AMHUSE: a multimodal dataset for HUmour SEnsing

Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Raffaella Lanzarotti

AMHUSE: A Multimodal dataset for HUmour SEnsing

Giuseppe Boccignone
PHuSe Lab - Dipartimento di Informatica
Università degli Studi di Milano
Milano, Italy
giuseppe.boccignone@unimi.it

Donatello Conte
Université de Tours
Tours, France
donatello.conte@univ-tours.fr

Vittorio Cuculo
PHuSe Lab - Dipartimento di Informatica
and Dipartimento di Matematica
Università degli Studi di Milano
Milano, Italy
vittorio.cuculo@unimi.it

Raffaella Lanzarotti
PHuSe Lab - Dipartimento di Informatica
Università degli Studi di Milano
Milano, Italy
lanzarotti@di.unimi.it

ABSTRACT

We present AMHUSE (A Multimodal dataset for HUmour SEnsing) along with a novel web-based annotation tool named DANTE (Dimensional ANnotation Tool for Emotions). The dataset is the result of an experiment concerning amusement elicitation, involving 36 subjects in order to record the reactions in presence of 3 amusing and 1 neutral video stimuli. Gathered data include RGB video and depth sequences along with physiological responses (electrodermal activity, blood volume pulse, temperature). The videos were later annotated by 4 experts in terms of valence and arousal continuous dimensions. Both the dataset and the annotation tool are made publicly available for research purposes.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Graphical user interfaces*;

KEYWORDS

Emotion recognition; Affective computing; Dataset; Multimodal dataset; Amusement evaluation; Annotation tool

ACM Reference Format:
Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2021. AMHUSE: A Multimodal dataset for HUmour SEnsing. In . ACM, New York, NY, USA, 8 pages. https://doi.org/

1 INTRODUCTION

Emotions lie behind human-human and human-computer interaction [34, 38, 42] and the resulting behavioural landscape is complex and hardly decomposable into its affect and cognition dimensions [37]. However, to bring affective computing [38] into the interaction game, the issue of gathering good affective data is crucial for learning models and subsequent benchmarking [39].

Early research in affective computing was mostly focused on facial expression recognition and, by and large, datasets were built on a single modality, namely images or videos [6, 47]. To truly capture the multidimensionality of emotion, in the last fifteen years the number of public repositories has grown larger, where behavioral data gathered in realistic and natural setting experiments have been recorded by multiple modalities [6, 12, 20, 32, 39, 41].

In such perspective, the main contributions of this study can be summarised as follows. First, we present a novel public multimodal dataset focusing on a positive emotion, namely amusement, which has not been covered in other datasets; signals have been gathered from multiple signal sources: RGB image sequences; depth sequences from a Kinect 3D scanner system; physiological signals, namely, temperature, electrodermal activity and blood volume pulse.

Second, we provide DANTE (Dimensional ANnotation Tool for Emotions), a simple and efficient web-based tool for annotating emotions in the arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant) dimensions [46].

Data and annotation tool are freely available at http://phuselab.di.unimi.it/resources.php.

In this study, we use the term "amusement" for referring to the affective state evoked by humorous material [21]. The rationale behind the choice of amusement is that a large body of research has investigated negative emotions, but, surprisingly enough, much less is known about positive emotions, markedly their autonomic response. As Shiota et al. [49] put it , "*this lack of attention is consistent with the historical underrepresentation of positive emotions in psychological research, and with the still common perception among theorists that positive emotions have fewer implications for evolutionary fitness, are less differentiated, and have less distinct impact on motivation and behaviour than is true of the negative emotions.*" However, positive emotions have broad implications for cognition, physiology, behaviour, and more generally for human well-being [17, 21, 49, 51]. Further, they serve important social interactions, facilitating approach behaviour, motivating social engagement, promoting new social connections, and reversing the physiological activation caused by negative emotions. They represent a hard task, since inducing more subtle autonomic responses, as opposed to negative emotions that are usually characterized by greater physiological activation. Apart from these theoretical points, positive emotions are likely to play a key role for a number of human-computer applications (e.g, videogames [22], HCI [9]), gaining a great interest for a range of fields, notably the gaming industry and advertisement.

To the best of our knowledge, this is the first multimodal dataset focusing on this specific emotion.

