# Depression Recognition based on Dynamic Facial and Vocal Expression Features using Partial Least Square Regression

Hongying Meng*
Brunel University, UK

Di Huang*
Beihang University, China

Heng Wang
Beihang University, China

Hongyu Yang
Beihang University, China

Mohammed Al-Shuraifi
Brunel University, UK

Yunhong Wang
Beihang University, China

## ABSTRACT

Depression is a typical mood disorder, and the persons who are often in this state face the risk in mental and even physical problems. In recent years, there has therefore been increasing attention in machine based depression analysis. In such a low mood, both the facial expression and voice of human beings appear different from the ones in normal states. This paper presents a novel method, which comprehensively models visual and vocal modalities, and automatically predicts the scale of depression. On one hand, Motion History Histogram (MHH) extracts the dynamics from corresponding video and audio data to represent characteristics of subtle changes in facial and vocal expression of depression. On the other hand, for each modality, the 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. The proposed approach is evaluated on the AVEC2013 dataset and experimental results clearly highlight its effectiveness and better performance than baseline results provided by the AVEC2013 challenge organiser.

## Categories and Subject Descriptors

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

## Keywords

Affective computing, emotion recognition, speech, facial expression, challenge

*These authors equally contribute to this work.


*AVEC'13,* October 21, 2013, Barcelona, Spain.

ACM 978-1-4503-2395-6/13/10 ...$15.00.
http://dx.doi.org/10.1145/2512530.2512532.

## 1. INTRODUCTION

In recent years, the study on mental health problems has been given increasing attention from various domains in the modern society. According to EU Green Papers established in 2005 [19] and 2008 [20] respectively, mood disorders, to some extent, affect one in four citizens of working age, causing significant losses and burdens to the economic, social, educational, as well as justice systems [44]. Among these mood disorders already defined, 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 persons' thoughts, behavior, feelings, and sense of well-being [33]. 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]. Most of existing clinical techniques, e.g. the Scale for Assessment of Negative Symptoms (SANS [3]), are subjectively rated and therefore provide qualitative measurements. In addition, they also require extensive human expertise and interpretation. In such case, 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.

In spite of very limited progress currently achieved on depression recognition or other mental disorders, recent technological revolutions of basic 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 between facial expressions and emotional states [14]. Initial work aims at describing expressions using static features extracted from still face images [31, 37]. However, to some expressions, especially the subtle ones like anger and disgust, these features prove incompetency. More 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. [30] [41] [9] [21] [27]. 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 [7] [49].

Similarly, in depression recognition, 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 to depression recognition using both the visual and vocal clues, aiming to combine their advantages. Motion History Histogram (MHH) is firstly exploited to extract dynamic features from videos and audios, comprehensively describing emotional fluctuation in the two channels. The Partial Least Square (PLS) regression algorithm is then used to model the mapping between dynamic features and the depression state on each modality respectively. The fusion is finally made at decision level for prediction. Experimental results achieved on the AVEC2013 dataset illustrate the effectiveness of the proposed method.

Our main contributions remain three fold: (1) MHH based dynamic features are introduced to describe the behavioural characteristics of facial expressions and naturalistic vocal expressions in depression recognition; (2) EOH and LBP are adopted to highlight the edge and texture variations of temporal details of the HMM respectively, and their feature level fusion leads to a more comprehensive dynamic representation; (3) Partial Least Square (PLS) regression is employed to predict the depression states by learning the relationship between the dynamic features and depression scale in training.

The rest part 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 AVEC2013 dataset [44]. Section 5 concludes the paper.

## 2. RELATED WORK

To the best of our knowledge, the progress that has been achieved so far on affective computing mental disorder analysis is remarkably little. The primary published study [45] towards such an issue dates in 2008 from University of Pennsylvania. Wang et al. 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. They also pointed out that temporal information is essential to understand facial expressions. Later, they enhanced the approach by combining Gabor based facial texture features with the previous AAM based geometry ones, claiming that the features are complementary and their fusion thus improves the performance of Action Unit (AU) detection [18].

