem with these databases is that the facial expressions consciously posed by the participants may not reflect their true emotions. This is an obvious restriction for these databases to be applied to real-world scenarios. Although laboratory-controlled databases seem impractical, they still form the foundation of FER research and are reported to be referred to in the majority of FER methods [33], [34].

Different from the laboratory-controlled databases, the in-the-wild databases consist of facial images captured from real-world scenarios. Such databases are characterised by spontaneous facial expressions and unconstrained conditions where there are various changes in head pose, illuminations and occlusions. Related databases include FER-Wild [35], FER2013 [36], RAF-DB [37], EmotioNet [38], AffectNet [13], Static Facial Expression in the Wild (SFEW) [39] and Acted Facial Expression in the Wild (AFEW) [40]. To tackle the in-the-wild challenges, a series of facial image preprocessing and mid-level feature processing methods have been proposed, such as view-invariant methods [41], [29], [42], characterized expression enhancement (expressionlets) [43], illumination normalization [44], de-occlusion [45] and time alignment [46].

Small vs large: The typical small-scale FER databases include CK+, SFEW and BU-3DFE, which contains only hundreds or thousands of images. The main advantage of small-scale databases is that the emotion label is relatively more accurate and reliable since all the images can be well annotated by human experts. On the contrary, EmotioNet [38] and AffectNet [13], currently the largest FER databases, may include incorrect labels. This is because the images are obtained from the Internet by using emotion-related keywords to query search engines. Due to the large data volume, it is nearly impossible to clean and annotate all the data by human experts. Therefore, the noises may occur due to the potential mistakes caused by the search engines. Although there are obvious drawbacks, these large-scale databases still become the mainstream and have promoted a series of high-quality end-to-end FER methods with valuable outputs [47], [48].

Categorical model vs continuous dimensional model: Based on different FER tasks, there are generally three expression models for quantifying facial expression distributions: basic emotional model, AU model and continuous dimensional model. The basic emotional model and AU model belong to the categorical model where the facial expressions are classified into several meaningful categories.

The basic emotional model describes emotions by six basic emotional states (Happy, Sad, Angry, Fear, Disgust and Surprise). The majority of the existing databases follow this benchmarking. AU model is the encoded facial actions that directly represent the small facial muscle movements. The AU and basic emotions are often jointly labelled in many databases, such as CK+, DISFA and EmotioNet.

In contrast to categorical models, the dimensional model describes the continuous affective response reflecting the changes of both the emotional state and intensity [51]. The most popular dimensional model is a Valence-Arousal (VA) pair where valence is the emotional scale ranging from negative to positive and arousal represents the intensity level ranging from calm to exciting. VA is a fine-grained model that requires professional researchers to continuously observe and capture the subtle expression changes and carefully locate each frame in the VA space. The related database include AFEW-VA [15], AffectNet [13] and Aff-Wild2 [49].

Recently, the dimensional VA model has attracted significantly increasing attention. However, annotating VA levels is challenging due to the lack of objective metrics. The annotators have to rely on their own experience and provide a subjective assessment of VA levels when labelling data [58]. Therefore, it is difficult to obtain commonly accepted VA evaluation according to visual modality alone. The high-resolution electrical signals recording biological changes provide a proper way to characterize the fine-grained subtle VA changes. Recently, multimodal emotional databases have shown the advantages of VA evaluation. Many existing multimodal databases are annotated with VA levels.

B. Physiological Signal-based Emotion Recognition

Apart from facial expressions, human emotions can also be reflected by physiological changes. Compared to vision-based facial expression databases, physiological signal-based emotional databases have been applied to the broader emotion distributions, such as basic emotions and VA scales [20], [52], wellbeing [17], [53], stress / tension / panic [56], [6], like / dislike [17], and personality [54]. The commonly used physiological signals include Electroencephalography (EEG), Electromyography (EMG), Electrocardiography (ECG), Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), respiration amplitude, Skin Temperature (SKT) and Electrooculography (EOG). The popular multimodal databases are shown in Table II.

EEG records the brain electrical signal measuring neurons synaptic excitation characterized by its amplitude and frequency. EEG signal typically reflects cognitive processing in the human brain[59]. Quantitative EEG data within a physiological database is normally obtained through non-invasion Brain-Computer Interfaces (BCI) which capture multi-channel brain signals. There have been several works on developing affect monitoring methods using EEG alone [60], [61], [62]. The relation between facial expression actions and EEG is investigated in [18] and promising emotion recognition results have been obtained by using both visual facial expression features and EEG. Although visual features fused with EEG signals do not always boost performance, the researchers still elaborate that EEG signals contain valuable information of affects. Most existing EEG-based research focuses on EEG feature extraction. Popular EEG feature representations include statistics features (e.g. standard deviation and mean) [60], neural networks [63] and Power Spectral Density (PSD) [64].

Facial EMG records the electrical signals of facial muscle actions in which a tiny electric impulse can be detected and amplified. EMG can easily capture the changes in the baseline muscle tone and record electrical activity directly. The existing EMG acquisition systems normally attach two electrodes to the Zygomaticus Major muscle that is located near the corner of the mouth below the cheekbones [17], [65]. The sensors can detect the cheek raising and lip corner stretching movements that are highly associated with a smile expression. However, the main drawback of the EMG sensors is that they are directly attached to the human face, which may cause uncomfortable feelings to the participants. The early methodologies of EMG feature extraction often applied discrete wavelet transform method to extract discriminative features [66]. There are also a number of works using statistical features for EMG-based emotion analysis [65], [67].

ECG captures an electrical signal of the heart through tracking heart rhythms. There are many different ECG sensors. Most of them are chest straps consisting of a group of Carbon electrodes that are directly placed on the chest. Modern products equipped with ECG sensors, such as smart watches and armbands, are designed to be wearable and able to monitor the heart rate in real-time. The ECG wave directly reflects cardiac cycles which are associated with Heart Rate Variability (HRV) that measures the variations of time intervals between adjacent heartbeats (RR intervals). In [68], 56.9% accuracy of 5-class (happy, sad, angry, fear, relax) emotion recognition is reported by using ECG signals alone. The relatively low accuracy indicates that the ECG may not contain rich information on the 6 basic emotional states. The emotion classification experiments in [69] show that ECG prefers discriminating negative emotions rather than positive emotions. There are also some other works using ECG for negative emotion recognition [70], [71]. A more reasonable application of ECG is arousal prediction as human may have exciting / calm responses with obvious changes in heart beat rhythms. Much effort has been made to develop the arousal scoring system based on ECG [72], [73]. The fundamental ECG features are the measurement of the difference between adjacent RR intervals (e.g. Root Mean Square of the Successive Differences–RMSSD) [55]. Other means of ECG data processing include frequency domain analysis [68], [69], [55], time domain analysis [71] and statistical analysis [74].

GSR is also referred to as Electrodermal Activity (EDA) which detects sweat gland activity by measuring skin conductance. The wearable GSR sensor is in a ring shape whose electrodes are often placed on the proximal part (below the joint) of the index and middle fingers. Similar to ECG, GSR is often used for negative emotion (deception, stress, anxiety) recognition [75], [76], even if no evidence shows that GSR prefers differentiating negative emotions. The GSR signals have also been applied to 6-class emotion recognition [77] and VA estimation [78].