Note that the dataset is the result of an experiment concerning amusement elicitation using italian spoken video clips. A previous effort adopting italian language was described in [15], though in a different setting and without considering physiological signals. Indeed, multimodal datasets, such as the one presented here, allow to explore human emotional experience more in depth, but leave some problems open. Among them, cogently when analysing the continuous dimensions of emotion, there is the necessity of continuous annotations. As summarized in [47]: "*Labelling data is a challenging and laborious task, particularly for spontaneously displayed expressions and emotions*." Moreover, combining multiple annotations is even harder [30]. In this respect the DANTE tool is friendly and intuitive, allowing the user to seamlessly annotate valence and arousal separately, during the video reproduction of either the stimuli and the subject's recorded behaviour.

2 RELATED WORKS

In the last decade, several benchmarks for emotions detection have been presented. Yet, new datasets are proposed to cope with novel challenges: multiple data sources, spontaneous facial expressions, continuous emotional space, and so on.

The Kanade *et al.* dataset was put forward in 2000 [23] and subsequently extended in 2010 [28]. It is widely known with the acronym CK (CK+ for the extended version). Its main characteristic is the annotation of images by means of FACS [14]. CK+ includes 593 images sequences from 123 subjects. The image sequences vary in duration and incorporate the onset to peak formation of the facial expressions. Full FACS coding of peak frames is provided. These frames are also labeled with subject's impression of each of the 7 basic emotion categories: Anger, Contempt, Disgust, Fear, Happy, Sadness and Surprise. Therefore, this benchmark is used for testing emotion recognition methods measuring their accuracies. The main problem of this dataset is that the images contain unnatural and exaggerated expressions.

A first step towards a more realistic dataset was taken by Pantic et al. [36] in 2005 with the MMI benchmark. It is composed by 2900 videos for 79 subjects. As well as CK+, MMI is annotated in terms of Facial Action Coding System by FACS experts. MMI shows more naturalistic view of faces and subjects are captured in frontal and in profile views. Similarly to CK, images are annotated with emotion labels, and algorithms on this dataset are evaluated with respect to detection accuracy.

In CK+ the actors were expected to mime a requested affective expression, while in MMI the expression was triggered by asking the subjects to watch a video that was supposed to activate the desired emotion. Research in psychology and social sciences has shown that presence or absence of motor mimicry behaviour can serve as an indicator for emotion inference [1, 53]. Thus, Bilakhia et al. [3] proposed a database (MAHNOB mimicry database) suitable for investigation of mimicry and negotiation behaviour. The dataset consists of 54 recordings of face-to-face interactions. It is a multimodal dataset capturing audio and visual data, from different points of view. The data have been fully annotated for 15 out of 54 sessions, in terms of gestures (hand gestures, head movements, etc.). The evaluation of this database is done in terms of cross-correlation between the annotated and the detected gestures.

All videos of the previous benchmarks were produced in a "lab-controlled" recording environment. In [11] the authors proposed a new facial expression database (AFEW, Acted Facial Expressions in the Wild) consisting in clips of videos obtained from movies (SFEW, Dhall et al. [10], is the static subset of AFEW). By extracting data from public movies, the authors were able to annotate them with dense information about the subjects (theme of the scene, emotion of the actors, information about the context, etc.).

OPEN EmoRec II [45] is an open multimodal corpus as the result of an HCI-experiment, concerning the solving of 6 sequences of a mental trainer designed to induce different emotions. The corpus contains sensory signal of video, audio, physiology (SCL, respiration, BVP, EMG Corrugator supercilii and Zygomaticus Major) and facial reaction annotations.

Soleymani et al. [50] proposed a database, named MAHNOB-HCI, of multimodal recordings of participant responses to affectively stimulating movie excerpts, images and videos. The recordings of this database are precisely synchronized and multimodality allows to study the simultaneous emotional responses using different channels. The videos are annotated by each participant, by means of a self annotation form. In particular annotations concern an emotional label (neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear), and a measure (on a nine point scale) for arousal, valence, dominance, and predictability ([16]).

Nowadays, we are moving more and more towards a continuous emotional space [16] and the measure of emotion is often shaped in terms of arousal and valence. The Audio Visual Emotion Recognition Challenge (AVEC [43]) has been proposed, focusing on affect analysis as a regression problem. The RECOLA database deployed for this challenge [44] addresses multimodal data recordings in the context of remote collaborative work. Emotion has to be detected in terms of continuous time and continuous valued dimensional affect in the two dimensions of arousal and valence; the Concordance Correlation Coefficient (CCC) [27] has been chosen as evaluation measure.

