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

Asim Jan, Hongying Meng,Yona Falinie A. Gaus, 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 (BDI-II) from vocal and visual expressions. Firstly, different visual features are extracted from facial expression images. Deep Learning method is utilized to extract key visual features from the facial expression frames. Secondly, Spectral Low-level Descriptors (LLDs) and Mel-frequency cepstral coefficients (MFCCs) features are extracted from short audio segments to capture the vocal expressions. Thirdly, 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 AVEC2014 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, Depression, Beck Depression Inventory, Facial Expression, Vocal Expression, Regression, Deep Learning**

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 major depressive disorders. 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], [9], [10], [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, EEG signals and brain activity using MRI imaging, are modalities recent to computer vision [15], [16]. Electrocardiogram (ECG) and electro-dermal activity (EDA) 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 AVEC16 [21], [22], [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 non-controlled 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 towards 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 pre-trained networks.

Pre-trained CNNs can be exploited for artificial depression analysis, mainly using the visual modality. However, the pre-trained 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 the following: 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 the paper is organized as follows. Section 2 briefly reviews related work in this area. Section 3 provides a detailed description of the proposed method, and Section 4 displays and discusses the experimental results on the AVEC2014 dataset [18]. Section 5 concludes the paper.

II. Related Works

Recent years have witnessed an increase of research for clinical and mental health analysis from facial and vocal expressions [30], [31], [32], [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 the 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 semi-structured 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 questionnaire which indicated that the mean BDI II score of 20.7 (borderline clinically 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 is a distinct difference between the highest and lowest distressed participants, in terms of automatically detected behaviors.

The depression recognition sub-challenge of AVEC2013 [1] and AVEC2014 [18]; had proposed some good methods which achieved good results [19], [36], [37], [38], [39], [40], [41], [42], [43], [44], [11], [10]. From this, Williamson et al. [19], [20] was the winner of the depression sub-challenge for the AVEC2013 and AVEC2014 competitions. In 2013, they exploited the effects that reflected changes in coordination of vocal tract motion associated with Major Depressive Disorder. 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

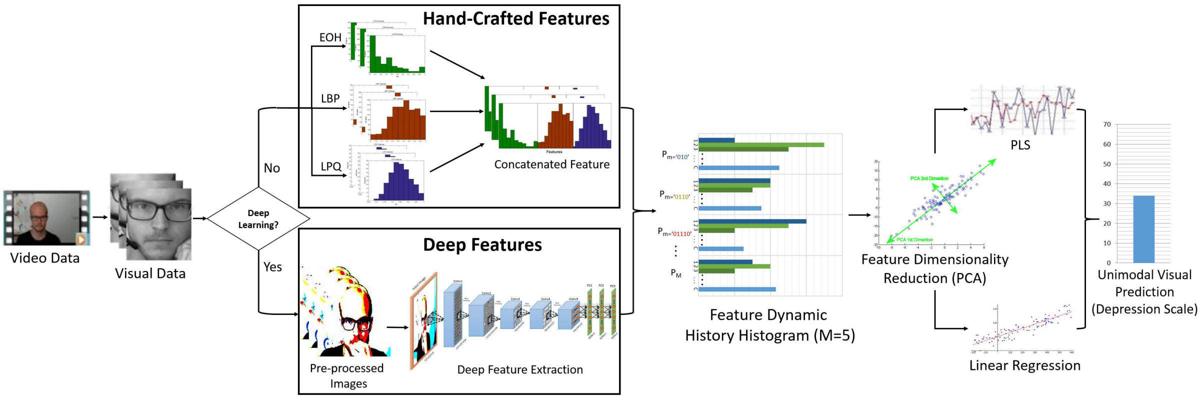

Fig. 1. Overview of the proposed automatic depression scale recognition system from facial expressions. The Video Data is broken down into visual frames. If deep learning is utilized, then deep features are extracted from the frames, otherwise a set of hand-crafted features are extracted. This is followed 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.