Physiological data acquisition systems require a complicated laboratory setup with many different sensors attached to different parts of the human body and the interfaces of the sensors need to connect to a computer with high specifications for real-time signal processing and recording. DEAP [17] is one of the earliest multimodal emotion databases. It covers nearly all known physiological signals including 32-channel EEG, 4-channel EMG, 4-channel EOG, GSR, SKT, BVP and respiration. The MAHNOB-HCI [6] database focuses more on visual signal acquisition where six web cameras are used to capture the facial expressions and body gestures of the participants. The SEED-V [20], DECAF [53], Amigos [54] and DREAMER [55] databases follow similar laboratory-setups as DEAP and MAHNOB-HCI.

The BP4D+ [21] and BU-EEG [57] are currently the only two benchmarks including AU annotations. The recent research [79], [80] on these databases shows that fused physiological signals indeed improve the performance of basic emotion and pain recognition. However, the AU recognition methods are still based on visual information alone. The brief analysis shows that the dynamic AU activities are somewhat associated with the physiological signals [79],

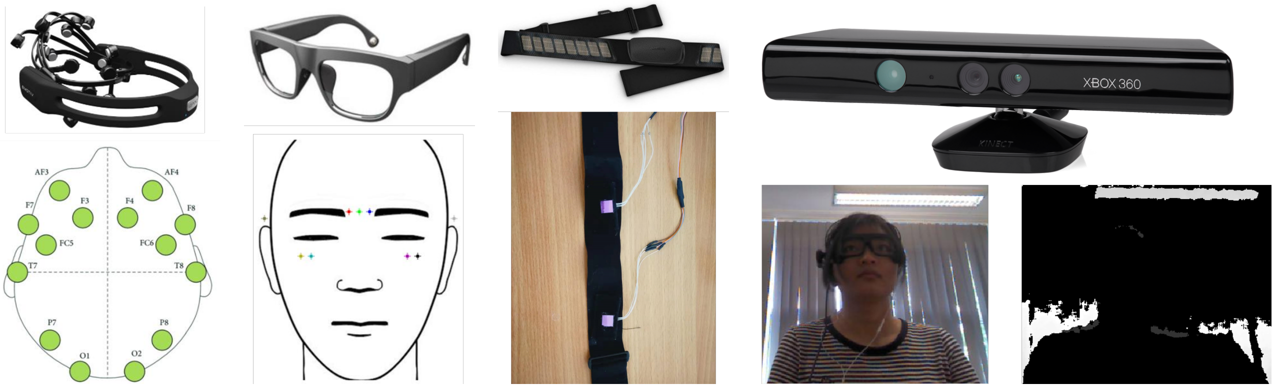
TABLE I
ATTRIBUTES OF SOME REVIEWED VISION-BASED FACIAL EXPRESSION DATABASES

Database	**Database information**	**Condition**	**Expression model**
CK+ [12]	● 593 video sequences, 123 subjects ● Faces are captured in frontal view ● Annotation of facial landmarks	Controlled, posed	7 classes + contempt, AU
Multi-PIE [31]	● More than 750,000 images, 337 subjects ● 15 facial viewpoints including frontal view, 19 illumination conditions	Controlled, posed	Smile, surprise, squint, etc
DISFA [32]	● 27 videos including 130,788 images ● Facial expressions of subjects are aroused by video stimuli ● Smile accounts for a large number of frames	Controlled, spontaneous	Smile, AU
SFEW 2.0 [39]	● Train (958 images), Validation (436 images) and Test (372 images) ● Strictly person-independent evaluation protocol	In-the-wild	7 classes
AFEW 7.0 [40]	● 1809 video clips selected from movies ● Strictly person-independent evaluation protocol	In-the-wild	7 classes
AFEW-VA [15]	● 600 videos selected from AFEW ● Accurate annotation of ladmarks ● Discrete VA levels from -10 to 10	In-the-wild	VA (discrete)
EmotioNet [38]	● 1,000,000 images queried from Internet ● 25,000 images labelled manually	In-the-wild	23 emotions, AU
Aff-Wild2 [49]	● 564 Youtube videos including around 2.8 millions frames ● First large scale in-the-wild database containing annotations of all the three expression models	In-the-wild	7 classes, AU, VA
AffectNet [13]	● 450,000 images (labelled) queried from Internet ● Annotation of facial landmarks	In-the-wild	7 classes + contempt, VA
RAF-DB [37]	● 29672 images queried from Internet ● Annotation of 5 facial landmarks	In-the-wild	7 basic + 12 compound
BU-3DFE [50]	● 2500 3D facial models, 100 subjects ● Annotation of 5 facial landmarks	Controlled, posed	7 classes
SAMM [14]	● 159 facial micro-movements, 32 subjects ● A high-resolution facial micro-expression database	Controlled. spontaneous	6 classes + contempt, AU

TABLE II
ATTRIBUTES OF REVIEWED PHYSIOLOGICAL SIGNAL-BASED EMOTION RECOGNITION DATABASES

Database	**Database information**	**Signals**	**Emotion annotations**
DEAP [17]	● 32 participants ● 40 one-minute music video stimuli	EEG, EOG, EMG, GSR, BVP, SKT, respiration, face video	VA, liking, dominance, familiarity
MAHNOB-HCI [6]	● 27 participants ● 20 video stimuli	EEG, ECG, GSR, SKT, EOG, respiration, face and body video, audio	7 classes + amusement + anxiety, VA, dominance, predictability
SEED-V [20]	● 20 participants ● 15 video (2-4 minutes movie clips) stimulus	EOG, EEG	5 classes (surprise excluded)
RECOLA [52]	● 46 participants attending a video conference ● Participants will complete a task requiring affective interactions	ECG, GSR, audio, face video	VA
DECAF [53]	● 46 participants ● 40 one-minute music video and 30 movie clips as stimuli	MEG, EOG, ECG, EMG, face video	VA, dominance
Amigos [54]	● 40 participants ● 16 short and 4 long video stimuli ● RGB and depth videos recording full body of an individual or a group	EEG, ECG, GSR, audio, video, depth	VA, personality traits
DREAMER [55]	● 23 participants ● 18 video stimuli	EEG, ECG	VA, dominance
Schneegass et al [56]	● 10 participants in real-world driving condition ● five different road types	ECG, GSR, SKT, face video	Assessing driver's workload
BP4D+ [21]	● 140 participants ● 10 tasks (interview, game, video stimuli) for emotion elicitation	ECG, GSR, BVP, respiration, dynamic 3D face model, face video, thermal	10 emotions (6 classes, pain, embarrassment, startle, skeptical), AU
BU-EEG [57]	● 29 participants with 2320 experiments trails ● Posed expressions and spontaneous pain by cold-pressor	EEG, face video	7 classes, AU, pain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TAFFC.2023.3286351 IEEE IEEE Transactions on Affective Computing



(a) Epoc+ and the position of 14 electrodes

(b) Smart glasses and the position of 9 sensors

(c) HRM strap

(d) Kinect and its output

Fig. 1. Deceives for data acquisition

[57].

