AVEC 2014 – 3D Dimensional Affect and Depression Recognition Challenge*

Michel Valstar
University of Nottingham
School of Computer Science

Björn Schuller†
TU München
MISP Group, MMK

Kirsty Smith
University of Nottingham
School of Computer Science

Timur Almaev
University of Nottingham
School of Computer Science

Florian Eyben
TU München
MISP Group, MMK

Jarek Krajewski
University of Wuppertal
Schumpeter School of
Business and Economics

Roddy Cowie
Queen's University
School of Psychology

Maja Pantic‡
Imperial College London
Intelligent Behaviour
Understanding Group

ABSTRACT

Mood disorders are inherently related to emotion. In particular, the behaviour of people suffering from mood disorders such as unipolar depression shows a strong temporal correlation with the affective dimensions valence and arousal. In addition to structured self-report questionnaires, psychologists and psychiatrists base their evaluation of a patient's level of depression on the observation of facial expressive and vocal cues.It is in this context that we present the fourth Audio-Visual Emotion recognition Challenge (AVEC 2014). This edition of the challenge uses a subset of the AVEC 2013 data, to allow for more focussed study. In addition, labels for a third dimension (Dominance) has been added and the number of annotators per clip has been increased to a minimum of three, with most clips annotated by 5. The challenge has two goals logically organised as sub-challenges: the first is to predict the continuous values of the affective dimensions valence, arousal and dominance at each moment in time. The second sub-challenge is to predict the value of a single self-reported depression indicator for each recording in the dataset. This paper presents the challenge guidelines, the common data used, and the performance of the baseline system on the two tasks.

Categories and Subject Descriptors

J [**Computer Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Affective Computing, Emotion Recognition, Speech, Facial Expression, Challenge

1. INTRODUCTION

The 2014 Audio-Visual Emotion Challenge and Workshop (AVEC 2014) will be the fourth competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, video and audio-visual emotion analysis, with all participants competing under strictly the same conditions. The goal of the Challenge is to compare the relative merits of the two approaches (audio and video) to emotion recognition and severity of depression estimation under well-defined and strictly comparable conditions and establish to what extent fusion of the approaches is possible and beneficial. A second motivation is the need to advance emotion recognition for multimedia retrieval to a level where behaviomedical systems are able to deal with large volumes of non-prototypical naturalistic behaviour in reaction to known stimuli, as this is exactly the type of data that diagnostic tools and other applications would have to face in the real world.

According to European Union Green Papers dating from 2005 [15] and 2008 [16], mental health problems affect one in four citizens at some point during their lives. As opposed to many other illnesses, mental ill health often affects peo-

*This is a preliminary version of the baseline paper, released early for the convenience of the AVEC 2014 participants. While we have tried our utmost best to ensure all information in this manuscript is correct, some modifications may become necessary.

†The author is further affiliated with Imperial College London, Department of Computing, London, U.K.

‡The author is further affiliated with Twente University, EEMCS, Twente, The Netherlands.

ACM-Multimedia 2014 Orlando, Florida, USA

http://dx.doi.org/10.1145/2512530.2512533.

ple of working age, causing significant losses and burdens to the economic system, as well as the social, educational, and justice systems. It is therefore somewhat surprising that despite the scientific and technological revolutions of the last half century remarkably little innovation has occurred in the clinical care of mental health disorders in general, and unipolar depression in particular.

Affective Computing and Social Signal Processing are two developing fields of research that promise to change this situation. Affective Computing is the science of automatically analysing affect and expressive behaviour [21]. By their very definition, mood disorders are directly related to affective state and therefor affective computing promises to be a good approach to depression analysis. Social Signal Processing addresses all verbal and non-verbal communicative signalling during social interactions, be they of an affective nature or not [26]. Depression has been shown to correlate with the breakdown of normal social interaction, resulting in observations such as dampened facial expressive responses, avoiding eye contact, and using short sentences with flat intonation. Although the assessment of behaviour is a central component of mental health practice it is severely constrained by individual subjective observation and lack of any real-time naturalistic measurements. It is thus only logical that researchers in affective computing and social signal processing, which aim to quantify aspects of expressive behaviour such as facial muscle activations and speech rate, have started looking at ways in which their communities can help mental health practitioners.

In the case of depression, which is the focus of AVEC 2013, the clinician-administered Hamilton Rating Scale for Depression [13] is the current gold standard to assess severity [3, 30], whereas the gold-standard for diagnosis is the Structured Clinical Interview for DSM-IV (SCID) [10]. The Hamilton scale is not free to use, but other self report measures are. The frequently-used Beck Depression Inventory-II [4] is one of them, and is the one used to obtain the ground truth measure for AVEC. All of these instruments pay little or no attention to observational behaviour. In part for that reason, social signal processing and affective computing could make significant contribution by achieving an objective, repeatable and reliable method to incorporate measurable behaviour into clinical assessment.