they looked at the change in motor control that can effect the mechanisms for controlling speech production and facial expression. They derived a multi-scale correlation structure and timing feature from vocal data. Based on these two feature sets, they designed a novel Gaussian mixture model (GMM) 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 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 (SVR) 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 (CCA) and Moore-Penrose Generalized Inverse (MPGI) were utilized for regression in a multimodal framework. Jain et al. [44] proposed using Fisher Vector (FV) to encode the LBP-TOP and Dense Trajectories visual features, and LLD audio features. Perez 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 seconds and less than 20 seconds*’ 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 multi-task learning based on audio and visual data. They used Long-Short Term Memory (LSTM) modules with features extracted from a pre-trained CNN, where the CNN was trained on a small facial expression dataset FER2013 by Kaggle. This dataset contained a total of 35,886 48x48 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 the AVEC2014 dataset. We will improve previous work from feature extraction using deep learning, regression as well as 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.

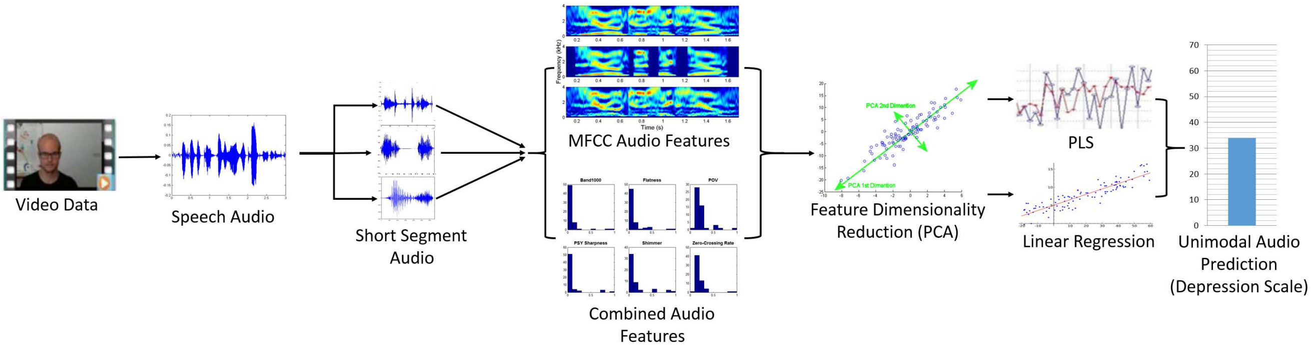

Fig. 2. Overview of the proposed automatic depression scale recognition system from vocal expressions. The Speech Audio is extracted from the Video Data, where short segments are produced. Then a bunch of audio features are extracted from each segment and averaged. These are reduced in dimensionality and a fusion of PLS and Linear regression is used to predict the Depression Scale.

A. *System Overview*

Figure 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. Figure 2 follows a similar architecture as Figure 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 pre-processed 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 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 Local Binary Patterns (LBP); Edge Orientation Histogram (EOH) and Local Phase Quantization (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 (HOG) [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 pre-trained 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 pre-trained deep models, including their *Very Deep Networks*. These networks are VGG-S, VGG-F, 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 pre-trained 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 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] pre-trained 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 3x3 and as the layers increase, so does the filter depth which varies from 64 to 512. The fully connected layers are of 1x1 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 Figure 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

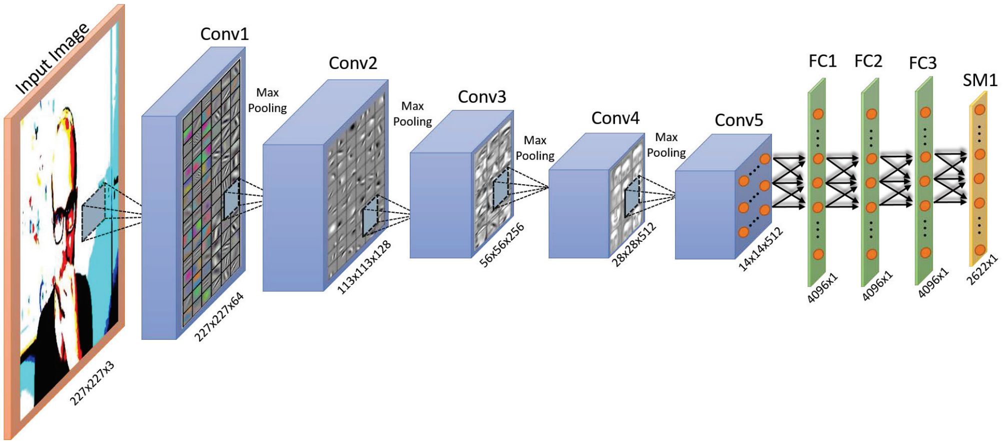

Fig. 3. The VGG-Face architecture visualizing low to high level features captured as a facial expression is propagated through-out the network, stating the dimensions produced by each layer.