Physiological signals have extended the modalities and applications of emotional research. However, none of the above-mentioned signals can be used alone for emotion recognition. The most remarkable cues of emotions are still conveyed by facial expressions. The existing research of multimodal emotion analysis generally utilizes physiological signals as auxiliary features integrated with the facial expression features for emotion recognition.

III. Experimental Setup

A. Participants

The experiment is designed to monitor facial behaviours of participants and record their physiological signals. We use a group of video stimuli that attempt to induce the emotions of the participants. The emotional videos used in this experiment are from the DISFA dataset[81]. It consists of 9 independent video clips extracted from online resources and TV programmes. The duration of the video is around 4 minutes.

The ethical approval of this study has been granted by the ethical committee. Before the experiment, we firstly released the information sheet and introduced this study to the participants. Then, by signing the consent agreement, the participants gave consent to recording, sharing and publishing their physiological signals and facial images for academic purposes. The participant should be aware that their participation is voluntary and they are free to withdraw at any time without giving any reasons.

In this experiment, 17 voluntary participants aged from 18 to 40 are recruited. During the experiment, participants are asked to wear the physiological sensors and watch the stimuli videos. The 9 video clips play one-by-one continuously with a 5-second between-clip interval. The participants watch all the video clips at once. Finally, the participants are asked to self-rate their emotions for the time when watching each video clip.

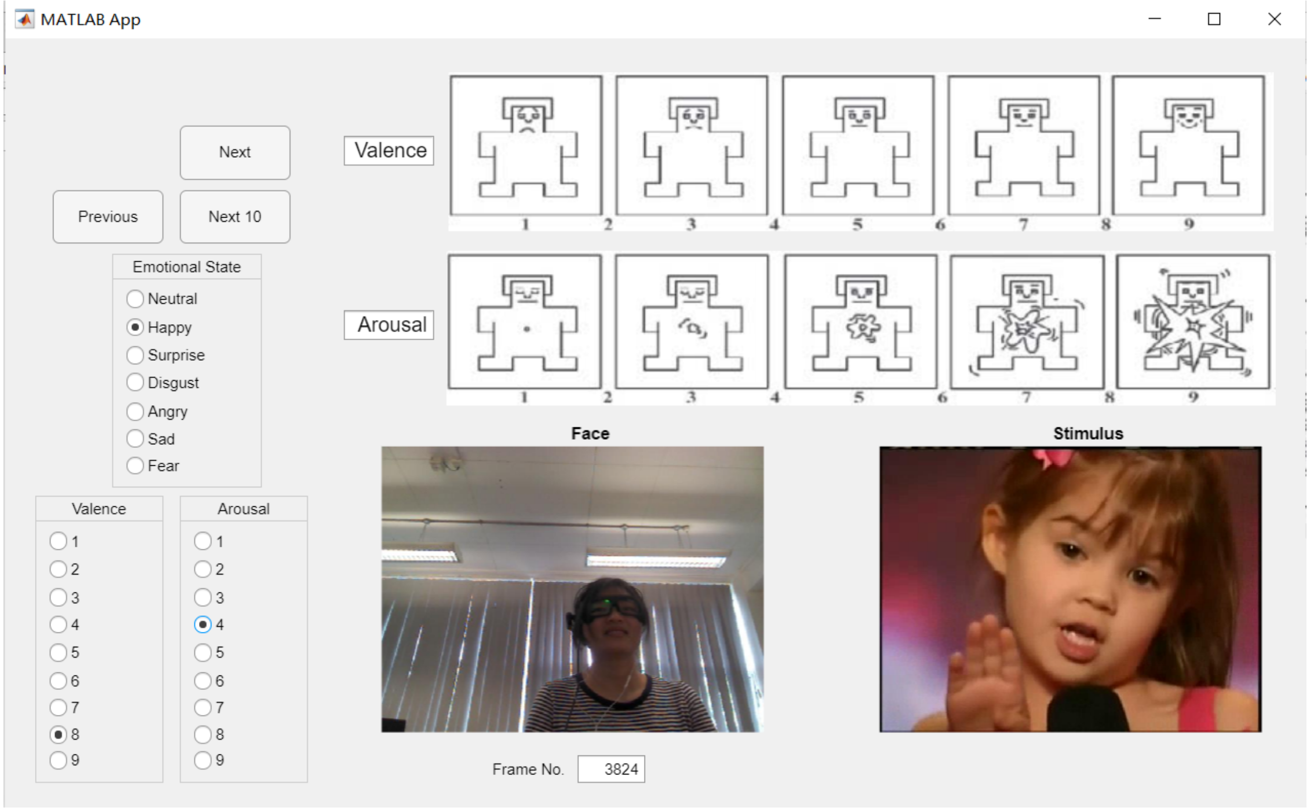
B. Device

There are four sensing devices used in this data acquisition system for capturing the physiological and visual signals of the participants as shown in Fig. 1.

1). EPOC+: We use EMOTIVE EPOC+ to collect EEG signals during the experiment. EMOTIVE EPOC+ is a neuron-headset with an electrode placement system which is designed for contextualised research and advanced brain-computer interface (BCI) applications. It has 14 electrodes at a 128Hz sampling rate (16-bit resolution). The 14 sensors are attached to the corresponding positions of the head (scalp), shown in Fig. 1a, capturing electric brain activities in terms of various frequencies. These raw EEG electric signals are converted to the PSD using the Fast Fourier Transform (FFT) method. The PSD features of EEG signals are identified as four main types of brainwaves located at different frequencies, named Beta (14Hz to 30Hz), Alpha (7Hz to 13Hz) theta(4Hz to 7Hz) and Delta(less than 4Hz). EPOC+ further divides the Beta wave into high-frequency Beta wave and low-frequency Beta waves. Each electrode can detect the brainwaves in 5 bands, which leads to a total number of $14 \times 5 = 70$ EEG features. These features are available at an 8Hz sampling rate.

2) Smart Glasses: The Emteq smart glasses are used in our experiment for collecting facial OMG signals. Fig. 1b illustrates the smart glasses and the position of sensors. The smart glasses use the latest non-contact sensing techniques that can detect local facial movements without contacting the human face. There are 9 optical sensors measuring momentary position changes (and output the accumulation along the time axis) and 9 corresponding proximity sensors measuring the distance from the face. Three sensors distribute between eyebrows, named respectively sensor 1, sensor 2 and sensor 3 from the right side to the left side of the face. Sensor 4, sensor 5, sensor 6 and sensor 7 are located under the right eye and left eye. Sensor 8 and sensor 9 are facing towards the right side and the left side of the face from the side arms of the glasses. The raw output of the smart glasses is 27-channel signal data representing the movements of the nine points along X, Y and Z directions. Empirically, we remove the signals from the centred-eyebrows sensors and the Z-direction values from the eyebrow sensors (sensor 1 and 3) and side-face sensors (sensor 8 and 9), as these values are not useful for facial emotion analysis. The final recording of OMG signals contains 20 channels. The recorded OMG signals are available at a 50 HZ sampling rate.

3) Kinect: Kinect is used to capture depth maps and RGB images. A depth map is a single-channel image in which the pixel value reflects the distance between the sensor and the object point. The core technology of Kinect is the depth and motion sensing system whose hardware incorporates an RGB camera, an infrared projector and a detector. Kinect can record a sequence of RGB images and the corresponding sequence of depth maps. But Kinect does not always output an RGB image and a depth map in pairs. Therefore, we record the timestamp of each frame of RGB image and depth map in preparation for synchronisation.



