Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

Multimodal Assistive Technologies for Depression Diagnosis and Monitoring

Jyoti Joshi[1] · Roland Goecke[1,2] · Abhinav Dhall[2] · Sharifa Alghowinem[2] · Michael Wagner[1] · Julien Epps[3] · Gordon Parker[3] · Michael Breakspear[4]
[1]University of Canberra, [2]Australian National University, [3]University of New South Wales, [4]Queensland Institute of Medical Research
jyoti.joshi@canberra.edu.au, roland.goecke@ieee.org, abhinav.dhall@anu.edu.au, sharifa.m.f@gmail.com, j.epps@unsw.edu.au, g.parker@blackdog.org.au, mjbreaks@gmail.com

Received: date / Accepted: date

Abstract Depression is a severe mental health disorder with high societal costs. Current clinical practice depends almost exclusively on self-report and clinical opinion, risking a range of subjective biases. The long-term goal of our research is to develop assistive technologies to support clinicians and sufferers in the diagnosis and monitoring of treatment progress in a timely and easily accessible format. In the first phase, we aim to develop a diagnostic aid using affective sensing approaches. This paper describes the progress to date and proposes a novel multimodal framework comprising of audio-video fusion for depression diagnosis. We exploit the proposition that the auditory and visual human communication complement each other, which is well-known in auditory-visual speech processing; we investigate this hypothesis for depression analysis. For the video data analysis, intra-facial muscle movements and the movements of the head and shoulders are analysed by computing spatio-temporal interest points. In addition, various audio features (fundamental frequency f0, loudness, intensity and mel-frequency cepstral coefficients) are computed. Next, a bag of visual features and a bag of audio features are generated separately. In this study, we compare fusion methods at feature level, score level and decision level. Experiments are performed on an age and gender matched clinical dataset of 30 patients and 30 healthy controls. The results from the multimodal experiments show the proposed framework's effectiveness in depression analysis.

Keywords Depression Analysis, Multimodal, LBP-TOP, Bag of Words

1 Introduction

Affect, meaning emotions and mood, is an essential, integral part of human perception and communication. As research in the last two decades has shown, emotions and the display of affect play an essential role not only in cognitive functions such as rational decision making,

Corresponding author: Jyoti Joshi
University of Canberra, ACT 2601 AUSTRALIA

perception and learning, but also in interpersonal communication [33]. Affective sensing – the sensing of affective states – plays a key role in emerging transformational uses of IT, such as healthcare, security and next generation user interfaces. Recent advances in affective sensing, e.g. automatic face tracking in videos, measuring facial activity, recognition of facial expressions, analysis of affective speech characteristics and physiological effects that occur as a result of affective state changes, paired with the decreasing cost and increasing power of computing, have led to an arsenal of prototypical affective sensing tools now at our finger tips. We can employ these to tackle higher problems, e.g. supporting clinicians in the diagnosis of mental health disorders.

Depression is one of the most common and disabling mental disorders, and has a major impact on society. The landmark WHO 2004 Global Burden of Disease report by Mathers *et al.*[24] quantified depression as the leading cause of disability worldwide (an estimated 154 million sufferers). The lifetime risk for depression is reported to be at least 15% [19]. People of all ages suffer from depression, which is also a major risk factor for suicide. Fortunately, depression can be ameliorated through the provision of suitable objective technology for diagnosing depression to health professionals and patients [34]. Disturbances in the expression of affect reflect changes in mood and interpersonal style, and are arguably a key index of a current depressive episode. This leads directly to impaired interpersonal functioning, causing a range of interpersonal disabilities, functioning in the workforce, absenteeism and difficulties with a range of everyday tasks (such as shopping). Whilst these are a constant source of distress in affected subjects, the economic impact of mental health disorders through direct and indirect costs has long been underestimated. Despite its severity and high prevalence, there currently exist no laboratory-based measures of illness expression, course and recovery. This compromises optimal patient care, compounding the burden of disability. As healthcare costs increase worldwide, the provision of effective health monitoring systems and diagnostic aids is highly important. Affective sensing technology can and will play a major role in this. With the advancement of affective sensing and machine learning, computer aided diagnosis can and will play a major role in providing an objective assessment.

In a close collaboration of computer scientists and psychologists, we aim to develop multimodal assistive technologies that support clinicians during the diagnosis, and both clinicians and sufferers in the monitoring of treatment progress. The development of an objective diagnostic measure for a leading cause of disability worldwide would represent a major diagnostic breakthrough with significant future medical possibilities. The proposed multimodal approach will underpin a new generation of objective laboratory-style markers of illness expression. In the first phase, we investigate multimodal affective sensing technologies, in particular face and voice analysis techniques, for a prototypical system that is tested at the Black Dog Institute – a clinical research institute focussing on depression and other mental health disorders – in Sydney, Australia, and at the Queensland Institute of Medical Research, Australia. In the medium term, we plan to translate the developed approaches into an assistive laptop system, so that clinicians and patients can assess response to treatment in a timely and easily accessible format. In the long term, we hope to assist patients with depression to monitor the progress of their illness in a similar way that a patient with diabetes monitors their blood sugar levels with a small portable device, e.g. a smartphone. In mental healthcare, our approach also lends itself to expansion into other disorders (schizophrenia, autism, bipolar disorder), where laboratory-style diagnoses are also lacking.

The aim of this study is to investigate the utility of affective sensing methods for automated depression analysis, which can assist clinicians in depression diagnosis and monitoring. The proposed framework is based on extracting audio-video features and comparing various fusion approaches at different stages.

2 Related Work

Inferring emotions from facial expression analysis is a well-researched problem [31,41]. Over the past two decades, various geometric, texture, static and temporal visual descriptors have been proposed for various expression analysis related problems (e.g. [41,23,3,42]). Emotion analysis methods can be broadly divided into three categories based on the type of feature descriptor used. Shape feature based methods such as [3,9] are based on facial geometry only. The second class are the appearance features based emotion analysis methods [42,10,11], which are based on analysing the skin texture. The third category are the hybrid methods [23], which used both shape and appearance features. [42] show that appearance based features are more effective in emotion analysis as they are able to capture subtle facial movements, which are difficult to capture otherwise using shape based features.

This knowledge can also be used for depression analysis and it is, therefore, no surprise that computer-based depression analysis research to date has been drawing inspirations from this mature research field [7]. Various audio and video-based methods have been proposed in the past, of which we can only list some here. In one of the first seminal works for automatic depression analysis, Cohn *et al.* [7] explored the relationship between Facial Action Coding System (FACS)-based facial and vocal features and clinical depression detection. They learnt subject-dependent Active Appearance Models (AAM) [12,35] to automatically track facial features. The shape and appearance features after AAM fitting are further used to compute parameters such as the occurrence of so called FACS Action Units (AU, associated with depression), mean duration, ratio of onset to total duration and ratio of offset to onset phase. However, the audio and video features were not fused. To the best of our knowledge, our proposed framework is the first multimodal attempt at depression analysis.