In the first published efforts towards this, the University of Pennsylvania has already applied a basic facial expression analysis algorithm to distinguish between patients with Schizophrenia and healthy controls [27, 14]. Besides diagnosis, affective computing and social signal processing would also allow quantitative monitoring of the progress and effectiveness of treatment. Early studies that addressed the topic of depression are e.g. [27, 5].

More recently, Girard et al. [12] performed a longitudinal study of manual and automatic facial expressions during semi-structured clinical interviews of 34 clinically depressed patients. They found that for both manual and automatic facial muscle activity analysis, participants with high symptom severity produced more expressions associated with contempt, smile less, and the smiles that were made were more likely to be related to contempt. Yang et al [29] analysed the vocal prosody of 57 participants of the same study. They found moderate predictability of the depression scores based on a combination of F_0 and switching pauses. Both studies used the Hamilton Rating Scale for Depression, which is a multiple choice questionnaire filled in by a clinician and used to provide an indication of depression, and as a guide to evaluate recovery. Scherer et al. [22] studied the correlation between automatic gaze, head pose, and smile detection and three mental health conditions (Depression, Post-Traumatic Stress Disorder and Anxiety). Splitting 111 participants into three groups based on their self-reported distress, they found significant differences for the automatically detected behavioural descriptors between the highest and lowest distressed groups.

Dimensional affect recognition aims to improve the understanding of human affect by modelling affect as a small number of continuously valued, continuous time signals. Compared to the more limited categorical emotion description (e.g. six basic emotions) and the computationally intractable appraisal theory, dimensional affect modelling has the benefit of being able to: a. encode small changes in affect over time, and b. distinguish between many more subtly different displays of affect, while remaining within the reach of current signal processing and machine learning capabilities. The disadvantage of dimensional affect is the way in which annotations are obtained: inter-rater reliability can be notoriously low, caused by interpersonal differences in the interpretation of expressive behaviour in terms of dimensional affect but also issues surrounding reaction time, attention, and fatigue of the rater [20].

Depression severity estimation aims to provide an event-based prediction of the level of depression. Different from the continuous dimensional affect prediction, event-based recognition provides a single label over a pre-defined period of time rather than at every moment in time. In essence, continuous prediction is used for relatively fast-changing variables such as valence, arousal or dominance, while event-based recognition is more suitable for slowly varying variables such as mood or level of depression. One important aspect is that agreement must exist on what constitutes an event in terms of a logical unit in time. In this challenge, an event is defined as a participant performing a single human-computer interaction task from beginning to end.

We are calling for teams to participate in emotion and depression recognition from video analysis, acoustic audio analysis, linguistic audio analysis, or any combination of these. As benchmarking database the Depression database of naturalistic video and audio of participants partaking in a human-computer interaction experiment will be used, which contains labels for the three target affect dimensions arousal, valence and dominance, and Beck Depression Index-II (BDI-II), a self-reported 21 multiple choice inventory [4].
Two Sub-Challenges are addressed in AVEC 2014:

- The *Affect Recognition Sub-Challenge (ASC)* involves fully continuous affect recognition of three affective dimensions: Valence, Arousal, and Dominance (VAD), where the level of affect has to be predicted for every moment of the recording.
- The *Depression Recognition Sub-Challenge (DSC)* requires participants to predict the level of self-reported depression as indicated by the BDI for every experiment session, that is, one continuous value per multimedia file.

For the ASC, three regression problems need to be solved for Challenge participation: prediction of the continuous dimensions Valence, Arousal, and Dominance. The ASC

competition measure is the Pearson's product-moment correlation coefficient taken over the concatenation of labels over all tasks and averaged over all three dimensions. For the DSC, a single regression problem needs to be solved. The DSC competition measure is root mean square error over all tasks.

Both Sub-Challenges allow contributors to find their own features to use with their regression algorithm. In addition, standard feature sets are provided (for audio and video separately), which participants are free to use. The labels of the test partition remain unknown to the participants, and participants have to stick to the definition of training, development, and test partition. They may freely report on results obtained on the development partition, but are limited to five trials per Sub-Challenge in submitting their results on the test partition.

To be eligible to participate in the challenge, every entry has to be accompanied by a paper presenting the results and the methods that created them, which will undergo peer-review. Only contributions with a relevant accepted paper will be eligible for Challenge participation. The organisers reserve the right to re-evaluate the findings, but will not participate in the Challenge themselves.

We next introduce the Challenge corpus (Sec. 2) and labels (Sec. 3), then audio and visual baseline features (Sec. 4), and baseline results (Sec. 5), before concluding in Sec.6.

