# Artificial Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions

Asim Jan, Hongying Meng, *Senior Member, IEEE*, Yona Falinie Binti A. Gaus, *Student Member, IEEE*, and Fan Zhang

***Abstract*—A human being's cognitive system can be simulated by artificial intelligent systems. Machines and robots equipped with cognitive capability can automatically recognize a humans mental state through their gestures and facial expressions. In this paper, an artificial intelligent system is proposed to monitor depression. It can predict the scales of Beck depression inventory II (BDI-II) from vocal and visual expressions. First, different visual features are extracted from facial expression images. Deep learning method is utilized to extract key visual features from the facial expression frames. Second, spectral low-level descriptors and mel-frequency cepstral coefficients features are extracted from short audio segments to capture the vocal expressions. Third, feature dynamic history histogram (FDHH) is proposed to capture the temporal movement on the feature space. Finally, these FDHH and audio features are fused using regression techniques for the prediction of the BDI-II scales. The proposed method has been tested on the public Audio/Visual Emotion Challenges 2014 dataset as it is tuned to be more focused on the study of depression. The results outperform all the other existing methods on the same dataset.**

***Index Terms*—Artificial system, Beck depression inventory (BDI), deep learning, depression, facial expression, regression, vocal expression.**

## I. Introduction

MENTAL health issues such as depression have been linked to deficits of cognitive control. It affects one in four citizens of working age, which can cause significant losses and burdens to the economic, social, educational, as well as justice systems [1], [2]. Depression is defined as *"a common mental disorder that presents with depressed mood, loss of interest or pleasure, decreased energy, feelings of guilt or low self-worth, disturbed sleep or appetite, and poor concentration"* [3].

Among all psychiatric disorders, major depressive disorder (MDD) commonly occurs and heavily threatens the mental health of human beings. 7.5% of all people with disabilities suffer from depression, making it the largest contributor [4], exceeding 300M people. Recent study [5] indicates that having a low income shows an increased chance of having MDDs. It can also affect the major stages in life such as educational attainment and the timing of marriage. According to [6], majority of the people that obtain treatment for depression do not recover from it. The illness still remains with the person. This may be in the form of insomnia, excessive sleeping, fatigue, loss of energy, or digestive problems.

Artificial intelligence and mathematical modeling techniques are being progressively introduced in mental health research to try and solve this matter. The mental health area can benefit from these techniques, as they understand the importance of obtaining detailed information to characterize the different psychiatric disorders [7]. Emotion analysis has shown to been an effective research approach for modeling depressive states. Recent artificial modeling and methods of automatic emotion analysis for depression related issues are extensive [8]–[11]. They demonstrate that depression analysis is a task that can be tackled in the computer vision field, with machine-based automatic early detection and recognition of depression is expected to advance clinical care quality and fundamentally reduce its potential harm in real life.

The face can effectively communicate emotions to other people through the use of facial expressions. Psychologists have modeled these expressions in detail creating a dictionary called the facial action coding system (FACS). It contains the combination of facial muscles for each expression [12], and can be used as a tool to detect the emotional state of a person through their face. Another approach to classify emotion through facial expressions is using local and holistic feature descriptors, such as in [13]. Unlike FACS, these techniques treat the whole face the same and look for patterns throughout, and not just for certain muscles. However, the depression disorder is not limited to be expressed by the face. The perception of emotional body movements and gestures has shown it can be observed through a series of controlled experiments using patients with and without MDD [14]. Furthermore, electroencephalogram (EEG) signals and brain activity using magnetic resonance imaging, are modalities recent to computer vision [15], [16]. Electrocardiogram (ECG) and electro-dermal

Manuscript received January 14, 2017; revised April 28, 2017 and June 9, 2017; accepted June 13, 2017. Date of publication July 31, 2017; date of current version September 7, 2018. The work of A. Jan was supported in part by the School of Engineering and Design, and in part by the Thomas Gerald Gray PGR Scholarship. The work of H. Meng was supported by the Joint Project through the Royal Society of U.K. and the National Natural Science Foundation of China under Grant IE160946. The work of Y. F. B. A. Gaus was supported by the Majlis Amanah Rakyat Scholarship. *(Corresponding author: Asim Jan.)*

The authors are with the Department of Electronic and Computer Engineering, Brunel University London UB8 3PH, U.K. (e-mail: asim.jan@brunel.ac.uk; hongying.meng@brunel.ac.uk; yonafalinie.abdgaus@brunel.ac.uk; fan.zhang@brunel.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2017.2721552

activity are also considered for depression analysis alongside the audio-visual modality [17].

All of this research is evidenced by the series of International Audio/Visual Emotion Recognition Challenges (AVEC2013 [1], AVEC2014 [18], and most recently AVEC2016 [17]). Each challenge provides a dataset that has rich video content containing subjects that suffer from depression. Samples consist of visual and vocal data, where the facial expressions and emotions through the voice have been captured carefully from the cognitive perspective. The objective is to communicate and interpret emotions through expressions using multiple modalities. Various methods have been proposed for depression analysis [11], [19], [20], including most recent works from AVEC2016 [21]–[23].

In order to create a practical and efficient artificial system for depression recognition, visual and vocal data are key as they are easily obtainable for a system using a camera and microphone. This is a convenient data collection approach when compared to data collection approaches that requires sensors to be physically attached to the subject, such as EEG and ECG data. For machines and systems in a noncontrolled environment, obtaining EEG and ECG can therefore be difficult to obtain. The depression data from the AVEC2013 and AVEC2014 datasets provide both visual and vocal raw data. However, AVEC2016 provides the raw vocal data but no raw visual data, for ethical reasons. Instead is provided a set of different features obtained from the visual data by the host. For this reason, the AVEC2014 dataset has been chosen in order to run experiments using raw visual and vocal data.

Deep learning is also a research topic that has been adopted toward visual modality, especially in the form of a convolutional neural network (CNN). It has significantly taken off from its first big discovery for hand digit recognition [24]. Recently, the effectiveness of deep networks have been portrayed in different tasks such as face identification [25], image detection, segmentation and classification [26], [27], and many other tasks. The majority of these applications have only become achievable due to the processing movement from CPU to GPU. The GPU is able to provide a significantly higher amount of computational resources versus a CPU, to handle multiple complex tasks in a shorter amount of time. Deep networks can become very large and contain millions of parameters, which was a major setback in the past. Now there are a variety of deep networks available such as AlexNet [28] and the VGG networks [29]. These networks have been trained with millions of images based on their applications, and are widely used today as pretrained networks.

Pretrained CNNs can be exploited for artificial image processing on depression analysis, mainly using the visual modality. However, the pretrained CNN models such as VGG-Face provide good features at the frame level of videos, as they are designed for still images. In order to adapt this across temporal data, a novel technique called feature dynamic history histogram (FDHH) is proposed to capture the dynamic temporal movement on the deep feature space. Then partial least square (PLS) and linear regression (LR) algorithms are used to model the mapping between dynamic features and the depression scales. Finally, predictions from both video and audio modalities are combined at the prediction level. Experimental results achieved on the AVEC2014 dataset illustrates the effectiveness of the proposed method.

The aim of this paper is to build an artificial intelligent system that can automatically predict the depression level from a user's visual and vocal expression. The system is understood to apply some basic concepts of how parts of the human brain works. This can be applied in robots or machines to provide human cognitive like capabilities, making intelligent human–machine applications.

The main contribution of the proposed framework are as follows:

1) a framework architecture is proposed for automatic depression scale prediction that includes frame/segment level feature extraction, dynamic feature generation, feature dimension reduction, and regression;
2) various features, including deep features, are extracted on the frame-level that captured the better facial expression information;
3) a new feature (FDHH) is generated by observing dynamic variation patterns across the frame-level features;
4) advanced regressive techniques are used for regression.

The rest of this paper is organized as follows. Section II briefly reviews related work in this area. Section III provides a detailed description of the proposed method, and Section IV displays and discusses the experimental results on the AVEC2014 dataset [18]. Section V concludes this paper.

## II. Related Works

Recent years have witnessed an increase of research for clinical and mental health analysis from facial and vocal expressions [30]–[33]. There is a significant progress on emotion recognition from facial expressions. Wang *et al.* [30] proposed a computational approach to create probabilistic facial expression profiles for video data. To help automatically quantify emotional expression differences between patients with psychiatric disorders, (e.g., Schizophrenia) and healthy controls.