In the same vein, our multimodal AMHUSE dataset is conceived following such research trend, but focusing on the positive emotion of amusement. We also cope with the labelling problem, by exploiting DANTE to produce video annotations continuously and not on a frame-by-frame basis. The goal is to provide a suitable tool for continuous affect state modelling and benchmarking.

3 PROPOSED DATASET

Participants. AMHUSE collects the data of 36 different subjects, who agreed with the scientific use of the recorded material. The participants were 9 females and 27 males, with an age varying from 18 to 54 years old ($\mu = 26.7$ and $\sigma = 8.8$). All participants were Italian speaking and did not receive any monetary contribution for the experiment.

Procedure. The experiment was conducted on different days, with the same acquisition protocol and in the same room, namely the PHuSe Laboratory at the department of Computer Science of the Università degli Studi di Milano, Italy. Each participant was welcomed and did receive instructions on the experiment without

Dataset	Subjects	Audio	RGB	Depth	EEG	ECG	BVP	EDA	EMG	R	T	G	FP	Pose	AU
MAHNOB-HCI [50]	30	✓	✓	-	✓	✓	-	✓	-	✓	✓	✓	-	-	-
RECOLA [44]	46(18)	✓	✓	-	-	✓	-	✓	-	-	-	-	-	✓	✓
OPEN_EmoRec_II [45]	30	✓	✓	-	-	-	✓	✓	✓	✓	-	-	-	-	-
AMHUSE	36	-	✓	✓	-	-	✓	✓	-	-	✓	-	✓	-	✓

Table 1: Recorded signals in each dataset. In brackets the number of complete data. EEG = electroencephalography, ECG = electrocardiogram, BVP = blood volume pulse, EDA = electrodermal activity, R = respiration, T = temperature, G = eye gaze tracking. Extracted visual features in each dataset. FP = Fiducial Points, AU = Action Units

Dataset	Annotator	Type	Emotion space
MAHNOB-HCI [50]	S	D	PAD + 9 tags
RECOLA [44]	E(6) + S	C	VA + 5 tags
OPEN_EmoRec_II [45]	E(4) + S	D	VA + 6 tags
AMHUSE	E(4) + S	C	VA

Table 2: Emotional annotations provided in each corpus. S = Self report, E = External. In brackets the number of external annotators. C = Continuous, D = Discrete. PAD = Pleasure, Arousal, Dominance, VA = Valence, Arousal

specific details on the purpose of the study, in order to avoid biased reactions. In particular, they were told that they would have to watch four videos, for the total duration of about 10 minutes. At the end of each video, they were asked to annotate the emotional state they felt when watching the movie. Each participant was asked to sit in front of a screen equipped with stereo speakers and with their backs toward a neutral background, as shown on the left of Fig. 1.

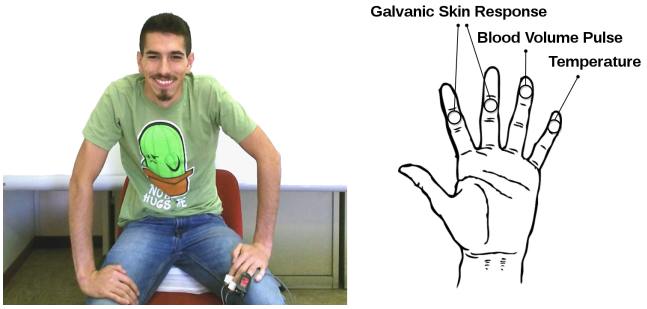

Figure 1: Experimental setup (left) and detail of the physiological sensors used during the acquisition process (right).

Physiological data were acquired via e-Health Sensor Platform together with an Arduino UNO, connected to three wired sensors worn by participants. Sensors were placed all on the left hand, leaving the option to annotate the videos with the right hand. Specifically the sensors are: 1) a body temperature sensor, placed on the little finger; 2) a Galvanic Skin Response sensor (GSR), recording the electrical resistance between the medial phalanges of the middle and index fingers; 3) a pulse oximeter sensor, placed on the ring finger. A detail of such setting is shown in Fig. 1, on the right. We encountered artefacts, especially for the GSR signal, in some subjects affected by “cold hands”: hence, we waited for the subject’s adaptation. A few subjects were excluded since post-analysis revealed artefacts due to excessive movement or adjustment of the device.