According to Ellgring's hypothesis [13], depression leads to a remarkable drop in facial activity, while facial activity increases with the improvement of subjective well-being. Considering Ellgring's hypothesis as a starting point, McIntyre *et al.* [25] analysed the facial response of the subjects when shown a short video clip. Like Cohn *et al.* [7], subject-specific AAMs were learned and shape features were computed from every fifth video frame. The shape features were combined and classified at the frame level by the means of Support Vector Machine (SVM). However, facial activity is dynamic in nature. It has been shown in the literature that temporal facial dynamics provide more information than using static information only [2].

A limitation of both [7] and [25] is their use of subject-specific AAM models. For a new subject, a new AAM model needs to be trained, which is both complex and time consuming. In contrast, the video analysis in our proposed framework is subject-independent. It has been shown in the literature that temporal texture features perform better than geometric features only for dynamic facial expression analysis [42]. Simple temporal features, such as mean duration of AUs [7], have been used. However, in this paper, sophisticated spatio-temporal descriptors (Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) and Space-Time Interest Points (STIP), see Section 4.1), which have been successfully used for incorporating temporal information in earlier facial expression recognition approaches [42], have been applied.

In our recent work for automatic depression recognition, [17] a vision-based framework is proposed which is based on analysing facial dynamics using LBP-TOP and body movements using STIP in a Bag-of-Words (BoW) framework. Various classifiers were also compared on these features. In this paper, a similar vision-based pipeline is used. In our another work [18], a thorough comparison of the discriminative power of facial dynamics and the remaining body parts for depression detection is provided. Further, a histogram of head movements is proposed and results show that head movements alone are a powerful cue for depression detection.

In general emotion recognition from speech information, Mel-scale Frequency Cepstral Coefficients (MFCC) are considered one of the more relevant features [4]. For example, MFCCs were investigated in [8], who found that the classification results were statistically significant for detecting depression. The MFCCs are a compact representation of the short-time power spectrum of speech after weighting the frequency scale in accordance with the frequency sensitivity of human hearing. Pitch features, which have been widely investigated in the literature for prosody analysis, show a lower range of fundamental frequency f0 in depressed subjects [28,29,20,14], which increases after treatment [30]. The lower range of f0 indicates monotone speech [26] and its low variance indicates a lack of normal expression in depressed subjects [27]. f0 estimation, often also referred to as pitch detection, has been a popular approach used in speech processing in general and lately for speech-based emotion recognition. f0 is the lowest frequency of a periodic waveform. Several methods are used to estimate the f0 values, mostly based on the Auto-Correlation Function (ACF). In this paper, f0-raw is used as it results in better depression recognition [1].

There is convincing evidence that sadness and depression are associated with a decrease in loudness [37], showing lower loudness values for depressed subjects. Since the loudness is intimately related to sound intensity, both features are investigated (see Section 4.2). Sound intensity I is measured as the sum over a short time frame of the squared signal values. Loudness L is directly related to sound intensity, describing the magnitude of the auditory sound intensity sensation. A gain in performance is reported by [38] in speech-based emotion recognition by fusing several acoustic features as compared to single features only. Therefore, the effect of fusing individual audio features is also investigated in this paper. Pitch, intensity, loudness and MFCC are experimented on as audio features (Section 4.2).The results discussed in the experiment section confirm the finding of [38].

Researchers have also explored various multimodal approaches for improved affect recognition. Zeng *et al.* [41] presented a thorough survey on existing approaches and outlined some of the challenges. In one of the works by Busso *el al.* [5], they describe a multimodal framework and show that the fusion of facial expression with speech information performs better than unimodal systems for emotion recognition. A comparison of various fusion methods for multimodal emotion analysis is presented in [22]. They also show that multimodal information provides more discriminative information for various classification problems, which serves as an inspiration for our study here. This paper explores the fusion of audio and video features for depression analysis.

The contributions of this paper are as follows:

- We propose a multimodal fusion framework for affective sensing, which is evaluated on the real-world example of developing assistive technologies for depression diagnosis and monitoring.
- We show the increase in performance for depression detection when multiple signals are used as compared to unimodal signals only.
- We compute STIP-based visual descriptors on upper body videos and compare their performance with intra-face based visual descriptors (i.e. without the upper body).

- In order to handle the large amount of interest points generated from the upper body videos in the depression dataset, a key interest point selection method is proposed for learning a Bag-of-Words model.
- LBP-TOP is computed on subsequences in a piece-wise manner so as to compute spatio-temporal words for learning the visual BoW model.
- An audio BoW model is learned form the audio features (pitch, intensity, loudness and MFCCs).
- This paper explores various approaches (feature, score and decision fusion) for the fusion of audio and video features for depression analysis and compares the performance with that of audio and video features alone.
- Finally, this study compares the performance of these methods with that of a Support Vector Machine (SVM) added as second-stage classifier on the output of the individual classifiers.

3 Data Collection

The clinical database used in this study was collected at the Black Dog Institute, a clinical research institute focussing on mood disorders, including depression and bipolar disorder.[1] 60 subjects (30 males and 30 females) with an age range of $19 - 72$yr participated. Subjects included 30 healthy controls (mean age 33.9 ± 13.6yr) as well as 30 patients (mean age 44.3 ± 12.4yr) who had been diagnosed with severe depression (but no other mental disorders or co-morbid conditions).

Participants in the Black Dog research program first complete a computerised mood assessment program (MAP), which generates diagnostic decisions and a profile of personality, co-morbid conditions such as anxiety disorders, current functioning assessments, as well as current and past treatments, and a section on the aetiology of their depressive episode (e.g. family history; stressful life events). Following the MAP, the participants undergo a structured interview (MINI) that assesses current and past depression as well as hypo(manic) episodes and psychosis (both current and past) as per the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). If they are currently depressed and are deemed eligible for the ongoing study (unipolar depression and no history of psychosis), they will also be rated on the CORE measure of psycho-motor disturbance [32]. In the present study, only severely depressed patients (HAMD > 15) were included. The recordings were made after their initial diagnosis and before the start of any treatment. Control subjects were carefully selected to have no history of mental illness and to broadly match the depressed subjects in age and gender.

