Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression

Asim Jan
Department of Electronic and Computer Engineering
Brunel University London, UK
asim.jan@brunel.ac.uk

Hongying Meng
Department of Electronic and Computer Engineering
Brunel University London, UK
hongying.meng@brunel.ac.uk

Yona Falinie A. Gaus
Department of Electronic and Computer Engineering
Brunel University London, UK
yonafalinie.abdgaus@brunel.ac.uk

Fan Zhang
Department of Electronic and Computer Engineering
Brunel University London, UK
1322726@my.brunel.ac.uk

Saeed Turabzadeh
Department of Electronic and Computer Engineering
Brunel University London, UK
saeed.turabzadeh@brunel.ac.uk

ABSTRACT

Depression is a state of low mood and aversion to activity that can affect a person's thoughts, behavior, feelings and sense of well-being. In such a low mood, both the facial expression and voice appear different from the ones in normal states. In this paper, an automatic system is proposed to predict the scales of Beck Depression Inventory from naturalistic facial expression of the patients with depression. Firstly, features are extracted from corresponding video and audio signals to represent characteristics of facial and vocal expression under depression. Secondly, dynamic features generation method is proposed in the extracted video feature space based on the idea of Motion History Histogram (MHH) for 2-D video motion extraction. Thirdly, Partial Least Squares (PLS) and Linear regression are applied to learn the relationship between the dynamic features and depression scales using training data, and then to predict the depression scale for unseen ones. Finally, decision level fusion was done for combining predictions from both video and audio modalities. The proposed approach is evaluated on the AVEC2014 dataset and the experimental results demonstrate its effectiveness.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Vision and Scene Understanding; J.3 [**Computer Applications**]: Life and Medical Science—*Health*

Keywords

Affective computing; depression recognition; Beck Depression Inventory; facial expression; challenge

AVEC'14, November 7, 2014, Orlando, FL, USA.

ACM 978-1-4503-3119-7/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661806.2661812.

1. INTRODUCTION

The study on mental health problems has been given increasing attention from various domains in the modern society in recent years. Among mood disorders, depression commonly occurs and heavily threatens the mental health of human beings. Generally, depression is a state of low mood and aversion to activity that can affect a person's thoughts, behavior, feelings, and sense of well-being [22]. Depressed people may feel sad, anxious, hopeless, empty, worried, helpless, worthless, hurt, irritable, guilty, or restless. They may lose interest in activities that once were pleasurable, experience loss of appetite or overeating, suffer trouble in concentrating, remembering details, or making decisions, and may contemplate or even attempt suicide. Insomnia, excessive sleeping, fatigue, loss of energy, aches, pains, or digestive problems that are resistant to treatment may be present as well [1].

In spite of very limited progress currently achieved on automatic depression scale prediction or other mental disorders, recent technological revolutions of automatic emotion analysis in the field of affective computing and social signal processing are extensive, which can be regraded as a good place to start. People express their emotions through the visual (i.e. facial expressions and bodily gestures), vocal, and physiological modalities, and among these modalities, facial expression plays a primary role in representing emotional states of human beings. Much effort has hence been dedicated by psychologists to model the mapping from facial expressions to emotional states [10]. Initial work aims at describing expressions using static features extracted from still face images [21] [24]. However, to some expressions, especially the subtle ones like anger and disgust, these features prove insufficient. Recently, the focus gradually veers to facial expression analysis in video data, since they convey richer information than images do. From videos, dynamic features can be obtained, which are also critical to represent facial expressions formed by periodical muscular movements, e.g. [6] [13] [18]. Meanwhile, in video data, there often exists vocal information which is corresponding to the visual, and it is another important channel to emotion recognition. Therefore, a natural trend appears to combine the vocal

modality with the visual one, claiming that both the clues are complementary to each other and the joint use improves system performance [5].

Similarly, in depression scale prediction, visual and vocal features (included in video data) are both indispensable, since depressed people tend to behave disorderly in facial expression, gesture, verbal communication, etc. For example, they may seem unable to relax, quicker to anger, or full of restless energy, which can be reflected by changes in their facial expressions; while they may make their speech slurred, slow, and monotonous, and this can be represented by variations of their voices. Based on such consideration, this paper proposes a novel approach for depression scale predcition using both the visual and vocal clues, aiming to combine their advantages. Based on the idea of Motion History Histogram (MHH) for 2-D motion capture, dynamic features are extracted based on the features from videos. Feature selection is done to reduce the dimension of the feature. Then Partial Least Square (PLS) and Linear regression algorithm is then used to model the mapping between dynamic features and the depression scales. Finally, predictions from both video and audio modalities were combined on decision level. Experimental results achieved on the AVEC2014 dataset illustrate the effectiveness of the proposed method.