Emotional State
○ Neutral
● Happy
○ Surprise
○ Disgust
○ Angry
○ Sad
○ Fear

Valence
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
● 8
○ 9

Arousal
○ 1
○ 2
○ 3
● 4
○ 5
○ 6
○ 7
○ 8
○ 9

Frame No. 3824

Fig. 2. User application for emotion annotation

4) HRM soft strap: The ECG device we used in the experiment is Garmin soft strap for Heart Rate Monitoring (HRM). The HRM strap is worn by the participants around their chests. The ECG signals are available at a 1000Hz sampling rate. The collected ECG data contain the single-channel ECG signals and the time stamps.

C. Emotional Response Ratings

The ratings of emotional responses mainly rely on the self-rating of the participants. In this experiment, the self-rating requires the participants to rate three types of emotional responses: emotional states, valence and arousal. The emotional state can be justified according to the experience of the participants. The VA responses of the participants are rated based on the Self-Assessment Manikins (SAM) [82]. SAM describes three types of emotional response: valence (pleasure), dominance and arousal, and rates them by 9-score scales, respectively. Valence ranging from 1 to 9 represents the emotional response gradually changing from negativity to positivity. Arousal rates the level of excitation ranging from calm (at lower scores) to excitation (at higher scores).

To help the participants self-rating their emotional responses, a user application is developed with a friendly graphic user interface, as is shown in Fig. 2. The two figures illuminating 9-score VA scales is displayed in this application, as a reference for the participant to rate VA scales. With the help of this friendly software tool, the participants can easily choose 1) which emotional state is (by ticking one option from Happy, Surprise, Neutral, Disgust, Anxiety, Sad and Fear); and 2) what levels of valence and arousal are (both ranging from 1 to 9). By clicking the “Next” (or “Previous”) button, the participants can simultaneously observe the stimuli video and their recorded images of facial expression, and then frame-by-frame provide self-rated emotional annotation correspondingly. It should be noted that the emotion changes are normally very slow and the adjacent 10 frames may share the same levels of emotional responses. This application provides an option for quick annotation by clicking the “Next 10” button to skip over 10 frames and assign these frames with the same annotation. With this convenient function, the participants can complete the whole process of self-rating within 5 minutes.

After obtaining all the data, two researchers who have done professional training, are then responsible for checking the annotation frame-by-frame. Both researchers should independently rate all the frames and provide feedback if they strongly disagree with the original annotations. For those frames where one researcher agrees with the original annotations while the other disagrees, the original annotations will have remained. For those frames where both researchers disagree with their original annotations, the average scores of the rating results from the two researchers will be used as the corrected annotations. Both the original and the researcher-corrected annotations are provided in the MGEED database, so that the researchers who use this database can conduct comprehensive experiments and make the correct conclusions.

IV. Baseline Method for Emotion Analysis

Given the collected dataset, a multimodal emotion analysis method is developed including the following steps: synchronisation, feature extraction, feature fusion and emotion recognition.

A. Synchronisation

In this experiment, all the sensing devices are connected to one computer and we have developed a program to assign each sample point with a timestamp. For the visual signals, the program assigns each frame with a timestamp. For the other sensors whose sample frequencies are very high, the program only records the timestamp

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TAFFC.2023.3286351 IEEE IEEE Transactions on Affective Computing

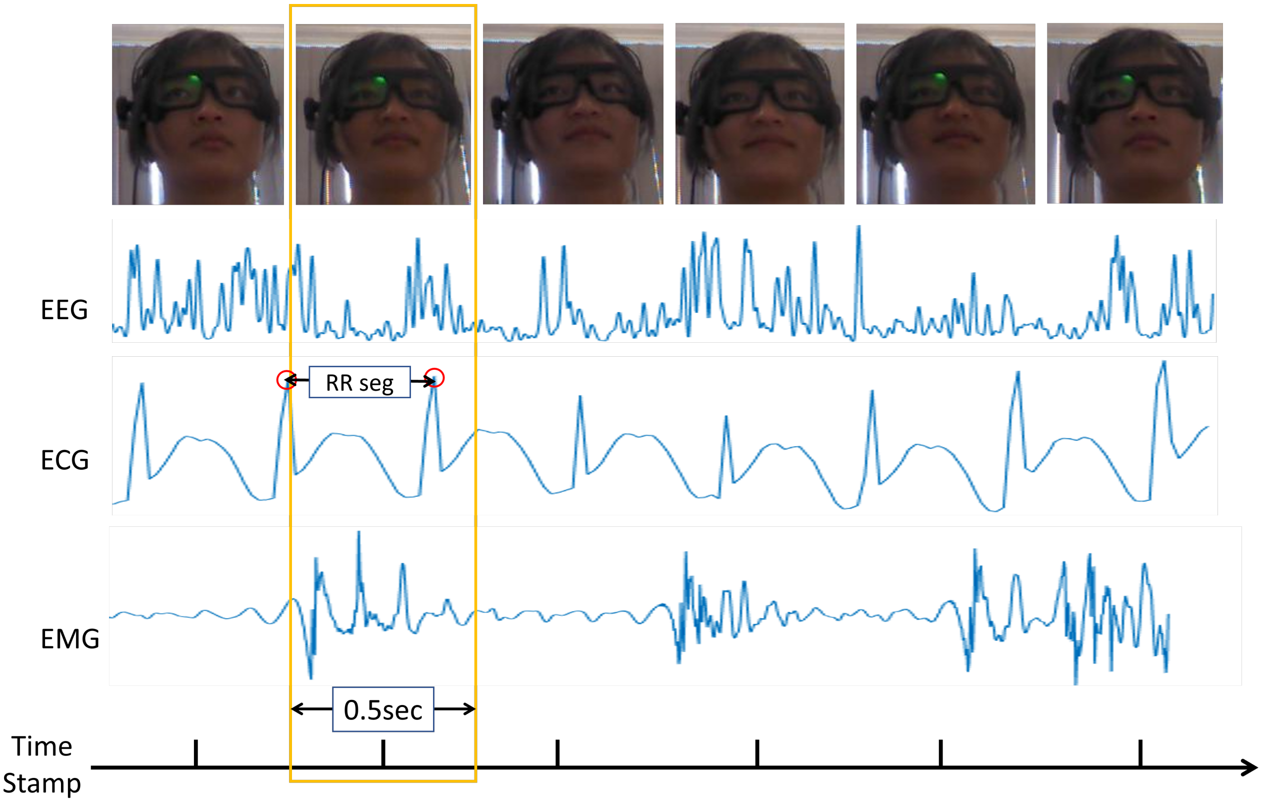


Fig. 3. Multimodal data synchronisation

of the first sample point and the remaining samples are calculated afterwards according to the sampling rate. It can be seen that timestamps are not aligned since the EEG, OMG, ECG and visual signals are captured in different sample rates and starting times. Therefore, a synchronisation operation is needed to align all the signals to the same time axis.

