Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text

Anastasia Pampouchidou[*] Olympia Simantiraki[†] Amir Fazlollahi[‡]
Matthew Pediaditis[†] Dimitrios Manousos[†] Alexandros Roniotis[§]
Georgios Giannakakis[†] Fabrice Meriaudeau[*¶] Panagiotis Simos[‖]
Kostas Marias[†] Fan Yang[*] Manolis Tsiknakis[†§]

ABSTRACT

Depression is a major cause of disability world-wide. The present paper reports on the results of our participation to the depression sub-challenge of the sixth Audio/Visual Emotion Challenge (AVEC 2016), which was designed to compare feature modalities (audio, visual, interview transcript-based) in gender-based and gender-independent modes using a variety of classification algorithms. In our approach, both high and low level features were assessed in each modality. Audio features were extracted from the low-level descriptors provided by the challenge organizers. Several visual features were extracted and assessed including dynamic characteristics of facial elements (using Landmark Motion History Histograms and Landmark Motion Magnitude), global head motion, and eye blinks. These features were combined with statistically derived features from pre-extracted features (emotions, action units, gaze, and pose). Both speech rate and word-level semantic content were also evaluated. Classification results are reported using four different classification schemes: i) gender-based models for each individual modality, ii) the feature fusion model, ii) the decision fusion model, and iv) the posterior probability classification model. Proposed approaches outperforming the reference classification accuracy include the one utilizing statistical descriptors of low-level audio features. This approach achieved f1-scores of 0.59 for identifying depressed and 0.87 for identifying not-depressed individuals on the development set and 0.52/0.81, respectively for the test set.

Keywords

AVEC 2016; image processing; speech processing; pattern recognition; affective computing; depression assessment; multimodal fusion

1. INTRODUCTION

Depression is a common mood disorder, which burdens many people around the globe at an alarmingly increasing rate. Objective measures of depressive symptomatology could be advantageous for clinicians, in the context of a decision support system. Thus, automatic depression assessment has been drawing increasing attention in the past few years to a certain extend due to the AVEC. However, AVEC 2013 [30] and AVEC 2014 [29] differed from the present in a number of aspects. The provided dataset is different, in terms of the data collection protocol: in the previous, volunteers were performing tasks in response to presented stimuli, while in the present the volunteers interact with a virtual human in the form of an interview. Previously both video recordings, as well as pre-extracted features were provided, while this year only the features are available, increasing the difficulty of the challenge. The aim of the depression sub-challenge used to be continuous, while now it is categorical. In the present work, the participation of our team *FORTH-TEIC*, for the depression sub-challenge of the AVEC 2016 [27], is being described. Section 2 covers related work, while Section 3 describes the feature extraction methods designed and implemented by our team. Section 4 presents the experimental results of the participation, while Section 5 concludes with the discussion and conclusions.

2. RELATED WORK

Motion History Images (MHI) is an algorithm widely used in the field of human action recognition [3]. Valstar et al. proposed the use of MHI for the recognition of facial action from videos [28]. Ptucha and Savakis [23], among else, proposed an Active Shape Model (ASM) based history image, which encoded the motion of each landmark by a black

[*]Laboratoire Electronique, Informatique et Image, Université de Bourgogne, Le Creusot, France. Emails: anastasia.pampouchidou@gmail.com, fanyang@u-bourgogne.fr

[†]Computational BioMedicine Laboratory, Institute of Computer Science, Foundation for Research & Technology - Hellas (FORTH), Heraklion, Crete, Greece. Emails:{osimantir, mped, mandim, ggian, kmarias, tsiknaki}@ics.forth.gr

[‡]The Australian E-Health Research Centre, CSIRO, Brisbane, Australia. Email:fazlollahi@gmail.com

[§]Biomedical Informatics & eHealth Laboratory, Technological Educational Institute of Crete (TEIC), Heraklion, Crete, Greece. Emails: {roniotis, tsiknaki}@ie.teicrete.gr

[¶]CISIR, Electrical Engineering Department, Universiti Teknologi Petronas, Malaysia. Email:fmeriau@u-bourgogne.fr

[‖]Division of Psychiatry, School of Medicine, University of Crete, Heraklion, Crete, Greece. Email:akis.simos@gmail.com

AVEC'16, October 16 2016, Amsterdam, Netherlands

ISBN 978-1-4503-4516-3/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2988257.2988266

motion vector with a blue tail and red tip on white background. Therefore only the beginning and the end of the movement was indicated by the colored pixels, while the in-between motion was all black. Meng et al. [17] in their participation to AVEC 2013 [30], used MHIs for depression recognition, by extracting Local Binary Patterns (LBP) and Edge of Oriented Histograms from them. Optical flow, first introduced as a concept by Gibson in 1950 [9] to describe the visual stimulus provided to animals when they move, has been widely used in the area of computer vision for motion detection and recognition, object segmentation, motion compensation etc. Histogram of Optical Flow (HOF) has been proposed for depression estimation in AVEC 2014 [11] [14]. HOF estimates the motion within visual representations, by using vectors for each pixel in order to describe a dense motion field. Head motion is another widely used feature in depression assessment [8] [11]. Blinking rate has also been proposed, among others, by Zhou et al. in [31]. Text features, finally, have not been used as widely as the visual for depression assessment, with the only relevant approaches reported in [8] and [11].

