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, arousal and dominance. In addition to structured self-report questionnaires, psychologists and psychiatrists use in their evaluation of a patient's level of depression the observation of facial expressions 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 tasks used in a previous challenge, allowing for more focussed studies. In addition, labels for a third dimension (Dominance) have 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 is to predict the value of a single self-reported severity of 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.

*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 978-1-4503-3119-7/14/11…$15.00.
http://dx.doi.org/10.1145/2661806.2661807 .

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

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 [26] 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 [16] and 2008 [17], 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 people 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 technological 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 creating emotionally aware technology, including automatically analysing affect and expressive behaviour [22]. 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 [28]. 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 [14] is the current gold standard to assess severity [3, 32], whereas the gold-standard for diagnosis is the Structured Clinical Interview for DSM-IV (SCID) [10]. In this challenge the Beck Depression Inventory-II (BDI-II, [4]) is used, which is another frequently used self-report scheme comprised of a 21 multiple choice inventory. 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 [29, 15]. 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. [29, 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 [31] 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. Scherer et al. [23] 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: (i) encode small changes in affect over time, and (ii) 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 [21]. The solution to this problem is to increase the number of raters.

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 BDI-II scores for depression.

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 HCI experiment session.

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 HCI experiment sessions.

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 [27], 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 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 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 the traces of all tasks. We first calculated inter-rater correlations between each rater and the rest by comparing their ratings with that obtained

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

X-vs-All	Arousal		Valence		Dominance		Average	
	r	RMSE	r	RMSE	r	RMSE	r	RMSE
A1	0.508	0.148	0.475	0.094	0.432	0.165	0.445	0.139
A2	0.474	0.159	0.624	0.083	0.422	0.178	0.319	0.146
A3	0.142	0.284	0.460	0.144	0.257	0.321	0.150	0.263
A4	0.505	0.190	0.627	0.150	0.400	0.211	0.474	0.186
A5	0.456	0.159	0.661	0.090	N/A	N/A	0.501	0.132

by averaging the ratings of all other raters. One-vs-rest ICCs are shown in Table 1. Correlations were highest for VALENCE, followed by AROUSAL and then DOMINANCE. This can be explained as VALENCE being the most easily understood concept for naive raters, with clearly associated behaviour. AROUSAL and then DOMINANCE are increasingly less clearly understood. Correlation levels were uniform for all raters except A3. Interestingly, A3 represents the ratings created for AVEC 2013.

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.

Fig. 2 shows for each BDI score (0-45) the overall mean and standard deviation of the affect labels, where the average was taken over each relevant clip.

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 can use these feature sets exclusively or in addition to their own features.

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)

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) [24] and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) [25] 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. AVEC 2014 uses basically the same features as AVEC 2013.

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 coefficients 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

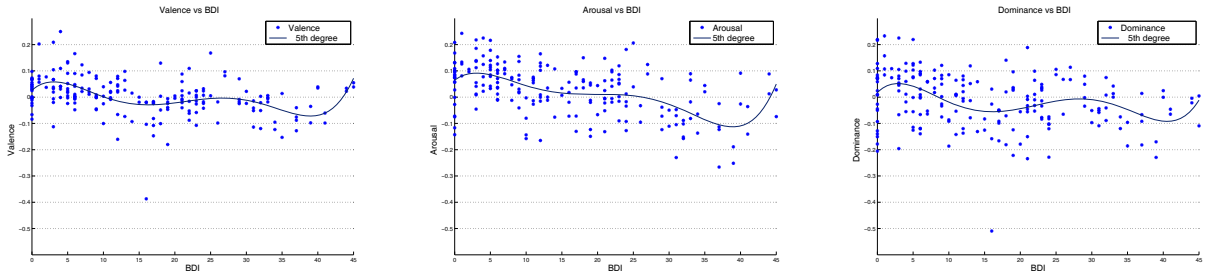

Figure 1: Ground-truth Valence, Arousal and Dominance vs BDI, for each recording

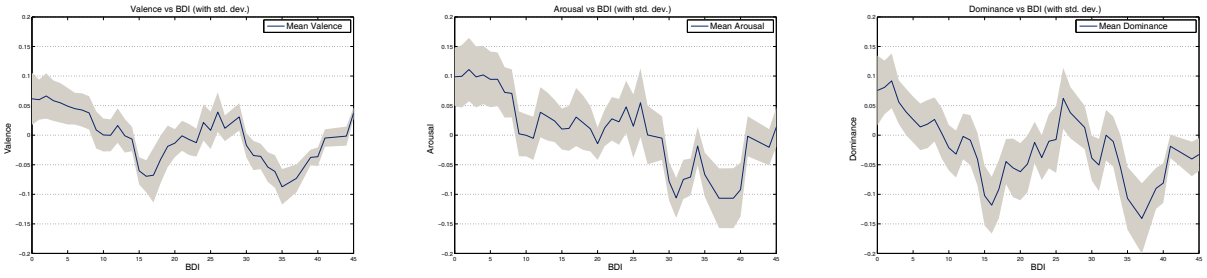

Figure 2: Mean Valence, Arousal and Dominance label per BDI score, shown with standard deviation

Table 3: Set of all 42 functionals. [1]Not applied to delta coefficient contours. [2]Delta coefficients are calculated using only positive values. [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

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 two seconds. 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 implementation is publicly available as part of the publicly available 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. 3).

Preprocessing of video frames includes face localisation and segmentation by means of the 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.