As particularly concerning on depression analysis, Cohn et al. [10] 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. Several more recent research [34] [48] [17] show a growing focus on this topic using either the facial expression or verbal modality. Specifically, Yang et al. [48] 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. [17] analyzed both manual and automatic facial expressions during semi-structured clinical interviews of clinically depressed patients, concluding that participants with high symptom severity behave more expressions associated with contempt, smile less, and their smiles were more likely to be related to contempt. Scherer et al. [34] 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.

Compared to the quite few studies for mental health analysis, especially on depression detection, the research within the domain of affective computing and social signal processing on basic emotion (including anger, disgust, fear, happiness, sadness, and surprise) recognition is pervasive, where the key problems, i.e. representing and measuring the changes in the modality of facial expression and acoustic audio, are widely discussed.

The recent years have witnessed the achievements in Automatic Facial Expression Recognition (AFER), both in the aspect of methodology and data. The AFER approaches in the literature can be distinguished according to the way they describe the facial expression and roughly categorized as feature based and model based. Inspired by the Facial Expression Coding System (FACS) proposed by Ekman [14], which codes a facial expression using the patterns of facial muscle movement, the former ones concentrate on extraction of expression sensitive features and represent a facial expression as a set of local descriptions, such as geometry features [40], Local Binary Patterns [37], and Gabor Wavelets [50]. While the latter ones build a generic model based on which common expressions can be indicated by measuring the feature vector composed by the parameters of shape deformation or texture variation achieved in the fitting process. One typical method falling in this category is Active Appearance Model [13] that allows for the decoupling of the shape of the face from its appearance. The performance can be improved by integrating both two types of methods. To predict the facial expression label, extracted features are further fed to classifiers, e.g. the Nearest Neighbor classifier, Neural Networks, Support Vector Machine (SVM), AdaBoost and Sparse Representation Classifier (SRC).

The evolution of techniques and the update of databases in facial expression analysis have always been staggered. With the development of camera devices, facial expressions can be recorded by videos (or continuous image sequences). Besides static information provided by separate still images, videos