3. FEATURE EXTRACTION

The current DAIC-WOZ dataset [10] includes audio recordings, audio features, interview transcripts, video features, pixel coordinates for 68 2D facial landmarks, world coordinates for 68 3D facial landmarks, gaze vector, head-pose vector, Histogram of Oriented Gradients (HOG) features, emotion, and Action Unit (AU) labels. The challenge provides three data subsets: training, development, and test. Depression labels - one for each recording - are provided only for the training and development sets, and are withheld for the test set. Participants' gender is provided for all three subsets.

The provided baseline features are all frame-based, derived from static images, as opposed to features employed in the proposed work which are all video-based, both of high and of low level. As a results, none of the baseline features have been used herein without undergoing further processing in order to derive video-based features. The choice of video-based against frame-based features is in accordance to literature [20] [21], which supports that signs of depression are of temporal nature [7]. Thus information derived over the course of an interview session is expected to be much more informative than any single frame. In fact the Diagnostic and Statistical Manual of Mental Disorders [2] suggests that symptoms must have a minimum duration of two weeks to qualify for depression diagnosis.

Features selected for the proposed work were the ones considered as clinically relevant [2] [7], and which are further supported by the presented related work. High level features are those which can be translated to common sense knowledge; for instance head motion, blinking, facial expressions, AUs, and text related features, can be annotated with a high degree of certainty by a human expert. Low level features, on the other hand, are derived from image processing algorithms, which extract descriptors from an image, but cannot be directly translated to human knowledge. Such features extracted here are the Landmark Motion History Images combined with LBP and HOG, Landmark Motion Magnitude, and most of the audio-based features.

In the following paragraphs, feature extraction algorithms are explained in detail for each modality (video, audio, and text). Features used in the classification approaches were computed over the entire interview period excluding the initial and final portions during which the participants were positioned and given instructions. Audio and text features were further constrained to the participants' own vocal responses, disregarding periods when the virtual human is speaking. Most of the feature extraction algorithms were implemented in MATLAB using embedded libraries, with the following exceptions: i) the LBPs that were extracted using the library provided by Ojala et al. [19], ii) the Landmark Motion Magnitude (LMM) which employed methods from OpenCV [12] and the Boost C++ libraries [24], and iii) the Linguistic Inquiry and Word Count (LIWC) features that were extracted by the corresponding software [15].

3.1 Visual Features

Although visual features were limited by the unavailability of raw video recordings, several meaningful (both high and low level) features could be extracted from the set of numbered 2D facial landmarks illustrated in Fig.1. Statistics derived from the provided features of emotion, AUs, gaze, and pose were also used as complementary features.

3.1.1 Landmark Motion History Images

Landmark Motion History Images (LMHI) were computed on the provided 2D landmarks, in the absence of the actual intensities of the video frames. LMHI encodes the motion of the facial features into a grayscale image, with the most recent movement corresponding to white pixels, the earliest corresponding to the darkest gray, and temporally intermediate movements indexed by corresponding gray values. The extension of the proposed work in comparison to that of Ptucha and Savakis [23] is that the in-between motion is also preserved with the use of respective gray-scales, which is important for the descriptors applied later on the LMHI. The landmarks used for LMHI, according to the numbering of Fig.1, were those corresponding to eyebrows {18-27}, eyes {37-48}, nose-tip {32-36}, and mouth {49-68}. Before computing the LMHI all landmarks were coregistered, using affine transformation, by aligning the points corresponding to the temples, chin, and inner and outer corners of the eyes {1, 9, 17, 37, 40, 43, 46}.

The gray value was defined by a step s, which corresponded to the maximum pixel value (255) divided by the total number of frames. Thus in every frame the gray value was computed by s multiplied by the frame count (i.e. for the 4th frame the gray value was 4 x s). The morphological operation of erosion was applied in order to remove outliers (very distant movements), and the image was cropped to the non-black pixels (non-zero). The resulting LMHI was further resized to fit the average size of the LMHI. An example of the resulting LMHI is illustrated in Fig.2, where the amount of movement is indexed by the thickness of the pixels.