In addition to the physiological sensors, a Microsoft Kinect v2 camera, was placed on top of the screen to record the participant reactions in terms of color and depth streams, together with facial landmarks both in two and three dimensional spaces. Once inserted anonymous information such as sex, age and nationality, the stimulus video started playing together with the data recording.

Stimuli. The stimuli (video clips) have been chosen being aware that people have different sense of humour. Thus, aiming at stimulating amusement in each subject at least once, each clip features a different kind of comicality. All videos were in Italian language, all participants were Italian speaking, and no one had never seen the stimuli before.

The four videos chosen as stimuli and shown consecutively are:

(1) A fragment of a documentary[1] made by an Italian TV program on the topic of Hammond organ instruments. This video was chosen to induce in the subjects a neutral state and to create a baseline for the following measurements (duration: 56 seconds).
(2) A fake movie trailer[2] by the Italian comedian Marcello Macchia. This video is a parody of "The Sixth Sense" drama movie. Marcello Macchia’s parodies are always very biting and are mostly appreciated among a younger audience. For this reason, the target of this clip are people between 18 and 30 years old (duration: 87 seconds).
(3) A satirical gag of Maurizio Crozza[3], one of the most appreciated comedians in Italy, specialized in political satire. The target of this video are people who enjoy political satire (duration: 25 seconds).
(4) A snippet of a sketch of Aldo, Giovanni & Giacomo[4]. This trio of comedians exploits a quite classical type of comicality, and they are appreciated by people of all ages. For this reason, the target of this video is the widest of all three (duration: 117 seconds).

Gathered data. As mentioned earlier, several spontaneous responses of the participants were gathered during the experiment. One is their facial expressions. For this purpose we adopted a low cost Microsoft Kinect Sensor for Xbox One, equipped with the Kinect Adapter for Windows. This sensor presents a color camera with a full HD resolution of 1920×1080 pixels, a frequency of 30 fps

[1] https://youtu.be/-CZWXgPFj6A (from 1:17 to 2:13).
[2] https://youtu.be/Nwc1kRJdtYw (full length).
[3] https://youtu.be/C7xZYfxC8k4 (from 0:29 to end).
[4] https://youtu.be/tb_2TJjsq4g (from 4:53 to 6:50).

and a FOV of 84.1×53.8 degrees resulting in an average of about 22×20 pixels per degree (Fig 2a). The depth camera, instead, is able to record images with a resolution of 512×424 pixels, a frequency of 30 fps and a field of view (FOV) of 70.6×60 degrees, resulting in about 7×7 pixels per degree. Depth frames are stored in binary format, each value is a 16-bit unsigned integer (distance in mm). Nevertheless, also a grayscale version of each frame is provided (cfr. Fig. 2b).

Even if the camera captures only the frontal view of the participant, the depth stream permits the extraction of 3D head and body movements. Moreover, the wide angle lens of the color camera captures part of the body, arms and hands, which also carry important cues about the affective state of the participant. Furthermore, the Microsoft Kinect SDK extracts in real-time and for each frame a pool of 1347 face points. Such points rely on the camera's 3D space, as shown in Fig. 2c, resulting in a very dense mesh of participant's face. The landmarks are translated in color and depth camera space, making it easy to work with such spaces. As to data compression, we aggregated all the color frames in a video for each session, varying the frame rate and the video duration to pair the stimulus within an AVI container using MPEG-2 codec. The videos were cropped based on the white background, in order to give more focus on subject's reactions, eventually resulting in a resolution of 1024×768 pixels. We also extracted Action Units (AU) activations, relying on a freely available AU detector [2]. The AU detector provides, at each frame, the activation level of the following AUs $AU_k, k = 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 45$, plus the binary presence of $AU28$.

Along with the visible cues, a set of physiological signals was acquired during the experiment. Electrodermal activity (EDA) was measured via a Galvanic Skin Response (GSR) sensor, gauging conductance between two points. The level of conductivity is provided in terms of microSiemens (μS) and Volts (V), as well as its reciprocal resistance, expressed in Ohms (Ω). Put simple, a GSR sensor measures the electrical skin resistance in presence of sweat produced by the body: when a high condition of sweating occurs, the electrical skin resistance drops down. Emotions with a prominent presence of positive or negative arousal, such as excitement, stress or fear can induce fluctuations of skin conductivity [26, 31]. Skin temperature (in Celsius degrees) was also recorded since it changes in different emotional states [40]. Eventually, a blood volume pulse (BVP) sensor was used as a non invasive mean to obtain an indirect measure of the heart rate, via the arterial oxygen saturation of hemoglobin. This signal is strictly correlated with the heart rate measured via ECG and increases in presence of pleasant stimuli [48]. Heart rate values are provided as the number of contractions of the heart per minute (bpm). For completeness, values of blood oxygenation are also included in the dataset, though there is not a direct relationship between this kind of signal and emotional states.