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 [27]. Section 5 concludes the paper.

2. RELATED WORK

Recent years have witnessed the growing study for clinical and mental health analysis from facial and vocal expressions [29] [32] [23] [12] due to significant progress on emotion recogntion from facial expression. Wang et al. [29] proposed a computational approach, which creates probabilistic facial expression profiles for video data and helps to automatically quantify emotional expression differences between patients with neuropsychiatric disorders (Schizophrenia) and healthy controls.

As particularly concerning on depression analysis, Cohn et al. [7] pioneered at seeking for solutions in the view of affective computing. They fused both clues of facial actions and vocal prosody, attempting to investigate systematic and efficient ways of incorporating behavioral observations that are strong indicators of psychological disorders, much of which may occur outside the awareness of either individual. Their findings suggest the feasibility of automatic detection of depression, and possess exciting implications for clinical theory and practice. Yang et al. [32] explored variations in 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. [12] analyzed both manual and automatic facial expressions during semi-structured clinical interviews of clinically depressed patients, concluding that participants with high symptom severity behave with more expressions associated with contempt, smile less, and their smiles were more likely to be related to contempt. Scherer et al. [23] studied the correlation between the properties of gaze, head pose, and smile and three mental disorders (i.e. depression, post-traumatic stress disorder and anxiety), and discovered that significant differences of automatically detected behaviors appear between the highest and lowest distressed participant groups.

Depression recognition sub-challenge of AVEC2013 [28] was held in 2013 and some good methods were proposed with good results [30] [14]. Williamson et al. [30] 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 formant frequencies and also across channels of the delta-mel-cepstrum. Both feature domains provide measures of coordination in vocal tract articulation while reducing effects of a slowly-varying linear channel, which can be introduced by time-varying microphone placements. Based on these two complementary feature sets, they designed a novel Gaussian mixture model (GMM)-based multivariate regression scheme, referred to as Gaussian Staircase Regression, that provided very good prediction on the standard Beck depression rating scale. Meng et al. [14] presents a novel method, which comprehensively models visual and vocal modalities with dynamic features to automatically predicts the scale of depression. Motion History Histogram (MHH) was used firstly to extract the dynamics from corresponding video and audio data to represent characteristics of subtle changes in facial and vocal expression of depression. And then, Partial Least Square (PLS) regression algorithm is applied to learn the relationship between the dynamic features and depression scales using training data, and then predict the depression scale for an unseen one. Predicted values of visual and vocal clues are further combined at decision level for final decision.

From these two methods, it can be seen that feature extraction is the key for depression scale prediction. In this paper, we proposed a new way for dynamic feature extraction. Firstly, standard feature extraction methods are used on each frames or vocal segments. Then the dynamic variations in the feature space is extracted based on the idea of MHH of the image. The reason is that the feature space contains more useful information than the raw images that were used in [14]. Then, fusion process on advanced regression methods is investigated and proposed to improve the overall performance.

3. DEPRESSION SCALE PREDICTION

Since human facial expressions and voices in depression are theoretically different from those under normal mental states, we attempt to address the problem of depression scale prediction by combining dynamic descriptions within naturalistic facial and vocal expressions. This paper proposes a novel method that comprehensively models the variations in visual and vocal clues and automatically predicts the Beck Depression Inventory scale of depression.

3.1 System Overview

Figure 1 illustrates the process of how the features are extracted from both visual and audio data, which is then combined, reduced and used with machine learning. With the visual data, a sequence of steps are taken to combine different extraction methods that are used for dynamic and static data. The data itself is in a video format from which Edge Orientation Histogram (EOH), Local Phase Quantization (LPQ) and Local Binary Patterns (LBP) features are extracted of each frame from the video, extracting the different descriptors to capture different characteristics of