In depression analysis, Cohn *et al.* [34], who is a pioneer in the affective computing area, performed an experiment where he fused both visual and audio modality together in an attempt to incorporate behavioral observations, from which are strongly related to psychological disorders. Their findings suggest that building an automatic depression recognition system is possible, which will benefit clinical theory and practice. Yang *et al.* [31] explored variations in the vocal prosody of participants, and found moderate predictability of the depression scores based on a combination of $F_0$ and switching pauses. Girard *et al.* [33] analyzed both manual and automatic facial expressions during semistructured clinical interviews of clinically depressed patients. They concluded that participants with high symptom severity tend to express more emotions associated with contempt, and smile less. Yammine *et al.* [35] examined the effects caused by depression to younger patients of both genders. The samples ($n = 153$) completed the Beck depression inventory II (BDI-II) questionnaire which indicated that the mean BDI-II score of 20.7 (borderline clinically

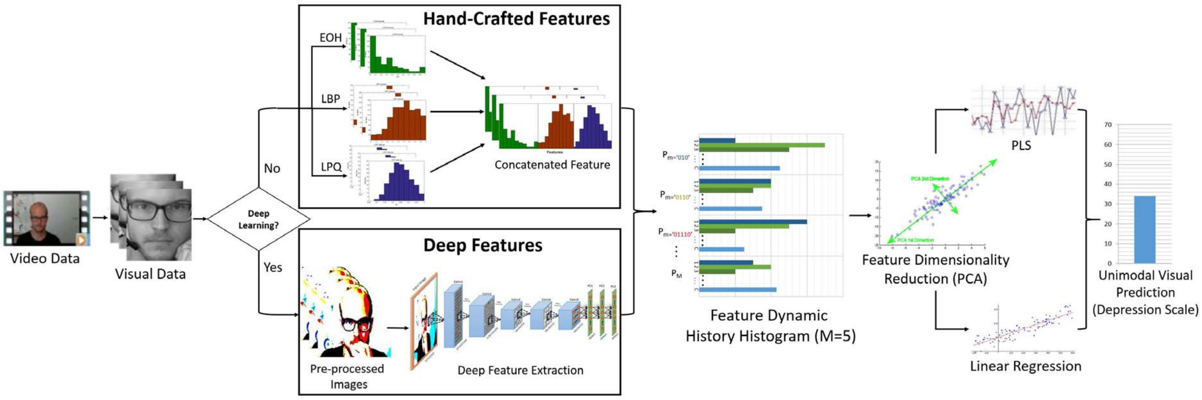

Fig. 1. Overview of the proposed automatic depression scale recognition system from facial expressions. The video data is broken down into visual frames. If deep learning is utilized, then deep features are extracted from the frames, otherwise a set of hand-crafted features are extracted. This is followed by FDHH to produce a dynamic descriptor. The dimensionality is reduced and a fusion of PLS and LR is used to predict the depression scale.

depressed), from the patients that were feeling depressed in the prior year. Scherer *et al.* [32] studied the correlation between the properties of gaze, head pose, and smile of three mental disorders (i.e., depression, post-traumatic stress disorder, and anxiety). They discovered that there was a distinct difference between the highest and lowest distressed participants, in terms of automatically detected behaviors.

The depression recognition subchallenge of AVEC2013 [1] and AVEC2014 [18]; had proposed some good methods which achieved good results [10], [11], [19], [36]–[44]. From this, Williamson *et al.* [19], [20] were the winner of the depression subchallenge (DSC) for the AVEC2013 and AVEC2014 competitions. In 2013, they exploited the effects that reflected changes in coordination of vocal tract motion associated with MDD. Specifically, they investigated changes in correlation that occur at different time scales across dormant frequencies and also across channels of the delta-mel-cepstrum [19]. In 2014, they looked at the change in motor control that can effect the mechanisms for controlling speech production and facial expression. They derived a multiscale correlation structure and timing feature from vocal data. Based on these two feature sets, they designed a novel Gaussian mixture model-based multivariate regression scheme. They referred this as a Gaussian staircase regression, that provided very good prediction on the standard Beck depression rating scale.

Meng *et al.* [11] modeled the visual and vocal cues for depression analysis. Motion history histogram (MHH) is used to capture dynamics across the visual data, which is then fused with audio features. PLS regression utilizes these features to predict the scales of depression. Gupta *et al.* [42] had adopted multiple modalities to predict affect and depression recognition. They fused together various features such as local binary pattern (LBP) and head motion from the visual modality, spectral shape, and mel-frequency cepstral coefficients (MFCCs) from the audio modality and generating lexicon from the linguistics modality. They also included the baseline features local Gabor binary patterns—three orthogonal planes (LGBP-TOP) [18] provided by the hosts. They then apply a selective feature extraction approach and train a support vector regression machine to predict the depression scales.

Kaya *et al.* [41] used LGBP-TOP on separate facial regions with local phase quantization (LPQ) on the inner-face. Correlated component analysis and Moore–Penrose generalized inverse were utilized for regression in a multimodal framework. Jain *et al.* [44] proposed using Fisher vector to encode the LBP-TOP and dense trajectories visual features, and low-level descriptor (LLD) audio features. Pérez Espinosa *et al.* [39] claimed; after observing the video samples; that subjects with higher BDI-II showed slower movements. They used a multimodal approach to seek motion and velocity information that occurs on the facial region, as well as 12 attributes obtained from the audio data such as "*number of silence intervals greater than 10 s and less than 20 s*" and "*percentage of total voice time classified as happiness*."

The above methods have achieved good performance. However, for the visual feature extraction, they used methods that only consider the texture, surface, and edge information. Recently, deep learning techniques have made significant progress on visual object recognition, using deep neural networks that simulate the humans vision-processing procedure that occurs in the mind. These neural networks can provide global visual features that describe the content of the facial expression. Recently, Chao *et al.* [43] proposed using multitask learning based on audio and visual data. They used long short-term memory modules with features extracted from a pretrained CNN, where the CNN was trained on a small facial expression dataset FER2013 by Kaggle. This dataset contained a total of 35 886 48 × 48 grayscale images. The performance they achieved is better than most other competitors from the AVEC2014 competition, however, it is still far away from the state-of-the-art. A few drawbacks of their approach are the image size they adopted is very small, which would result in downsizing the AVEC images and reducing a significant amount of spatial information. This can have a negative impact as the expressions they wish to seek are very subtle, small and slow. They also reduce the color channels to grayscale, further removing useful information.

In this paper, we are targeting an artificial intelligent system that can achieve the best performance on depression level prediction, in comparison with all the existing methods on

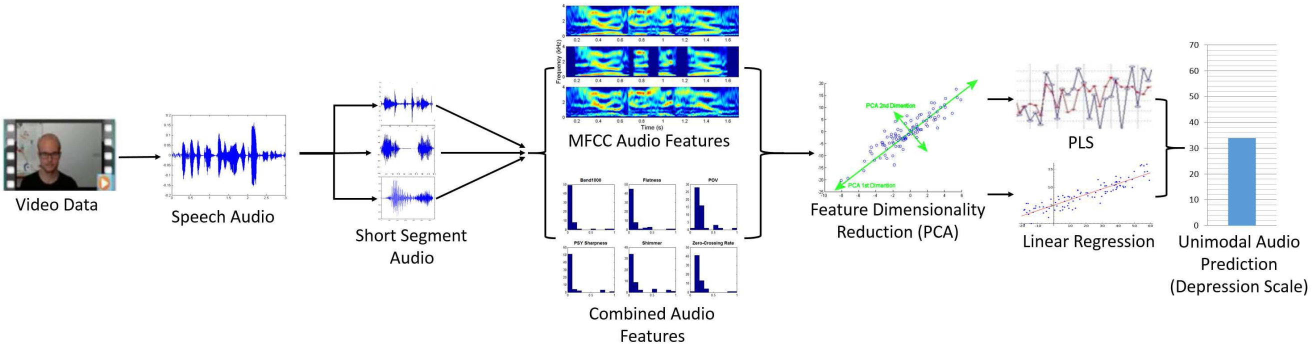

Fig. 2. Overview of the proposed automatic depression scale recognition system from vocal expressions. The speech audio is extracted from the video data, where short segments are produced. Then a bunch of audio features are extracted from each segment and averaged. These are reduced in dimensionality and a fusion of PLS and LR is used to predict the depression scale.

the AVEC2014 dataset. Improvement will be made on the previous work from feature extraction by using deep learning, regression with fusion, and build a complete system for automatic depression level prediction from both vocal and visual expressions.