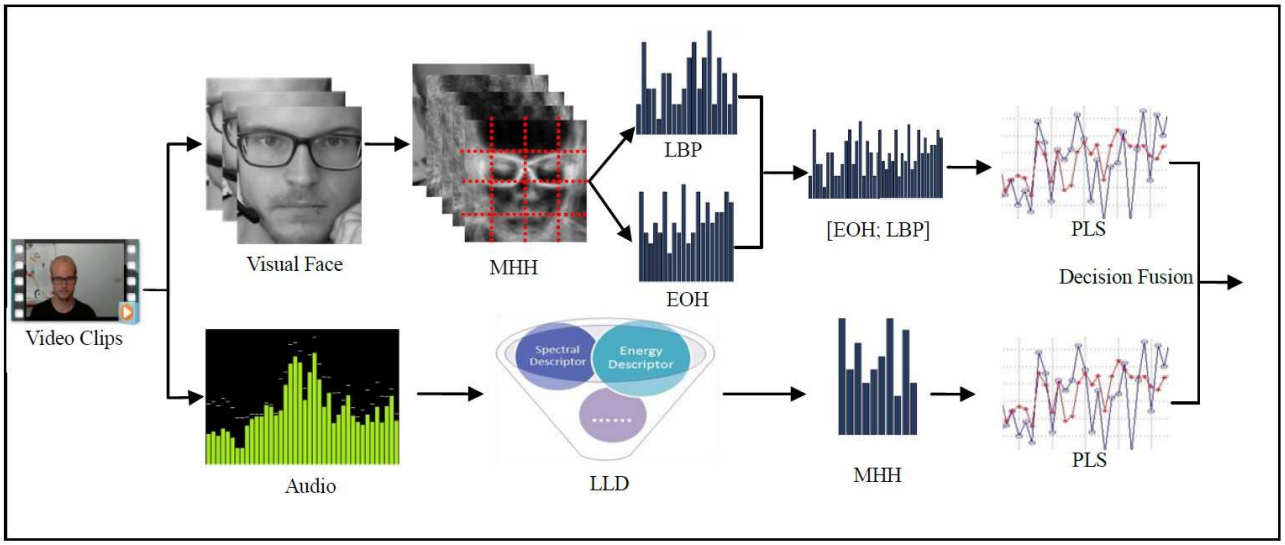


**Figure 1: Framework of the proposed automatic depression recognition system.**

convey temporal information, which is crucial to detect more subtle facial expression changes. Yacoob and Davis [46] divided the video of each facial expression into three segments: the beginning, apex and ending, and ad-hoc rules were defined to classify the temporal model of the facial expression. Cohen et al. [9] introduced Motion-Units (MU) to describe facial expressions in terms of magnitudes, and two-stage approach was proposed to combine the analysis at both levels of static and dynamic features. Furthermore, as most current studies focused on acted dataset (e.g. FERA2011 [43]), there is an increasing demand to work on more naturalist expressions (e.g. AVEC2011 [36], AVEC2012 [35]) in order to improve the performance of such technology in a naturalist setting of practical application [22].

Regrading the audio modality, one of most important issues is also the extraction of speech features that efficiently characterize the emotional content of speech and at the same time are irrelevant to the speaker and the lexical content [4]. Current Speech features can be taxonomically grouped into: continuous, spectral, qualitative, and TEO (Teager Energy Operator)-based. Continuous features include the pitch, energy, formant etc., largely related to the arousal state of the speaker [32]. Qualitative features contain main attributes of voice quality such as harsh, tense, breathy, having a strong relation with the the perceived emotion [8]. Spectral features are usually selected as a short-time representation for speech signals, and Nwe et al. [28] demonstrated that the emotional content of an utterance has an impact on the distribution of spectral energy across the speech range of frequency. Considering that the speech is produced by non-linear air flow in vocal systems, Teager [39] proposed the TEO-based feature with the supporting evidence that hearing is the process of detecting energy, followed by several recent variants [51]. Among these features, which one is the best still remains unsolved, and the choice of proper features for speech emotion recognition highly depends on the task of classification being considered [4]. In addition to these classifiers that are mentioned above for facial expression analysis, Hidden Markov Models (HMM) [23], Gaussian Mixture Models (GMM) are popular tools for this modality as well.

# 3. DEPRESSION RECOGNITION

Since human facial expression and voice in depression are theoretically different from those under normal mental states, we attempt to address the problem of depression recognition by combining dynamic descriptions of the facial expression as well as naturalistic oral expressions. This paper proposes a novel method that comprehensively models the variations in visual and vocal clues and automatically predicts the scale of depression.

## 3.1 Approach Overview

Figure 1 depicts the framework of the proposed approach to depression recognition. For each video clip, we deal with the channel of visual and vocal signal respectively, and the system hence involves in two independent steps. In the step of video process, Motion History Histograms (MHH) [24] is introduced to capture the movement of each pixel (texture variation) within the face area, describing temporal information during facial expressions. The details in MHH based dynamic information are then highlighted by Edge Orientation Histogram (EOH) and Local Binary Patterns (LBP), whose features are further concatenated for more better representation. The Partial Least Square (PLS) regression is finally applied to predict the depression scale. While in the step of audio process, a set of spectral Low-Level Descriptors (LLD) features are employed to encode the characteristics of the audio. MHH is then used to extract change information of the vocal expression, followed by the PLS based repression as does in video process

## 3.2 Dynamic Feature Extraction

Based on the modality of facial expression and vocal expression recorded in video data, two dynamic features are extracted respectively, which are detailedly presented in the subsequent.

### *3.2.1 Dynamic Video Feature*

Dynamic video feature is extracted in a serial way by first computing MHH from the continuous image sequences, then

highlighting its temporal details by EOH as well as LBP, and finally concatenating the two achieved feature vectors.

MHH is a descriptive representation of temporal template for visual motion recognition, originally proposed and applied to human action recognition [25]. MHH records gray scale value changes for each pixel in the video. In comparison with other well-known motion features, such as Motion History Image (MHI) [6], it contains more dynamic information of the pixels and achieves better performance in the application of human action recognition. Furthermore, MHH still remains computationally inexpensive [26].