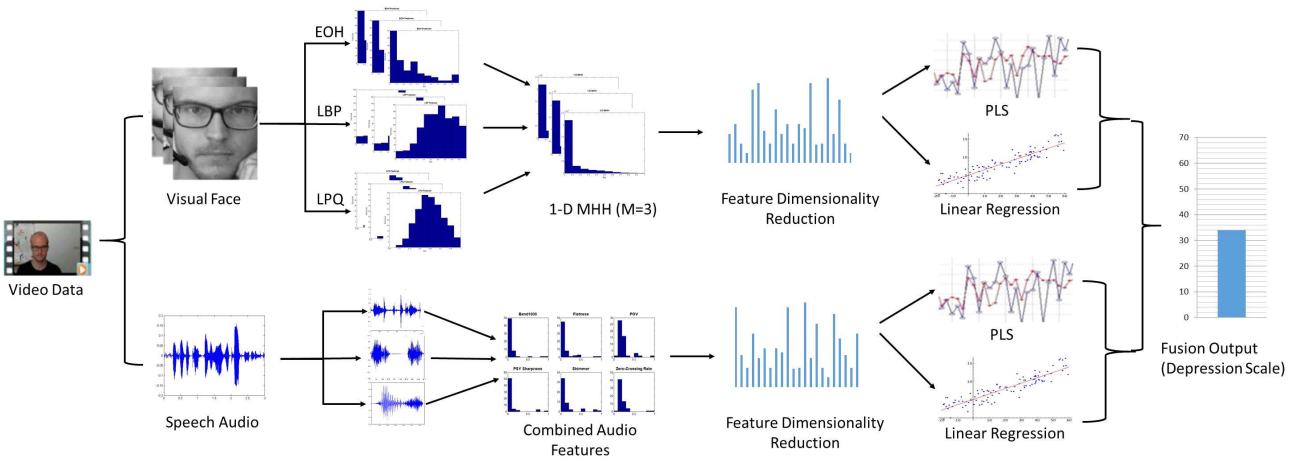

Figure 1: Overview of the proposed automatic depression scale recognition system.

the facial expression images. The idea of Motion History Histograms (MHH) is then applied to capture the dynamic movement of the features, describing temporal information of different extracted facial descriptors. It is effective in summarizing the facial feature movements. All the features are then concatenated for a better representation and reduced in dimensionality to make the system more accurate and efficient. Partial Least Square (PLS) regression and Linear regression are then combined to predict the depression scale.

3.2 Image Feature Extraction

Based on the frames of facial expression in video data, three different texture features are extracted.

3.2.1 Local Binary Patterns

Local Binary Patterns (LBP) describes the local texture structure of an image by creating patterns for the pixels and its surroundings. It is namely invariance to monotonic gray level changes [19] and computational efficiency making it useful for face recognition [2], which would be useful for extracting the facial structure variance across the expressions.

Each pixel is compared with its surrounding 8 pixels, using a threshold to compare if they are higher or lower. From this a pattern of 8 bits is generated and converted to decimal format from which a histogram is created. This will be applied on the textured facial images from the database.

3.2.2 Edge Orientation Histogram

EOH, an efficient and powerful operator, is regarded as a simpler version of Histogram of Oriented Gradients (HOG) [8] that captures the edge or the local shape information of an image. It has been widely investigated in a variety of applications in computer vision such as hand gesture recognition [11], object tracking [31] and facial expression recognition [18].

3.2.3 Local Phase Quantization

Local Phase Quantization (LPQ) is widely used in face recognition and texture classification [20] because of its ability to extract local phase information using Discrete Fourier Transform on windows of M-by-M across the whole image, and checks for the phase difference to produce a histogram for each window. The histograms are combined to produce one overall histogram for the image.

3.3 Audio Feature Extraction

The spectral low-level descriptors (LLDs) and MFCCs 11-16 are included in the AVEC2014 baseline audio feature set and adopted as the basic representation of the vocal clue. The baseline feature set consists of 2268 features, composed of 32 energy and spectral related×42 functionals, 6 voicing related LLD×32 functionals, 32 delta coefficients of the energy/spectral LLD×19 functionals, 6 delta coefficients of the voicing related LLD×19 functionals, and 10 voiced/unvoiced durational features.

LLD features are extracted from 25 ms and 60 ms overlapping windows which are shifted forward at a rate of 10 ms. Among these features, pitch (F0) based LLD are extracted from 60 ms windows, while all other LLD are extracted from 25 ms windows. For detailed information, please see the baseline paper [27].