2. DEPRESSION DATABASE

The challenge uses a subset of the AVEC 2013 audio-visual depression corpus [25], which is formed of 150 videos of task-oriented depression data recorded in a human-computer interaction scenario. It includes recordings of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone. There is only one person in every recording and the total number of subjects is in our dataset is 84, i.e. some subjects feature in more than one recording. The speakers were recorded between one and four times, with a period of two weeks between the measurements. 18 subjects appear in three recordings, 31 in 2, and 34 in only one recording. The length of the full recordings is between 50 minutes and 20 minutes (mean = 25 minutes). The total duration of all clips is 240 hours. The mean age of subjects was 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 63 years. The recordings took place in a number of quiet settings.

The behaviour within the clips consisted of different human-computer interaction tasks which were Power Point guided. The recordings in the AVEC 2014 subset consist of only 2 of the 14 tasks present in the original recordings, to allow for a more focussed study of affect and depression analysis. Both tasks are supplied as separate recordings, resulting in a total of 300 videos (ranging in duration from 6 seconds to 4 minutes 8 seconds).

The 2 tasks were selected based on maximum conformity (i.e. most participants completed these tasks). The set of source videos is largely the same as that used for AVEC 2013, however 5 pairs of previously unseen recordings were introduced to replace a small number of videos which were deemed unsuitable for the challenge. The two tasks selected are as follows:

- NORTHWIND - Participants read aloud an excerpt of the fable "Die Sonne und der Wind" (The North Wind and the Sun), spoken in the German language
- FREEFORM - Participants respond to one of a number of questions such as: "What is your favourite dish?";"What was your best gift, and why?"; "Discuss a sad childhood memory", again in the German language

The original audio was recorded using a headset connected to the built-in sound card of a laptop at a variable sampling rate, and was resampled to a uniform audio bitrate of 128kbps using the AAC codec. The original video was recorded using a variety of codecs and frame rates, and was resampled to a uniform 30 frames per second at 640 x 480 pixels. The codec used was H.264, and the videos were embedded in an mp4 container.

For the organisation of the challenge, the recordings were split into three partitions: a training, development, and test set of 150 Northwind-Freeform pairs, totalling 300 task recordings. Tasks were split equally over the three partitions. Care was taken to have similar distributions in terms of age, gender, and depression levels for the partitions. There was no session overlap between partitions, i.e. multiple task recordings taken from the same original clip would be assigned to a single partition. The audio and audio-visual source files and the baseline features (see section 4) can be downloaded for all three partitions, but the labels are available only for the training and development partitions. All data can be downloaded from a special user-level access controlled website (`http://avec2013-db.sspnet.eu`).

3. CHALLENGE LABELS

The affective dimensions used in the challenge were selected based on their relevance to the task of depression estimation. These are the dimensions VALENCE, AROUSAL, and DOMINANCE (VAD) which form a well-established basis for emotion analysis in the psychological literature [11].

VALENCE is an individual's overall sense of "weal or woe": Does it appear that, on balance, the person rated feels positive or negative about the things, people, or situations at the focus of his/her emotional state? AROUSAL (Activity) is the individual's global feeling of dynamism or lethargy. It subsumes mental activity, and physical preparedness to act as well as overt activity. DOMINANCE is an individual's sense of how much they feel to be in control of their current situation.

A team of 5 naive raters annotated all human-computer interactions. The raters annotated the three dimensions in continuous time and continuous value using a tool developed especially for this task. The annotations are often called traces after the early popular system that performed a similar function called FeelTrace [6]. Instantaneous annotation value is controlled using a two-axis joystick. Every video was annotated by a minimum of three raters, and a maximum of five, due to time constraints. To reduce annotators' cognitive load (and hence improve annotation accuracy) each dimension was annotated separately. The annotation process resulted in a set of trace vectors$\{\mathbf{v}_i^v, \mathbf{v}_i^a, \mathbf{v}_i^d\} \in \mathbb{R}$ for every rater i and dimension v (VALENCE), a (AROUSAL), and d (DOMINANCE).

Sample values are obtained by polling the joystick in a tight loop. As such, inter-sample spacing is irregular (though

minute). These original traces are binned in temporal units of the same duration as a single video frame (i.e., 1/30 seconds). The raw joystick data for Arousal, Valence and Dominance lies in the range [-1000, 1000] labels, which is scaled by a factor 1/1000 to the range [-1, 1].

Inter-rater correlation coefficients (ICC) have been calculated using a combination of Pearson's r and RMSE. Since a number of annotation traces naturally contain zero variance, each rater's annotations were concatenated into a single "master trace" that contained traces of all tasks. We first calculated pair-wise inter-rater correlations. Not all raters annotated the same data. In all comparisons, only the files common to both raters were included in this process, and the ordering of the concatenations remained consistent throughout. Pairwise ICCs are shown in Table 1. Combinations in which no common files where available (and thus no comparison took place) are noted as "Not Applicable" (N/A) in the table.