To synchronise different modalities, the recorded time is firstly segmented at every 0.5 second into non-overlapping time intervals along the time axis. For each segmentation, the centred images and depth maps are selected to represent the visual modality of this segment. Then the EEG and OMG signals are synchronised by simultaneously aligning and segmenting the starting time and ending time within this segment. Finally, the ECG signal segmented by the RR interval is synchronised by selecting an RR interval whose time period has the largest overlapping ratio with the current time segment. This three-step synchronisation is illuminated in Fig. 3. With this synchronisation operation, all the modalities are well aligned and thus can be used for further feature extraction and emotion recognition.

B. Physiological Feature Extraction

1) OMG features: The OMG signals can be easily interfered with. There are two main problems in the raw OMG data. 1) The optical sensors of the smart glasses capture momentary spatial changes and accumulate their values along the time axis. It can be observed that the OMG signals tend to involve several discrete monotonous changes. 2) The time series signals may occasionally drift due to external factors, such as illumination changes. Owing to these two problems, a three-step pre-processing is undertaken to clean the OMG signals. Firstly, we apply a least square fitting of a straight line to remove the linear trend. Secondly, the Notch filter is used to block 50Hz components and their harmonics up to 350 Hz. Finally, the bandpass filter is applied to retain the components from 30 Hz to 450 Hz.

The cleaned OMG signals are then available for feature extraction. Following our previous work on [83], we consider reducing the dimensionality by segmenting and compressing the signals. The signals are divided into non-overlapping segments of the same length that lasts for 0.5 seconds. Each segment contains 25 sample points.

In each segment, the Root Mean Square (RMS) value is calculated as the time-domain feature representation. The RMS is computed as:

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{1}$$

where x_n is a sample point within a segment and N is the length of this segment (25 for OMG). Generally, the 20 OMG sensing channels produce a 20-dimensional vector representing the compressed OMG feature within 0.5 second.

2) EEG features: There is no need to perform the EEG pre-processing. The output EEG signals from EPOC+ have been filtered and cleaned already. The EEG feature extraction is consistent with the recorded OMG signals that are segmented every 0.5 seconds and then compressed using RMS. The derived EEG feature is a 70-dimensional vector.

3) ECG features: The raw ECG signals also suffer from the same problem as OMG signals where the values of sample points are accumulated along the time axis. To remove this linear trend, the least square fitting is used to filter the raw ECG signals. Then, the filtered ECG signal is segmented into individual RR intervals. In each segment, three types of ECG features are extracted: difference of RR intervals, frequency domain features and statistical features.

RR interval is the duration of an individual heartbeat. It is obtained by measuring the distance between two adjacent R-peak. In the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TAFFC.2023.3286351 IEEE IEEE Transactions on Affective Computing

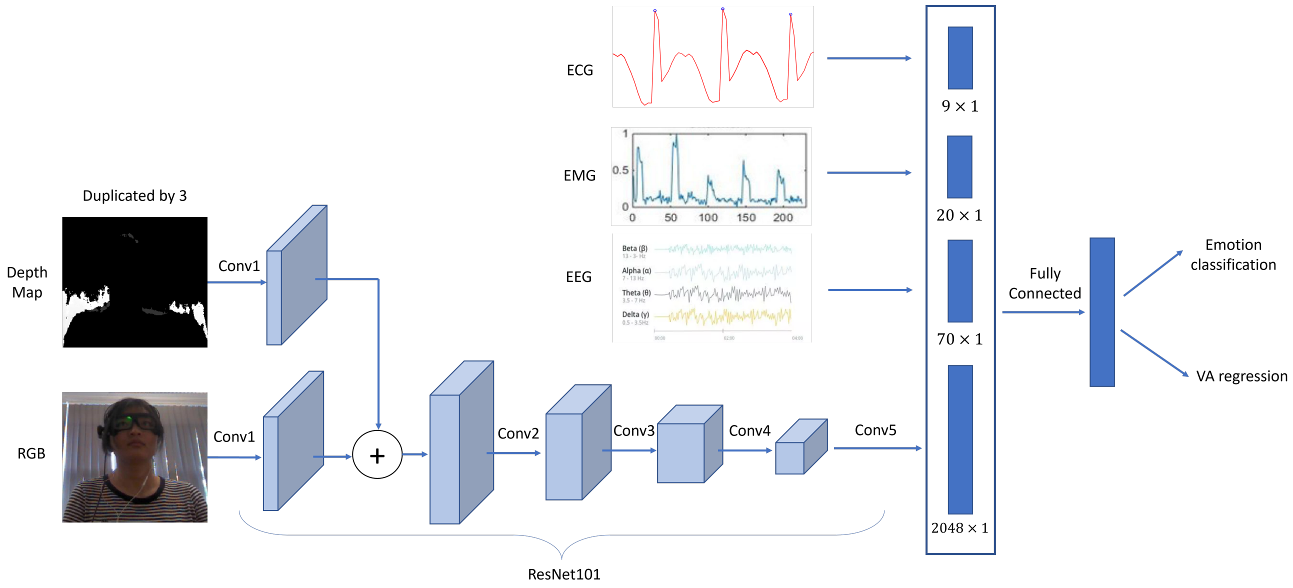


Fig. 4. Feature extraction and fusion

ECG wave, the R-peak refers to the maximum amplitude within an individual heartbeat. In order to detect the R-peak, we find out all the local maximum points and set an appropriate threshold to block those points a the relatively low level of amplitude. Fig. 5 illuminates the R-peak and RR interval. Given the consecutive RR intervals, the feature of RR interval changes is obtained by calculating the difference between the current RR interval and the mean value of the RR interval.

For the frequency domain feature extraction, Fast Fourier Transform is firstly used to convert each RR interval into a band with different spectral frequencies. Then, we select two bands of the power spectrum: the low-frequency band (LF) ranging from 0.04 to 0.15Hz and the high-frequency band (HF) ranging from 0.15 to 0.4Hz. Finally, the frequency domain features are extracted including the PSD for LF, PSD for HF, the ratio of LF to HF, and the total power.

The statistical feature extraction is simple and straightforward. The mean value, standard deviation, minimum value and maximum value are extracted from the ECG segment. Consequently, all the features are concatenated to form a 9-dimensional feature vector.

C. Image and Depth Feature Extraction

The visual analysis of facial expressions starts with facial region detection. In our dataset, the depth information can be used to accurately crop the facial region from an image. Given a depth map, a reasonable threshold is set to distinguish foreground and background. The obtained foreground region can be seen as the facial region which will be cropped for the next step process of the facial visual feature extraction.

The field of visual feature extraction and recognition has been dominated by CNN [84], [85] in recent years. Various CNN architectures have been proposed and proved to be effective for vision-based facial expression recognition. Among numerous CNN architectures, ResNet is a popular method and has achieved impressive results in the ImageNet competition [86], [87]. ResNet has been recognised as the gold-standard architecture in computer vision applications and served as the default method in many existing studies [88]. ResNet addresses the problem of accuracy saturation in which the accuracy of a very deep CNN model may get saturated, and then degrade with increasing depth. ResNet is designed to alleviate this problem by introducing the so-called shortcuts to jump over some layers. The shortcut connection learns to adaptively block some layers so that even a very deep model can be suitably trained without degradation. Moreover, the model size of ResNet is relatively smaller, compared with the popular CNN architectures (e.g. VGG and AlexNet) [89].