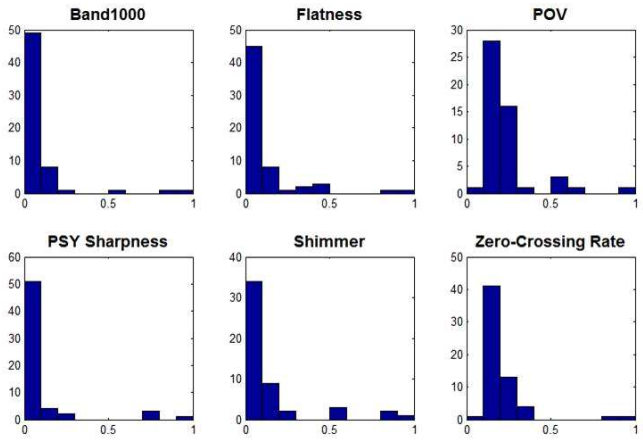

Figure 2: The selected audio feature from audio features provided in AVEC2014.

Firstly, the 2268 features are split to several vectors containing features from each individual LLD extraction method which are investigated separately. The process of selecting

the best audio descriptors is to test each individual feature vector with the development data set, then take the top 8 performing descriptors. These are: Flatness, Energy in Bands 1k-4kHz (Band1000), Entropy, MFCC, Probability of Voicing (POV), PSY Sharpness, Shimmer and Zero Crossing Rate (ZCR). Then, further investigation is done on all 256 possible combination of these 8 vectors, which are also tested with the development data set. From different combinations, Flatness, Band1000, PSY Sharpness, POV, Shimmer and ZCR performed the best, which are concatenated together to produce the final descriptor for the audio modality as shown in Figure 2.

3.4 1-D Motion History Histogram

MHH is a descriptive temporal template motion representation for visual motion recognition. It was originally proposed and applied in human action recognition [16]. The detailed information can be found in [15] and [17]. 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) [4], it contained more dynamic information of the pixels and provides better performance in human action recognition[15]. MHH not only provides rich motion information, but also remains computationally inexpensive [17].

In the previous research, MHH is used for extracting the motion on each pixel in the 2D image sequences and then creating the multiscale histograms (M is the number of scales) for the whole video. Here, we proposed an 1-D MHH that extracts the changes on each component in a feature vector sequence (instead of one pixel from a image sequence) and then create the histogram for all the components of the feature vector in one video. So the dynamic of facial movement are replaced by the feature movements. Figure 3 shows the process of computing 1-D MHH on the sequence of feature vector. Firstly, on one component of the feature vector, consecutive two values are compared and thresholded to produce a binary value. If there is change bigger than threshold, it will be '1', otherwise, it will be '0'. Along the time line, a binary sequence will be generated. Within the binary sequence, M different pattern will be counted. So for each component, M counts will be generated. All these counts together make the 1-D MHH features of the feature vector.

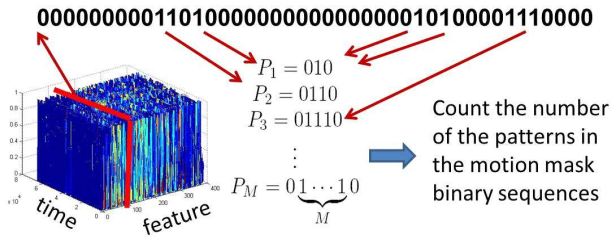

Figure 3: Process of computing 1-D MHH on the sequence of feature vectors.

3.5 Feature Combination and Fusion

The LBP, LPQ and EOH feature extraction methods are applied on each of the frames as each algorithm highlights different patterns, all of which are considered to be useful. This then effectively converts a video with frames of images to three sequences of extracted feature vectors. 1-D MHH algorithm is applied to each feature vector sequence and it produces M feature vectors from each sequence, that contains the dynamic details of the motion occurred across the features. The resulting histograms are then concatenated to provide one feature vector.

Combining all the features produces a large feature vector, which can cause the training period of the machine learning to be very slow. Two approaches have been adapted to reduce the dimensionality of the large feature set in order to improve the training speed and performance. Firstly, a novel method was used to analyze each feature column and decide whether it can be discarded to increase the performance of the system. Secondly, Principal Component Analysis (PCA) is applied to obtain the coefficients for each feature set.