LBP, as well as HOG, were extracted based on LMHI. LBP was computed for two sets of parameters: radius and neighborhood for {1, 8} and {2, 16}, resulting in a total of 28 features covering the entire face. The LMHI was further partitioned using corresponding half ratios to represent three composite regions: "nose+mouth", "left eye+eyebrow", and "right eye+eyebrow". These regions are partitioned by red dashed-lines in Fig.2. For each subregion 28 additional features were computed. Further, HOG was computed for

Figure 1: 2D Facial Landmarks numbering

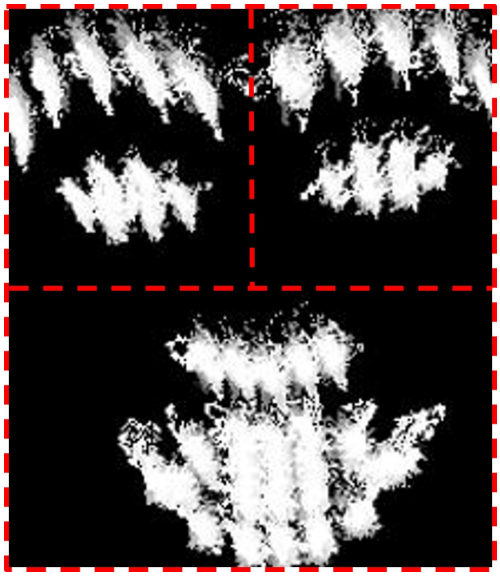

Figure 2: Landmark Motion History Image

the entire face area with 1080 features, as well as for the remaining sub-regions of "nose+mouth", giving 360 features, "left eye+eyebrow" with 144, and "right eye+eyebrow" with 144 features as well. Additionally, the LMHI histogram bins were computed resulting in 255 additional features (black was excluded). Mean and standard deviation of the pixel values constitute the final two features, resulting in a total of 2097 LMHI features.

3.1.2 *Landmark Motion Magnitude*

Landmark Motion Magnitude (LMM), inspired by HOF, was applied here to the 2D landmarks. For the extraction of the landmark activity features, the vectors that displace each landmark from one frame to the next were calculated based on the landmark coordinates. From these vectors, unwanted global head motion was removed by subtracting the average motion vector of a landmark group representing the nose {28:36}. Subsequently the magnitude of the displacement vectors was calculated as in (1)

$$M = \sqrt{(x^2 + y^2)} \tag{1}$$

and the maximum magnitude from each of the following 5 landmark groups was selected for each frame: right eyebrow {18-22}, left eyebrow {23-27}, mouth center {51-53, 62-64, 66-68, 57-59}, right mouth corner {49, 61} and left mouth corner {55, 65}. A sample of LMM vectors is shown for the mouth region in Fig.3. The five LMM time-signals, corresponding to the five regions defined above, were used for the feature extraction. Statistical and spectral features were extracted for the whole duration of the interview. These features were selected based on our previous positive experience for stress/anxiety assessment [22], as well as for activity recognition from accelerometer data [5] which describe motion patterns similar to LMM. The resulting 70 features in total are described next.

The *variance* of the time intervals between any two subsequent spikes or transients was used as an index of movement based on the assumption that rhythmic movements would be associated with near-zero variance. The *energy ratio* of the

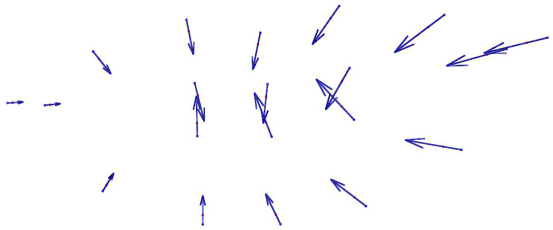

Figure 3: Landmark Motion Magnitude for the mouth region between 2 consecutive frames.

autocorrelation sequence was calculated as the relationship of the energy contained in the last 75% of the samples of the autocorrelation sequence and the energy contained by the first 25%, and was used as a measure of the motion manifested as quasiperiodic spikes (randomness). The *median* of the signal based on the P2 algorithm [13]. The *standard deviation* of the signal. The *interquartile range* (i.e. the difference between the 75th and 25th quartiles), based on the P2 algorithm [13]. The *skewness* of the sample distribution, defined as the ratio of the 3rd central moment to the 3/2-th power of the 2nd central moment of the samples. The *kurtosis* of the sample distribution, defined as the ratio of the 4th central moment and the square of the 2nd central moment of the samples, minus 3. The *Shannon entropy* of the energy in bins, $H(x_i)$ calculated from the normalized energy of 10 equally sized consecutive bins, taken from the signal, as in (2).

$$H(x_i) = -\sum_{i=1}^{N} x_i log_2 x_i \tag{2}$$

The *25% spectral power frequency*, which corresponds to the upper bound of the frequency band starting at 0 Hz that contains 25% of the total spectral power. The *dominant frequency* which corresponds to the signal frequency associated with the highest power. The *spectral roll-off*: the frequency value at which 80% of the spectral power is below that point. The *spectral centroid* as in (3) corresponding to the average