## III. Framework

Human facial expressions and voices in depression are theoretically different from those under normal mental states. An attempt to find a solution for depression scale prediction is achieved by combining dynamic descriptions within naturalistic facial and vocal expressions. A novel method is developed that comprehensively models the variations in visual and vocal cues, to automatically predict the BDI-II scale of depression. The proposed framework is an extension of the previous method [10] by replacing the hand-crafted techniques with deep face representations as a base feature to the system.

### A. System Overview

Fig. 1 illustrates the process of how the features are extracted from the visual data using either deep learning or a group of hand-crafted techniques. Dynamic pattern variations are captured across the feature vector, which is reduced in dimensionality and used with regression techniques for depression analysis. Fig. 2 follows a similar architecture as Fig. 1, but is based on audio data. The audio is split into segments and two sets of features are extracted from these segments. Then one of these sets are reduced in dimensionality and used with regression techniques to predict the depression scale.

For the deep feature process, the temporal data for each sample is broken down into static image frames which are preprocessed by scaling and subtracting the given mean image. These are propagated forward into the deep network for high level feature extraction. Once the deep features are extracted for a video sample, it is rank normalized between 0 and 1 before the FDHH algorithm is applied across each set of features per video. The output is transformed into a single row vector, which will represent the temporal feature of one video.

Both frameworks are unimodal approaches. The efforts are combined at feature level by concatenating the features produced by each framework just before principal component analysis (PCA) is applied. This gives a bimodal feature vector, which is reduced in dimensionality using PCA and is rank normalized again between 0 and 1. It is applied with a weighted sum rule fusion of regression techniques at prediction level, to give the BDI-II prediction.

### B. Visual Feature Extraction

This section looks at the different techniques and algorithms used to extract visual features from the data.

*1) Hand-Crafted Image Feature Extraction:* Previously [10], the approach was based on investing in hand-crafted techniques to represent the base features. These were applied on each frame, similar to the deep face representation, with three different texture features LBP; edge orientation histogram (EOH) and LPQ.

LBP looks for patterns of every pixel compared to its surrounding 8 pixels [45]. This has been a robust and effective method used in many applications including face recognition [46]. EOH is a technique similar to histogram of oriented gradients [47], using edge detection to capture the shape information of an image. Applications include hand gesture recognition [48], object tracking [49], and facial expression recognition [50]. LPQ investigates the frequency domain, where an image is divided into blocks where discrete Fourier transform is applied on top to extract local phase information. This technique has been applied for face and texture classification [51].

*2) Architectures for Deep Face Representation:* In this section, different pretrained CNN models are introduced, detailing the architectures and its designated application. Two models are then selected to be testing within the system for the experiments.

*3) VGG-Face:* Visual Geometry Group have created a few pretrained deep models, including their *very deep networks*. These networks are VGG-S, VGG-F, and VGG-M [52] networks which represent slow, fast, and medium, respectively. VGG-D and VGG-E are their very deep networks, VGG-D containing 16 convolutional layers and VGG-E containing 19 [29]. These networks are pretrained based on the ImageNet dataset for the object classification task [53]. VGG-Face is a network which they train on 2.6M facial images for the application of face recognition [25]. This network is more suited

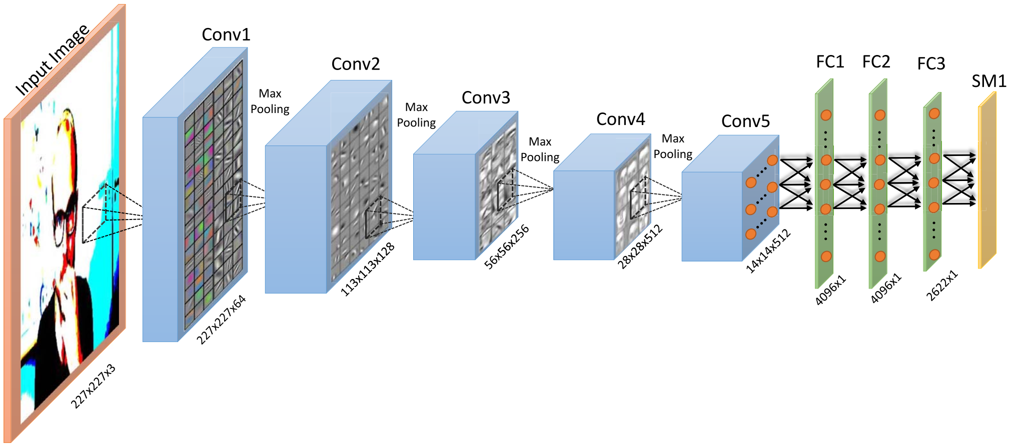

Fig. 3. VGG-Face architecture visualizing low to high-level features captured as a facial expression is propagated through-out the network, stating the dimensions produced by each layer.

for the depression analysis task as it is trained mainly on facial images, as opposed to objects from the ImageNet dataset.

The VGG-Face [25] pretrained CNN contains a total of 36 layers, where 16 are convolution layers and 3 are fully connected layers. The filters have a fixed kernel size of 3×3 and as the layers increase, so does the filter depth which varies from 64 to 512. The fully connected layers are of 1×1 kernel and have a depth of 4096 dimensions, with the last layer having 2622. The remaining 20 are a mixture of rectified linear activation layers and max pooling layers, with a softmax layer at the end for probabilistic prediction. The full architecture is shown in Fig. 3, along with how the high and low level features look like throughout the network. It can be seen how certain filter responses are activated to produce edges and blobs that over the network they combine to remake the image.

This network is designed to recognize a given face between 2622 learned faces, hence the 2622 filter responses in the softmax layer. However, the task is more to observe the facial features that are learned by the convolutional layers. The early to later convolution layers (Conv1–Conv5) contain spatial features from edges and blobs, to textures and facial parts, respectively. Their filter responses can be too big to be used directly as a feature vector to represent faces. Therefore, the fully connected layers are looked upon to obtain a plausible feature vector to describe the whole input facial image. In these experiments, the feature at the three fully connected layers FC1–FC3 were acquired and used.

*4) AlexNet:* AlexNet, created by Krizhevsky *et al.* [28], is another popular network, which was one of the first successful deep networks used in the ImageNet challenge [53]. This pretrained CNN contains 21 layers in total. The architecture of AlexNet varies from the VGG-Face network in terms of the depth of the network and the convolution filter sizes. The targeted layers for this experiment are 16 and 18, which are represented as FC1 and FC2, respectively. This network is designed for recognizing up to 1000 various objects, which may result in unsuitable features when applied with facial images. However, it will be interesting to see how it performs against VGG-Face, a network designed specifically for faces.

### C. Audio Feature Extraction

For audio features, the descriptors are derived from the set provided by the host of the AVEC2014 challenge. They include spectral LLDs and MFCCs 11–16. There are a total of 2268 features, with more details in [18]. These features are further investigated to select the most dominant set by comparing the performance with the provided audio baseline result. The process includes testing each individual feature vector with the development dataset, where the top eight performing descriptors are kept. Then, each descriptor is paired with every other in a thorough test to find the best combination. This showed Flatness, Band1000, PSY Sharpness, POV, Shimmer, and ZCR to be the best combination, with MFCC being the best individual descriptor. Fig. 2 shows the full architecture using the selected audio features, where two paths are available, either selecting the MFCC feature or the combined features.

### D. Feature Dynamic History Histogram

MHH is a descriptive temporal template of motion for visual motion recognition. It was originally proposed and applied for human action recognition [54]. The detailed information can be found in [55] and [56]. It records the grayscale value changes for each pixel in the video. In comparison with other well-known motion features, such as motion history image [57], it contains more dynamic information of the pixels and provides better performance in human action recognition [55]. MHH not only provides rich motion information, but also remains computationally inexpensive [56].