An MHH feature extracted from a video clip is $M$ grayscale images of the same size as the video frame. $M$ is the level of motion in the video, in which the bigger value means faster movements and the smaller value means slower movement. It represents the patterns of movement (pixel value change) on each pixel. For example, $M$ is set at 3, indicating that the values of a pixel are consecutively changed for 3 times over 4 frames. According to some preliminary experiments, $M = 5$ is sufficiently large to capture the majority of movement information on a pixel in a video for depression recognition.

Figure 2 shows some examples of MHH features from the AVEC2013 dataset. They are actually grayscale images although they look like binary images. More detailed technical information of can be found in [24] and [26].

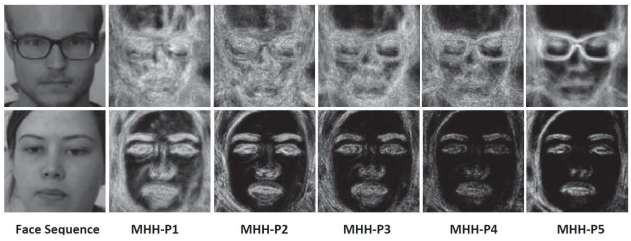


**Figure 2: Examples of MHH features extracted from samples on the AVEC2013 dataset.**

To highlight the details of the dynamic features encoded by MHH, both the Edge Orientation Histogram (EOH) and Local Binary Patterns (LBP) operators are utilized on the $M$ MHH images respectively.

EOH, an efficient and powerful operator, is regarded as a simpler version of Histogram of Oriented Gradients (HOG) [11] 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 [16], object tracking [47] and facial expression recognition [27].

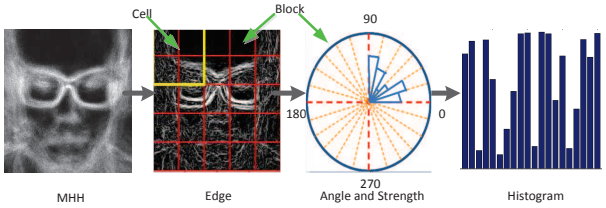


**Figure 3: The process for computing EOH.**

The process for computing EOH is demonstrated in Fig. 3. For an image $f(u, v)$, edges are detected using the horizontal and vertical Sobel operators [15], $K_u$ and $K_v$ as:

$$G_u(u, v) = K_u * f(u, v) \quad (1)$$

$$G_v(u, v) = K_v * f(u, v) \quad (2)$$

The strength $S$ and the orientation $\theta$ of the edges are

$$S(u, v) = \sqrt{G_u^2(u, v) + G_v^2(u, v)} \quad (3)$$

$$\theta = arctan(G_u(u, v)/G_u(u, v)) \quad (4)$$

The angle interval is divided into $N$ bins and the strengths in the same bin are summed to build a histogram. The whole image was divided into cells and each cell was divided into blocks. The histograms from each blocks were linked together to generate the EOH feature.

LBP, a non-parametric descriptor, summarizes local texture structures of images into a set of patterns. It is highly discriminative and its key advantages, namely invariance to monotonic gray level changes and computational efficiency, make it successful and popular in many topics, e.g. texture classification [29] and face recognition [2].

The basic LBP operator labels the pixels of an image with decimal numbers, called LBP codes, which encode the local structure around each pixel. It proceeds hence, as illustrated in Fig.4: Each pixel is compared with its $P$ neighbors on a circular neighborhood with the radius $R$, and these resulting strictly negative values are assigned to 0 and the others are assigned to 1. A binary number is obtained by concatenating all the $P$ binary codes in a clockwise direction starting from the top-left one and its corresponding decimal value is used for labeling. The operator denoted as $\text{LBP}_{(P,R)}$ produces $2^P$ different output values, corresponding to $2^P$ different binary patterns. Because certain patterns convey more information than others [29], it is possible to make use of only a subset of $2^P$ binary patterns to describe the texture of images, namely uniform patterns. A pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the corresponding bit string is considered circular. Uniform LBP proves compact and more effective than the basic LBP, and we hence apply it to highlight the details of MHH based temporal information.