The experimental paradigm contains several parts similar to [25]: (a) watching movie clips, (b) watching and rating International Affective Picture System (IAPS) pictures, (c) reading sentences containing affective content, and (d) an interview between the participants and a research assistant. In this study, we are interested in analysing the changes in facial expressions, head and shoulder movements, and variations in speech pattern in response to the interview questions. There are a total of eight groups of questions asked in the interview in order to induce emotions in the participants. Questions are designed to arouse both positive and negative emotions, for instance ideographic questions such as, “Can you recall some recent good news you had and how did that make you feel?” and “Can you recall news of bad or negative nature and how did you feel about it?”. The length of the

[1] http://www.blackdoginstitute.org.au/

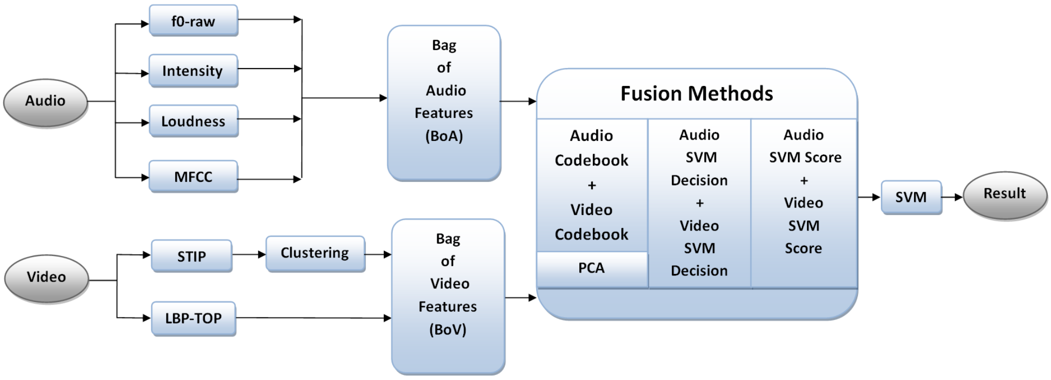

Audio
f0-raw
Intensity
Loudness
MFCC
Bag of Audio Features (BoA)
Fusion Methods
Audio Codebook + Video Codebook
PCA
Audio SVM Decision + Video SVM Decision
Audio SVM Score + Video SVM Score
SVM
Result
Video
STIP
Clustering
LBP-TOP
Bag of Video Features (BoV)

Fig. 1 Flow of the proposed system: Audio and video data are processed individually and respective features are computed. All audio features are combined in a Bag of Audio Features (BoA), while video features are combined in a Bag of Visual Features (BoV). Different fusion methods are then experimented on.

video recordings of the interviews lies in the range of $208 - 1672s$. In an ideal situation, one would wish to have a larger dataset. However, this project is part of an ongoing study and more data is being collected. Similar limitations with the sample size have been reported by Ozdas *et al.* [30] and Moore *et al.* [27].

4 Method

Given an input audio-video clip $\mathcal{AV}$ containing N video frames $\{\mathcal{V}_1, \mathcal{V}_2 ... \mathcal{V}_N\}$ and M audio frames $\{\mathcal{A}_1, \mathcal{A}_2 ... \mathcal{A}_M\}$, STIPs are computed on the video frames. Due to the relatively large number of video frames, the number of interest points generated is very high. Therefore, key interest point selection is applied in a two-level clustering phase for computing the bag of video features. LBP-TOP features are computed piece-wise temporally to capture intra-face movements. A visual dictionary is learnt from these spatio-temporal LBP-TOP based words. For the audio frames, multiple features (f0-raw, intensity, loudness, MFCC) are computed. Further clustering is applied on the combined audio features to create bag of audio features. Three different types of fusion approaches are then experimented on (see Figure 1).

4.1 Video Features

Two descriptors are computed for capturing the visual spatial and temporal information. The framework starts by computing the Viola-Jones object detector [40] for detecting a face blob $\mathcal{F}$, which is used as a seed for facial feature extraction. Chew *et al.* argue that subject-specific AAM perform better than subject independent constrained local models (CLM) [36], however the use of an efficient feature descriptor can compensate for the error induced by subject-independent methods. Taking a similar approach, a pictorial structure based approach [15] is used to extract 9 facial points, which describe the location of the left and right corners of both eyes, the nose tip, the left and right corners of the nostrils, and

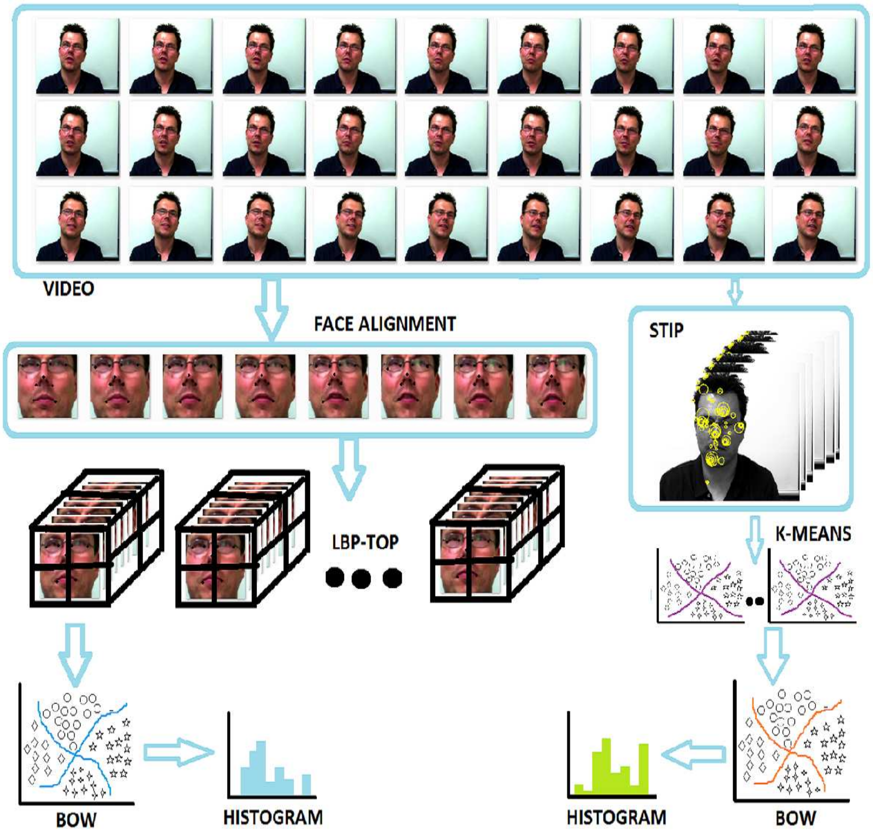


Fig. 2 Video processing pipeline: STIPs are computed over the unaligned raw video frames. Key interest points are detected by video level clustering. A BoW dictionary is learnt from the key interest points of all the videos. Note that the STIPs capture head and shoulder movements along with the facial dynamics, as they are applied to the entire video frame. Faces are then detected and aligned. For capturing the intra-face motion, the LBP-TOP descriptor is computed in a piece-wise manner over sub clips and a BoW is learnt.

the left and right corners of the mouth. This approach is based on part-based models and has been applied successfully to facial feature localisation [15]. Part-specific detectors are applied to the facial blob and the facial parts are localised using dynamic programming on the response of the part specific detectors. The power of pictorial structures stem from its representation of an object (face in this case) as an undirected graph, which has recently been shown performing better than AAM and CLM [43] in both subject dependent and independent settings. For aligning the faces, an affine transform based on these points is computed. Figure 2 describes the visual processing pipeline.