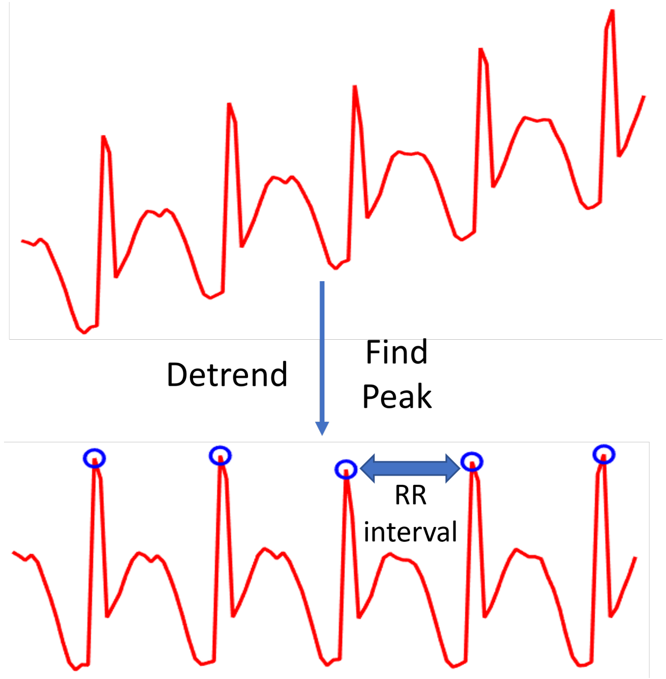


Fig. 5. ECG signal pre-processing and RR interval detection

Given so many advantages of ResNet, we employed ResNet101, a 101-layer network, as the visual feature extraction method. To improve the generalization ability, the transfer learning technique [90] is also used where the ResNet101 model is pre-trained on ImageNet and fine-tuned on the MGEED dataset. To extract the depth features, we modify the ResNet101 architecture by adding an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TAFFC.2023.3286351 IEEE IEEE Transactions on Affective Computing

TABLE III
DIVISION OF THE DATABASE BASED ON SUBJECTS

	Set 1	Set 2
	1,2,5,9,13,14,16,17	3,4,6,7,8,10,11,12,15
Neutral	1522	1300
Happy	1180	1150
Sadness	349	404
Angry	43	126
Fear	748	981
Disgust	311	469
Surprise	592	751

additional convolutional layer to receive and process the depth map. The weights of this additional convolutional layer are initialized by the first convolutional layer of the pre-trained ResNet101 model. The depth map is duplicated three times and stacked so that a three-channel map is generated to meet the input requirement of the convolutional operation. The generated depth map passes through the additional convolutional layer and its output is accumulated to the first convolutional layer of the main CNN architecture. This modified network can extract both feature and depth features, and then fuse them at the feature level. The architecture of this network is shown in Fig. 4. The output of this network is a 2048-dimensional feature vector.

D. Fusion and Recognition

Given the synchronised signals, their features can be extracted according to section IV-B. Then, the EEG, OMG, ECG and visual features are fused to make a joint prediction of the emotions.

The architecture of the feature fusion and recognition method is shown in Fig. 4. The main branch of this architecture is the ResNet101 dedicated to extracting the visual features represented by a 2048-dimensional feature vector. The other branches receive the feature vectors of EEG (70 dimensions), OMG (20 dimensions) and ECG (9 dimensions), respectively. To allow effective feature fusion, the EEG, OMG and ECG features are firstly normalized by the commonly used L2 normalization. Then, all the features are concatenated to form a 2147-dimensional (2048+70+20+9) feature vector. The concatenated features are fed to a fully connected network for the joint prediction of emotions.

The output layer of the network includes 9 prediction nodes regarding seven emotional states and a pair of VA levels. For the loss in the training stage, the cross-entropy (with softmax activation) is calculated as 7-emotion classification loss and the mean square error is used as VA intensities regression loss. The final loss function is obtained by adding these two loss functions.

To address the problem of class imbalance, we investigate two commonly used strategies to deal with the imbalanced classes: over-sampling and weighted loss function. The implementation of over-sampling is to randomly duplicate some samples from the minority classes and add them to the training sets. By over-sampling, classes can be balanced within a mini-batch although the training time may increase due to the increasing size of the training set.

Another commonly used solution for the class imbalanced problem is the weighted loss function, which highlights the importance of minority classes by extending the classification loss to its weighted counterpart. For each class, the weighted cross-entropy loss function is expressed by:

$$\mathcal{L}_i = -w_i \log \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

where $i, j = 1, ..., K$ is the class index, $[z_1, ..., z_k]$ is the output vector (class score) of the prediction model and w_i is the weight of the i-th class.

To tackle the imbalanced classes, dynamic weighting scheme [91] is used to adjust the weights based on the class frequency computed over the training set. The class weight can be designed as follow:

$$w_i = \log(\frac{max(n_j|j = 1, ..., K)}{n_i}) + 1 \tag{3}$$

where n_i is the total number of samples of the i-th class.

V. DATA ANALYSIS

A. Evaluation Protocol

The MGEED dataset includes 17 subsets with regard to 17 different subjects. Each contains around 8000 sequential images, around 7000 sequential depth maps, 20-channel OMG signals, single-channel ECG signals and 70-channel EEG signals. As mentioned above, the timestamps are also recorded for each frame or sample point. The frames from the image sequences are designed as fiducial points where each frame is labelled as one of 7 emotional categories (6 basic emotions and a neutral category), a valence level and an arousal level.

To enable evaluating the performance of an algorithm using MGEED dataset, we recommend using the proposed synchronisation method for data preprocessing, as is described in section IV-A. With this strategy, a simplified dataset is generated containing a total of 9926 sample points. Each sample consists of an image, a depth map and the synchronised signals of OMG, EEG and ECG.

The evaluation protocol follows strict person-independent validation. In this protocol, all images are rearranged according to the identity of subjects where all the images from an arbitrary subject can only be assigned to either the training set or the testing set. To enable the person-independent evaluation protocol, the dataset is divided into two subsets following two principles: 1) the two subsets contain non-overlapping subjects; and 2) for each emotional category and each level of the 9-scale VA, the number of frames should be balanced in both subsets. Consequently, 8 subjects including 4745 frames are assigned to one subset and the remaining 9 subjects with 5181 frames placed in the other subset. The detailed information can be found in Table III.

With the two subsets, researchers can evaluate the performance of their methods/algorithms by training the models on one set and testing on the other, and vice versa. This protocol could guarantee that the experiment is strictly person-independent without subject overlapping.

For the evaluation metric, the recognition rate is a commonly used option for measuring the seven-class emotion classification. However, this is not a good choice for tackling imbalanced data. From Table III, it is obvious that the data is class imbalanced. There are a large number of images with neutral and happy while only a few images with angry. To account for the imbalanced problem, we follow the protocol from [92] by adopting a balanced accuracy:

$$Acc = 0.5 \times (\frac{n_p}{N_p} + \frac{n_n}{N_n}) \tag{4}$$

where n_p and n_n denote the numbers true positive and true negative samples, while N_p and N_n are the numbers of positive and negative samples.