MHH normally consists of capturing motion data of each pixel from a string of 2-D images. Here, a technique is proposed to capture dynamic variation that occurs within mathematical representations of a visual sequence. Handcrafted descriptors such as EOH, LBP, and LPQ model the mathematical representations from the still images, which can be interpreted as a better representation of the image. Furthermore, fusion of these technical features can provide a combination of several mathematical representations, improving the feature as demonstrated in [13]. Several techniques

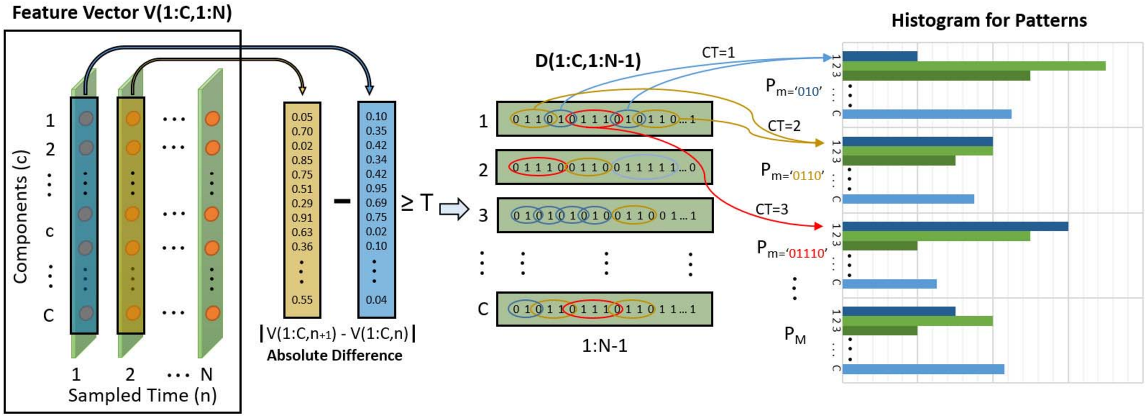

Fig. 4. Visual process of computing FDHH on the sequence of feature vectors. The first step is to obtain a new binary vector representation based on the absolute difference between each time sample. From this, binary patterns $P_m$ are observed throughout each components of the binary vector and a histogram is produced for each pattern.

have been proposed to move these descriptors into the temporal domain in [58]–[61]. They simply apply the hand-crafted descriptors in three spatial directions, as they are specifically designed for spatial tasks. This ideally extends the techniques spatially in different directions rather than dynamically taking the time domain into account.

A solution was proposed to obtain the benefits of using hand-crafted techniques on the spatial images, along with applying the principals of temporal-based motion techniques. This was achieved by capturing the motion patterns in terms of dynamic variations across the feature space. This involves extracting the changes on each component in a feature vector sequence (instead of one pixel from an image sequence), so the dynamic of facial/object movements are replaced by the feature movements. Pattern occurrences are observed in these variations, from which histograms are created. Fig. 4 shows the process of computing FDHH on the sequence of feature vector, the algorithm for FDHH can be implemented as follows.

Let $\{V(c,n), c = 1, \ldots, C, n = 1, \ldots, N\}$ be a feature vector with $C$ components and $N$ frames, and a binary sequence $\{D(c,n), c = 1, \ldots, C, n = 1, \ldots, N-1\}$ of feature component $c$ is generated by comprising and thresholding the absolute difference between consecutive frames as shown in (1). $T$ is the threshold value determining if dynamic variation occurs within the feature vector. Given the parameter $M = 5$, we can define the pattern sequences $P_M$ as $P_m (1 \le m \le M)$, where $m$ represents how many consecutive "1"s are needed to create the pattern, as shown in Fig. 4. The final dynamic feature can be represented as $\{\text{FDHH}(c,m), c = 1, \ldots, C, m = 1, \ldots, M\}$

$$D(c,n) = \begin{cases} 1, & \text{if } \{|V(c,n+1) - V(c,n)| \ge T\} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Equation (1) shows the calculation for the binary sequence $D(c,n)$. The absolute difference is taken between the sample $n+1$ and $n$, which is then compared with a threshold to determine if the sequence should be a 1 or "0." A counter is then initialized to CT $= 0$, which is used to search for patterns of 1s in a sequence $D(1:C, 1:N-2)$

$$\text{CT} = \begin{cases} \text{CT}++, & \text{if } \{D(c,n+1) = 1\} \\ 0, & \text{a pattern } P_{1:M} \text{ found, reset CT} \end{cases} \quad (2)$$

$$\text{FDHH}(c,m) = \begin{cases} \text{FDHH}(c,m) + 1, & \text{if } \{P_m \text{ is found}\} \\ \text{FDHH}(c,m), & \text{otherwise.} \end{cases} \quad (3)$$

When observing a component from a sequence $D(c, 1:N)$, a pattern of $P_m (1 \le m \le M)$ is detected by counting the number of consecutive 1s, where CT is updated as shown in (2). This continues to increment for every consecutive 1 until a 0 occurs within the sequence, and for this case the histogram FDHH is updated as shown in (3), followed by the counter CT being reset to 0.

Equation (3) shows the FDHH of pattern $m$ is increased when a pattern $P_m$ is found. This is repeated throughout the sequence for each component until all the FDHH histograms are created for the desired patterns $1:M$. There are two special cases that have been dealt with. These are the case where CT $= 0$ (consecutive 0s), none of the histograms are updated and where CT $> M$, the histogram for $P_M$ is incremented. The full algorithm can be seen in Fig. 5.

### *E. Feature Combination and Fusion*

Once the deep features are extracted, FDHH is applied on top to create $M$ feature variation patterns for each deep feature sequence. The resulting histograms provide a feature vector that contains the information of the dynamic variations that occur throughout the features. The audio features are then fused with the dynamic deep features by concatenating them together, producing one joint representational vector per sample, which is normalized between [0, 1]. For testing purposes, the normalization is based on the training set. Equation (4) shows how the features are ranked within the range of the training data

$$\hat{X}_i = \frac{X_i - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}} \quad (4)$$

```
Algorithm (FDHH)
Input: Feature Vector $V(c,n)$, component $c=1,...,C$, frame $n=1,...,N$
Initialisation: Patterns $m=1,...,M$,
                $FDHH(1:C,1:M) = 0$,
                $D(1:C,1:N-1) = 0$,
                $CT = 0$
For $n$=1 to $N$-1   (For N1)
    Compute $D(1:C,n) = |V(c,n+1) - V(c,n)| \geq T$ (Binary Output)
End   (For N1)
For $c$=1 to $C$   (For C)
    For $n$=1 to $N$-2   (For N2)
        If (D($c$,$n$+1)==1)   (If 1)
            Update: $CT$++
        ElseIf($CT>0$ & $CT\leq M$)
            Update: $FDHH(c,CT)$++
            Update: $CT$=0
        ElseIf($CT>M$)
            Update: $FDHH(c,M)$++
            Update: $CT$=0
        End   (If 1)
    End   (For N2)
End   (For C)
Output: $FDHH(1:C,1:M)$
```

Fig. 5. FDHH algorithm.

where $X_i$ is the training feature component, $\alpha_{\min}$ and $\alpha_{\max}$ are the training minimum and maximum feature vector values, and $\hat{X}_i$ is the normalized training feature component. PCA is performed on the fused feature vector to reduce the dimensionality and decorrelate the features. Doing this reduces the computational time and memory for some regression techniques such as LR. The variance within the data is kept high to about 94% by retaining dimensions $L = 49$. After PCA is applied, the feature is once again normalized between [0, 1] using the training data.

### F. Alternate Feature Extraction Approach

For comparison purposes, another approach (APP2) is undergone to apply the original MHH [54] on the visual sequences, that was used similarly in the previous AVEC2013 competition by Meng *et al.* [11]. Their approach has been extended here by using deep features. MHH is directly applied on the visual sequences to obtain $M$ motion patterns $\text{MHH}(u, v, m)$ and $\{u = 1, \ldots, U, v = 1, \ldots, V, m = 1, \ldots, M\}$, where $\{u, v\}$ are the frames for $1 : M$ motion patterns. The frames are treated as individual image inputs for the deep CNNs and are forward propagated until the softmax layer. This approach closely resembles the main approach to allow for fair testing when evaluating and comparing them together.

The 4096-dimensional features are extracted from similar layers to the main approach, resulting in a deep feature vector of $M \times 4096$ per video sample. These features are then transformed to a single vector row from which it is fused with the same audio features used in the main approach. They are then rank normalized between [0, 1] using the training data range before the dimensionality is reduced using PCA to $L = 49$, and finally the reduced feature vector is rank normalized again between [0, 1] using the training data range.

### G. Regression

There are two techniques adopted for regression. PLSs regression [62] is a statistical algorithm which constructs predictive models that generalize and manipulates features into a low-dimensional space. This is based on the analysis of relationship between observations and response variables. In its simplest form, a linear model specifies the linear relationship between a dependent (response) variable, and a set of predictor variables.

