AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge [*]

Michel Valstar
University of Nottingham
Mixed Reality Lab

Björn Schuller
Technische Universität München
Institute for Human-Machine Communication

Kirsty Smith
University of Nottingham
Mixed Reality Lab

Florian Eyben
Technische Universität München
Institute for Human-Machine Communication

Bihan Jiang
Imperial College London
Intelligent Behaviour Understanding Group

Sanjay Bilakhia
Imperial College London
Intelligent Behaviour Understanding Group

Sebastian Schnieder
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, psychologists and psychiatrists base their evaluation of a patient's condition to a large extent on the observation of facial expressive and vocal cues, such as dampened facial expressive responses, avoiding eye contact, and using short sentences with flat intonation. It is in this context that we present the third Audio-Visual Emotion recognition Challenge (AVEC 2013). The challenge has two goals logically organised as sub-challenges: the first is to predict the continuous values of the affective dimensions valence and arousal at each moment in time. The second sub-challenge is to predict the value of a single 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

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

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

1. INTRODUCTION

According to EU Green Papers dating from 2005 [?] and 2008 [?], mental health problems affect one in four citizens at some point during their lives and too often lead to suicide. 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 innovation has occurred in the clinical care of mental health disorders.

Affective Computing and Social Signal Processing are two of the more recent technological revolutions that promise to change this. Affective Computing is the science of automatically analysing affect and expressive behaviour [?]. By their very definition, mood disorders are directly related to affective state. Social Signal Processing addresses all verbal and non-verbal communicative signalling that goes on during social interactions, be they of an affective nature or not [?]. And although the measurement and assessment of behaviour is a central component of mental health practice

[*] This is a pre-publication, and the content may change until the camera ready deadline of AVEC 2013

[†] The author is further affiliated with Twente University, EEMCS, Twente, The Netherlands.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM-Multimedia 2013 Barcelona, Spain

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

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 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 [?, ?]. Besides diagnosis, affective computing and social signal processing would also allow quantitatively monitor the progress and effectiveness of treatment. Preliminary studies into this have already been published on the topic of depression and autism [?, ?].

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 for contemporary computational modelling techniques 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 2013 Audio-Visual Emotion Challenge and Workshop (AVEC 2013) will be the third competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, video and audiovisual emotion analysis, with all participants competing under strictly the same conditions. The goal of the AVEC Challenges series is to provide a common benchmark test set for individual multimedia processing and to bring together the audio and video emotion recognition communities, to compare the relative merits of the two approaches to emotion recognition under well-defined and strictly comparable conditions and establish to what extent fusion of the approaches is possible and beneficial. In addition, AVEC 2013 has the goal to accelerate the development of behavio-medical technologies that can aid the mental health profession in their aim to help people with mood disorders.

Following up from AVEC 2011 [?] and AVEC 2012 [?], which respectively used a categorical description of affect and automatic continuous affect recognition from audio and video on the SEMAINE database of natural dyadic interactions [?], we now aim to extend the analysis of affective behaviour to infer a more complex mental state, to wit, depression. Both the dimensional affect and the depression recognition problems are posed as a regression problem, and can thus be considered to be both challenging and rewarding. A major difference between this AVEC and the previous two is that the first two had as task making only very short-term predictions (either for every video frame or per spoken word), whereas AVEC 2013 extends this to include event detection in the form of inferring a measure of depression for every recording.

Different from the continuous dimensional affect prediction, event-based recognition provides a single label over some 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 or arousal, 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 experiment 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 two target affect dimensions arousal and valence, and Beck Depression Index (BDI), a self-reported 21 multiple choice inventory [?]. [MFV: WHICH BDI DO WE USE? I'M ASSUMING BDI-II HERE, AS IT'S THE LATEST (1996)]

Two Sub-Challenges are addressed in AVEC 2013:

- The *Affect Recognition Sub-Challenge (ASC)* involves fully continuous affect recognition of the dimensions valence and arousal (VA), 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, two regression problems need to be solved for Challenge participation: prediction of the continuous dimensions AROUSAL and VALENCE. The ASC competition measure is cross correlation averaged over all sessions and both dimensions. For the DSC, a single regression problem needs to be solved. The DSC competition measure is root mean square error over all sessions.

Both Sub-Challenges allow contributors to find their own features to use with their regression algorithm. However, 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 an accepted paper will be eligible for Challenge participation. The organisers preserve 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 audio-visual depressive language corpus (AVDLC), which includes 340 video clips of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone. There is only one person per clip and the total number of subjects is 292, i.e. some subjects feature in more than one clip.

The speakers were recorded between one and four times, with a period of two weeks between the measurements. Five subjects appears in four recordings, 93 in 3, 66 in 2, and 128 in only one sessions. The length of the clips 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 tasks which were Power Point guided: i.e., sustained vowel phonation, sustained loud vowel phonation, and sustained smiling vowel phonation; speaking out loud while solving a task; Counting from 1 to 10; Read speech: excerpts of the novel "Homo" Faber by Max Frisch and the fable "Die Sonne und der Wind" (The North Wind and the Sun); singing: a German nursery rhyme "Guten Abend, gute Nacht" and "Aber bitte mit Sahne" by Udo Jörgens; telling a story from the subject's own past: best present ever and sad event in the childhood; Telling an imagined story applying the Thematic Apperception Test (TAT), containing e.g. pictures of a women and a man in a bed, a housewife and children who are trying to reach the cookies.