For each dimension trace of every recording, the mean trace over all raters was calculated to form the ground truth affect labels for the Affect recognition Sub-Challenge.

The level of depression is labelled with a single value per recording using a standardised self-assessed subjective depression questionnaire, the Beck Depression Inventory-II (BDI-II, [4]). BDI-II contains 21 questions, where each is a forced-choice question scored on a discrete scale with values ranging from 0 to 3. Some items on the BDI-II have more than one statement marked with the same score. For instance, there are two responses under the Mood heading that score a 2: (2a) I am blue or sad all the time and I can't snap out of it and (2b) I am so sad or unhappy that it is very painful. The final BDI-II scores range from 0 – 63. Ranges can be interpreted as follows: 0–13: indicates no or minimal depression, 14–19: indicates mild depression, 20–28: indicates moderate depression, 29–63: indicates severe depression.

The average BDI-level in the AVEC 2014 partitions was 15.0 and 15.6 (with standard deviations of 12.3 and 12.0) for the Training and Development partitions, respectively. For every recording in the training and development partitions a separate file with a single value is provided for the DSC, together with three files containing the ground truth labels for each of the affective dimensions. The original traces from each rater were also provided for use within the ASC.

Similarly to AVEC 2013, we observed a non-linear correlation between the depression and affect labels. Graphs in Figure 1 demonstrate the mean emotional state for the entire duration of each clip, compared with the participants' BDI score at time of recording.

For each BDI score (0-45) in the graphs in Figure 2, an overall mean label has been calculated from the mean emotional state of each relevant clip. These figures also show the 95% confidence intervals, the width of which increases significantly where BDI is 15+. This is partially due to the distribution of the depression levels in the available data, which is shown in figures **??-??**.

4. BASELINE FEATURES

In the following sections we describe how the publicly available baseline feature sets are computed for either the audio or the video data. Participants could use these feature sets exclusively or in addition to their own features.

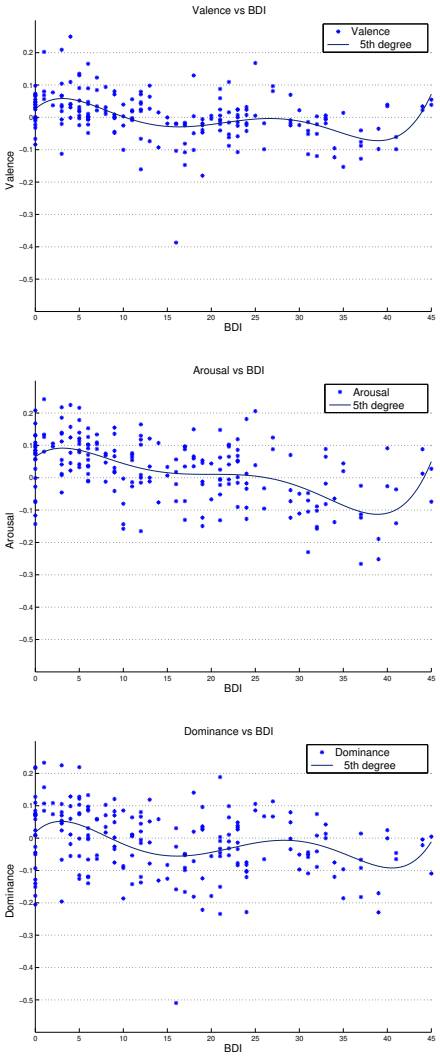

Figure 1: Ground-truth Valence, Arousal and Dominance vs BDI, for each recording

4.1 Audio Features

In this Challenge, as was the case for AVEC 2011-2013, an extended set of features with respect to the INTERSPEECH 2009 Emotion Challenge (384 features) [23] and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) [24] is given to the participants, again using the freely available open-source Emotion and Affect Recognition (openEAR) [8] toolkit's feature extraction backend openSMILE [9]. In contrast to AVEC 2011, the AVEC 2012 feature set was reduced by 100 features that were found to carry very little information, as they were zero or close to zero most of the time. In the AVEC 2013 feature set bugs in the extraction of jitter and shimmer were corrected, the spectral flatness was added to the set of spectral low-level descriptors (LLDs) and the MFCCs 11–16 were included in the set.

Thus, the AVEC 2014 audio baseline feature set consists of 2 268 features , composed of 32 energy and spectral related low-level descriptors (LLD) x 42 functionals, 6 voicing related LLD x 32 functionals, 32 delta coefficients of the energy/spectral LLD x 19 functionals, 6 delta coeffi-

Table 1: Pairwise inter-rater correlation coefficients, measured as Pearson's r across all trace combinations.