This method reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components, instead of on the original data. PLS regression is especially useful when the predictors are highly collinear, or when there are more predictors than observations and ordinary least-squares regression either produces coefficients with high standard errors or fails completely. PLS regression fits multiple response variables in a single model. PLS regression models the response variables in a multivariate way. This can produce results that can differ significantly from those calculated for the response variables individually. The best practice is to model multiple responses in a single PLS regression model only when they are correlated. The correlation between feature vector and depression labels is computed in the training set, with the model of PLS as

$$\begin{aligned} S &= KG^K + E \\ W &= UH^K + F \end{aligned} \tag{5}$$

where $S$ is an $a \times b$ matrix of predictors and $W$ is an $a \times g$ matrix of responses. $K$ and $U$ are two $n \times l$ matrices that are, projections of $S$ (scores, components or the factor matrix) and projections of $W$ (scores); $G$, $H$ are, respectively, $b \times l$ and $g \times l$ orthogonal loading matrices; and matrices $E$ and $F$ are the error terms, assumed to be independent and identical normal distribution. Decompositions of $S$ and $W$ are made so as to maximize the covariance of $K$ and $U$.

LR is another approach for modeling the relationship between a scalar dependent variable and one or more explanatory variables in statistics. It was also used in the system along with PLS regression for decision fusion. The prediction level fusion stage aims to combine multiple decisions into a single and consensus one [63]. The predictions from PLS and LR are combined using prediction level fusion based on the weighted sum rule.

## IV. Experimental Results

### A. AVEC2014 Dataset

The proposed approaches are evaluated on the AVEC2014 dataset [18], a subset of the audio-visual depressive language corpus. This dataset was chosen over the AVEC2013 dataset as it is a more focused study of affect on depression, using only 2 of the 14 related tasks from AVEC2013. The dataset contains 300 video clips with each person performing the two human–computer interaction tasks separately whilst being recorded by a web-cam and microphone in a number of quiet settings. Some subjects feature in more than one clip. All the participants are recorded between one and four times, with a period of two weeks between each recording. Eighteen subjects appear in three recordings, 31 in 2, and 34 in only one recording. The length of these clips are between 6 s to 4 min and 8 s. The mean age of subjects is 31.5 years, with a standard deviation of 12.3 years and a range of 18–63 years. The range of the BDI-II depression scale is [0, 63], where 0–10 is considered normal, as ups and downs; 11–16 is mild mood disturbance; 17–20 is borderline clinical depression; 21–30 is moderate depression; 31-40 is severe depression; and over 40 is extreme depression. The highest recorded score within the AVEC14 dataset is 45, which indicates there are subjects with extreme depression included.

### B. Experimental Setting

The experimental setup has been followed by the AVEC2014 guidelines which can be found in [18]. The instructions are followed as mentioned in the DSC, which is to predict the level of self-reported depression; as indicated by the BDI-II that ranges of from 0 to 63. This concludes to one continuous value for each video file. The results for each test are evaluated by its the mean absolute error (MAE) and root mean squared error (RMSE) against the ground-truth labels. There are three partitions to the dataset, these are training, development, and testing. Each partition contains 100 video clips, these are split 50 for "Northwind" and 50 for "Freeform." However, for the experiments the *Northwind* and *Freeform* videos count as a single sample, as each subject produces both videos with what should be the same depression level.

The MatConvNet [64] toolbox has been used to extract the deep features. This tool has been opted for the experiments as it allows full control over deep networks with access to data across any layer along with easy visualization. They also provide both AlexNet and VGG-FACE pretrained networks.

*1) Data Preprocessing:* In order to obtain the optimal features from the pretrained networks, a set of data preprocessing steps were followed, as applied by both Parkhi *et al.* [25] and Krizhevsky *et al.* [28] on their data. For each video, each frame was processed individually to extract its deep features. Using the meta information, the frames were resized to 227×227×3 representing the image height, width, and color channels. AlexNet has the same requirement of 227×227×3 image as an input to the network. The images were also converted into single precision as required by the MatConvNet toolbox, followed by subtracting the mean image provided by each network. The next stage was to propagate each preprocessed frame through the networks and obtain the features produced by the filter responses at the desired layers.

*2) Feature Extraction:* For each video clip, the spatial domain is used as the workspace for both approaches. With AlexNet, the 4096-dimensional feature vector is retained from the 16th and 18th fully connected layers. The decision to take the features at the 16th layer is in order to observe if the first 4096 dimension fully connected layer produces better features than the second (layer 18). For the VGG-Face network, the 4096-dimensional feature vectors are extracted at the 35th, 34th, and 32nd layers. The 34th layer is the output directly from the fully connected layer, the 35th is the output from the following rectified linear unit (ReLU) activation function layer and the 32nd layer is the output from the first fully connected layer.

The initial convolution layers are bypassed as the parameter and memory count would have been drastically higher if they were to be used as individual features. After observing the dimensions for AlexNet, there were around 70K versus 4096 when comparing the initial convolution layer versus the fully connected layers, and a staggering 802K versus 4096 for VGG-Face. The connectivity between filter responses are responsible for the dramatic decrease in dimensions at the fully connected layers. The fully connected layer is observed using (6), where $y_j$ is the output feature by taking the function $f(x)$ of the given input $x_i$ from the previous layer. The function calculates the sum over all inputs $x_i$ multiplied by each individual weight ($j = 1 : 4096$) of the fully connected layer plus the bias $b_j$

$$y_j = f\left(\sum_{i=1}^{m} x_i \cdot w_{i,j} + b_j\right). \tag{6}$$

The role of a ReLU layer can be described with (7), where $x_i$ is the input filter response and $y(x_i)$ is the output

$$y_j = \max(0, x_i). \tag{7}$$

For testing purposes, the decision to investigate the effects of a feature vector before and after a ReLU activation layer (layers 34 and 35) had been taken into account. As the activation function kills filter responses that are below 0, it was assumed that the resulting feature vector will become sparse with loss of information.

When extracting the dynamic variations across the deep features, the parameter $M$ is set to $M = 5$, capturing five binary patterns across the feature space. Based on a sample feature visualization of the binary pattern histograms, $M = 5$ was chosen as beyond this results to histograms with a low count. Given that the deep feature data ranges from [0, 1], the optimized threshold value for FDHH has been set to $1/255 = 0.00392$, after optimization on the training and development partitions. This will produce five resulting features with 4096 components each, making a total of feature dimension count of $5 \times 4096 = 20\,480$ per video sample. As there are two recordings per ground truth label, (*Northwind* and *Freeform*), the 20 480 features are extracted from both recordings and concatenated together to make a final visual feature vector of 40 960 dimensions.

The features that are extracted using AlexNet and FDHH are denoted as A16_FD and A18_FD, representing the deep features extracted from the 16th and 18th layer, respectively. For VGG-Face, the feature vectors are denoted as V32_FD, V34_FD, and V35_FD, representing the deep features extracted from the 32nd, 34th, and 35th layer, respectively.

Due to the nature of feature extractors used, it is difficult to pinpoint which parts of the face contributes the most. The movement of these facial parts play a big role in the system, and the FDHH algorithm is designed to pick up these facial movements that occur within the mathematical representations. This approach has been denoted as APP1. The whole system was tested on a Windows machine using MATLAB 2017a with an i7-6700K processor @ 4.3 GHz, and a Titan X (Pascal) GPU. For 6-s video clip, it will take less than 3.3 s to process.

*3) Alternate Approaches for Feature Extraction:* An Alternate approach, denoted as APP2, started by extracting MHH of each visual sequence, for both Northwind and Freeform. The parameter $M$ is set to $M = 6$ to capture low to high movement across the face, this results in six motion pattern frames. Each of these frames are then propagated through AlexNet and VGG-Face, however, as there are no color channel for the pattern frames, each frame is duplicated twice making three channels in total to imitate the color channels. The output of the CNNs will produce $6 \times 4096$ features, which is transformed into a single row to make 24 576 features, and 49 152 features when both Northwind and Freeform are concatenated. These features will be denoted as MH_A16, MH_A18, MH_V32, MH_V34, and MH_V35.

Previous research [10] worked in the spatial domain to produce local features using EOH, LBP and LPQ. These features are extracted frame by frame to produce 384-, 944-, and 256-dimensional histograms, respectively, for each frame. FDHH was used to capture the dynamic variations across the features to produce $M = 3$ vectors of temporal patterns. The features are denoted as EOH_FD, LBP_FD, and LPQ_FD and are reshaped producing 1152, 2835, and 768 components, respectively, which were concatenated to produce a vector of 4755 components. These components are produced for both Northwind and Freeform videos and were also concatenated together producing a total of 9510 components per video sample, which is denoted as (MIX_FD). The experiments were run on the concatenated features MIX_FD, as well as their individual feature performance. The vectors EOH_FD, LBP_FD, and LPQ_FD had been tested with the development set before they were concatenated, to provide a comparison from its individual and combined benefits.