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 to 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 3 fully connected layers FC1, FC2 and FC3 were acquired and used.

4) AlexNet: AlexNet [28], created by Alex Krishevsky, is another popular network, which was one of the first successful deep networks used in the ImageNet challenge [53]. This pre-trained 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 low-level descriptors (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 8 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. Figure 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 grey scale value changes for each pixel in the video. In comparison with other well-known motion features, such as Motion History Image (MHI) [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 2D images. Here, a technique is proposed to capture dynamic variation that occurs within mathematical representations of a visual sequence. Hand-Crafted 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 have been proposed to move these descriptors into the temporal domain in [58], [59], [60], [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

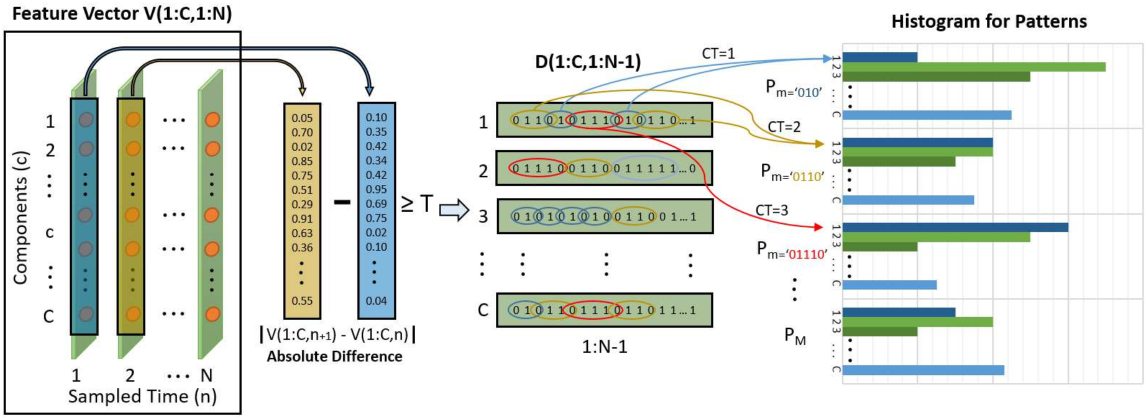

Fig. 4. Visual process of computing FDHH on the sequence of feature vectors. The first step is to obtain a new binary vector representation based on the absolute difference between each time sample. From this, Binary patterns P_m are observed throughout each components of the binary vector and a histogram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

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. Figure 4 shows the process of computing FDHH on the sequence of feature vector, the algorithm for FDHH can be implemented as follows:

We let $\{V(c,n), c = 1, \cdots, C, n = 1, \cdots, N\}$ be a feature vector with C components and N frames, and a binary sequence $\{D(c,n), c = 1, \cdots, C, n = 1, \cdots, N-1\}$ of feature component c is generated by comprising and thresholding the absolute difference between consecutive frames as shown in Equation 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 $P_m \, (1 \leq m \leq M)$, where m represents how many consecutive '1's are needed to create the pattern, as shown in Figure 4. The final dynamic feature can be represented as $\{FDHH(c,m), c = 1, \cdots, C, m = 1, \cdots, M\}$.

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

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'. We then initialize a counter CT to 0, which is used to search for patterns of 1's in a sequence $D(1:C, 1:N-2)$.