Pairs		Arousal		Valence		Dominance		Average	
Rater 1	Rater 2	r	RMSE	r	RMSE	r	RMSE	r	RMSE
A1	A2	0.424	0.170	0.371	0.062	0.260	0.179	0.352	0.137
A1	A3	0.261	0.213	0.362	0.139	0.248	0.303	0.290	0.218
A1	A4	0.442	0.211	0.396	0.153	0.302	0.220	0.380	0.195
A1	A5	0.180	0.198	0.319	0.125	N/A	N/A	0.249	0.161
A2	A3	0.225	0.200	0.262	0.142	0.067	0.326	0.184	0.223
A2	A4	0.342	0.229	0.492	0.148	0.349	0.184	0.394	0.187
A2	A5	0.541	0.157	0.607	0.099	N/A	N/A	0.574	0.128
A3	A4	0.296	0.232	0.397	0.176	0.173	0.352	0.289	0.253
A3	A5	0.151	0.208	0.309	0.158	N/A	N/A	0.230	0.183
A4	A5	0.285	0.181	0.480	0.152	N/A	N/A	0.382	0.166

Table 2: 32 low-level descriptors.

Energy & spectral (32)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, flatness, MFCC 1-16
Voicing related (6)
F_0 (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

cients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features. Details for the LLD and functionals are given in tables 2 and 3 respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition.

The audio features are computed on short episodes of audio data. As the data in the Challenge contains long continuous recordings, a segmentation of the data had to be performed. A set of baseline features is provided for three different versions of segmentation: First, a voice activity detector [7] was applied to obtain a segmentation based on speech activity. Pauses of more than 200 ms are used to split speech activity segments. Functionals are then computed over each detected segment of speech activity. These features can be used both for the emotion and depression tasks. The second segmentation method considers overlapping short fixed length segments (3 seconds) which are shifted forward at a rate of one second. These features are intended for the emotion task. The third method also uses overlapping fixed length segments shifted forward at a rate of one second, however, the windows are 20 seconds long to capture slow changing, long range characteristics. These features are expected to perform best in the depression task.

4.2 Video Features

For AVEC 2014 the local dynamic appearance descriptor LGBP-TOP has been adopted as video features. The imple-

Table 3: Set of all 42 functionals. [1]Not applied to delta coefficient contours. [2]For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]Not applied to voicing related LLD.

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 %–99 %, percentage of frames contour is above: minimum + 25%, 50%, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[1,3], standard deviation of segment length[1,3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude range of minima, amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1 – 5

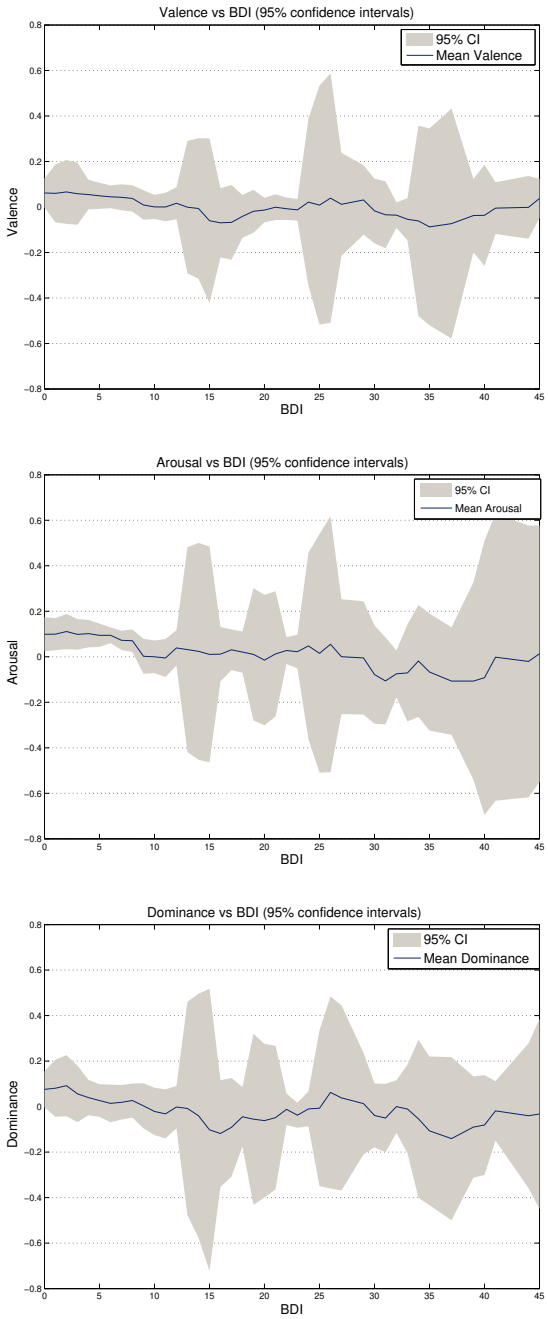

Figure 2: Mean Valence, Arousal and Dominance label per BDI score, shown with 95% Confidence Intervals