4.1.1 Space-Time Interest Points

In recent years, the STIP concept [21] has found much attention in computer vision and video analysis research. It successfully detects useful and meaningful interest points in videos by extending the idea of the Harris spatial interest point detector to local structures in the spatio-temporal domain. Salient points are detected where image values have sufficient local variation in both the space and time dimensions. Two histograms, the Histogram of Gradients (HOG) and the Histogram of Flow (HOF), are calculated around an interest point in a fixed sized spatial and temporal window. These volumes around the interest point are used to learn visual dictionaries and have shown robust performance for computer vision and video analysis problems such as human action recognition [21]. The video frames in our dataset typically contain the upper body of the subjects as well as the head. Therefore, it is worthwhile to investigate the movement patterns of all upper body parts. The STIPs reflect the spatio-temporal changes, which account for movements inside the facial area and elsewhere (e.g. hands, shoulders and head movements).

Key-Interest Point Selection: To reduce the complexity due to the large number of frames, a keyframe selection method was used for emotion analysis by [10]. The authors apply clustering over aligned facial landmark points computed using the Constrained Local Model approach [36]. The cluster centres' nearest neigbour frames are chosen as the keyframes. The video clips in the depression dataset are relatively long and there is a large amount of motion due to the presence of the upper body in the frame, so that a key-interest point selection scheme is advisable.

A video $\mathcal{V}$ gives K interest points. A total of 4.8×10^7 interest points are computed from the 60 video clips. This is both computationally and memory wise non-trivial, as a leave-one-subject-out protocol is followed in the experiments. To reduce the feature set size, inspired by [10], the K-Means algorithm is employed to each $\mathcal{V}$. K interest points give K_c cluster centres. These K key-interest points are then the representative interest points of a video sample. The value K_c is chosen empirically.

4.1.2 Local Binary Patterns Three Orthogonal Planes

Recently, Local Binary Patterns (LBP) have become popular in computer vision. Their power stems from their simple formulation and dense texture information. For computing the intra-face muscle movements in subjects, we computed a temporal variant of LBP, LBP-TOP [42]. It considers patterns in three orthogonal planes: XY, XT and YT, and concatenates the pattern co-occurrences in these three directions. The local binary pattern part of the LBP-TOP descriptor assigns binary labels to pixels by thresholding the neighborhood pixels with the central value. Therefore, for a centre pixel $\mathcal{O}_p$ of an orthogonal plane $\mathcal{O}$ and its neighbouring pixels N_i, a decimal value d is assigned

$$d = \sum_{\mathcal{O}}^{XY,XT,YT} \sum_{p} \sum_{i=1}^{k} 2^{i-1} I(\mathcal{O}_p, N_i) \quad . \tag{1}$$

In dynamic facial expression analysis, the apex frame shows the peak intensity of an expression. The XY plane in LBP-TOP ideally should be the apex frame of the video. However, given the complex nature of the videos in the depression dataset, it is non-trivial to label the apex frames. To overcome this limitation, rather than computing LBP-TOP on the video in a temporally holistic manner, the descriptor is computed temporally 'piece-wise'. These piece-wise LBP-TOP units form spatio-temporal words for the BoW dictionary. Formally, for a video $\mathcal{V}$ of length l, uniformly timed sub-clips are segmented of length t. The LBP-TOP descriptor is computed on these sub-clips individually. Therefore, there are l/t sub-clips and their corresponding LBP-TOP based spatio-temporal descriptors $d^{l/t}$.

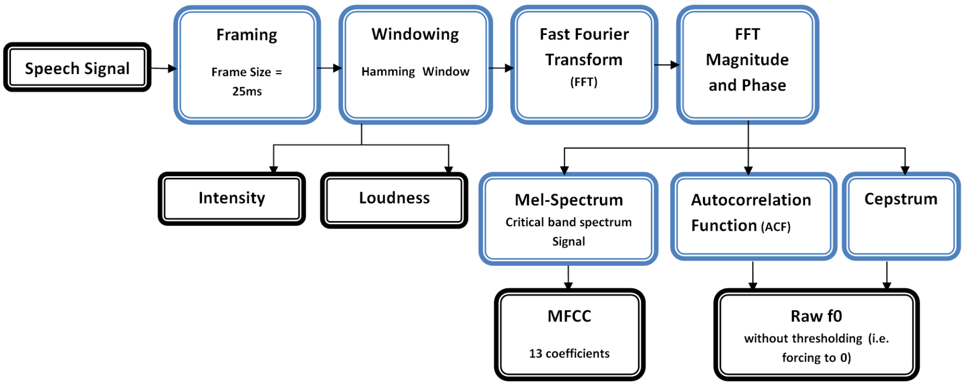


Fig. 3 Flow of the speech processing subsystem to extract audio features: Intensity, Loudness, f0 and MFCC.

4.2 Audio Features

Investigations of depressed speech by psychologists have found several distinguishable prosodic features (see Section 2). Four different audio features – Fundamental frequency f0, loudness, intensity and Mel-frequency cepstral coefficients – are computed in this study. Figure 3 presents the audio processing subsystem.

Each subject's speech data is first segmented into frames. The frame size is set to 25ms, with 10ms overlap between two adjacent frames. As a result, there will be high frequency noise at the beginning and the end of each frame. To reduce this boundary effect, a Hamming window is applied to each frame

$$w_{Ham}[n] = 0.54 + 0.46cos\frac{(2\pi n)}{N-1} \tag{2}$$

where N is the number of samples per frame, and $n = 1 \cdots N$. After applying a Fast Fourier Transform (FFT) on each frame, the magnitudes and phases are computed. Intensity is calculated as the mean of the squared frame multiplied by a Hamming window, while loudness is computed from intensity as

$$L = (\frac{I}{I_0})^{0.3} \tag{3}$$

where I is the intensity and $I_0 = 0.000001$. To extract f0, the auto-correlation function (ACF) and the cepstrum are computed. The ACF is calculated by squaring the magnitude spectrum and applying an inverse FFT. The cepstrum is computed by applying a log function to the magnitude spectra. The difference between f0 and f0-raw is that with f0-raw, there is no thresholding, i.e. there is no forcing to 0 in unvoiced frames. To generate the MFCC, the Mel-spectrum is computed by applying overlapping triangular filters equidistantly on the Mel-frequency scale

$$Mel(f) = 1127ln(1 + \frac{f}{700}) \tag{4}$$

to the FFT magnitude spectrum.