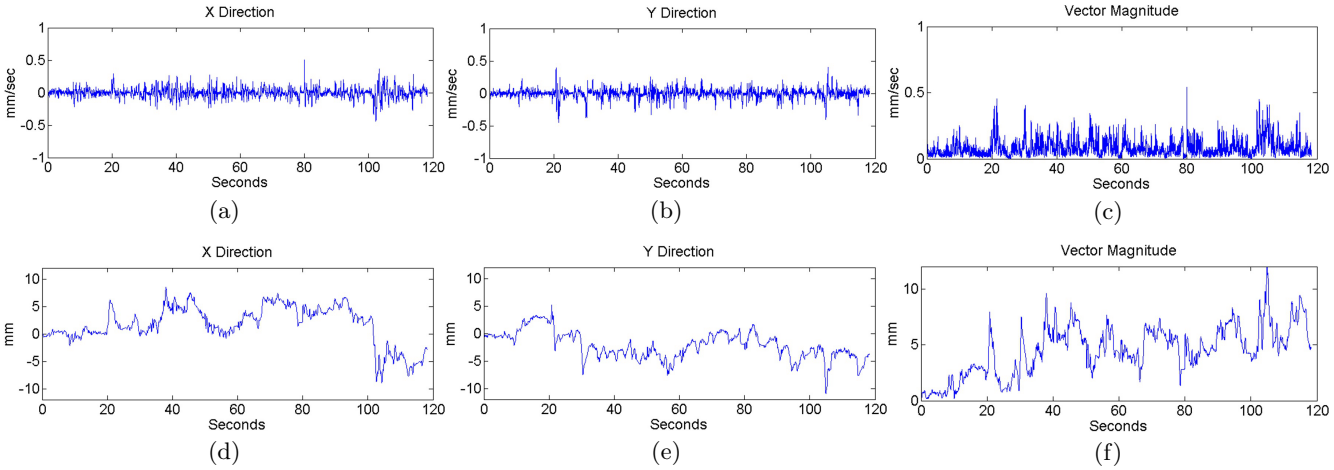

Figure 4: Head motion time series for velocities in first row, and for displacements in the second.

frequency of the spectrum, according to the formula:

$$SpectralCentroid = \frac{\sum_{i=1}^{N} f_i p_i}{\sum_{i=1}^{N} p_i} \quad (3)$$

where p_i is the power at frequency f_i.

3.1.3 Head Motion

The set of head-pose features provided by the baseline is frame-based, just as the landmarks. Both consist of coordinates and do not incorporate temporal and motion related information. Hereby, head motion was computed based on the horizontal and vertical deviations of specific reference points (landmarks 2, 4, 14, 16) between consecutive frames. The reference points selected were those characterized by minimal intra-facial movement, such as mouth movements, eye blinks, and other facial expressions, in order to emphasize global head motion.

The assumption was that the region between the eyes and mouth is the most appropriate region, as it is minimally involved in facial expressions and occlusions. The positions of the reference landmarks in each frame were stored in four different vectors that represent the trajectory of each landmark [16]. The most stable reference points were selected for further analysis by calculating the maximum distance travelled by each point between consecutive frames; points with a distance exceeding the mode of the distribution were considered as unstable and therefore discarded. Finally, the trajectories of the points were analyzed, producing six different time series related to the frame to frame movement and velocity: the horizontal and vertical scalar components and the vector magnitude. Fig.4 displays an example of unfiltered time series.

Several statistical indices were subsequently derived from the six time series in the form of the mean, median, and standard deviation of a) velocity and displacement separated on the X and Y axes, and b) velocity and displacement magnitude, resulting in a total of 18 features.

3.1.4 Blinking Rate

Blinking rate was extracted using the 2D landmarks in order to segment and mark out the eyeball perimeter. The landmark points used for each eye were numbered {37-42} and {43-48} in Fig.1. The area defined by each set of landmarks was computed over the entire recording. Time series were filtered to remove spikes and smooth out highly variable segments. Sharp decreases were considered as blinks. Detection of downward peaks was performed using a gradient peak detection algorithm utilizing the following parameters: minimum peak distance, peak duration, derivative amplitude, and derivative peak distance. The resulting feature was blink frequency, i.e. the number of blinks per minute.

3.1.5 Emotions, AUs, Gaze & Pose

Emotion variability features were computed based on the emotion and AU labels embedded in the AVEC dataset. The statistical measures chosen to represent this variability were: minimum, maximum, mean, mode, median, range, mean deviation, variance, standard deviation, skewness, and kurtosis. These indices were calculated for each of the 10 prelabeled emotions, and for each of 19 AUs, resulting in 110 and 209 feature sets, respectively. The same set of statistical indices were computed for the provided gaze and pose.