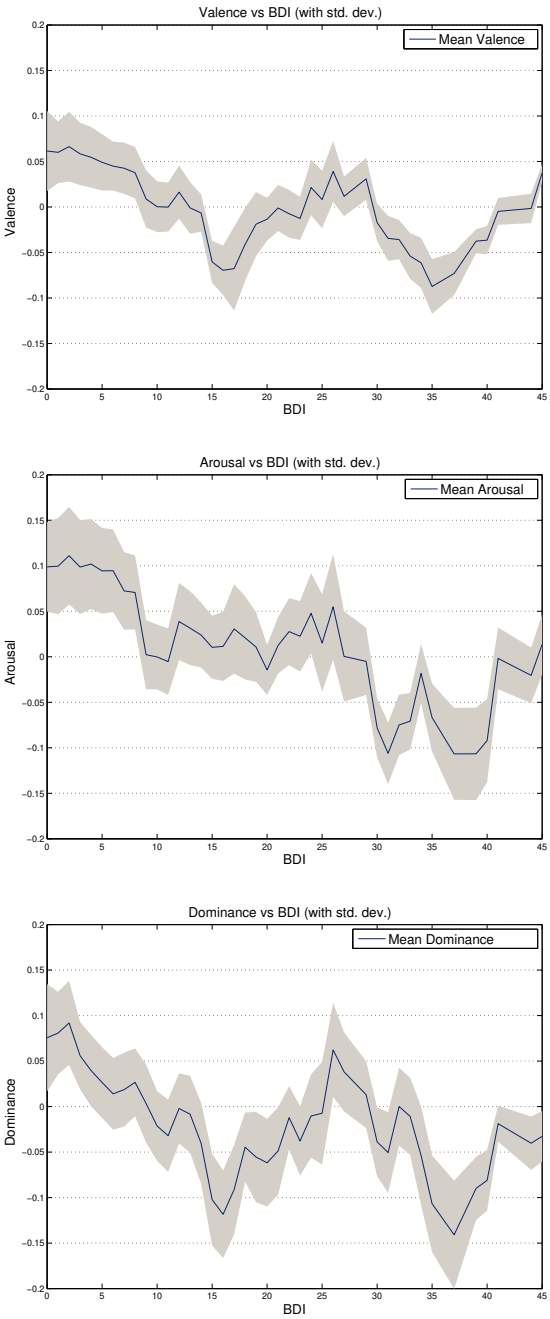

Figure 3: Mean Valence, Arousal and Dominance label per BDI score, shown with standard deviation

mentation is publicly available as part of the eMax face analysis toolbox [1]. LGBP-TOP takes a block of consecutive input video frames which are first convolved with a number of Gabor filters to obtain Gabor magnitude response images for each individual frame. This is followed by LBP feature extraction from the orthogonal XY, XT and YT slices through the set of Gabor magnitude response images. The resulting binary patterns are histogrammed for the three orthogonal slices separately, and concatenated into a single feature histogram (see Fig. 4).

Preprocessing of video frames includes face localisation and segmentation by means of the publicly available Viola & Jones face detector prior to LGBP-TOP feature extraction. Fast and easy to use, it sometimes struggles to correctly detect a face on noisy data such as that used in this challenge. To keep the dimensionality of all feature vectors constant and the number of instances per video consistent with the number of frames, in this paper frames where the face detector failed to locate a face are marked with a feature vector of all zeros.

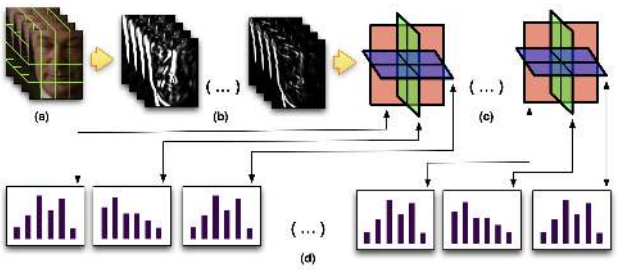

Figure 4: LGBP-TOP feature extraction procedure: a) original block of frames, b) Gabor magnitude responses, c) XY, XT and YT mean slices of each response and d) LBP histograms concatenated into LGBP-TOP histogram

Gabor filtering is a popular filter bank approach, proven to be robust against misalignments and illumination differences for facial expression recognition. A filter is represented with a complex Gabor function composed of a sinusoidal carrier and a Gaussian modulation. A combination of filters of different wavelet parameters applied prior to feature extraction allows to remove unwanted noise and highlight edges valuable for facial expression recognition. In this paper 18 filters with variable orientations and frequencies, but constant amplitude and phase have been used. Each image of the every input block was therefore convolved with 18 different Gabor wavelets, which resulted in 18 Gabor magnitude responses for each of the blocks (Figure 4, b).