Furthermore, modeling the temporal features of facial expressions in the dynamic feature space was explored, similar to [11]. First, MHH was applied on the video to produce five ($M = 5$) frames. Then, the local features (EOH, LBP, and LPQ) was extracted from each motion histogram frame. Finally, all the feature vectors were concatenated and denoted as (MH_MIX).

The baseline audio features (2268) are provided by the dataset. The short audio segments (short) were used, which were a set of descriptors that extracted features every 3 s of audio samples. The mean of the segments were taken to provide a single vector of $1 \times 2268$ per sample and it was denoted it (audio). The combined audio features of Flatness, Band1000, POV, PSY Sharpness, Shimmer, and ZCR were used, containing 285 of the 2268 features and was denoted as (Comb). MFCC was also investigated as a sole feature vector, and was denoted as (MFCC). For all the dynamic features from visual and vocal modalities, the dimensionality was reduced with PCA to $L = 49$ components, and the depression analyzed by the PLS and LR.

TABLE I
PERFORMANCE OF DEPRESSION SCALE PREDICTION USING THE DYNAMIC VISUAL FEATURE FDHH (FD) MEASURED BOTH IN MAE AND RMSE AVERAGED OVER ALL SEQUENCES IN THE DEVELOPMENT SET

| Partition | Methods | MAE | RMSE |
|---|---|---|---|
| Develop | EOH_FD_PLS | **8.87** | **11.09** |
| Develop | EOH_FD_LR | 9.92 | 12.39 |
| Develop | EOH_FD_(PLS+LR) | 9.14 | 11.39 |
| Develop | LBP_FD_PLS | 9.34 | 11.16 |
| Develop | LBP_FD_LR | 9.86 | 12.68 |
| Develop | LBP_FD_(PLS+LR) | **9.18** | **11.15** |
| Develop | LPQ_FD_PLS | 8.79 | 10.88 |
| Develop | LPQ_FD_LR | 9.73 | 11.49 |
| Develop | LPQ_FD_(PLS+LR) | **8.70** | **10.63** |

TABLE II
PERFORMANCE OF DEPRESSION SCALE PREDICTION USING FDHH (FD) AFTER MIX (EOH, LBP, LPQ, AND DEEP) VISUAL FEATURES ARE SHOWN UNDER APP1 AND MHH (MH) BEFORE MIX (EOH, LBP, LPQ, AND DEEP) VISUAL FEATURES ARE SHOWN IN APP2, MEASURED BOTH IN MAE AND RMSE AVERAGED OVER ALL SEQUENCES IN THE DEVELOPMENT SET

| Methods | Develop | | Methods | Develop | |
|---|---|---|---|---|---|
| APP1 | **MAE** | **RMSE** | APP2 | **MAE** | **RMSE** |
| MIX_FD_PLS | 7.72 | 9.68 | MH_MIX_PLS | 8.91 | 10.78 |
| MIX_FD_LR | 7.52 | 10.05 | MH_MIX_LR | 10.59 | 12.71 |
| A16_FD_PLS | **6.66** | **9.02** | MH_A16_PLS | **7.42** | **9.58** |
| A16_FD_LR | **6.96** | **9.52** | MH_A16_LR | 7.41 | 9.73 |
| A18_FD_PLS | 7.19 | 9.36 | MH_A18_PLS | **7.33** | **9.46** |
| A18_FD_LR | 7.23 | 9.43 | MH_A18_LR | **7.41** | **9.56** |
| V32_FD_PLS | 7.25 | 9.52 | MH_V32_PLS | 8.06 | 10.13 |
| V32_FD_LR | **6.90** | **9.32** | MH_V32_LR | 7.75 | 9.70 |
| V34_FD_PLS | 7.08 | 9.52 | MH_V34_PLS | 8.53 | 10.46 |
| V34_FD_LR | 7.09 | 9.53 | MH_V34_LR | 8.56 | 10.55 |
| V35_FD_PLS | 7.44 | 9.43 | MH_V35_PLS | 9.47 | 13.17 |
| V35_FD_LR | 7.50 | 9.44 | MH_V35_LR | 9.48 | 12.86 |

### C. Performance Comparison

Starting with the hand-crafted features LBP, LPQ, and EOH, Table I demonstrates the individual performance of the three hand-crafted feature extraction methods that are combined with FDHH. The depression scales were predicted using the two regression techniques separately and fused. It is clear that using PLS for regression was better than LR in all tests. However, when they were fused with a weighting more toward PLS, the results were improved further. LBP was shown to be the weakest amongst the three and LPQ the strongest.

Table II contains results of both approaches, with APP1 combining the efforts of the individual hand-crafted features, and demonstrates the effectiveness of the deep features using the FDHH algorithm. APP2 applies MHH before the hand-crafted and deep features. Three of the best results from each part have been highlighted in bold. MIX_FD has shown a significant improvement over the individual performances in Table I. However, it is clear from this that the deep features perform consistently better than the individual and combined hand-crafted features. The AlexNet deep features with FDHH (A16_FD) have shown a good performance on the development subset, closely followed by VGG-Face deep features with FDHH (V32_FD). The overall performance of APP2 can be viewed as inferior when compared to our main approach APP1, with all performances projecting a worse result than

TABLE III
PERFORMANCE OF DEPRESSION SCALE PREDICTION USING COMPLETE AUDIO, COMB FEATURES, AND MFCC. MEASURED BOTH IN MAE AND RMSE AVERAGED OVER ALL SEQUENCES IN THE DEVELOPMENT AND TEST SUBSETS

| Methods | Develop | | Test | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| Comb_short_PLS | 9.31 | 11.52 | 8.52 | 10.49 |
| Comb_short_LR | 9.42 | 11.64 | 10.33 | 12.99 |
| Audio_short_PLS | **8.25** | **10.08** | 8.42 | 10.46 |
| Audio_short_LR | **8.39** | **10.21** | 8.45 | 10.73 |
| MFCC_short_PLS | 8.86 | 10.70 | **8.04** | **10.42** |
| MFCC_short_LR | 8.86 | 10.92 | **8.07** | **10.28** |

TABLE IV
PERFORMANCE OF DEPRESSION SCALE PREDICTION USING FDHH ON VARIOUS SPATIAL FEATURES. MEASURED BOTH IN MAE AND RMSE AVERAGED OVER ALL SEQUENCES IN THE DEVELOPMENT AND TEST SUBSETS

| Methods | Develop | | Test | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| MIX_FD+MFCC_PLS | 7.41 | 9.30 | 7.28 | 9.15 |
| MIX_FD+MFCC_LR | 7.69 | 9.57 | 7.11 | 8.98 |
| A16_FD+MFCC_PLS | 7.40 | 9.21 | 6.58 | 8.19 |
| A16_FD+MFCC_LR | 7.14 | 8.99 | 6.87 | 8.45 |
| A18_FD+MFCC_PLS | **6.92** | **8.79** | 6.44 | 7.96 |
| A18_FD+MFCC_LR | **6.66** | **8.67** | 7.10 | 8.57 |
| V32_FD+MFCC_PLS | 7.35 | 9.54 | **6.17** | **7.44** |
| V32_FD+MFCC_LR | **7.07** | **9.34** | **6.31** | **7.59** |
| V34_FD+MFCC_PLS | 7.08 | 9.35 | **6.14** | **7.56** |
| V34_FD+MFCC_LR | 7.16 | 9.44 | 6.51 | 7.80 |
| V35_FD+MFCC_PLS | 7.20 | 9.24 | 8.34 | 10.43 |
| V35_FD+MFCC_LR | 6.90 | 9.06 | 8.12 | 10.15 |

TABLE V
SYSTEM PERFORMANCE USING WEIGHTED FUSION OF REGRESSION TECHNIQUES, PLS AND LR, ON SELECTED FEATURES FOR THE DEVELOPMENT AND TEST SUBSETS

| Partition | Methods | MAE | RMSE |
|---|---|---|---|
| Develop | V32_FD+MFCC_(PLS+LR) | 7.06 | 9.35 |
| Test | V32_FD+MFCC_(PLS+LR) | **6.14** | **7.43** |
| Develop | A18_FD+MFCC_(PLS+LR) | **6.58** | **8.65** |
| Test | A18_FD+MFCC_(PLS+LR) | 6.52 | 8.08 |

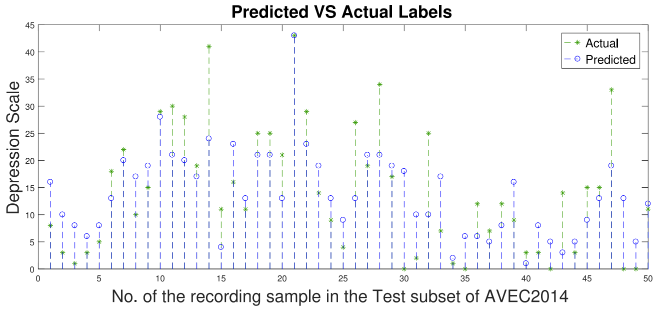

Fig. 6. Predicted and actual depression scales of the test subset of the AVEC2014 dataset based on audio and video features with regression fusion.