3.2 Audio Features

The audio data provided by AVEC consisted of a series of pre-extracted features using the COVAREP toolbox at 10-ms intervals over the entire recording (F_0, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd conf, MCEP 0-24, HMPDM 1-24, HMPDD 1-12, and Formants 1-3). For the purposes of the present work, the resulting timeseries data were submitted to additional preprocessing steps as follows: First, the participant's voice was isolated using the time stamps in the TRANSCRIPT files. Segments with TRANSCRIPT values of "<synch>", "<laughter>", "[laughter]", "<sigh>" and "scrubbed entry" were ignored as non-informative segments. Next, segments containing unvoiced segments (VUV=0) were removed from the final concatenated time series. Furthermore, to correct for instances of apparently inaccurate annotation analyses were restricted to continuous voiced segments lasting > 5 ms. To control for speaker dependency, the F_0 was normalized to a scale of 0 to 1, and the deltas and delta-deltas were extracted for F_0 and MFCCs.

The main analyses consisted in computing three sets of features to be used in subsequent classification approaches.

Table 1: Statistical descriptors calculated from the pre-extracted audio features

Low Level Descriptors	Statistical descriptors
normalized F_0	mean, min, skewness, kurtosis, standard deviation, median, peak-magnitude to root-mean-square ratio, root mean square level, interquartile range, spectral flatness
delta F_0 , delta-delta F_0	mean, standard deviation
NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd conf	mean, standard deviation, peak-magnitude to root-mean-square ratio, root mean square level, interquartile range, spectral flatness
MCEP 0-24	mean, standard deviation, peak-magnitude to root-mean-square ratio, root mean square level, interquartile range, spectral flatness
delta MCEP 0-24, delta-delta MCEP 0-24	mean, standard deviation
HMPDM 1-24, HMPDD 1-12	mean, standard deviation, root mean square level, interquartile range
Formants 1-3	mean, standard deviation, peak-magnitude to root-mean-square ratio, root mean square level, interquartile range, spectral flatness

The first set of audio features consisted of a series of statistical descriptors (shown in Table 1) for each pre-extracted descriptor. The second set of audio features consisted of Discrete Cosine Transform (DCT) coefficients for each descriptor in the first column of Table 1. The first 10 values of the DCT were retained, reducing the number of parameters and therefore complexity. The third set of audio features consisted of 8 high level features which were computed for the entire duration of the concatenated time-series. The *Pause Ratio* measuring the frequency of pauses during the participant's speech. Pauses were detected automatically using a pause detector [26], which relied on a low loudness detection function based on the Perceptual Quality measure. The *Voiced Segment Ratio*, computed as the number of voiced segments divided by the length of the entire speech segment. The *Speaking Ratio*, computed as the number of speaking instances that there is participant's speech, was divided by the total number of selected recorded segments, as in (4)

$$SpeakingRatio = \frac{\#allinstants - \#pauses}{\#allinstants} \quad (4)$$

The *Mean Laughter Duration* was defined as the duration of laughter segments divided by the total number of laughter instances. The *Mean delay in response* to Ellie's questions. The *Mean duration of pauses*. The *Maximum duration of pauses*. The *Fraction of pauses in overall time*.

Finally, the former two sets of features were individually combined in feature level with the high level features into single feature vectors. The final set of statistical descriptors with high-level features was of size 494, and the set of DCTs with high-level features was of size 1278.

3.3 Text Features

We also extracted a number of features from participants' responses. Namely, total number of words and number of sentences, normalized by video duration, average number of words in each sentence, and ratio of laughters over the total number of words. The use of these features is supported by literature indicating that slowed and reduced amount of speech, elongated speech pauses, and short answers are nonverbal manifestations of depression [7]. A fifth text-based feature was extracted in the form of the ratio of depression-related words over the total number of words spoken. Depression-related words were identified from a dictionary of about 700 words that was manually constructed based on online resources such as [6] [18].

Additionally, the Affective Norms for English Words ratings (ANEW) [4] provided seven additional text features: Mean and Standard Deviation for pleasure, arousal, and dominance ratings, and word frequency.

Finally, features extracted by the LIWC software [15] provided 93 additional features. The LIWC software was developed to identify words referring to social processes (such as reference to family, friends, and social affiliation) and psychological states (such as overall negative/positive and specific emotions).

4. FEATURE SELECTION

The value of individual features was assessed by documenting the effect of removing each feature or set of features on the resulting f1 score on the development set.

The initial performance, that of the complete visual features' set, was 0.0 for identifying the depressed and 0.83 for identifying the non-depressed. When gaze was removed the performance improved to 0.36 (0.88) respectively. Additionally, with emotions and AUs removal the performance improved to 0.5 (0.9). Among visual features only LMHI-FaceHOG appeared to have a noticeable impact on the performance, since its removal resulted in a drastic drop of f1 to 0.13 (0.76). Removal of the remaining features had null impact on classification accuracy. Therefore, only LMHI-FaceHOG was included in the final visual feature set.

Audio feature statistical descriptors that did not appear to impact classification accuracy were the mean, Standard Deviation (SD) and RMS of mcep(20-24), iqr of mcep(20), sp flatness of mcep20, mean and SD of delta mcep(20-24), and SD of delta-delta mcep(20-24). DCT features that did not add to classification performance included mcep(20-24), delta mcep(20-24), and delta-delta mcep(20-24).