4.3 Bag of Words

The Bag of Words approach, originally developed in the natural language processing domain, has been successfully applied to image analysis [21] and depression analysis [17]. It represents documents based on the unordered word frequency. The power of the BoW framework stems from its tolerance to variation in the appearance of objects. Recently, [39] compared different BoW approaches for facial expression recognition as compared to object recognition. The authors achieved state-of-the-art performance for facial expression analysis.

In the problem described in this paper, a video clip (set of video frames) and an audio clip (set of audio frames) are documents in the BoW sense. The BoW computed from the videos are termed Bag of Visual (BoV) features and the BoW computed on the audio features are called Bag of Audio (BoA) features. BoA are computed for f0-raw, intensity, loudness and MFCC individually and also on selected combinations. The performance of these are computed and the best performing is used further for fusion. BoV are computed separately for LBP-TOP and STIP. For STIP, BoV are computed on the cluster centres of interest points of each video. This two-level clustering helps in dealing with the high number of interest points generated by the STIP. The size of the codebooks is decided empirically. The use of BoW gives two advantages in the framework. The interviews are of different duration, depending on how much the subject was saying. The use of codebooks makes it simpler to deal with such samples of different length. Secondly, BoW are computed for audio and video independently, which overcomes the problem of different sampling rates in the two modalities. This simplifies feature fusion.

5 Fusion

As discussed in the introduction (Section 1), depression analysis has been primarily limited to single channel/modal information. Multimodal analysis is a general extension. Three standard fusion techniques are investigated.

5.1 Feature Fusion

This is the simplest form of fusion. Raw features computed from the different modalities are concatenated to form a single feature vector. Despite the simplicity, feature fusion results in a performance increase compared to the performance of single modalities (see Section 6 for details). However, the downside of feature fusion is that it suffers from the curse of dimensionality. As more modalities are joined, this increases the dimension of the feature vectors. To overcome this issue, Principal Component Analysis (PCA) is applied to the combined features and then the classification is performed.

5.2 Score Fusion

In score level fusion, different scores such as probability estimates, likelihoods, etc. are combined, before making a classification decision. There are several popular methods for score fusion. In this paper, two techniques – score fusion by weighted sum and by learning a new SVM classifier on the scores – are investigated. The distance from the SVM hyperplane is calculated and used as a score.

5.3 Decision Fusion

In decision fusion, multiple classifiers are trained on different feature sets. The output of these classifiers is used to infer the final class result. Various techniques are used for decision fusion: weighted voting, algebraic combination rules and operators [22]. In this paper, the AND and OR operators are used to fuse the decisions from the separate audio and video SVM classifiers. Furthermore, we also experiment with decision fusion by learning a new, second-stage SVM classifier.

6 Experiments and Results

The original spatial resolution of the video frames was 800 × 600 pixels. The videos were downsampled to 320×240 pixels for computational efficiency. For STIP, the Harris 3D interest point detector was used. The spatial window size for computing HOG was set to 3 and the temporal window size for HOF to 9. Two values, $K = 2500$ and $K = 5000$, were experimented on for the number of clusters. LBP-TOP was computed for two different sub-clip sizes, $t = 6s$ and $t = 1s$. Moreover, the codebook C_s for BoV was computed on clusters from each video clip. The codebook C_l for BoV was computed on different LBP-TOP configurations. Various codebook sizes in the range of $200 - 750$ were experimented with C_s and C_l of BoV. From here on, $STIP1$ means STIP with level-one cluster size $K = 2500$ and $STIP2$ refers to STIP with level-one cluster size $K = 5000$. For LBP-TOP, $LBP1$ is the configuration with clip length $t = 6s$ and $LBP2$ with clip length $t = 1s$.

Furthermore, experiments combining codebooks C_s and C_l for all different codebook sizes, $200 - 750$ were also performed. The four possible descriptor combinations analysed were $STIP1 + LBP1$, $STIP1 + LBP2$, $STIP2 + LBP1$ and $STIP2 + LBP2$. Some of the combinations such as $STIP1 + LBP1$, where the C_s and C_l size was 200, result in a good increase to the individual feature performance, resulting in an accuracy of 81.7%, whereas the maximum accuracy given by individual video features was 76.7% from $STIP1$ and $STIP2$ as shown in Table 1(a).

For computing the audio descriptors, the publicly available open-source software "openSMILE" [16] was used to extract low-level voice features from the subject speech labelled intervals. The spontaneous speech from the dataset interview was manually labelled to extract pure subject speech, i.e. to remove voice inactive regions. The frame size was set to $25ms$ at a shift of $10ms$ and using a Hamming window. The number of MFCC coefficients used for the experiments was 13, where the deltas were not included. f0 was calculated using ACF, where f0-raw was calculated without threshold (i.e. without forcing to 0) in unvoiced segments.

BoA were learned for all the individual audio features and various codebook sizes were experimented. As reported in Table 1(a), the best detection accuracy obtained from the individual audio features was 75%. To further increase the performance using audio features only, other configurations for BoA were experimented, first by combining all the four audio features together: f0+I+L+M and then in the second case leaving out MFCC and combining the other three audio features, f0+I+L. For both of these codebooks, again different cluster sizes were tried and chosen empirically. The combined BoA, specifically f0+I+L, performed reasonably better than the individual ones, giving an accuracy of 83.3%.

Table 1(a) presents the classification performance of bag of features computed on individual features. Out of all, $STIP1$ and $STIP2$ performed the best, giving an accuracy of 76.8%. Here, the values of level-one cluster centres were $K = 2500$ and $K = 5000$

Table 1 Performance of the system at various stages.

(a) Comparison of classification accuracies for individual video and audio features. Here, STIP1 - Level One clusters $C = 2500$, STIP2 - Level One clusters $C = 5000$, LBP1 - LBP-TOP with clip length $t = 6s$, LBP2 - LBP-TOP with clip length $t = 1s$.

Individual Feature	**f0-raw**	**Loud.**	**Inten.**	**MFCC**	**STIP1**	**STIP2**	**LBP1**	**LBP2**
Accuracy	70.0%	73.3%	75.0%	63.3%	76.7%	76.7%	70.0%	66.7%

(b) Computed audio and video features were combined separately. Best audio and video only combinations are presented here. In the notation Feature_N, N refers to the codebook size.