The initial method normalizes each column of the feature vector, e.g. if the training set contains 50 videos with 2000 features each, a column vector would represent 1 feature from the 50 videos i.e. (1x50). The goal is to reduce the total number of columns whilst retaining key information. From each individual column, a histogram is taken to test how well distributed those features are. Equation 1 describes the process of selecting which columns of the feature vector to remove, where k is the number of bins, $m(i)$ is the number of samples in each individual bin, n is the number of total samples and T is the sampling threshold.

$$\sum_{i=1}^{k} m(i) > n \times T \tag{1}$$

If the number of bins (k) is 10, the 50 samples are then distributed between the 10 bins depending on their value. The number of samples in each of the 10 bins $m(i)$ are then checked against the threshold value which is calculated as $n \times T$. If $m(i)$ is greater then that column would be removed as this would mean that the contents of that column are too similar across the 50 samples. This method not only reduces the vector size, but for most cases also increases the system performance. The system is less confused by not having similar features across majority of the samples. Once the vector is reduced, PCA is applied to further reduce the vector size by calculating the coefficients which would best describe the features. The amount of coefficients we generate from the features are between 5-10, as this range gives the best performance when testing with the development data set. This new vector is then used at the regression stage to predict the depression values.

3.6 Regression

The Partial Least Squares (PLS) regression [9] is a statistical algorithm that bears some relation to principal components regression. Instead of finding hyperplanes of minimum variance between the response and independent variables, it builds a linear regression model by projecting the response and independent variables to another common space. Since both the response and independent variables are projected to a new space, the approaches in the PLS family are known as bilinear factor models.

More specifically, PLS tries to seek fundamental relations between two matrices (response and independent variables), i.e. a latent variable way to model the covariance structures in these two spaces. A PLS model aims to search the multidimensional direction in the independent variable space that explains the maximum multidimensional variance direction

in the response variable space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among independent variable values. By contrast, standard regression will fail in these cases.

In the case of depression recognition, it is necessary to reduce the dimension of the feature vector. In the AVEC2014 dataset, the first 50 samples are for training; another 50 for developing; and the left 50 for test. This is a relatively small number in comparison with the feature dimensionality, making the regression problem more redundant. For this reason, the feature selection technique is used to only concern the feature component that is relevant to the depression label. The correlation between feature vector and depression labels is computed in the training set and only the feature components with an absolute value bigger than a threshold are kept and others are discarded.

The general underlying model of multivariate PLS is

$$\begin{aligned} X &= TP^T + E \\ Y &= UQ^T + F \end{aligned} \tag{2}$$

where X is an $n \times m$ matrix of predictors and Y is an $n \times p$ matrix of responses. T and U are two $n \times l$ matrices that are, projections of X (scores, components or the factor matrix) and projections of Y (scores); P, Q are, respectively, $m \times l$ and $p \times l$ orthogonal loading matrices; and matrices E and F are the error terms, assumed to be independent and identical normal distribution. Decompositions of X and Y are made so as to maximize the covariance of T and U.

Linear Regression 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 decision fusion stage aims to combine multiple decisions into a single and consensus one [25]. The linear opinion pool method is used in this case due to its simplicity [3], and a weighted sum rule is defined to combine the predicted values from each decision as in [26].

4. EXPERIMENTAL RESULTS

4.1 AVEC2014 Dataset

The proposed approach is evaluated on the Audio/Visual Emotion Challenge (AVEC) 2014 dataset [27], a subset of the audio-visual depressive language corpus (AViD-Corpus). The dataset contains 340 video clips from 292 subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone in a number of quiet settings. There is only one person in each clip and some subjects feature in more than one clip. All the participants are recorded between one and four times, with an interval of two weeks. 5 subjects appears in 4 recordings, 93 in 3, 66 in 2, and 128 in only one session. The length of these clips is between 20 minutes and 50 minutes with the average of 25 minutes, and the total duration of all clips lasts 240 hours. 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.

4.2 Experimental Setting

AVEC2014 addresses two Sub-Challenges: the Affect Recognition Sub-Challenge (ASC) and the Depression Recognition Sub-Challenge (DSC). ASC concentrates fully on continuous affect recognition of the dimensions of dominance, valence and arousal, where the level of affect has to be predicted for each frame of the recording, while DSC requires to predict the level of self-reported depression as indicated by the Beck Depression Inventory (BDI) for every session, that is, one continuous value per video clip file. This study focuses on DSC, where a single regression problem needs to be solved. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) over all sessions are both used as measurements in the competition.

For each video clip, we work on 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. Then the algorithm on MHH was modified to be able to use for 1-D case and is operated on all the data to produce $M = 3$ vectors of temporal information. The vectors EOH_MHH, LBP_MHH and LPQ_MHH are reshaped producing 1152, 2835 and 768 components 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 sample, which is denoted as (MIX_MHH). Figure 4 shows MIX_MHH Feature vectors for 50 samples of the development set of the AVEC2014 dataset.