TABLE IV
COMPARISON OF BALANCED ACCURACY (%) OF EMOTION CLASSIFICATION

	Neutral	Happy	Sadness	Angry	Fear	Disgust	Surprise	Overall
RGB (weighted loss)	25.73	19.61	7.32	0	30.15	5.35	26.32	20.59
RGB (over-sampling)	27.13	24.49	7.39	1.60	27.98	15.30	25.21	22.00
RGB + Depth	25.34	23.07	7.27	1.60	28.59	8.89	25.21	20.69
RGB + ECG	22.18	14.48	20.55	13.44	32.70	8.84	25.08	19.58
OMG + EEG	20.56	22.14	**20.72**	**55.96**	**30.17**	**27.13**	25.95	26.77
RGB + EEG + OMG	**31.04**	**37.56**	20.36	25.29	28.52	22.34	**26.82**	**29.24**

TABLE V
COMPARISON OF RMS EVALUATION OF VA

	Valence	Arousal
RGB	0.212	0.355
RGB + Depth	0.354	0.411
RGB + OMG	0.226	0.347
RGB + EEG	**0.202**	**0.312**
RGB + ECG	0.234	0.326

For the evaluation evaluation of VA estimation, we measure the results by calculating Root Mean Square (RMS) differences:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{5}$$

where x and y are the predicted value and groundtruth, respectively, and N is the number of testing samplings.

B. Results

Table IV shows the evaluation result of emotion classification comparing the 6 methods based on different feature fusion. We firstly compare the two methods for solving the class imbalance problem based on the RGB feature alone. It can be seen that the accuracy of the minority classes (sadness, angry and disgust) are still very low. For the weighted loss function-based method, the accuracy on a minority class (fear) can even become zero, which suggests that the model completely ignores this class. Compared to the weighted loss function-based method, the oversampling method is obviously a better solution. We use oversampling strategy as the default option for solving the problem of imbalanced classes.

Compared with the single modality method (using RGB feature alone), there are two fusion strategies that perform inferior to the single modality method. The results show that depth maps and ECG signals do not provide useful features for emotion recognition. For the depth map data, we have set the Kinect to capture the depth map in a high frame rate (10 fps), which leads to poor performance on depth data acquisition. Although the depth features degrade the recognition rate, the depth map is still useful as it directly contributes to fast and reliable face detection, as described in section IV-C. For the ECG signals, the result indicates that ECG may not be a good feature for basic emotion classification.

The fusion of OMG and EEG features achieves remarkable results. The accuracy of fear is outstanding even though it is a minority class. The performance of negative emotions (sadness, angry ,fear, disgust) is also impressive. The best performance is achieved by the fusion of RGB, EEG and OMG features with over 7% superior accuracy than the RBG modality and 2% superior to the fusion of OMG and EEG. Generally, this result demonstrates that genuine emotions sometimes cannot be reflected by the visual information, whilst they can be better detected by the OMG and EEG signals.

Table V shows the results of VA estimation. The depth features still harm the feature-fusion method. The EEG feature achieves superior performance on both valence and arousal prediction but inferior performance on valence estimation. The VA levels are inherently difficult to be annotated. Due to the lack of agreement on how to determine the VA levels, the participants just annotate the VA values according to their own understandings. Therefore, the annotations of VA are not so convincing as emotional classes. However, the results of VA estimation still show that the physiological signals have great significance on emotion prediction.

VI. CONCLUSION

In this paper, we have presented a new MGEED database for multimodal genuine emotion recognition. This database consists of EEG, OMG, ECG and RGB-D signals of human participants. MGEED is the first public database containing OMG signals. The non-contact OMG sensors enable more naturalistic and high-resolution OMG data acquisition. With the MGEED database, a baseline method is developed for signal preprocessing, synchronisation, feature extraction, multimodal feature fusion and multi-dimensional emotion recognition. The data analysis results demonstrate the effectiveness of physiological signals in the emotion recognition task. In the future, we plan to extend the MGEED database by introducing more physiological signals and recruiting more participants to increase the scale of this database.

ACKNOWLEDGMENT

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) through the Project 4D Facial Sensing and Modelling under Grant EP/N025849/1.

REFERENCES

[1] M. B. Mariappan, M. Suk, and B. Prabhakaran, “Facefetch: A user emotion driven multimedia content recommendation system based on facial expression recognition,” in *2012 IEEE International Symposium on Multimedia*. IEEE, 2012, pp. 84–87.

[2] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, “Corpus development for affective video indexing,” *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.

[3] S. D. Pollak and D. J. Kistler, “Early experience is associated with the development of categorical representations for facial expressions of emotion,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 13, pp. 9072–9076, 2002.

[4] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, “Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit,” in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 3723–3726.

[5] M. Chen, Y. Jiang, Y. Cao, and A. Y. Zomaya, "Creativebioman: a brain- and body-wearable, computing-based, creative gaming system," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 1, pp. 14–22, 2020.

[6] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2011.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[8] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 25–32.

[9] Z. Guoying, L. Yante, and Q. Xu, "From emotion ai to cognitive ai," *International Journal of Network Dynamics and Intelligence*, vol. 1, no. 1, pp. 65–72, 2023.

[10] Y. Zhang, B. Wu, W. Dong, Z. Li, W. Liu, B.-G. Hu, and Q. Ji, "Joint representation and estimator learning for facial action unit intensity estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3457–3466.

[11] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-workshops*. IEEE, 2010, pp. 94–101.

[13] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[14] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2016.

[15] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[16] D. Kollias and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3652–3660.

[17] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[18] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2015.

[19] J. Kwon, D.-H. Kim, W. Park, and L. Kim, "A wearable device for emotional recognition using facial expression and physiological response," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 5765–5768.

[20] T.-H. Li, W. Liu, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from eeg and eye movement signals: Discrimination ability and stability over time," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 607–610.

[21] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438–3446.

[22] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, p. 293, 2020.

[23] J. C. Yau, B. Girault, T. Feng, K. Mundnich, A. Nadarajan, B. M. Booth, E. Ferrara, K. Lerman, E. Hsieh, and S. Narayanan, "Tiles-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit," *Scientific Data*, vol. 9, no. 1, p. 536, 2022.

[24] S. Saganowski, J. Komoszyńska, M. Behnke, B. Perz, D. Kunc, B. Klich, Ł. D. Kaczmarek, and P. Kazienko, "Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables," *Scientific Data*, vol. 9, no. 1, p. 158, 2022.

[25] H. H. Muhammed and J. Raghavendra, "Optomyography (omg): A novel technique for the detection of muscle surface displacement using photoelectric sensors," *measurements*, vol. 10, p. 13, 2015.

[26] C. Shan, S. Gong, and P. W. McOwan, "Recognizing facial expressions at low resolution," in *IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2005, pp. 330–335.

[27] J. Lou, X. Cai, J. Dong, and H. Yu, "Real-time 3d facial tracking via cascaded compositional learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 3844–3857, 2021.

[28] Y. Xia, H. Yu, X. Wang, M. Jian, and F.-Y. Wang, "Relation-aware facial expression recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1143–1154, 2021.