Audio Feature Combined	**Audio Only Accuracy**	**Video Feature Combined**	**Video Only Accuracy**
A1; f0+I+L_200	83.3%	**V1**; STIP1_200+LBP1_200	81.7%
A2; f0+I+L_500	83.3%	**V2**; STIP1_200+LBP2_750	78.8%
A3; f0+I+L_750	83.3%	**V3**; STIP1_750+LBP2_500	78.8%
A4; f0+I+L+M_500	78.3%	**V4**; STIP1_750+LBP2_750	80.0%

(c) Audio-Video Fusion Results: Top five classification accuracy for different fusion methods for various parameters of the features. Here, W.Sum - Weighted Sum, W.Prod. - Weighted Product, Concat. - Concatenated

A-V Combination	**Feature Fusion**		**Score Fusion**			**Decision Fusion**		
	Concat.	**PCA**	**W.Sum**	**W.Prod.**	**SVM**	**AND**	**OR**	**SVM**
A1+V2	81.7%	**91.7%**	85.0%	85.0%	86.7%	68.3%	93.3%	86.7%
A2+V3	81.7%	80.0%	86.7%	86.7%	**88.3%**	66.7%	95.0%	**91.7%**
A3+V1	85.0%	86.7%	85.0%	85.0%	86.7%	71.7%	93.3%	91.7%
A1+V4	81.7%	80.0%	85.0%	85.0%	85.0%	70.0%	93.3%	88.3%
A2+V2	81.7%	86.7%	85.0%	85.0%	85.0%	66.7%	95.0%	88.3%

for $STIP1$ and $STIP2$, respectively. The size of codebook $C_{s1} = 500$ for $STIP1$ and $C_{s2} = 200$ for $STIP2$. The increase in the size of K did not increase the performance as expected. It can be argued that the discriminating ability of the descriptor is well covered with $K = 2500$; anything more does not add to the discriminability.

For classification, a non-linear SVM [6] was used. The parameters were searched using an extensive grid search. A leave-one-subject-out experiment methodology was used for all of the classifications. From here on, individual classifiers means the SVM model trained for BoA (f0/I/L/MFCC/f0+I+L/f0+I+L+M) or BoV (STIP1/STIP2/LBP1/LBP2/STIP+LBP) individually.

Figure 4 shows the effect of changing codebook size of BoA with different fusion methods. For a clear comparison, the fusion of different configurations of BoA is shown for one selected BoV combination, i.e. V1. The choice of this visual configuration is based on the highest performance of this STIP and LBP-TOP combination. Figure 4 (a) clearly shows the performance increase due to PCA-based dimensionality reduction in feature fusion. Figure 4 (b) shows the difference in performance due to score fusion. In Figure 4 (c), SVM and OR based decision fusion clearly perform better than AND based decision fusion. Figure 5 shows the fusion of an audio codebook A3 with different combinations of various sizes of C_s and C_l. The observations are consistent with the graphs in Figure 4.

Table 1(c) describes the top five results for fusion methods on various descriptor parameters and Table 2 describes the confusion matrix for the best configuration of different fusion methods. For feature fusion, different combinations of BoA and BoV were created. As discussed earlier, the high dimensionality of the feature vector is a drawback of the feature fusion technique. Therefore, PCA was applied to the combined features and 98% of the

Table 2 Confusion Matrix for the best results for different fusion methods shown in Table 1(c)

(a) Feature Fusion (PCA)

	Patients (Predicted)	Controls (Predicted)
Patients (Actual)	25	5
Controls (Actual)	30	0

(b) Score Fusion (SVM)

	Patients (Predicted)	Controls (Predicted)
Patients (Actual)	25	5
Controls (Actual)	28	2

(c) Decision Fusion (SVM)

	Patients (Predicted)	Controls (Predicted)
Patients (Actual)	26	4
Controls (Actual)	29	1

variance was kept. A further SVM was trained on the new reduced dimensionality features. As expected, applying PCA post feature fusion increased the performance of the system. Moreover, the performance of the classifier trained on feature fused samples was higher than the performance of classifiers trained on individual feature based BoA or BoV. As shown in Table 1(b), the best accuracies for individual feature based BoA and BoV are 83.3% and 81.7%, respectively, whereas combining audio and video features via feature fusion boosts the accuracy to 91.7% (see Table 1(c)). To statistically validate the difference between the fused and individual features, a t-test was performed. Various individual features are compared with one combination, i.e. V3+A2, for $\alpha = 0.01$. The average p-value for the cohort of STIP1_500 was 0.0006, LBP_500 was 0.00007 and f0+I+L_500 was 0.00001.

For score fusion, the distances from the SVM hyperplane were computed for all the individual BoA and BoV. To fuse the scores, the weighted sum and weighted product was computed. Acknowledging that better weights optimisation may increase the recognition rate, our method is simply a linear search for the best weights, which gave a maximum accuracy rate of 86.7% in both cases. Also, a SVM classifier is trained on the scores of individual BoA and BoV, which gave a higher classification accuracy of 88.3%.

For decision fusion, the classification outputs from classifiers trained individually on BoA and BoV were combined via the AND and OR operators. In the AND operation, the final positive is based on the evidence of presence of positives from the classification accuracies of all the individual classifiers. The maximum classification achieved by using this fusion technique was 71.7%. This means that both the individual classifiers have a consensus on at least 71.7% of the samples. For the OR operator, which shows a correct recognition if at least one modality classifies a subject correctly, the maximum accuracy wss 95.0%. However, a word of caution is in order here. The OR fusion inherently runs the risk of creating a larger number of false positives than the other fusion methods, as no consensus of the individual classifiers is required and all classifiers are treated as having equal weight, with the acceptance threshold being such that a positive recognition in one classifier leads to a positive recognition overall, which is a comparatively low acceptance threshold. In other words, OR fusion assumes equal confidence in both classifiers, which may not be a true reflection of the real world. Feedback from psychologists indicates that they would not rely just on an OR fusion approach in the real world. Furthermore, an SVM classifier on top of the decision of the individual classifiers was learned. The maximum accuracy achieved is 91.7%, which shows that training classifiers via decision fusion gives robust performance for depression classification.

The best performance from all three fusion methods was 91.7%. There is an absolute increase of 8.4% over audio-only and 10% - 12.9% over video-only classification. The increase in system performance using different fusion methods is consistent with the results discussed by [22] for fusion-based multimodal emotion analysis.

7 Conclusions and Future Work

Depression is a severe mental health disorder with a high individual and societal cost. The study described in this paper proposes a multimodal framework for automatic depression analysis. The STIP detector is computed on the image frames and HOG and HOF histograms are calculated around the interest points in a spatio-temporal window. Further, in order to decrease the number of interest points, clustering is applied at the video level. The cluster centres from each video are used to train a BoV feature codebook. LBP-TOP is computed on sub-sequences in a piece-wise manner on aligned faces to analyse facial dynamics. Moreover, a separate BoV codebook is learned from it. For audio analysis, f0-raw, intensity, loudness and MFCC are computed and BoA features are derived from the fusion of these audio features. Audio-video fusion at three levels has been investigated: feature level, score level and decision level. The experimental results show that fusing audio and video channels is effective for training a depression classifier that can assist clinicians. As part of future work, extracting subject speech will be made fully automatic using the bag of words framework.