TABLE VI
PERFORMANCE COMPARISON AGAINST OTHER APPROACHES ON THE TEST PARTITION, MEASURED IN RMSE AND MAE. MODALITY FOR EACH IS MENTIONED AND GROUPED FOR COMPARISON (A = AUDIO AND V = VISUAL)

| Method | Modality | MAE | RMSE |
|---|---|---|---|
| **Ours** | Unimodal (V) | **6.68** | **8.01** |
| Kaya [41] | Unimodal (V) | 7.96 | 9.97 |
| Jain [44] | Unimodal (V) | 8.39 | 10.24 |
| Mitra [40] | Unimodal (A) | 8.83 | 11.10 |
| Baseline [18] | Unimodal (V) | 8.86 | 10.86 |
| Perez [39] | Unimodal (V) | 9.35 | 11.91 |
| **Ours** | Bimodal (A+V) | **6.14** | **7.43** |
| Williamson [20] | Bimodal (A+V) | 6.31 | 8.12 |
| Kaya [41] | Bimodal (A+V) | 7.69 | 9.61 |
| Chao [43] | Bimodal (A+V) | 7.91 | 9.98 |
| Senoussaoui [38] | Bimodal (A+V) | 8.33 | 10.43 |
| Perez [39] | Bimodal (A+V) | 8.99 | 10.82 |
| Kachele [37] | Multimodal | 7.28 | 9.70 |
| Gupta [42] | Multimodal | - | 10.33 |

its respective main approach feature, e.g., MH_V34_PLS versus V34_FD_PLS. Second, we can see that the deep learning approaches have performed better than hand-crafted features using both approaches.

A subexperiment was to investigate the features before and after a ReLU layer. This would supposedly introduce sparsity by removing negative magnitude features, which would result in a bad feature. This was tested by observing the features at the 34th and 35th layer of the VGG-Face network. From the individual performance evaluation on both approaches, it is clear that there is a higher RMSE and MAE for V35 using either regression techniques.

In Table III, the audio features for short segments were tested. From the 2268 audio features (audio), the combined features (Comb) and MFCC features have been taken out to be tested separately. The individual tests show the audio and MFCC features performing well on the development subset, with MFCC showing great performance on the test subset. When compared to visual features, they fall behind against most of them.

Features from both audio and visual modalities were combined as proposed in the approach, to produce bimodal performances that can be found in Table IV. This table demonstrates that the fusion of the two modalities boosts the overall performance further, especially on the test subset. VGG deep features have once again dominated the test subset, with AlexNet performing better on the development subset. A final test has been on fusing the performances of the regression techniques using the best features observed in Table IV. This involved using a weighted fusion technique on the PLS and LR predictions, the performance are detailed in Table V.

The best performing unimodal feature based on the test subset has been V32_FD, producing 6.68 for MAE and 8.04 for RMSE. Both achieving the state-of-the-art when compared against other unimodal techniques. The best overall feature uses the fusion of the audio and visual modalities, along with the weighted fusion of the regression techniques (V32_FD+MFCC)_(PLS+LR). This feature produced 6.14 for MAE and 7.43 for RMSE, beating the previous state-of-the-art produced by Williamson *et al.* [19], [20] who achieved 6.31 and 8.12, respectively. The predicted values of (V32_FD+MFCC)_(PLS+LR) and actual depression scale values on the test subset are shown in Fig. 6. Performance comparisons against other techniques including the baseline can be seen in Table VI.

## V. Conclusion

In this paper, an artificial intelligent system was proposed for automatic depression scale prediction. This is based on facial and vocal expression in naturalistic video recordings. Deep learning techniques are used for visual feature extraction on facial expression faces. Based on the idea of MHH for 2-D video motion feature, we proposed FDHH that can be applied to feature vector sequences to provide a dynamic feature (e.g., EOH_FD, LBP_FD, LPQ_FD, deep feature V32_FD, etc.) for the video. This dynamic feature is better than the alternate approach of MHH_EOH that was used in previous research [11], because it is based on mathematical feature vectors instead of raw images. Finally, PLS regression and LR are adopted to capture the correlation between the feature space and depression scales.

The experimental results indicate that the proposed method achieved good state-of-the-art results on the AVEC2014 dataset. Table IV demonstrates the proposed dynamic deep feature is better than MH_EOH that was used in previous research [11]. When comparing the hand-crafted versus deep features shown in Table II, deep features taken from the correct layer shows significant improvement over hand-crafted. With regards to selecting the correct layer, it seems that features should be extracted directly from the convolution filters responses. Generally the earliest fully connected layer will perform be the best, although the performances are fairly close to call. Audio fusion contributed in getting state-of-the-art results using only the MFCC feature, demonstrating that a multimodal approach can be beneficial.

There are three main contributions from this paper. First is the general framework that can be used for automatically predicting depression scales from facial and vocal expressions. The second contribution is the FDHH dynamic feature, that uses the idea of MHH on the deep learning image feature and hand-crafted feature space. The third one is the feature fusion of different descriptors from facial images. The overall results on the testing partition are better than the baseline results, and the previous state-of-the-art result set by Williamson *et al.* [19], [20]. FDHH has proven it can work as a method to represent mathematical features, from deep features to common hand-crafted features, across a temporal domain. The proposed system has achieved remarkable performance on an application that has very subtle and slow changing facial expressions by focusing on the small changes of pattern within the deep/hand-crafted descriptors. In the case that a sample contains other parts of the body; has lengthier episodes; or reactions to stimuli, face detection and video segmentation can adapt the sample to be used in our system.

There are limitations within the experiment that can affect the system performance. The BDI-II measurement is assessed on the response of questions asked to the patients. The scale of depression can be limited by the questions asked, as the responses may not portray their true depression level. The dataset contains patients only of German ethnicity; who are all Caucasian race. Their identical ethnicity may affect the robustness of a system when validated against other ethnicities. Another limitation can be the highest BDI-II recording within the dataset, which is 44 and 45 for the development and testing partitions, respectively. All these things can be considered for further improvement of the system.

Further ideas can be investigated to improve the system performance. The performance may improve if additional facial expression images are added into the training process of the VGG-Face deep network. The raw data itself can be used to retrain a pretrained network, which can be trained as a regression model. For the vocal features, a combination of descriptors have been tested. However, other vocal descriptors should also be considered to be integrated in the system, or even adapting a separate deep network that can learn from the vocal data. Other fusion techniques can also be considered at feature and prediction level that would improve the performance further.

## Acknowledgment

The authors would like to thank the AVEC2014 organizers for providing the depression dataset for this paper.

## References

[1] M. F. Valstar *et al.*, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. Int. Conf. ACM Multimedia Audio/Vis. Emotion Challenge Workshop*, 2013, pp. 3–10.

[2] "Mental health in the EU," Health Consum. Protect. Directorate Gener., Brussels, Belgium, Tech. Rep., 2008.

[3] M. Marcus *et al.*, "Depression: A global public health concern," *WHO Dept. Mental Health Substance Abuse*, vol. 1, pp. 6–8, 2012.

[4] "Depression and other common mental disorders: Global health estimates," World Health Org., Geneva, Switzerland, Tech. Rep. WHO/MSD/MER/2017.2, 2017.

[5] R. C. Kessler and E. J. Bromet, "The epidemiology of depression across cultures," *Annu. Rev. Public Health*, vol. 34, pp. 119–138, Mar. 2013. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-publhealth-031912-114409

[6] E. Jenkins and E. M. Goldner, "Approaches to understanding and addressing treatment-resistant depression: A scoping review," *Depression Res. Treat.*, vol. 2012, pp. 1–7, Feb. 2012, doi: 10.1155/2012/469680.