With respect to the text features, removal of ANEW features had the greatest impact on classification performance, while custom features had a lesser yet significant effect on performance, and were retained in the final classification schemes. Given that removal of LIWC features, separately and in various combinations, had either null or positive impact on classification performance, they were not included in the final text-feature set.

Table 2: Performance of visual features in the course of selection, tested on the development set. F1-score: *depressed* (*not-depressed*)

Feature-Setup	F1-Score	Impact
Full-set	0.00 (0.83)	-
Gaze-removed	0.36 (0.88)	↑
AUs-removed	0.50 (0.90)	↑
LMHIFaceHOG-removed	0.13 (0.76)	↓

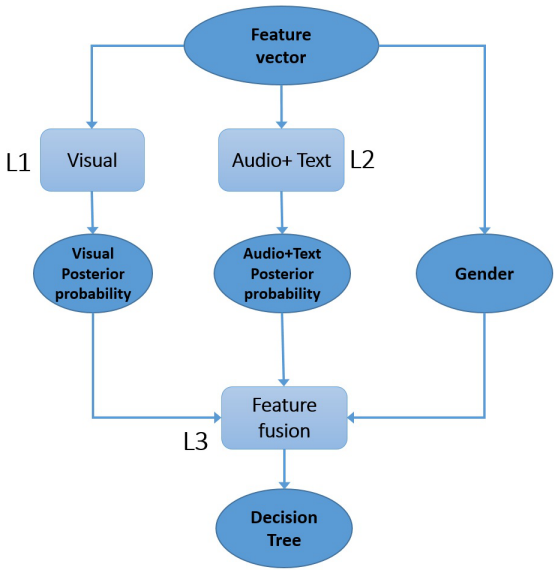

Figure 5: Posterior Probability Classification Model

5. CLASSIFICATION

Four different approaches were implemented for classification: Gender-based, Feature-Level Fusion, Decision Fusion, and the Posterior-Probability Model. A Decision Tree classification algorithm was applied in each case.

Gender-based classification for depression has been reported to substantially improve performance [1] [25]. In the present work, gender-based classification was implemented by building two separate classifiers, one for male and another for female participants. The classifier for male subjects was trained on feature-sets extracted from data of male participants and the female classifier with feature-sets extracted from data of female participants.

The Feature Level Fusion concatenates the individual modalities to produce a single feature vector, so that a single classifier is trained. The Decision Fusion model utilized labels that were produced from the individual classifiers per modality (either gender-independent or gender-based) and combined them through logical operations ("OR" and "AND") in all possible combinations.

The Posterior Probability Classification Model was based on the posterior probabilities resulting from the Decision Trees, and consisted of three layers. Layer 1 was the Decision Tree of the Visual features, while Layer 2 was the Decision Tree of the fused Audio+Text (feature fusion). The input to Layer 3 consisted of the posterior probabilities produced by Layers 1 and 2, plus the gender label, resulting in a three-item feature vector for the final decision. Fig. 5 illustrates this classification approach.

6. EXPERIMENTAL RESULTS

The aforementioned classification approaches were evaluated through training on the training set and subsequent testing with the development and test sets. In addition, the algorithms were assessed using the leave-one-out procedure on the joined training and development sets. Performance of each modality, and of the fused models, in gender-independent and gender-based models, as well as a comparison to the baseline performance, are reported in Tables 3, 4, and 5. The models that performed best on the development set were further evaluated on the test set.

As shown in Table 4, the gender-based approach outperformed gender-independent models with the audio statistical descriptors and text features. There was a slight improvement in performance of the gender-based over the gender-independent feature fusion approach as well.

The Decision Fusion models performed best on the development data set (c.f. Table 3 #7 and #8, and Table 5). Surprisingly, these models did not fare as well when applied to the test data set. The best-performing model in the Leave-one-Out scheme was the Posterior-Probability model which, however, performed rather poorly when applied to the development set.

Compared to the reference audio model provided by AVEC [27], three of the current models showed improved performance: the gender-based model utilizing statistical descriptors for both development and test sets, and the decision fusion (Table 3 #8) for the development set (c.f. Table 5).

7. CONCLUSIONS

Several conclusions can be drawn from the present work. First, the proposed gender-based model utilizing simple statistical descriptors of pre-extracted, low-level audio features or the audio gender-independent DCT features, outperformed reference classification accuracy (c.f. Tables 3 and 5). Improvement in performance could be attributed to the fact that the proposed audio-based features were computed over the entire recording as opposed to the reference model which was frame-based, as well as the fact that the classification model was gender-based. Additionally, the pre-processing of the audio features increased the performance significantly (c.f. Table 3 #2 in comparison to #3).