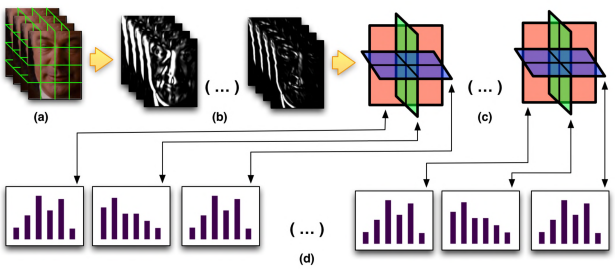

Figure 3: LGBP-TOP feature extraction procedure: a) original block of frames, b) Gabor magnitude responses, c) XY, XT and YT mean slices of each response and d) LBP histograms concatenated into LGBP-TOP histogram

Prior to feature extraction, each image is 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 the 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 subsequent 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. When 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 initiated immediately after it.

5. CHALLENGE BASELINES

For transparency and reproducibility, we use standard algorithms. We conducted three separate baselines: one using video features only, another one using audio features only and finally an audio-visual baseline that simply combines the results of the previous two.

For the video modality baseline, an epsilon-SVR with intersection kernel [19] 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.

SVR with linear kernel and SMO training as implemented in WEKA [13] was used for the audio baseline. For Arousal and Valence, slack-variable C = 0.00005 was found to be the best on the development partition from the set of C = [0.00005, 0.0001 0.0005 0.001 0.005] in combination with short (2s) audio segments. For Dominance, features from voiced segments trained with C = 0.001 found to perform the best on the development set. Task specific models were trained for all affect dimensions, i.e. separate models for Northwind and Freeform. These models were then only applied to the respective task data of the development and the test partition. For depression, the features computed over the long segments (20s) gave the lowest RMSE errors on the development set with C set to 0.005. WEKA's SMO implementation using a linear kernel was used. The Northwind data was ignored for the depression classification as it was found that using only the Freeform data attained superior performance. When no audio features were available baseline predictions were replaced with a mean value across all recordings for DSC challenge and zeros for ASC challenge.

In order to create the audio-visual baseline, predictions from both audio and video baselines were obtained by taking the mean value for audio and video. Thus for DSC challenge a single audio-visual prediction was created for every recording by taking mean of corresponding values from audio and video baselines. For ASC challenge the same procedure has been applied to predictions of every frame of each recording. When an audio prediction wasn't available (due to the absence of speech), only the video prediction was used. Note that no additional audio-visual feature sets were created and no dedicated classifiers trained for the audio-visual baseline.

Baseline scores for ASC are shown in Table 4, and DSC scores in Table 5. 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 0.355 for Valence and 0.540 for Arousal using the audio-only baseline. The winners of the AVEC 2013 ASC sub-challenge obtained scores of 0.155 and 0.127 for Valence and Arousal, respectively [20]. A recent paper by Kächele et al. reported scores of 0.150 and 0.170 for Valence and Arousal on the same set [18]. In addition, a Pearson product-moment correlation coefficient score of 0.360 was obtained for Dominance, which is similar to the other dimensions, indicating that Dominance can be used equally well.

In terms of the DSC sub-challenge, our audio-visual baseline obtained a RMSE error of 9.891 on the test set. This compares to an error of 13.61 for the AVEC 2013 baseline, and 8.50 for the winners of that sub-challenge [30]. 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

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	0.347	0.517	0.439	0.434
Development	Video	0.355	0.412	0.319	0.362
Development	Audio-Video	0.236	0.421	0.348	0.335
Test	Audio	0.355	0.540	0.360	0.419
Test	Video	0.188	0.206	0.196	0.197
Test	Audio-Video	0.282	0.478	0.324	0.361

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	8.934	11.521
Development	Video	7.577	9.314
Development	Audio-Video	6.680	8.341
Test	Audio	10.036	12.567
Test	Video	8.857	10.859
Test	Audio-Video	7.893	9.891

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

Given the simplistic fusion of audio and video results, it is not surprising that the ASC results do not benefit from an audio-visual approach. No temporal smoothness constraints are enforced and the correlation measure will suffer from simply taking the average value of the two predictions. On the other hand, for the DSC where only a single prediction has to be made per pair of tasks, taking the average prediction value *is* a suitable approach, as is clear from the results shown in Table 5.

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.

Acknowledgments

The work of Michel Valstar is partly funded by the NIHR-HTC 'MindTech' and Horizon Digital Economy Research, RCUK grant EP/G065802/1.The work of Jarek Krajewski is partly funded by the German Research Foundation (KR3698/4-1).The challenge in general has been generously supported by the Association for the Advancement of Affective Computing (AAAC, former HUMAINE association) and the EU network of excellence on Social Signal Processing SSPNet (EC's 7th Framework Programme [FP7/20072013] under grant agreement no. 231287). The authors further acknowledge funding from the EC and ERC (grants nos. 289021, ASC-Inclusion and 338164, iHEARu). Funding for preparatory work has been provided in part by the EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

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 *Proc. Int'l Conf. 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. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

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

[15] 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.

[16] 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.

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

[18] 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, *Proc. Int'l Conf. Pattern Recognition Applications and Methods (ICPRAM)*, pages 671–678. SciTePress, 2014.

[19] 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.

[20] 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 *Proc. Int'l Workshop Audio/Visual Emotion Challenge*, AVEC '13, pages 21–30, New York, NY, USA, 2013. ACM.

[21] 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.

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

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

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

[25] 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.

[26] M. Valstar. Automatic behaviour understanding in medicine. In *Proceedings ACM Int'l Conf. Multimodal Interaction*, 2014.

[27] 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 *Proc. 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.

[28] 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.

[29] 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.

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

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

[32] 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.