The audio was recorded using a wearable USB microphone connected to a computer, at a sampling rate of 41 KHz, 16 bit. 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 × 480 pixels, with 24 bits per pixels. The codec used was H.264, and the videos were embedded in an mp4 container.

[MFV: KIRSTY, WE NEED STATISTICS ON THE DATABASE]

For the organisation of the challenge, the recordings were split into three partitions: a training, development, and test set of 50 recordings each. 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://avec-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 AROUSAL and VALENCE, which form a well-established basis for emotion analysis in the psychological literature [?].

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. 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?

A team of 9 naive raters annotated all human-computer interactions. The raters annotated the two dimensions in continuous time and continuous value using a tool called [MFV: SANJAY, NEED INPUT HERE], and the annotations are often called *traces* after the early popular system that performed a similar function called FeelTrace [?]. Every video was annotated by only a single rater for every dimension, due to time constraints. The annotation process resulted in a set of trace vectors $\{\mathbf{v}_i^a, \mathbf{v}_i^v\} \in \mathbb{R}$ for every rater i and dimension a (AROUSAL) and v (VALENCE). The original traces are binned in temporal units of the same duration as a single video frame (i. e., 1/30 seconds). The labels for AROUSAL, and VALENCE lie in the range $[-1, 1]$.

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, [?]). BDI-II contains 21 questions, where each is a forced-choice question scored on a scale value of 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: 0–13: indicates minimal depression, 14–19: indicates mild depression, 20–28: indicates moderate depression, 29–63: indicates severe depression.

The average BDI-level in the AVEC 2013 partitions was ?? for the training partition and ?? for the test partition (standard deviations = ?? and ??, respectively). [MFV: KIRSTY, WE NEED SOME STATISTICS]

For every recording in the training and development partitions a separate file with a single value is provided for the SDC, together with two files containing the affective dimension labels.

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.

4.1 Video Features

The bulk of the features extracted from the video streams of the character interactions are dense local appearance descriptions. The descriptors that generate these features are most effective if they are applied to frontal faces of uniform size. Since the head pose and distance to the camera vary over time in the DEPRESSION recordings, we detect the locations of the eyes to help reduce this variance. The information describing the position and pose of the face and eyes are in themselves valuable for recognising the dimensional affect and are thus included with the set of video features together with the appearance descriptors.

To obtain the face position, we employ the open-source implementation of the Viola & Jones face detector that is included in OpenCV. This returns a four-valued descriptor of the face position and size. To wit, it provides the position of the top-left corner of the detected face area (f_x, f_y), followed by its width f_w and height f_h. The height and width output of this detector is rather unstable: Even in a video in which a face hardly moves the values for the height and width vary significantly (approximately 5 % standard deviation). The face detector also doesn't provide any information about the head pose.

To refine the detected face region, and allow the appearance descriptor to correlate better with the shown expression rather than variations in head pose and face detector output, we proceed with detection of the locations of the eyes. This is again done with the OpenCV implementation of a Haar-cascade object detector, trained for either a left or a right eye. Let us define the detected left and right eye locations as p_l respectively p_r, and the line connecting p_l and p_r as l_e. The angle between l_e and the horizontal is

Table 1: 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)

then defined as α. The registered image is now obtained by rotating it so that $\alpha = 0$ degrees, then scaled to make the distance between the eye locations $||p_l - p_r|| = 100$ pixels, and finally cropped to be 200 by 200 pixels, with p_r at position $\{p_r^x, p_r^y\} = \{80, 60\}$ to obtain the registered face image. The eye locations are included as part of the video features provided for candidates.

In AVEC 2011 and 2012, uniform Local Binary Patterns [?] were used as dense local appearance descriptors. For AVEC 2013, we chose instead to use Local Phase Quantisation (LPQ) as that was found to attain higher performance in facial expression recognition tasks [?]. The dynamic appearance descriptor LPQ-TOP was found to be even more accurate, but that descriptor depends on a near-perfect alignment of faces in subsequent frames, which is not possible in a near-real time automatic fashion on the DEPRESSION dataset.

LPQs have been used extensively for face analysis in recent years, e. g., for face recognition [?], emotion detection [?], or detection of facial muscle actions (FACS Action Units) [?]. [BIHAN PLEASE ADD LPQ DESCRIPTION HERE]

4.2 Audio Features

In this Challenge, as was the case for AVEC 2012 and AVEC 2011, an extended set of features with respect to the INTERSPEECH 2009 Emotion Challenge (384 features) [?] and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) [?] is given to the participants, again using the freely available open-source Emotion and Affect Recognition (openEAR) [?] toolkit's feature extraction backend openSMILE [?]. 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 2013 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 1 and 2 respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition.

Table 2: 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

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. For three different versions of segmentation a set of baseline features is provided: First, a voice activity detector [?] 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 capute slow changing, long range characteristics. These features are intended for the depression task.

5. CHALLENGE BASELINES

For transparency and reproducibility, we use

[MFV: ADD NICE TABLE/FIGURE FOR DSC]

6. CONCLUSION

We introduced AVEC 2013 – the first combined open Audio/Visual Emotion and Depression recognition Challenge. It addresses in two sub-challenges the detection of dimensional affect in continuous time and value, and the estima-

Table 3: Baseline results. Performance is measured in cross-correlation averaged over all sequences.

tion of self-reported depression. This manuscript describes AVEC 2013'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'.