[29] Y. Wang, X. Dong, G. Li, J. Dong, and H. Yu, "Cascade regression-based face frontalization for dynamic facial expression analysis," *Cognitive Computation*, pp. 1–14, 2021.

[30] M. Valstar, M. Pantic *et al.*, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. Paris, France., 2010, p. 65.

[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[32] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[33] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[34] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[35] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 58–65.

[36] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[37] S. Li, W. Deng, and J. Du, "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*, 2017, pp. 2852–2861.

[38] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "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*, 2016, pp. 5562–5570.

[39] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2106–2112.

[40] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 524–528.

[41] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 189–204, 2014.

[42] Y. Wang, H. Yu, J. Dong, B. Stevens, and H. Liu, "Facial expression-aware face frontalization," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 375–388.

[43] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749–1756.

[44] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 433–436.

[45] K. Li and Q. Zhao, "If-gan: Generative adversarial network for identity preserving facial image inpainting and frontalization," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 45–52.

[46] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition with atlas construction and sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 1977–1992, 2016.

[47] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5800–5809.

[48] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[49] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 637–643.

[50] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 211–216.

[51] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[52] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[53] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[54] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.

[55] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[56] S. Schneegass, B. Pfleging, N. Broy, F. Heinrich, and A. Schmidt, "A data set of real world driving to assess driver workload," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2013, pp. 150–157.

[57] X. Li, X. Zhang, H. Yang, W. Duan, W. Dai, and L. Yin, "An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 336–343.

[58] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[59] M. A. Bell and K. Cuevas, "Using eeg to study cognitive development: Issues and practices," *Journal of Cognition and Development*, vol. 13, no. 3, pp. 281–294, 2012.

[60] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Real-time eeg-based emotion monitoring using stable features," *The Visual Computer*, vol. 32, no. 3, pp. 347–358, 2016.

[61] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[62] S. Wu, X. Xu, L. Shu, and B. Hu, "Estimation of valence of emotion using two frontal eeg channels," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 1127–1130.

[63] R. Qiao, C. Qing, T. Zhang, X. Xing, and X. Xu, "A novel deep-learning based framework for multi-subject emotion recognition," in *2017 4th International Conference on Information, Cybernetics and Computational Social Systems (ICCSS)*. IEEE, 2017, pp. 181–185.

[64] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, 2017.

[65] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from facial emg signals using higher order statistics and principal component analysis," *Journal of the Chinese Institute of Engineers*, vol. 37, no. 3, pp. 385–394, 2014.

[66] B. Cheng and G. Liu, "Emotion recognition from surface emg signal using wavelet transform and neural network," in *Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering (ICBBE)*, 2008, pp. 1363–1366.

[67] O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 298–310, 2012.

[68] H.-W. Guo, Y.-S. Huang, C.-H. Lin, J.-C. Chien, K. Haraikawa, and J.-S. Shieh, "Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2016, pp. 274–277.

[69] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2011.

[70] B. Hwang, J. You, T. Vaessen, I. Myin-Germeys, C. Park, and B.-T. Zhang, "Deep ecgnet: An optimal deep learning framework for monitoring mental stress using ultra short-term ecg signals," *Telemedicine and e-Health*, vol. 24, no. 10, pp. 753–772, 2018.

[71] Z. Cheng, L. Shu, J. Xie, and C. P. Chen, "A novel ecg-based real-time detection method of negative emotions in wearable applications," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*. IEEE, 2017, pp. 296–301.

[72] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ecg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85–99, 2017.

[73] A. Goshvarpour, A. Abbasi, and A. Goshvarpour, "An accurate emotion recognition system using ecg and gsr signals and matching pursuit method," *Biomedical Journal*, vol. 40, no. 6, pp. 355–368, 2017.

[74] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics," *Scientific Reports*, vol. 4, no. 1, pp. 1–13, 2014.

[75] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," *arXiv preprint arXiv:1811.08854*, 2018.

[76] J. Lee and S. K. Yoo, "Recognition of negative emotion using long short-term memory with bio-signal feature compression," *Sensors*, vol. 20, no. 2, p. 573, 2020.

[77] G. Wu, G. Liu, and M. Hao, "The analysis of emotion recognition from gsr based on pso," in *2010 International Symposium on Intelligence Information Processing and Trusted Computing*. IEEE, 2010, pp. 360–363.

[78] C. A. Torres, Á. A. Orozco, and M. A. Álvarez, "Feature selection for multimodal emotion recognition in the arousal-valence space," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 4330–4333.

[79] S. Hinduja, S. Canavan, and G. Kaur, "Multimodal fusion of physiological signals and facial action units for pain recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 577–581.

[80] D. Fabiano and S. Canavan, "Emotion recognition using fused physiological signals," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 42–48.

[81] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–8.

[82] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[83] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu, "Realistic facial expression reconstruction for vr hmd users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 730–743, 2019.

[84] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[85] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[86] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[87] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[88] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

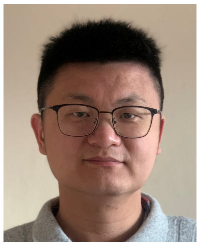
[89] J. Fu and Y. Rui, "Advances in deep learning approaches for image tagging," *APSIPA Transactions on Signal and Information Processing*, vol. 6, p. e11, 2017.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TAFFC.2023.3286351 IEEE IEEE Transactions on Affective Computing

[90] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 270–279.

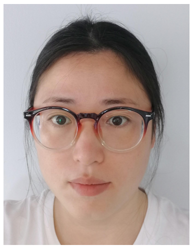
[91] K. R. M. Fernando and C. P. Tsokos, "Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 2940–2951, 2021.

[92] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

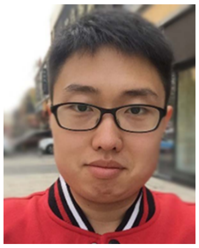


Yiming Wang is a research assistant with the Computer Science Department, Brunel University London, U.K. He completed his Ph.D. in the field of computer vision, from University of Portsmouth in 2018. His research interests include machine learning, computer vision and human machine interaction.

Hui Yu (Senior Member, IEEE) is currently a Professor with the University of Portsmouth, U.K. He worked at the University of Glasgow and Queen's University Belfast before joining the University of Portsmouth in 2012. His research interests include methods and practical development in visual computing, machine learning and AI with the applications focusing on human–machine interaction, virtual/augmented reality, robotics, as well as 4D facial expression generation, perception, and analysis. He serves as an Associate Editor for *IEEE Transactions on Human-Machine systems*, *IEEE Transactions on Computational Social Systems* and *Neurocomputing* journal.



Weihong Gao is a research associate at the University of Portsmouth, U.K. Her research interests include computer vision, sensor technology and facial expression analysis.



Yifan Xia received the M.Sc. degree from the Ocean University of China in 2017, and the Ph.D. degree from the University of Portsmouth, Portsmouth, U.K., in 2021. He is currently a Research Associate with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China. His research interests include facial expression analysis, computer vision and deep learning.

Charles Nduka is a fully accredited specialist in plastic surgery. He co-founded Emteq Labs, to develop a novel facial expression and emotion sensing platform with wide potential applications. His research interests include plastic and reconstructive surgery, facial paralysis, facial movement tracking, sensor technology, surgical technology.