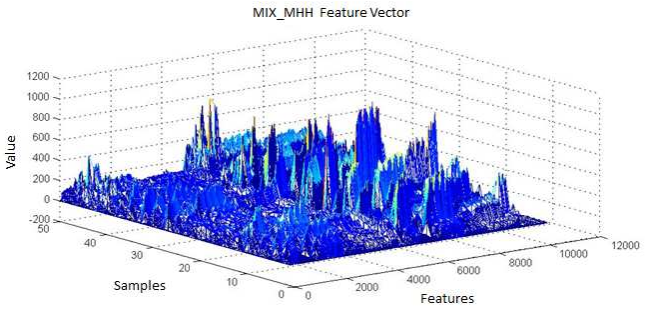

Figure 4: MIX_MHH Feature vectors for 50 samples of development subset of the AVEC2014 dataset.

To demonstrate the advantages of the proposed approach to depression recognition, we compare these achieved results with the ones of other configurations. The vectors EOH_MHH, LBP_MHH and LPQ_MHH have been tested with the development set before they are concatenated, to provide a comparison from its individual and combined benefits. Furthermore, we explore to model the temporal features of facial expressions in the dynamic feature space, similar to [14], i.e. first operating MHH on the video to produce 5 (M=5) frames, to then extract the local features (EOH, LBP and LPQ) from each of the 5 frames and then concatenate all the vectors, this is denoted as (MHH_MIX). The baseline audio features (2268) provided by data set for the short segments (short) and per instance (inst) have been used, denoted as (Audio). The combined audio features of Flatness, Band1000, POV, PSY Sharpness, Shimmer and ZCR are used, containing 285 of the 2268 features which is denoted as (Comb). For all the dynamic features including visual and vocal ones, the dimensionality is reduced with PCA and the results are provided by the fusion of PLS and Linear Regression.

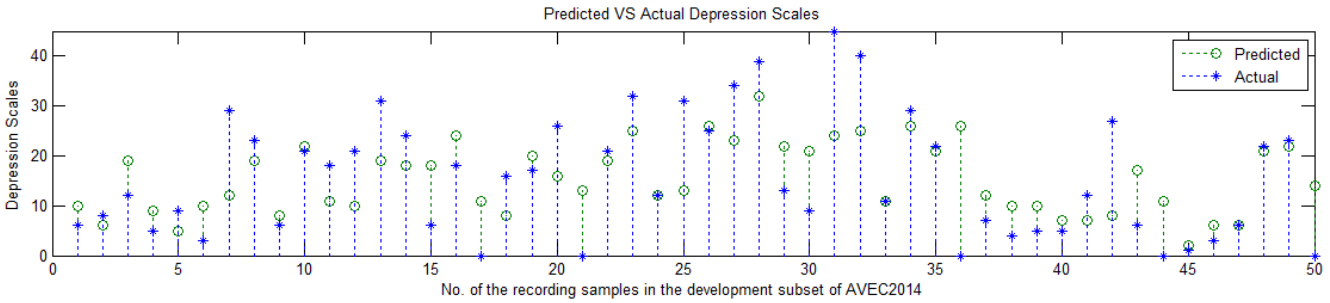

Figure 5: Predicted and actual depression scales of development subset of the AVEC2014 dataset based on audio and video features fusion at decision level.

4.3 Performance Comparison

Table 1 shows the individual performance of the three image feature extraction methods that are combined with 1-D MHH, from which the depression scales are predicted used two regression techniques separately and fused. We can see that using PLS for regression is better than Linear Regression in all tests. However, when they are fused with a weighting more towards PLS, the results are improved further. LBP is shown to be the weakest amongst the three and LPQ the strongest.

Table 1: Performance of depression scale prediction using the dynamic visual features measured both in MAE and RMSE averaged over all sequences in the development set.