All the physiological signals (Fig. 3) were recorded at 40 Hz during each session, and provided in CSV format. Acquisition software was developed adopting MATLAB Parallel Computing Toolbox. Each *worker* is responsible for a task: video playback, Kinect recording and physio recording. All *workers* are kept in sync and share messages by relying on the MATLAB built-in synchronization mechanism.

Annotations. At the end of each stimulus, all subjects were asked to evaluate their levels of pleasure, arousal and dominance (controlling vs controlled feeling, e.g. anger vs. fear). The annotations were made by using the AffectButton [5]. The latter was chosen for its intuitive usage: an emoticon representation of emotion, as shown in Fig. 4, where no prior training is required. The Affect Button measures the emotional states in three dimensions: pleasure, arousal and dominance, following the PAD emotional state model [29]. To employ AffectButton, the user moves the mouse cursor over the button and the emoticon-like face changes expression accordingly. Every position over the button maps to a corresponding point in the PAD space, with values ranging from -1 to 1, for each dimension. The user clicks when the face expression corresponds to his/her emotional state and the chosen value is saved in the dataset.

Along with self reports, a team of four annotators (3 males and 1 female, with previous experience) was engaged to annotate in the continuous valence/arousal space the reactions of all subjects to each stimulus. The annotators first performed the annotation of a couple of sequences to become familiar with the annotation interface, then started with the real data, observing the participant expressions only. Given the amount of data to be annotated, the annotation process was performed remotely via web browser, in order to split the workload in time. The subject videos were annotated along the arousal and valence dimensions, separately and time-continuously by using the annotation tool 'DANTE', developed for this purpose and presented in the next section. The frequency of annotations is 25 Hz, provided in CSV format, with values ranging from -1 to 1 and a step of 0.001.

DANTE - Dimensional ANnotation Tool for Emotions. Several emotional annotation tools were taken into account and tested. Some of these, such as 'Feeltrace' [13] and its successor 'Gtrace' were excluded since the beginning because they rely on software intended to be installed locally and do not allow remote annotations. Aiming at web-based annotation, we considered the 'Valence/Arousal Online Annotation Tool' released together with the AFEW-VA dataset [24] and the 'ANNEMO' [44] annotation tool.

The first requires people to annotate video clips frame by frame, providing continuous annotations for valence and arousal within the range $[-10, 10]$. We find that this tool could be adopted exclusively for very short videos, as occurs in the dataset AFEW-VA (average length of 50 frames). Moreover, we noticed that the process of *per-frame* annotation, even if it overcomes the problems related to the delays between the annotation and the video, introduces a bias resulting in sharp annotation signals that hardly follow the dynamics of an expression. In AFEW-VA it has been pointed out another drawback concerning continuous annotation tools: annotators could have a lapse of concentration and inaccuracy due to the sensitivity of the slider. However, we observe that this concern decreases as the annotator training increases.

The 'ANNEMO' tool, instead, is the one which best matches the desired features: separated annotations while playing video, and remote web-based framework. Nevertheless, we experienced some limitations, which may interfere with the annotation process and the usability of the platform. Some of the drawbacks are: i) the absence of indication for the videos already annotated and those to be done; ii) lack of recording with fixed rate; iii) possibility to save the annotations only on text files; iv) missing administration

(a) (b) (c)

Figure 2: Examples of video-based acquired data: color camera frame, depth camera frame and 3D facial landmarks.