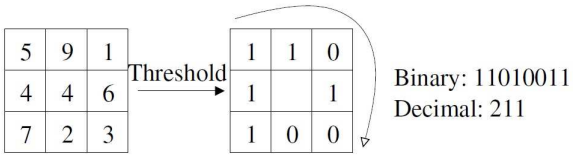


**Figure 4: The basic LBP operator.**

$\{\text{MHH}(:, :, i), i = 1, \cdots, M\}$ of a video clip is treated as $M$ grayscale images. In order to capture both local and spatial information, each MHH based image is divided into several blocks, from each of which EOH and LBP features are computed, and all these features are then concatenated into the MHH_EOH and MHH_LBP representation respectively. Finally, MHH_EOH and MHH_LBP are connected as a more comprehensive description of temporal details. An example of the MHH_LBP feature is shown in Fig. 5.

### 3.2.2 *Dynamic Audio Feature*

The spectral low-level descriptors (LLDs) and MFCCs 11-16 are included in the AVEC2013 baseline audio feature set

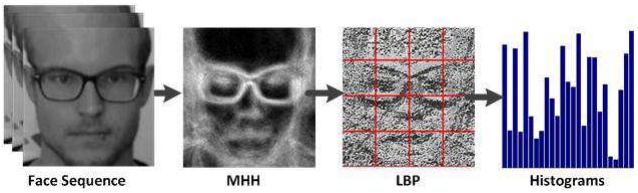


**Figure 5: The illustration of the process for MHH_LBP feature extraction.**

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 (all including and below F0final_sma) are extracted from 60 ms windows, while all other LLD are extracted from 25 ms windows. For detailed information, please see the baseline paper [44].

The dynamic property between consecutive audio segments is also critical. The vocal expression changes are considered to be discriminative between depression and other emotions using MHH. Instead of operating on each pixel in videos, each component of the audio baseline feature is used for dynamic modelling. For each component, a change sequence is created and its corresponding histogram is made based on the variation in patterns. Similarly, $M$ values are obtained for each component. The final dynamic audio feature vector is therefore $M$ times longer than the original one.

## 3.3 PLS Regression

The Partial Least Squares (PLS) regression [12] 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 AVEC2013 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 abstract value bigger than a threshold are kept and others are discarded.

The general underlying model of multivariate PLS is

$$\begin{aligned} X &= TP^T + E \\ Y &= UQ^T + F \end{aligned} \tag{5}$$

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 $U$.

## 3.4 Decision Fusion

The decision fusion stage aims to combine multiple decisions into a single and consensus one [38]. The linear opinion pool method is used in this case due to its simplicity [5], and a weighted sum rule is defined to combine the predicted values from each decision as in [42].

$$D_{\texttt{linear}}(\hat{x}) = \sum_{i=1}^{K} \alpha(i) D_i(\hat{x}) \tag{6}$$

where $\hat{x}$ is a testing sample and $D_i(\hat{x})$ is its $i_{th}$ decision value $(i = 1, 2, ..., K)$ while $alpha(i)$ is its corresponding weight which should satisfy $\sum_{i=1}^{K} \alpha(i) = 1$.

# 4. EXPERIMENTAL RESULTS

## 4.1 AVEC2013 Dataset

The proposed approach is evaluated on the Audio/Visual Emotion Challenge (AVEC) 2013 dataset, 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. Some examples of the AVEC2013 dataset are shown below in Fig.6.


**Figure 6: Video Samples on the AVEC2013 dataset.**

## 4.2 Experimental Setting

AVEC2013 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 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 Index (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 competition.

For each video, MHH produces 5 ($M = 5$) images of temporal information of each video clip (see Fig. 2 for an illustration). EOH then operates on each MHH image, leading to a 384-dimensional feature vector, and the final MHH_EOH representation concatenates all the 5 EOH features to make a vector of 1920 components. Similarly, $\text{LBP}_{(8,1)}$ is also applied to these MHH images, each of which corresponds to a histogram feature of 944 dimensions (59 bins $\times$ 16 blocks), and the final MHH_LBP representation thus possesses 4720 components.

To demonstrate the advantages of the proposed approach to depression recognition, we compare these achieved results with the ones of other configurations. In the video track, we also carry out experiments using Local Phase Quantization (LPQ) to highlight the details of MHH images (denoted as MHH_LPQ). Furthermore, we explore to model the temporal features of facial expressions in the local feature space, i.e. first computing different local features (LBP and LPQ) of each frame and then operating MHH for extracting dynamic information (denoted as LBP_MHH and LPQ_MHH). For all the dynamic features including visual and vocal ones, the results provided by Support Vector Regression (SVR) are displayed as well.