Partition	Method	MAE	RMSE
Development	EOH_MHH_PLS	**8.87**	**11.09**
Development	EOH_MHH_LR	9.92	12.39
Development	EOH_MHH_(PLS+LR)	9.14	11.39
Development	LBP_MHH_PLS	9.34	11.16
Development	LBP_MHH_LR	9.86	12.68
Development	LBP_MHH_(PLS+LR)	**9.18**	**11.15**
Development	LPQ_MHH_PLS	8.79	10.88
Development	LPQ_MHH_LR	9.73	11.49
Development	LPQ_MHH_(PLS+LR)	**8.70**	**10.63**

Table 2 compares the combined performance of the three image feature extraction methods which has MHH operated before and after them. MIX_MHH is for MHH after EOH, LBP and LPQ are applied to each frame and MHH_MIX is when MHH is applied before, on the original video clip. MIX_MHH has shown a significant improvement over the individual performances in Table 1, as well as out-performing MHH_MIX in all regression methods showing that combining the different individual features does provide a much more reliable performance. The benefit of having MHH after EOH, LBP and LPQ is that these three methods will highlight more features and MHH would then capture the frequency of the features across the frames. Whereas if MHH is applied first to the raw video, a lot of good information is compressed down-to six frames before the features are extracted. Since the AVEC2014 data set is different to the previous AVEC2013 data setdata set (AVEC2013 containing significantly more data), MHH_MIX, which is the method approached in [14], will not produce similar results to the AVEC2014 data set.

Table 2: Performance of depression scale prediction using MHH before and after MIX (EOH, LBP and LPQ) visual features, measured both in MAE and RMSE averaged over all sequences in the development set.

Partition	Methods	MAE	RMSE
Development	MIX_MHH_PLS	7.72	9.68
Development	MIX_MHH_LR	7.52	10.05
Development	MIX_MHH_(PLS+LR)	**7.36**	**9.49**
Development	MHH_MIX_PLS	**8.91**	**10.78**
Development	MHH_MIX_LR	10.59	12.71
Development	MHH_MIX_(PLS+LR)	9.00	10.95

Table 3: Performance of depression scale prediction using complete audio and Comb features measured both in MAE and RMSE averaged over all sequences in the development set.

Partition	Methods	MAE	RMSE
Development	Comb_inst_(PLS+LR)	8.35	10.56
Development	Comb_short_(PLS+LR)	**8.16**	**10.32**
Development	Audio_inst_PLS	9.45	11.25
Development	Audio_short_PLS	**8.92**	**10.69**

In Table 3, the audio features for short segments and per instance are taken into account. From the 2268 audio features, the Combined features (Comb) have been taken out to be tested separately. The Combined features have produced good results when using PLS and Linear regression fused together, whereas when all the audio features are used, PLS alone works best. The four tests only take audio features into account and have shown good results when compared to the individual feature results in Table 1, however they aren't as good as the combined visual features (MIX_MHH).

Table 4 combines the audio domain with the best features from the visual domain at decision level, to show the performance when fused together. The results, when compared to Table 2 and Table 3, show a significant improvement when having audio alone and some improvement than having visual alone. PLS and Linear Regression fused has been chosen for the testing, because in most cases it gives an increase to the performance. The Comb audio features provide a better outcome than using all the audio features provided. The predicted and actual depression scale values on the develop-

Table 4: Performance of depression scale prediction using the audio features combined with the best visual features (MIX_MHH), fusing them at decision level, measured both in MAE and RMSE averaged over all sequences in the development set.

Partition	Methods	MAE	RMSE
Development	(MIX+Audio)_(PLS+LR)	8.92	10.71
Development	(MIX+Comb)_(PLS+LR)	**7.34**	**9.09**

Table 5: System performance of proposed depression scale prediction method measured in MAE and RMSE averaged over all sequences in development and test set.

Partition	Modality	MAE	RMSE
Development	Audio	8.92	10.69
Development	Video	7.36	9.49
Development	Video&Audio	**7.34**	**9.09**
Test	Audio	9.10	11.30
Test	Video	**8.44**	**10.50**
Test	Video&Audio	**8.30**	**10.26**

ment dataset are shown in Figure 5 based on (MIX+Comb) (PLS+LR) method.

Table 5 displays the final performance of each modality as well as fusing both at decision level on both development and test sets. The audio/visual domain has outperformed the other modalities in the development and test set. The fusion process gives more of the weighting to the visual only labels than the audio, based on its performance. However the results show that even though the audio labels aren't as good compared to visual, it still helps in predicting an accurate set of labels overall.

The baseline results [27] are listed in Table 6 which are limited in both development and test sets. In comparison with the results in Table 5, our method has outperformed their development RMSE score and is better overall than their result for the test data set.