For each magnitude response three image planes are then composed, corresponding to XY, XT and YT slices of the response, where T refers to the time axis. Image planes are computed by taking mean of the response pixels values along a target axis (Figure 4, c) in 3-dimensional space. Finally, the LBP operator is applied to the every image plane, resulting in three LBP histograms per Gabor response, which are then concatenated into a single LGBP-TOP histogram across all image planes of all filter responses for the input block (Figure 4, d). Given the number of filters, 54 LBP histograms are composed for every block. Thus, to keep the amount of bins in the resulting histogram minimum, Uniform LBP has been employed instead of the classic LBP, implying 59 bins per histogram versus 256 in the conventional LBP.

Prior to feature extraction, each image is additionally split into 4x4 non-overlapping segments of equal size, each of which is processed independently from the others to maintain some local information captured by the features.

Due to its dynamic nature, LGBP-TOP can only be applied to blocks of frames and not to standalone images. The size of the blocks, typically called temporal window, can vary depending on the desired level of precision, computational cost as well as the framerate of a dataset. In this paper a fixed window of 5 overlapping frames has been used. The challenge however requires a feature vector to be composed for every frame of each video. For this reason, only features extracted from XY image planes have been used in this study thus making it possible to apply the descriptor to arrays of less than 5 images. In case no face is detected for a frame in a given block , a feature histogram is computed for all frames before the failing frame and a new block is started immediately after it.

5. CHALLENGE BASELINES

For transparency and reproducibility, we use standard algorithms. We conducted two separate baselines: one using video features only, and the other using audio-visual features where possible.

For the video modality baseline, an epsilon-SVR with intersection kernel [18] trained using LGBP-TOP features has been employed. In the ASC sub-challenge due to a high number of feature vectors (one per a video frame) the following sample selection has been applied to create the regressor training set for both training and development data partitions: since each feature vector apart from a few exceptions is composed by taking a mean of 5 frames, only every fifth feature vector from the original feature set has been used in the regressors training and testing procedures. For the DSC sub-challenge, where a single label is assigned for a recording, a single mean video feature vector has been taken across all feature vectors in the recording. Note that no additional feature selection and / or parameter optimisation have been applied. In our experiments, epsilon was set to 0.001, and the slack-variable C was set to 1.

In addition, a Pearson product-moment correlation coefficient score of 0.196 was obtained for Dominance, which is similar to the other dimensions, indicating that Dominance can be used equally well.

To put these results in context, we compare our baseline results to the results obtained during AVEC 2013. Those results were obtained on a set of 150 recordings that were almost the same as those used for AVEC 2014. The main differences are that this year's challenge uses only 2 out of 14 tasks per recording, and that annotation of dimensional affect is now the average value taken over a number of raters. The baseline result in 2013 using Video features obtained an ASC PCC score on the test partition of 0.076 for Valence, and 0.134 for Arousal. In contrast, this year we obtained a score of 0.188 for Valence and 0.206 for Arousal. The winners of the AVEC 2013 ASC sub-challenge obtained scores of 0.155 and 0.127 for Valence and Arousal, respectively [19]. A recent paper by Kächele et al. reported scores of 0.150 and 0.170 for Valence and Arousal on the same set [17].

In terms of the DSC sub-challenge, our current baseline obtained a RMSE error of 10.9 on the test set using Video features. This compares to an error of 13.61 for the AVEC 2013 baseline, and 8.50 for the winners of that sub-challenge [28]. The DSC baseline comparison is particularly relevant, as the goal of the task is to obtain a single BDI-II depression level per recording, irrespective of how many tasks were used to obtain this. So, whereas the ASC baselines are less comparable due to being assessed on different sets of tasks, the DSC comparison is a fairer one.

This is a large performance increase, in particular for the ASC baseline. We believe this may be attributed to three causes: firstly, the LGBP-TOP features have been shown before to outperform other descriptors for human behaviour analysis [1, 2]. Secondly, using an average dimensional affect label over multiple subjective ratings should remove some of the subjectivity of the interpretation of the affective behaviour, and remove rater errors caused by cognitive workload effects such as fatigue. In turn, this should lead to an easier machine learning task. Thirdly, the order of tasks in the AVEC 2013 recordings was not always exactly the same, and sometimes subjects skipped tasks entirely. AVEC 2014

Table 4: Baseline results for affect recognition. Performance is measured in Pearson's correlation coefficient averaged over all sequences.