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

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

When observing a component from a sequence $D(c, 1:N)$, a pattern of $P_m \, (1 \leq m \leq M)$ is detected by counting the number of consecutive 1's, where CT is updated as shown in Equation 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 Equation 3, followed by the counter CT being reset to 0.

```
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. The FDHH algorithm.

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 0's), none of the histograms are updated; and where $CT > M$, the histogram for P_M is incremented. The full algorithm can be seen in Figure 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.

$$\hat{X}_{\mathrm{i}} = \frac{X_i - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}} \tag{4}$$

Where X_i is the training feature component, $\alpha_{\min}$ and $\alpha_{\max}$ are the training minimum and maximum feature vector values and $\hat{X}_{\mathrm{i}}$ is the normalized training feature component. Principal Component Analysis (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 Linear Regression. 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 $MHH(u,v,m)$ and $\{u = 1,\cdots,U, v = 1,\cdots,V, m = 1,\cdots,M\}$, 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.

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. Partial Least Squares (PLS) 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.

Linear Regression (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 Audio/Visual Emotion Challenge (AVEC) 2014 dataset [18], a subset of the audio-visual depressive language corpus (AViD-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 2 Human-Computer Interaction tasks separately whilst being recorded by a webcam 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. 18 subjects appear in three recordings, 31 in 2, and 34 in only one recording. The length of these clips are between 6 seconds to 4 minutes and 8 seconds. The mean age of subjects is 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 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 Audio/Visual Emotion Challenge 2014 guidelines which can be found in [18]. The instructions are followed as mentioned in the Depression Sub-Challenge (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 pre-trained networks.

1) Data Pre-processing: In order to obtain the optimal features from the pre-trained networks, a set of data pre-processing steps were followed, as applied by both Krizhevsky and Parkhi 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 227x227x3 representing the image height, width and color channels. AlexNet has the same requirement of 227x227x3 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 pre-processed 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 vs. 4096 when comparing the initial convolution layer vs. the fully connected layers, and a staggering 802K vs. 4096 for VGG-Face. The connectivity between filter responses are responsible for the dramatic decrease in dimensions at the fully connected layers. We observe the fully connected layer using equation 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 equation 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 5 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 5 resulting features with 4096 components each, making a total of feature dimension count of $5 \times 4096 = 20480$ per video sample. As there are two recordings per ground truth label, ('Northwind' and 'Freeform'), the 20480 features are extracted from both recordings and concatenated together to make a final visual feature vector of 40960 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 @ 4.3GHz, and a Titan X (Pascal) GPU. For 6 second video clip, it will take less than 3.3 seconds 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 6 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 3 channels in total to imitate the color channels. The output of the CNNs will produce 6 × 4096 features, which is transformed into a single row to make 24576 features, and 49152 features when both Northwind and Freeform are concatenated. These features will be denoted as **MH_A16**, **MH_A18**, **MH_V32**, **MH_V34** and finally **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 are concatenated to produce a vector of 4755 components. These components are produced for both Northwind and Freeform videos and are also concatenated together producing a total of 9510 components per video sample, which is denoted as (MIX_FD). We experimented on the concatenated features MIX_FD, as well as their individual feature performance. The vectors EOH_FD, LBP_FD and LPQ_FD have been tested with the development set before they are concatenated, to provide a comparison from its individual and combined benefits.

Furthermore, we explored modeling the temporal features of facial expressions in the dynamic feature space, similar to [11]. First we operated MHH on the video to produce 5 ($M = 5$) frames, and then extract the local features (EOH, LBP and LPQ) from each frame. Finally, we concatenated all of the vectors and denoted it as (**MH_MIX**).

The baseline audio features (2268) are provided by the dataset. We used the short audio segments (**short**) which are a set of features extracted every 3 seconds of audio samples. We then take the Mean of the segments to provide a single vector of 1 × 2268 per sample and denote it as (**Audio**). The combined audio features of Flatness, Band1000, POV, PSY Sharpness, Shimmer and ZCR are used, containing 285 of the 2268 features which was denoted as (**Comb**). We also investigated using just the MFCC as a feature and denoted it 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 Linear Regression.