5. CONCLUSIONS AND PERSPECTIVES

In this paper, a novel approach is proposed for automatic depression scale prediction based on facial and vocal expression in naturalistic video recordings. Based on the idea of MHH for 2-D video motion feature, we proposed 1-D MHH that can be applied to feature vector sequences and provide a dynamic feature (e.g. EOH_MHH, LBP_MHH, LPQ_MHH) for the video. This dynamic feature is better than MHH_EOH that was used in previous research [14] because it is based on feature vectors instead of raw images. Finally, PLS regression and Linear regression are then adopted to capture the correlation between and feature space and depression scales.

From the experimental results, it can be seen that the proposed method achieved good results on the AVEC2014 dataset. From Table 2, it clearly demonstrates the proposed dynamic feature is better than MHH_EOH that was used in previous research [14]. In comparison with Table 1, fusion of the three image features produce better results than any of the individual features. For the audio feature, Comb achieved better results than whole audio feature provided by the organizer. The fusion on decision level solution achieved better results on the development set, and the same result for the testing set.

Table 6: Baseline performance of depression recognition measured both in MAE and RMSE over all sequences [27].

Partition	Modality	MAE	RMSE
Development	Audio	-	-
Development	Video	-	9.26
Test	Audio	-	-
Test	Video	8.86	10.86

There are two main contributions from this paper. First one is the dynamic feature extraction that use the idea of MHH on the feature space. Another one is the feature fusion from different features from images. The overall results on the testing set are better than baseline results.

There are still other image features such as Gabor features, these features can be used to extract dynamic features as well and easily be added into the system to improve the performance. There are also some dynamic features such as LBP_TOP, combination of these kind of feature with our dynamic features together will be an interesting topic. For the audio feature, the combination of descriptors were used, other features should also be considered to be integrated in the system. All this will be our future work.

6. ACKNOWLEDGMENTS

The work by Asim Jan was supported by School of Engineering & Design/Thomas Gerald Gray PGR Scholarship. The work by Hongying Meng and Saeed Turabzadeh was partially funded by the award of the Brunel Research Initiative and Enterprise Fund (BRIEF). The work by Yona Falinie Binti Abd Gaus was supported by Majlis Amanah Rakyat (MARA) Scholarship.

7. REFERENCES

[1] `http://www.nimh.nih.gov/health/publications/depression/index.shtml`, Retrieved 15 July 2013.

[2] 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*, 28(12):2037–2041, 2006.

[3] I. Bloch. Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 26(1):52–67, 1996.

[4] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, 2001.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *International Conference on Multimodal Interaction*, pages 205–211, 2004.

[6] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video

sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003. Special Issue on Face Recognition.

[7] 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*, pages 1–7, 2009.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[9] S. de Jong. Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

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

[11] 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*, pages 296–301, 1994.

[12] 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.

[13] R. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In B. Kisacanin, V. Pavlovic, and T. Huang, editors, *Real-Time Vision for Human-Computer Interaction*, pages 181–200. Springer US, 2005.

[14] 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*, AVEC '13, pages 21–30, New York, NY, USA, 2013. ACM.

[15] H. Meng and N. Pears. Descriptive temporal template features for visual motion recognition. *Pattern Recognition Letters*, 30(12):1049–1058, 2009.

[16] 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.

[17] H. Meng, N. Pears, M. Freeman, and C. Bailey. Motion history histograms for human action recognition. In B. Kisačanin, S. Bhattacharyya, and S. Chai, editors, *Embedded computer vision*, Advances in pattern recognition, pages 139–162. Springer, 2009.

[18] 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*, pages 854–859, 2011.

[19] 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*, 24:971–987, 2002.

[20] V. Ojansivu and J. Heikkila. Blur insensitive texture classification using local phase quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *Image and Signal Processing*, volume 5099 of *Lecture Notes in Computer Science*, pages 236–243. Springer Berlin Heidelberg, 2008.

[21] M. Pantic and L. J. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1424–1445, 2000.

[22] S. Salmans. *Depression: questions you have - answers you need*. People's Medical Society, 1995.

[23] 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.

[24] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27(6):803–816, May 2009.

[25] 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*, pages 1–6, 2006.

[26] N. Ueda. Optimal linear combination of neural networks for improving classification performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):207–215, 2000.

[27] 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.

[28] 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.

[29] 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*, 168(1):224–238, 2008.

[30] 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*, AVEC '13, pages 41–48, New York, NY, USA, 2013. ACM.

[31] C. Yang, R. Duraiswami, and L. Davis. Fast multiple object tracking via a hierarchical particle filter. In *IEEE International Conference on Computer Vision*, pages 212–219, Washington, DC, USA, 2005. IEEE Computer Society.

[32] 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.