Partition	Modality	Valence	Arousal	Dominance	Average
Development	Audio-Video	—	—	—	—
Development	Video	0.355	0.412	0.319	0.362
Test	Audio-Video	—	—	—	—
Test	Video	0.1879	0.2062	0.1959	0.1966

Table 5: Baseline results for depression recognition. Performance is measured in mean absolute error (MAE) and root mean square error (RMSE) over all sequences.

Partition	Modality	MAE	RMSE
Development	Audio	—	—
Development	Video	—	9.26
Test	Audio	—	—
Test	Video	8.857	10.859

uses only two tasks, and only recordings of which both tasks were completed were included in the data set.

6. CONCLUSION

We introduced AVEC 2014 – the second combined open Audio/Visual Emotion and Depression recognition Challenge. It addresses in two sub-challenges the detection of the affective dimensions arousal, valence and dominance in continuous time and value, and the estimation of a self-reported level of depression. This manuscript describes AVEC 2014's challenge conditions, data, baseline features and results. By intention, we opted to use open-source software and the highest possible transparency and realism for the baselines by refraining from feature space optimisation and optimising on test data. This should improve the reproducibility of the baseline results.

7. REFERENCES

[1] T. Almaev and M. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proc. Affective Computing and Intelligent Interaction*, 2013.

[2] T. R. Almaev, A. Yüce, A. Ghitulescu, and M. F. Valstar. Distribution-based iterative pairwise classification of emotions in the wild using lgbp-top. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 535–542, New York, NY, USA, 2013. ACM.

[3] M. R. Bagby, A. G. Ryder, D. R. Schuller, and M. B. Marshall. The hamilton depression rating scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, 161:2163–2177, 2004.

[4] A. Beck, R. Steer, R. Ball, and W. Ranieri. Comparison of beck depression inventories -ia and -ii in psychiatric outpatients. *Journal of Personality Assessment*, 67(3):588–97, December 1996.

[5] J. F. Cohn, S. Kreuz, I. Matthews, Y. Yang, M. H. Nguyen, M. Tejera Padilla, and et al. Detecting depression from facial actions and vocal prosody. In *Proc. Affective Computing and Intelligent Interaction*, pages 1–7, 2009.

[6] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *Proc. ISCA Workshop on Speech and Emotion*, pages 19–24, Belfast, UK, 2000.

[7] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Proc. of ICASSP, Vancouver, Canada*. IEEE, 2013. to appear.

[8] F. Eyben, M. Wöllmer, and B. Schuller. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII*, pages 576–581, Amsterdam, The Netherlands, 2009.

[9] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, Florence, Italy, 2010.

[10] M. First, R. Spitzer, M. Gibbon, and J. Williams. *Structured Clinical Interview for DSM-IV Axis I Disorders SCID-I: Clinician Version, Administration Booklet*. SCID-I: Clinician Version. American Psychiatric Press, 1997.

[11] J. Fontaine, S. K.R., E. Roesch, and P. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(2):1050 – 1057, 2007.

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

[13] M. Hamilton. Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psycology*, 8:278–296, 1967.

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

[15] Health & Consumer Protection Directorate General. Improving the mental health of the population:

Towards a strategy on mental health for the european union. Technical report, European Union, 2005.

[16] Health & Consumer Protection Directorate General. Mental health in the eu. Technical report, European Union, 2008.

[17] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In M. De Marsico, A. Tabbone, and A. Fred, editors, *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 671–678. SciTePress, 2014.

[18] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

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

[20] M. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic probabilistic cca for analysis of affective behaviour and fusion of continuous annotations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1, 2014.

[21] R. Picard. *Affective Computing*. MIT Press, 1997.

[22] S. Scherer, G. Stratou, J. Gratch, J. Boberg, M. Mahmoud, A. S. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2013.

[23] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 Emotion Challenge. In *Proc. INTERSPEECH 2009*, pages 312–315, Brighton, UK, 2009.

[24] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. INTERSPEECH 2010*, pages 2794–2797, Makuhari, Japan, 2010.

[25] M. Valstar, K. Smith, F. Eyben, S. Schnieder, and R. Cowie. Avec 2013 - the continuous audio / visual emotion and depression recognition challenge. In *Proc. 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.

[26] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'ericco, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Trans. Affective Computing*, 3:69–87, April 2012. Issue 1.

[27] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and et al. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of Neuroscience Methods*, 168(1):224 – 238, 2008.

[28] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 41–48, New York, NY, USA, 2013. ACM.

[29] Y. Yang, C. Fairbairn, and J. Cohn. Detecting depression severity from intra- and interpersonal vocal prosody. *IEEE Transactions on Affective Computing*, 4, 2013.

[30] M. Zimmerman, I. Chelminski, and M. Posternak. A review of studies of the hamilton depression rating scale in healthy controls: Implications for the definition of remission in treatment studies of depression. *Journal of Nervous & Mental Disease*, 192(9):595–601, 2004.