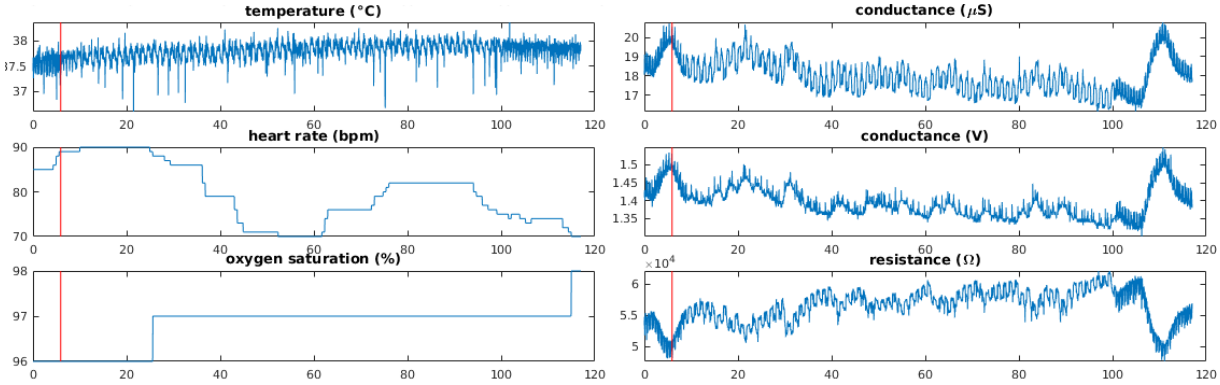

Figure 3: Visualization of raw data of the considered physiological signals (one session)

Figure 4: The self-report annotation tool showing 3 different emotional states. From left to right: neutral (PAD=0,0,0), amused (0.87,0.57,0.76) and happy (1,1,1).

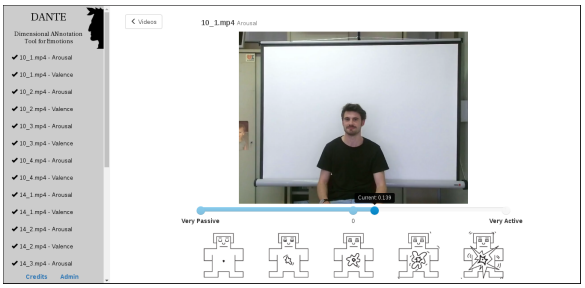

Figure 5: An annotation session using DANTE tool.

interface to manage annotators and videos; v) unable to differentiate annotators in groups, providing different videos.

For these and other minor UI-related reasons, we decided to develop DANTE (Dimensional ANnotation Tool for Emotions), by taking advantage of widely adopted programming languages such as PHP, JavaScript and HTML, backed with a mySQL database.

DANTE allows to create a new annotator account via a dedicated administration page and to assign him/her a random unique identifier corresponding to a private URL for accessing the personal annotation page. The web interface (Fig.5) presents two main parts: a sidebar on the left, which lists all the videos assigned to the specific annotator, marked with an icon to distinguish between the already annotated videos and those to be done. The center of the page is dedicated to annotation itself. Besides showing the video to be annotated and a sliding bar (with values ranging from −1 to +1 and a step of 0.001), we included a SAM (Self Assessment Manikin) visualization specific for the selected affective dimension (arousal or valence), to help annotators. The actual recording of annotations occurs when the video reaches the end, with a fixed rate specified in the configuration file (default is 25) and saved in CSV format files or directly in the database.

4 BASELINE RESULTS

Self reports As to the self-report annotations of the AMHUSE dataset, we collected discrete PAD annotations from all the 36 subjects over the four stimuli, as previously described. Table 3 summarises the mean and standard deviation of the annotations, for each stimulus and each emotional dimension. The ICC(3,k) figure

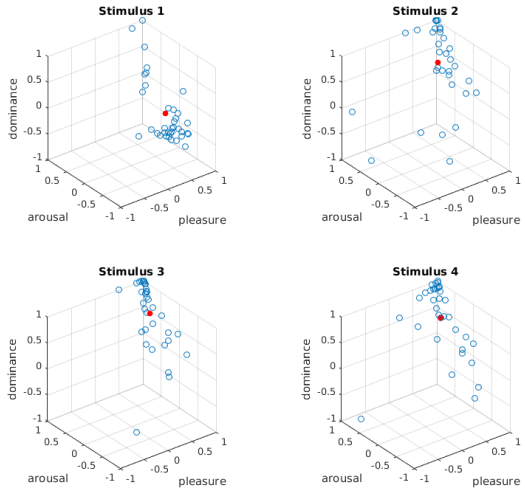

Figure 6: Distribution of the self-annotations in terms of pleasure, arousal and dominance for each of the four stimulus. The red filled points represent the mean value for each of the three dimensions.