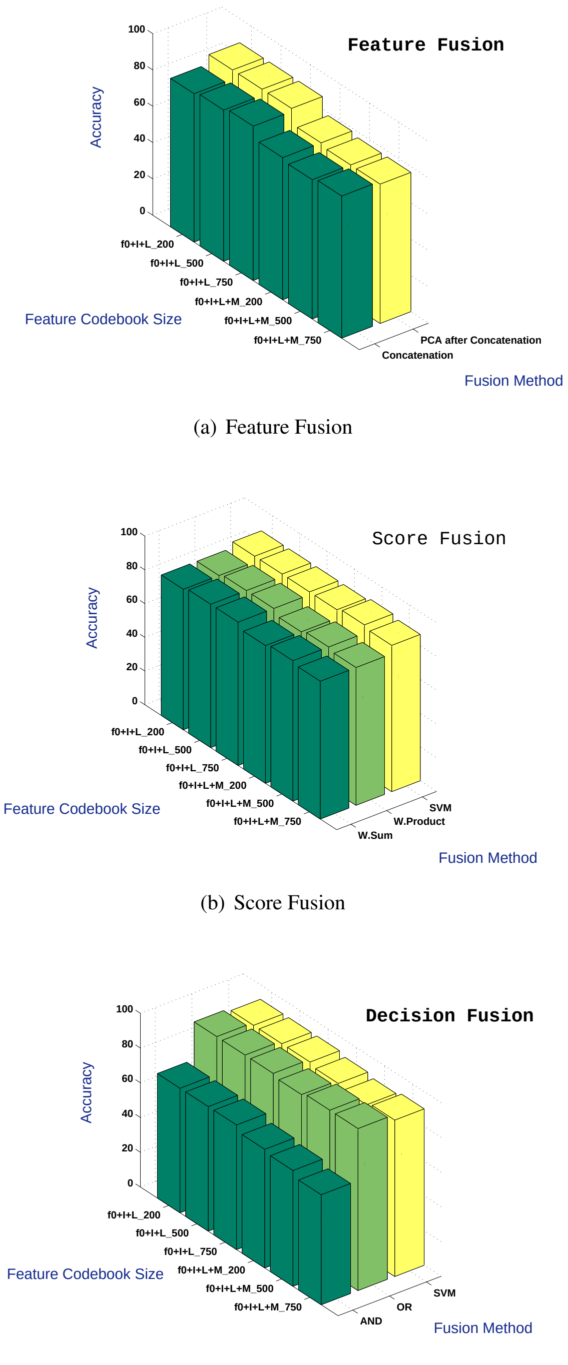
The study presented here forms part of the first phase of our ongoing research to develop a robust and objective diagnostic aid to support clinicians in their diagnosis of clinical depression, as current diagnosis suffers from a range of subjective biases. In this ongoing work, we investigate different modalities, features and classification approaches and experimentally validate them with our clinical partners. In the second phase, we will further clinically test the best performing diagnosis approaches in a prototypical assistive laptop system equipped with a video camera and microphone. In the third phase, we will explore how the affective sensing approaches can be implemented on smartphones and other mobile technology platforms, such as tablet computers, to assist doctors and patients in the monitoring of treatment progress, which requires robustness to a large variety of environmental conditions, such as different levels of illumination, occlusion and acoustic noise. We firmly believe that only a multimodal framework can truly deliver the robustness required for real-world applications.

References

1. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. In: Proc. FLAIRS-25 (2012). Accepted

2. Ambadar, Z., Schooler, J., Cohn, J.: Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. Psychological Science pp. 403–410 (2005)
3. Asthana, A., Saragih, J., Wagner, M., Goecke, R.: Evaluating AAM Fitting Methods for Facial Expression Recognition. In: Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction, ACII'09, pp. 598–605 (2009)
4. Batliner, A., Huber, R.: Speaker Classification I. chap. Speaker Characteristics and Emotion Classification, pp. 138–151. Springer-Verlag, Berlin, Heidelberg (2007)
5. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Sixth International Conference on Multimodal Interfaces ICMI 2004, pp. 205–211. ACM Press (2004)
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
7. Cohn, J.F., Kreuz, T.S., Matthews, I., Yang, Y., Nguyen, M.H., Padilla, M.T., Zhou, F., De la Torre, F.: Detecting Depression from Facial Actions and Vocal Prosody. In: Proc. Affective Computing and Intelligent Interaction, ACII'09, pp. 1–7 (2009)
8. Cummins, N., Epps, J., Breakspear, M., Goecke, R.: An Investigation of Depressed Speech Detection: Features and Normalization. In: Proc. Interspeech (2011)
9. Dhall, A., Asthana, A., Goecke, R.: Facial expression based automatic album creation. In: International Conference on Neural Information Processing (2), pp. 485–492 (2010)
10. Dhall, A., Asthana, A., Goecke, R., Gedeon, T.: Emotion recognition using PHOG and LPQ features. In: IEEE Automatic Face and Gesture Recognition 2011 (FG11) workshop Facial Expression Recognition and Analysis, FERA'11, pp. 878–883 (2011)
11. Dhall, A., Joshi, J., Radwan, I., Goecke, R.: Finding happiest moments in a social context. In: Asian Conference on Computer Vision, ACCV'12 (2012)
12. Edwards, G., Taylor, C., Cootes, T.: Interpreting Face Images Using Active Appearance Models. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition FG'98, pp. 300–305. IEEE, Nara, Japan (1998)
13. Ellgring, H.: Nonverbal communication in depression. Cambridge University Press (2008)
14. Ellgring, H., Scherer, K.R.: Vocal indicators of mood change in depression. Journal of Nonverbal Behavior **20**(2), 83–110 (1996)
15. Everingham, M., Sivic, J., Zisserman, A.: Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In: Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK, September 4-7, 2006, BMVC'06, pp. 899–908 (2006)
16. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proc. ACM Multimedia (MM'10) (2010)
17. Joshi, J., Dhall, A., Goecke, R., Breakspear, M., Parker, G.: Neural-net classification for spatio-temporal descriptor based depression analysis. In: Proceedings of the International Conference on Pattern Recognition, ICPR'12, pp. 2634–2638 (2012)
18. Joshi, J., Goecke, R., Breakspear, M., Parker, G.: Can body expressions contribute to automatic depression analysis? In: Proceedings of the International Conference on Automatic Face and Gesture Recognition, FG'13 (2013)
19. Kessler, R., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K., Rush, A., Walters, E., Wang, P.: The Epidemiology of Major Depressive Disorder: Results From the National Comorbidity Survey Replication (NCS-R). The Journal of the American Medical Association **289**(23), 3095–3105 (2003)
20. Kuny, S., Stassen, H.H.: Speaking behavior and voice sound characteristics in depressive patients during recovery. Journal of Psychiatric Research **27**(3), 289–307 (1993)
21. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'08, pp. 1–8 (2008)
22. Lingenfelser, F., Wagner, J., André, E.: A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In: Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, pp. 19–26. ACM (2011)
23. Lucey, S., Matthews, I., Hu, C., Ambadar, Z., de la Torre, F., Cohn, J.: AAM Derived Face Representations for Robust Facial Action Recognition. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, FG'2006, pp. 155–162 (2006)
24. Mathers, C., Boerma, T., Fat, D.M.: The global burden of disease: 2004 update. Tech. rep., WHO Press, Switzerland (2004)
25. McIntyre, G., Goecke, R., Hyett, M., Green, M., Breakspear, M.: An Approach for Automatically Measuring Facial Activity in Depressed Subjects. In: Proc. Affective Computing and Intelligent Interaction, ACII'09, pp. 223–230 (2009)