C. Performance Comparison

We started with the Hand-crafted features LBP, LPQ and EOH. Table I shows 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. We can see that using PLS for regression is better than LR in all tests. However, when they were fused with a weighting more towards PLS, the results are improved further. LBP was shown to be the weakest amongst the three and LPQ the strongest.

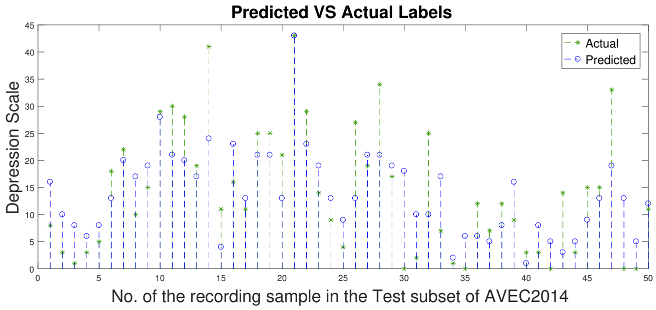

Fig. 6. Predicted and actual depression scales of the Test subset of the AVEC2014 dataset based on audio and video features with regression fusion.

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

Our prediction was that if we investigated the features before and after a ReLU layer, this would introduce sparsity by removing negative magnitude features, which would result in a bad feature. We tested this by observing the features at the 34th and 35th layer of the VGG-Face network. From the individual performance evaluation on both approaches, we can see that there is a higher RMSE and MAE for V35 using either regression techniques.

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

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**

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.

We have combined the features of the Audio and Visual modalities as proposed in our approach, to produce Bi-modal performances that can be found in Table IV. Here we can see 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.

Our best performing Uni-modal 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 Uni-modal 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. 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 Figure 6. Performance comparisons against other techniques including the baseline can be seen in Table VI.

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

V. CONCLUSIONS AND PERSPECTIVES

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, and 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 VS Deep features shown in Table II, Deep features taken from the correct layer shows significant improvement over hand-crafted.

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, 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

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 Multi-modal 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. 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 impact the system. 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. These are all things to consider to further improve 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 pre-trained 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 work by Asim Jan was supported by College of Engineering, Design & Physical Sciences/Thomas Gerald Gray PGR Scholarship. The work by Yona Falinie Binti Abd Gaus was supported by Majlis Amanah Rakyat (MARA) Scholarship. We would like to thank the AVEC2014 organizers for providing the depression dataset for our work.

REFERENCES

[1] M. F. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge," in *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop*, 2013.

[2] Health & Consumer Protection Directorate General, "Mental health in the EU," 2008.

[3] M. Marcus, M. T. Yasamy, M. van Ommeren, and D. Chisholm, "Depression, a global public health concern," pp. 1–8, 2012.

[4] World Health Organization, "Depression and other common mental disorders: global health estimates," Tech. Rep., 2017.

[5] R. C. Kessler and E. J. Bromet, "The epidemiology of depression across cultures." *Annual review of public health*, vol. 34, pp. 119–38, 2013.

[6] E. Jenkins and E. M. Goldner, "Approaches to understanding and addressing treatment-resistant depression: A scoping review," 2012.

[7] D. D. Luxton, *Artificial intelligence in behavioral and mental health care*. Academic Press, 2015.

[8] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De La Torre, "Detecting depression from facial actions and vocal prosody," in *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, 2009.

[9] H. Davies, I. Wolz, J. Leppanen, F. F. Aranda, U. Schmidt, and K. Tchanturia, "Facial expression to emotional stimuli in non-psychotic disorders: A systematic review and meta-analysis." *Neuroscience and biobehavioral reviews*, vol. 64, pp. 252–271, 2016.

[10] A. Jan, Y. Falinie, F. Zhang, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*. New York, New York, USA: ACM Press, 2014, pp. 73–80.

[11] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 2013, pp. 21–30.

[12] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978.

[13] A. Jan and H. Meng, "Automatic 3d facial expression recognition using geometric and textured feature fusion," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 05, May 2015, pp. 1–6.

[14] M. Kaletsch, S. Pilgramm, M. Bischoff, S. Kindermann, I. Sauerbier, R. Stark, S. Lis, B. Gallhofer, G. Sammer, K. Zentgraf, J. Munzert, and B. Lorey, "Major depressive disorder alters perception of emotional body movements," *Frontiers in Psychiatry*, vol. 5, no. JAN, 2014.

[15] K. Li, L. Shao, X. Hu, S. He, L. Guo, J. Han, T. Liu, and J. Han, "Video abstraction based on fMRI-driven visual attention model," *Information Sciences*, vol. 281, pp. 781–796, 2014.

[16] J. Han, J. Han, C. Chen, L. Shao, X. Hu, and T. Liu, "Learning Computational Models of Video Memorability from fMRI Brain Imaging," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1692–1703, 2015.