## 4.3 Performance Comparison

From Table 1, we can see that using both SVR and PLS, MHH_LBP and MHH_LPQ outperform LBP_MHH and LPQ_MHH respectively, showing that the dynamic feature encoded by LBP or LPQ from MHH images is more effective than that modeled by MHH from the feature space of LBP and LPQ. Meanwhile for all the configurations in this table, SVR achieves better results than PLS does. Additionally, MHH_LBP and MHH_LPQ are much less computationally expensive than LBP_MHH and LPQ_MHH, since they only apply LBP and LPQ to the 5 MHH images rather than to each frame in the video.

In Table 2, MHH_EOH achieves better accuracy than MHH_LBP in terms of MAE and RMSE based on both PLS and SVR. On the other hand, PLS outperforms SVR using MHH_EOH while SVR outperforms PLS using MHH_LBP. Furthermore, when we combine the representation of MHH_EOH and MHH_LBP, both SVR and PLS improve the performance compared with that of either of them along, indicating that EOH and LBP present complementary information when highlighting temporal details of MHH images. PLS obtains the best scores (MAE:7.08 and RMSE: 8.81) based on the visual modality. It should be noted that LPQ is not considered here because it captures texture information which is similar to that of LBP.

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

| Partition | Method | MAE | RMSE |
|---|---|---|---|
| Development | LBP_MHH_SVR | 8.74 | 10.89 |
| Development | LBP_MHH_PLS | 9.02 | 11.14 |
| Development | LPQ_MHH_SVR | 9.04 | 10.98 |
| Development | LPQ_MHH_PLS | 9.96 | 11.26 |
| Development | MHH_LBP_SVR | **7.83** | **9.67** |
| Development | MHH_LBP_PLS | 9.77 | 11.72 |
| Development | MHH_LPQ_SVR | 8.01 | 9.89 |
| Development | MHH_LPQ_PLS | 10.28 | 12.46 |

**Table 2: Performance of depression recognition using dynamic MHH_EOH and MHH_LBP visual features measured both in MAE and RMSE averaged over all sequences in the development set.**

| Partition | Methods | MAE | RMSE |
|---|---|---|---|
| Development | MHH_EOH_PLS | 7.16 | 8.86 |
| Development | MHH_LBP_PLS | 9.77 | 11.72 |
| Development | MHH_(EOH+LBP)_PLS | **7.08** | **8.81** |
| Development | MHH_EOH_SVR | 7.79 | 9.36 |
| Development | MHH_LBP_SVR | 7.83 | 9.67 |
| Development | MHH_(EOH+LBP)_SVR | 7.57 | 9.07 |

In Table 3, MHH is carried out on the three audio features (i.e. long, short and valid segmented) and PLS is then used for regression to predict the depression scale as does in the visual modality. From this table, the valid segmented audio feature achieves better accuracy than the other two although the difference is very small.

Table 4 displays the final performance of each modality as well as their combination with feature selections. We can see that when we fuse both the prediction values of the visual and vocal clues at decision level, the results of the proposed depression recognition system are further improved, showing that the joint use of both modalities is a promising way for amelioration

The baseline results [44] in both development and test sets are listed in Table 5. In comparison with the results in Table 4, it can be seen that significant improvement have been achieved by the proposed approach based on the visual modality, while the one based on the vocal modality is somewhat behind. Overall, the proposed approach based on both modalities outperform the best result in the baseline in both development and test sets.