[7] D. D. Luxton, *Artificial Intelligence in Behavioral and Mental Health Care*. London, U.K.: Academic Press, 2015.

[8] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interact. Workshops (ACII)*, Amsterdam, The Netherlands, 2009, pp. 1–7.

[9] H. Davies *et al.*, "Facial expression to emotional stimuli in non-psychotic disorders: A systematic review and meta-analysis," *Neurosci. Biobehav. Rev.*, vol. 64, pp. 252–271, May 2016.

[10] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2014, pp. 73–80.

[11] H. Meng *et al.*, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Barcelona, Spain, 2013, pp. 21–30.

[12] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Palo Alto, CA, USA: Consulting Psychol. Press, 1978.

[13] A. Jan and H. Meng, "Automatic 3D facial expression recognition using geometric and textured feature fusion," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 5. May 2015, pp. 1–6.

[14] M. Kaletsch *et al.*, "Major depressive disorder alters perception of emotional body movements," *Front. Psychiatry*, vol. 5, p. 4, Jan. 2014.

[15] J. Han *et al.*, "Video abstraction based on fMRI-driven visual attention model," *Inf. Sci.*, vol. 281, pp. 781–796, Oct. 2014.

[16] J. Han *et al.*, "Learning computational models of video memorability from fMRI brain imaging," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1692–1703, Aug. 2015.

[17] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[18] M. F. Valstar *et al.*, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. Int. Conf. ACM Multimedia Audio/Vis. Emotion Challenge Workshop*, Orlando, FL, USA, 2014, pp. 3–10.

[19] J. R. Williamson *et al.*, "Vocal biomarkers of depression based on motor incoordination," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Barcelona, Spain, 2013, pp. 41–48.

[20] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. 4th ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2013, pp. 41–47.

[21] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, 2016, pp. 35–42.

[22] M. Nasir, A. Jati, P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, 2016, pp. 43–50.

[23] R. Weber, V. Barrielle, C. Soladié, and R. Séguier, "High-level geometry-based features of video modality for emotion prediction," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Amsterdam, The Netherlands, 2016, pp. 51–58.

[24] Y. L. Cun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 1990, pp. 396–404.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2014, pp. 3431–3440.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[30] P. Wang *et al.*, "Automated video-based facial expression analysis of neuropsychiatric disorders," *J. Neurosci. Methods*, vol. 168, no. 1, pp. 224–238, 2008.

[31] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 142–150, Apr. 2013, doi: 10.1109/T-AFFC.2012.38.

[32] S. Scherer *et al.*, "Automatic behavior descriptors for psychological disorder analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Shanghai, China, 2013, pp. 1–8.

[33] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Shanghai, China, 2013, pp. 1–8.

[34] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *Proc. Int. Conf. Affective Comput. Intell. Interact. Workshops*, Amsterdam, The Netherlands, 2009, pp. 1–7.

[35] L. Yammine, L. Frazier, N. S. Padhye, J. E. Sanner, and M. M. Burg, "Two-year prognosis after acute coronary syndrome in younger patients: Association with feeling depressed in the prior year, and BDI-II score and Endothelin-1," *J. Psychosom. Res.*, vol. 99, pp. 8–12, Aug. 2017.

[36] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1432–1441, Jul. 2015.

[37] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2014, pp. 41–48.

[38] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, Orlando, FL, USA, 2014, pp. 57–63.

[39] H. Pérez Espinosa *et al.*, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: INAOE-BUAP's participation at AVEC'14 challenge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2014, pp. 49–55.

[40] V. Mitra *et al.*, "The SRI AVEC-2014 evaluation system," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 93–101.

[41] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2013, pp. 19–26.

[42] R. Gupta *et al.*, "Multimodal prediction of affective dimensions and depression in human–computer interactions categories and subject descriptors," in *Proc. 4th ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2014, pp. 33–40.

[43] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi task sequence learning for depression scale prediction from video," in *Proc. Int. Conf. Affective Comput. Intell. Interact. (ACII)*, Xi'an, China, 2015, pp. 526–531.

[44] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and Fisher vector encoding," in *Proc. 4th ACM Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Orlando, FL, USA, 2014, pp. 87–91.

[45] T. Ojala, M. M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[46] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[47] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, 2005, pp. 886–893.

[48] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. IEEE Int. Workshop Autom. Face Gesture Recognit.*, Zürich, Switzerland, 1994, pp. 296–301.

[49] C. Yang, R. Duraiswami, and L. Davis, "Fast multiple object tracking via a hierarchical particle filter," in *Proc. IEEE Int. Conf. Comput. Vis.*, Beijing, China, 2005, pp. 212–219.

[50] H. Meng, B. Romera-Paredes, and N. Bianchi-Berthouze, "Emotion recognition by two view SVM_2K classifier on dynamic facial expression features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Santa Barbara, CA, USA, 2011, pp. 854–859.

[51] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. 3rd Int. Conf. Image Signal Process. (ICISP)*, Cherbourg-Octeville, France, Jul. 2008, pp. 236–243, doi: 10.1007/978-3-540-69905-7_27.

[52] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.

[53] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 248–255.

[54] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," in *Proc. CVPR Workshop Embedded Comput. Vis.*, Minneapolis, MN, USA, 2007, pp. 1–6.

[55] H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition," *Pattern Recognit. Lett.*, vol. 30, no. 12, pp. 1049–1058, 2009.

[56] H. Meng, N. Pears, M. Freeman, and C. Bailey, "Motion history histograms for human action recognition," in *Embedded Computer Vision* (Advances in Pattern Recognition), B. Kisačanin, S. S. Bhattacharyya, and S. Chai, Eds. London, U.K.: Springer, 2009, pp. 139–162.

[57] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[58] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interact. (ACII)*, Geneva, Switzerland, 2013, pp. 356–361.

[59] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, Santa Barbara, CA, USA, 2011, pp. 314–321.

[60] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2008, pp. 995–1004.

[61] R. Mattivi and L. Shao, "Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, Münster, Germany, 2009, pp. 740–747.

[62] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, 1993.

[63] A. Sinha, H. Chen, D. G. Danu, T. Kirubarajan, and M. Farooq, "Estimation and decision fusion: A survey," in *Proc. IEEE Int. Conf. Eng. Intell. Syst.*, 2006, pp. 1–6.

[64] A. Vedaldi and K. Lenc. (2015). *MatConvNet—Convolutional Neural Networks for MATLAB*. [Online]. Available: http://www.vlfeat.org/matconvnet/

**Asim Jan** received the M.Eng. degree in electronic and computer engineering from Brunel University London, U.K., in 2013, where he is currently pursuing the Ph.D. degree.

His current research interests include machine learning, artificial intelligence, pattern recognition, deep learning, digital signal processing, electronics design and development, field-programmable gate array design, hardware coding, computer-aided design, electronic model simulation, and software development.

**Hongying Meng** (M'10–SM'17) received the Ph.D. degree in communication and electronic systems from Xi'an Jiaotong University, Xi'an, China, in 1998.

He is currently a Senior Lecturer with the Department of Electronic and Computer Engineering, Brunel University London, U.K., where he is also a member of the Institute of Environment, Health and Societies, the Human Centred Design Institute, and the Wireless Networks and Communications Research Center. His current research interests include digital signal processing, affective computing, machine learning, human–computer interaction, computer vision, and embedded systems with over 90 publications and 2700 citations in these areas.

Dr. Meng was a recipient of the International Audio/Visual Emotion Challenges AVEC2011 and AVEC2013 Prizes for his audio-based and video-based emotion recognition systems. He is a fellow of the Higher Education Academy, and a member of the Engineering Professors Council in U.K.

**Yona Falinie Binti A. Gaus** (S'17) received the B.E. and M.Eng. degrees in electrical and electronic engineering from University Malaysia Sabah, Kota Kinabalu, Malaysia, in 2009 and 2013, respectively. She is currently pursuing the Ph.D. degree in electronic and computer engineering with Brunel University London, U.K.

Her current research interests include modeling continuous emotion recognition, integrating machine learning, deep learning, and artificial intelligence in order to detect emotion in continuous time domain.

**Fan Zhang** received the B.Sc. degree in electronic and information engineering from the Jicheng College, Sichuan University, Chengdu, China, in 2012, and the M.Sc. degree in advance electronic and electrical engineering from Brunel University London, U.K., in 2014, where he is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering.

His current research interests include signal processing and machine learning, especially on emotion detection and recognition from music and audio signals.