26. Moore, E., Clements, M., Peifer, J., Weisser, L.: Comparing objective feature statistics of speech for classifying clinical depression. Proc. 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2004)
27. Moore, E., Clements, M., Peifer, J., Weisser, L.: Critical analysis of the impact of glottal features in the classification of clinical depression in speech. In: IEEE Transactions on Biomedical Engineering, vol. 55, pp. 96–107 (2008)
28. Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S.: Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of Neurolinguistics **20**(1), 50–64 (2007)
29. Nilsonne, A.: Speech characteristics as indicators of depressive illness. Acta Psychiatrica Scandinavica **77**(3), 253–263 (1988)
30. Ozdas, A., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D.: Analysis of fundamental frequency for near term suicidal risk assessment. SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. pp. 1853–1858 (2000)
31. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12), 1424–1445 (2000)
32. Parker, G., Hadzi-Pavlovic, D.: Melancholia: A disorder of movement and mood. Cambridge University Press (1996)
33. Picard, R.: Affective Computing. MIT Press, Cambridge (MA), USA (1997)
34. Prendergast, M.: Understanding Depression. Penguin, Australia (2006)
35. Saragih, J., Goecke, R.: Learning AAM fitting through simulation. Pattern Recognition (2009)
36. Saragih, J.M., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: Proceedings of the IEEE International Conference of Computer Vision, ICCV'09, pp. 1034–1041 (2009)
37. Scherer, K.R.: Vocal assessment of affective disorders. In: J.D. Maser (ed.) Depression and Expressive Behavior, pp. 57–82. Lawrence Erlbaum Associates (1987)
38. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L.: The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In: Proc. Interspeech (2007)
39. Sikka, K., Wu, T., Susskind, J., Bartlett, M.S.: Exploring bag of words architectures in the facial expression domain. In: European Conference on Computer Vision (ECCV) Workshops (2), pp. 250–259. Springer (2012)
40. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'01, pp. 511–518 (2001)
41. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 39–58 (2009)
42. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transaction on Pattern Analysis and Machine Intelligence (2007)
43. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (2012)

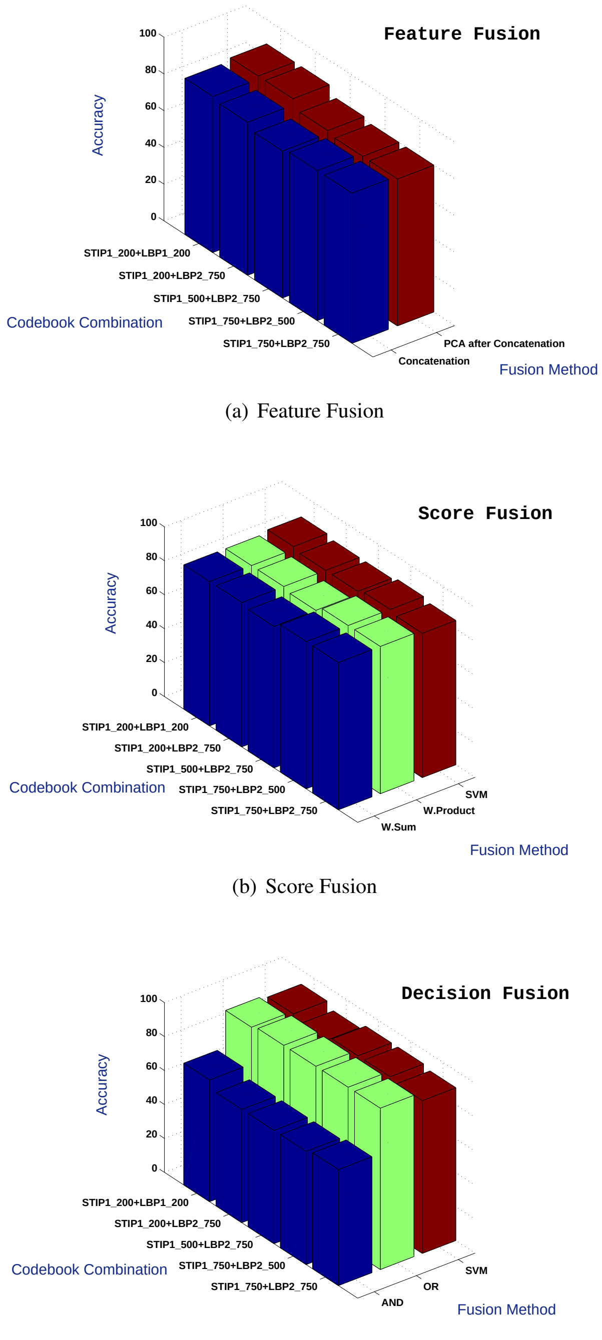


(a) Feature Fusion

(b) Score Fusion

(c) Decision Fusion

Fig. 4 The three graphs show the accuracy of the system and the effect of choosing different codebook sizes of BoA while fusing it with a selected BoV codebook combination STIP1_200+LBP1_200 for different fusion methods: a) Feature Fusion, b) Score Fusion, c) Decision Fusion.



(a) Feature Fusion

(b) Score Fusion

(c) Decision Fusion

Fig. 5 The three graphs show the accuracy of the system and the effect of choosing different combinations of C_s and C_l, while fusing with a selected audio feature f0+I+L_750 for different fusion methods: a) Feature Fusion, b) Score Fusion, c) Decision Fusion.