Another contribution of the present work lies in the novel implementation of the LMHI and LMM, as well as in fusing high and low level features. Results pertaining to the visual features extracted from the provided data set were rather surprising given that analyses highlighted a single visual feature (LMHIFaceHOG) has been the most significant. This could be explained by the fact that LMHIFaceHOG incorporates motion information, and by being registered and resized, it minimizes appearance-based variation. The remaining visual descriptors, as well as the sub-region HOGs, appeared not to contribute to the classification performance. However this conclusion is based on an empirical feature selection, as described in Section 4. Given that the performance of visual features, in any combination, did not achieve to outperform the respective baseline score, further future investigation is required.

Also, it is worth pointing out that some features reported to be significant for depression classification, both in clinical literature as well as in previous related work, did not emerge as significant in the current analyses, including AUs, emo-

Table 3: Comparison of f1-score for the single modalities classification, and the fused models, during testing with Leave-One-Out, as well as by training with the Training, and testing on the development, and test splits. F1-scores are reported for both classes *depressed* / *not-depressed*, with *not depressed* in brackets.

#	textbfMethod	**Leave-One-Out**	**Development**	**Test**
1	Vis	0.35 (0.81)	0.50 (0.9)	0.18 (0.75)
2	AudGen without pre-processing	0.43 (0.74)	0.42 (0.78)	-
3	AudGen	0.45 (0.85)	0.59 (0.87)	**0.52 (0.81)**
4	Audio-DCT	0.19 (0.71)	0.47 (0.83)	
5	TextGen	0.23 (0.79)	0.46 (0.88)	-
6	Feature Level Fusion {Vis+Aud+Text}	0.35 (0.79)	0.50 (0.76)	-
7	Decision Fusion {Vis OR AudGen}	0.44 (0.75)	0.63 (0.86)	0.43 (0.63)
8	Decision Fusion {(Vis OR AudGen) AND (Vis OR TextGen)}	0.42 (0.81)	**0.62 (0.91)**	0.23 (0.71)
9	Decision Fusion {(VisGen OR Text) AND (AudGen OR Text)}	0.47 (0.86)	0.33 (0.86)	-
10	Posterior Probability Classification Model	**0.70 (0.95)**	0.32 (0.46)	-

*Note: **Vis**= Visual Feature (LMHIFaceHOG); **VisGen**= Gender-based Vis; **Aud**= Audio Statistical Descriptors; **AudGen**= Gender-based Aud; **Text**=Text features (custom+ANEW); **TextGen** = Gender-based Text.

Table 4: Gender-independent versus gender-based classification. F1-score: *depressed* (*not-depressed*)

	Gender-independent	**Gender-based**
Visual	**0.5 (0.9)**	0.17 (0.83)
Audio	0.24 (0.75)	**0.59 (0.87)**
Text	0.36 (0.88)	**0.46 (0.88)**
Feature Fusion	0.35 (0.79)	**0.36 (0.83)**

Table 5: Comparison of the proposed approaches to the baseline paper [27], in respect to modality. F1-score: *depressed* (*not-depressed*)

Partition	**Modality**	**Baseline**	**Proposed**
Development	Audio	0.41 (0.58)	**0.59 (0.87)**
Development	Video	**0.58 (0.86)**	0.50 (0.90)
Development	Ensemble	0.58 (0.86)	**0.62 (0.91)**
Test	Audio	0.46 (0.68)	**0.52 (0.81)**
Test	Video	**0.50 (0.90)**	0.18 (0.75)
Test	Ensemble	**0.50 (0.90)**	0.23 (0.71)

tions, gaze, pose, head motion, and blinks. In part, this may be attributed to certain dataset characteristics. Performance of the pre-processing applications, which provided the baseline-visual features could not be verified in the absence of raw video recordings, thus landmarks or emotion/AU labels may not have been 100% accurate. Occasional inaccuracies of the time-stamps on the transcripts, could have an impact on the selection of the relevant segments/frames. Additionally, the unbalance in the number of participants originally rated as depressed versus not-depressed, as well as the fact that the categorization was based on self-report scales, as opposed to clinical diagnosis, could be important for the outcome.

The fact that Decision Fusion methods were the best performing on the development set, but did not perform accordingly on the test set (c.f. Table 2 #6 and #7), could be the result of overfitting. A similar explanation may apply to the poor performance of the Posterior Probability Classification Model on the development set, despite its very high performance in the Leave-One-Out validation.

In future work improvement of feature selection methods is probably the best avenue to enhance classification performance. Inspection of the bivariate and partial correlation matrix between individual features and using probability-based statistic indices (such as Fisher's z) to identify significant associations may help optimize feature selection. Extracting video-based features on shortened time windows (e.g., 1 sec) may further improve the sensitivity of visual features.

This year's dataset presents a better interpersonal context for depression assessment in view of the extant literature supporting the better suitability of interviews for detecting signs of depression. However, given the importance of symptom/sign stability for depression diagnosis, repeated recordings over several days or weeks would render results more clinically relevant. In a similar vein data recorded from persons with a clinical diagnosis of depression would be desirable and should be considered in future challenges.