**Table 3: Performance of depression recognition using MHH on audio features measured both in MAE and RMSE averaged over all sequences in the development set.**

| Partition | Methods | MAE | RMSE |
|---|---|---|---|
| Development | MHH_PLS_long | 9.75 | 11.72 |
| Development | MHH_PLS_short | 9.77 | 12.09 |
| Development | MHH_PLS_vad_seg | **9.47** | **11.63** |

**Table 4: System performance of proposed depression recognition method measured in MAE, RMSE and cross-correlation (CORR) averaged over all sequences in development and test set.**

| Partition | Modality | MAE | RMSE | CORR |
|---|---|---|---|---|
| Development | Audio | 9.78 | 11.54 | 0.42 |
| Development | Video | 7.09 | 8.82 | 0.67 |
| Development | Video&Audio | **6.94** | **8.54** | **0.70** |
| Test | Video | 9.14 | 11.19 | - |
| Test | Video&Audio | **8.72** | **10.96** | - |

**Table 5: Baseline performance of depression recognition measured both in MAE and RMSE over all sequences [44].**

| Partition | Modality | MAE | RMSE |
|---|---|---|---|
| Development | Audio | 8.66 | 10.75 |
| Development | Video | 8.74 | 10.72 |
| Test | Audio | 10.35 | 14.12 |
| Test | Video | 10.88 | 13.61 |

# 5. CONCLUSIONS AND PERSPECTIVES

In this paper, a novel approach is proposed for automatic depression recognition based on facial expression and vocal expression recordings. To model temporal changes conveyed in both the video and audio modalities, MHH based dynamic features are extracted. The PLS regression is then adopted to capture the correlation between and feature space and depression label. The method is validated on the AVEC2013 dataset and experimental results clearly demonstrate its effectiveness in comparison with the baseline performance. The performance can be even better if the dynamic feature on audio modality is further improved.

With the encouraging performance achieved in DSC of the AVEC2013 challenge, we are currently working on ASC as well. However, due to time limitation, the detailed information can not be presented in this paper, but probably shown in the near future.

# 6. ACKNOWLEDGMENTS

The work by Hongying Meng was partially funded by the award of the Brunel Research Initiative and Enterprise Fund (BRIEF) and research exchange major award of UK Royal Academy of Engineering.

This work by Di Huang was partially supported by the National Basic Research Program of China under Grant 2010CB327902, the National Natural Science Foundation of China under Grant 61202237, the international joint project by the LIA2MCSI Laboratory between Écoles Centrales and Beihang University, and the Fundamental Research Funds for the Central Universities.

# 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] N. C. Andreasen. *Scale for the Assessment of Negative Symptoms*. Department of Psychiatry, College of Medicine, the University of Iowa, 1984.

[4] M. E. Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[5] 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.

[6] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[7] 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.

[8] C.Gobl and A. N. Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2):189–212, 2003.

[9] 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.

[10] 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.

[11] 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.

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

[13] G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *European Conference on Computer Vision*, pages 581–595, London, UK, 1998. Springer-Verlag.

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

[15] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, us ed edition, August 2002.

[16] 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.

[17] 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.

[18] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods*, 200(2):237–256, 2011.

[19] Health & Consumer Protection Directorate General. Improving the mental health of the population: Towards a strategy on mental health for the european union, 2005.

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

[21] 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.

[22] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 41:1027–1038, 2011.

[23] H. Meng and N. Bianchi-Berthouze. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Cybernetics*, 2013.

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

[25] H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop on Embedded Computer Vision*, pages 1–6, 2007.

[26] 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.

[27] 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.

[28] T. Nwe, S. Foo, and L. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(1-2):603–623, 2003.

[29] T. Ojala, M. 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(7):971–987, 2002.

[30] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(2):433–449, 2006.

[31] 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.

[32] R.Cowie and R. R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, 2003.

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

[34] 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.

[35] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. AVEC 2012: The continuous audio/visual emotion challenge. In *ACM International Conference on Multimodal Interaction*, pages 449–456, 2012.

[36] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011: The first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, 2011.

[37] 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.

[38] 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.

[39] H. Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Process*, 28(5):599–601, 1990.

[40] Y. L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[41] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, Oct. 2007.

[42] 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.

[43] M. F. Valstar, B. Jiang, M. Méhu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011.

[44] 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.

[45] 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.

[46] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequeces using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[47] 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.

[48] 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.

[49] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson. Audio-visual affect recognition through multi-stream fused HMM for HCI. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 967–972, 2005.

[50] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 454–459, 1998.

[51] G. Zhou, J. Hansen, and J. Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3):599–601, 2001.