[17] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 3–10.

[18] M. F. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014 - 3d dimensional affect and depression recognition challenge," in *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop*, 2014.

[19] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 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," *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pp. 41–47, 2013.

[21] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 35–42.

[22] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 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 *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 51–58.

[24] L. D. Le Cun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, B. L. Cun, J. Denker, and D. Henderson, "Handwritten Digit Recognition with a Back-Propagation Network," *Advances in Neural Information Processing Systems*, pp. 396–404, 1990.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *Procedings of the British Machine Vision Conference 2015*, no. Section 3, pp. 41.1–41.12, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2014.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

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

[30] P. Wang, F. Barrett, E. Martin, M. Milanova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma, "Automated video-based facial expression analysis of neuropsychiatric disorders," *Journal of Neuroscience Methods*, vol. 168, no. 1, pp. 224–238, 2008.

[31] Y. Yang, C. Fairbairn, and J. Cohn, "Detecting depression severity from intra- and interpersonal vocal prosody," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[32] S. Scherer, G. Stratou, J. Gratch, J. Boberg, M. Mahmoud, A. S. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[33] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[34] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*, 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," *Journal of Psychosomatic Research*, vol. 99, pp. 8–12, 2017.

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

[37] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: ACM, 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," *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 57–63, 2014.

[39] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: ACM, 2014, pp. 49–55.

[40] V. Mitra, R. Ave, M. Park, E. Shriberg, R. Ave, M. Park, M. Mclaren, A. Kathol, R. Ave, M. Park, C. Richey, and M. Graciarena, "The SRI AVEC-2014 Evaluation System," *ACM International Conference on Multimedia*, pp. 93–101, 2014.

[41] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for Continuous Emotion Prediction," *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pp. 19–26, 2013.

[42] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions Categories and Subject Descriptors," *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pp. 33–40, 2014.

[43] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi Task Sequence Learning for Depression Scale Prediction from Video," pp. 526–531, 2015.

[44] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression Estimation Using Audiovisual Features and Fisher Vector Encoding," *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, no. 3, pp. 87–91, 2014.

[45] T. Ojala, M. Matti Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, 2002.

[46] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[47] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[48] W. T. Freeman, W. T. Freeman, M. Roth, and M. Roth, "Orientation histograms for hand gesture recognition," in *IEEE International Workshop on Automatic Face and Gesture Recognition*, 1994, pp. 296–301.

[49] C. Yang, R. Duraiswami, and L. Davis, "Fast multiple object tracking via a hierarchical particle filter," in *IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 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 *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, pp. 854–859.

[51] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, ser. Lecture Notes in Computer Science, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds. Springer Berlin Heidelberg, 2008, vol. 5099, pp. 236–243.

[52] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," *Proceedings of the British Machine Vision Conference*, pp. 1–11, 2014.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[54] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," in *CVPR workshop on Embeded Computer Vision*, 2007.

[55] H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition," *Pattern Recognition Letters*, 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*, ser. Advances in pattern recognition, B. Kisačanin, S. Bhattacharyya, and S. Chai, Eds. 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, 2001.

[58] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 356–361, 2013.

[59] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pp. 314–321, 2011.

[60] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, sep 2008, pp. 995–1004.

[61] R. Mattivi and L. Shao, "Human action recognition using LBP-TOP as sparse Spatio-temporal feature descriptor," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5702 LNCS, 2009, pp. 740–747.

[62] S. de Jong, "Simpls: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, 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 *IEEE International Conference on Engineering of Intelligent Systems*, 2006, pp. 1–6.

[64] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," 2015. [Online]. Available: http://www.vlfeat.org/matconvnet/