8. ACKNOWLEDGMENTS

Pampouchidou Anastasia was funded by the "In memory of Maria Zaousi" scholarship from the Greek State Scholarships Foundation (I.K.Y). Additionally, part of this work was performed in the framework of the FP7 Specific Targeted Research Project SEMEOTICONS, partially funded by the European Commission under Grant Agreement 611516.

9. REFERENCES

[1] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. Head Pose and Movement Analysis as an Indicator of Depression. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Publishing, 2013.

[3] A. Bobick and J. Davis. Real-time Recognition of Activity Using Temporal Templates. In *3rd IEEE Workshop on Applications of Computer Vision*, 1996.

[4] M. M. Bradley and P. J. Lang. Affective Norms for English words (ANEW): Instruction Manual and Affective Ratings. Technical report, Technical report C-1, University of Florida, 1999.

[5] C. Chatzaki, M. Pediaditis, G. Vavoulas, and M. Tsiknakis. Estimating Normal and Abnormal Activities Using Smartphones. In *13th International Conference on Wearable, Micro and Nano Technologies for Personalised Health*, 2016.

[6] Depression Vocabulary, Depression Word List. https://myvocabulary.com/word-list/depression-vocabulary/.

[7] H. Ellgring. *Non-verbal Communication in Depression*. Cambridge University Press, 2007.

[8] S. Ghosh, M. Chatterjee, and L.-P. Morency. A Multimodal Context-Based Approach for Distress Assessment. In *16th International Conference on Multimodal Interaction*, 2014.

[9] J. J. Gibson. *The Perception of the Visual World*. Houghton Mifflin, 1950.

[10] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2014.

[11] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan. Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions. In *4th ACM International Workshop on Audio/Visual Emotion Challenge*, 2014.

[12] Itseez. Open Source Computer Vision Library. https://github.com/itseez/opencv, 2015.

[13] R. Jain and I. Chlamtac. The P2 Algorithm for Dynamic Calculation of Quantiles and Histograms Without Storing Observations. *Communications of the ACM*, 28(10):1076–1085, 1985.

[14] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux. Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. In *4th ACM International Workshop on Audio/Visual Emotion Challenge*, 2014.

[15] Linguistic Inquiry and Word Count (LIWC). http://liwc.wpengine.com/.

[16] D. Manousos, G. Iatraki, E. Christinaki, M. Pediaditis, F. Chiarugi, M. Tsiknakis, and K. Marias. Contactless Detection of Facial signs Related to Stress: A Preliminary Study. In *International Conference on Wireless Mobile Communication and Healthcare*, 2014.

[17] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang. Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression. In *3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 2013.

[18] Negative Feeling Words. http://eqi.org/fw neg.htm.

[19] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Ratterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[20] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, G. Lemaître, and F. Meriaudeau. Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes. In *38th International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016.

[21] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, and F. Meriaudeau. Designing a Framework for Assisting Depression Severity Assessment from Facial Image Analysis. In *IEEE International Conference on Signal and Image Processing Applications*, 2015.

[22] M. Pediaditis, G. Giannakakis, F. Chiarugi, D. Manousos, A. Pampouchidou, E. Christinaki, G. Iatraki, E. Kazantzaki, P. Simos, K. Marias, and M. Tsiknakis. Extraction of Facial Features as Indicators of Stress and Anxiety. In *37th International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3711–3714. IEEE, 2015.

[23] R. Ptucha and A. Savakis. Towards the Usage of Optical Flow Temporal Features for Facial Expression Classification. In *International Symposium on Visual Computing*. Springer, 2012.

[24] B. Schäling. The boost C++ libraries. *XML Press*, 2014.

[25] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic Nonverbal Behavior Indicators of Depression and *PTSD*: Exploring Gender Differences. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[26] Y. Stylianou, V. Hazan, V. Aubanel, E. Godoy, S. Granlund, M. Huckvale, E. Jokinen, M. Koutsogiannaki, P. Mowlaee, M. Nicolao, T. Raitio, A. Sfakianaki, and Y. Tang. P8-Active Speech Modifications. In *International Summer Workshop on Multimodal Interfaces*, 2012.

[27] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016-Depression, Mood, and Emotion Recognition Workshop and Challenge. *arXiv preprint arXiv:1605.01600*, 2016.

[28] M. Valstar, M. Pantic, and I. Patras. Motion History for Facial Action Detection in Video. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 635–640. IEEE, 2004.

[29] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *4th International Workshop on Audio/Visual Emotion Challenge*, 2014.

[30] M. 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 *3rd International Workshop on Audio/Visual Emotion Challenge*, 2013.

[31] D. Zhou, J. Luo, V. M. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz. Tackling Mental Health by Integrating Unobtrusive Multimodal Sensing. In *29th AAAI Conference on Artificial Intelligence*, 2015.