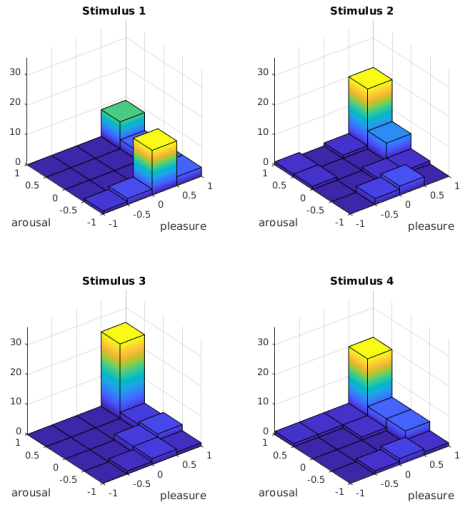

Figure 7: Histogram of frequencies for the self-annotations, in terms of pleasure and arousal for each of the four stimuli.

of merit (intraclass correlation coefficient, two-way randomized) was used to calculate the agreement between the subjects, where $k = 36$ is the number of the only raters of interest. The coefficient leads to $ICC = 0.844$, obtained as the mean of the ICC for pleasure (= 0.664), arousal (= 0.921) and dominance (= 0.948). This shows that on average, all the subjects had remarkable agreement on their ratings, validating stimulus' effectiveness. The visualization in Fig. 6 shows, as expected, a neutral state with very low arousal for the first stimulus, and a general high amusement state for the other videos. Interestingly enough, as shown in Fig. 7 and differently from what expected, the satirical video clip (Stimulus 3) received most annotations with a value of arousal and pleasure higher than 0.5. We surmise that selected subjects were politically interested and, on average, on the same wing.

Table 3: Mean and standard deviation of the self report annotations for each stimulus and each emotional dimension.

	P		A		D	
	μ	σ	μ	σ	μ	σ
Stim 1	0.47	0.40	-0.31	0.80	-0.02	0.44
Stim 2	0.63	0.50	0.49	0.62	0.54	0.53
Stim 3	0.76	0.36	0.50	0.64	0.69	0.36
Stim 4	0.67	0.39	0.47	0.59	0.64	0.43

External annotations. In addition to the time-continuous annotations, we derived an aggregated annotation by means of an estimator that weights the annotations of each rater by his/her respective agreement with the others, defining an individual evaluator confidence score. Such technique is called Evaluator Weighted Estimator (EWE) [19]. The individual evaluator score for the annotator k and emotional dimension $i \in \{V, A\}$ over N annotations is defined as

$$r_k^i = \frac{\sum_{n=1}^{N} \left(x_{n,k}^i - \mu_k^i\right)\left(x_n^{MLE,i} - \mu^{MLE,i}\right)}{\sqrt{\sum_{n=1}^{N}\left(x_{n,k}^i - \mu_k^i\right)^2}\sqrt{\sum_{n=1}^{N}\left(x_n^{MLE,i} - \mu^{MLE,i}\right)^2}}, \quad (1)$$

where μ_k^i is the mean annotation of the evaluator k, $\mu^{MLE,i}$ is the mean value considering all the evaluators. This score considers noise in the individual annotation as well as his/her grade of experience, with $r_k^i = 0$ interpreted as a completely unreliable evaluator. Using these measures as weights, we obtain for each annotation x_n^i, the corresponding "gold standard":

$$\hat{x}_n^i = \frac{1}{\sum_{k=1}^{K} r_k^i} \sum_{k=i}^{K} r_k^i x_{n,k}^i. \quad (2)$$

Figure 8 shows an example of annotation together with the weights for each annotator and the resulting gold standard. In order to show the temporal dynamics of annotations, in Fig. 9 we visualize the gold standard of a subject in terms of valence, arousal and time. Consistently with theoretical predictions, it shows a behaviour typical of a mean-reverting random walk [25, 35].

Other analyses were performed on the annotations to assess the agreement between annotators. Namely, the Cronbach's α [8], the mean Pearson's correlation coefficient and the Concordance Correlation Coefficient (CCC) [27]. Results indicate a good inter-rater reliability for the valence, whilst, as expected, a poorer value for arousal is observed. Indeed, it is well known that the level of arousal is more difficult to distinguish than valence, resulting in a lower agreement between the annotators.

In particular, the Cronbach's α is an estimate of the consistency between annotations. The valence's value is 0.842 and falls in the range $0.8 \leq \alpha < 0.9$, which is to be considered a good internal consistency, while is equal to 0.325 for the arousal. The mean correlation coefficient, instead, is equal to 0.310 for arousal and 0.742 for

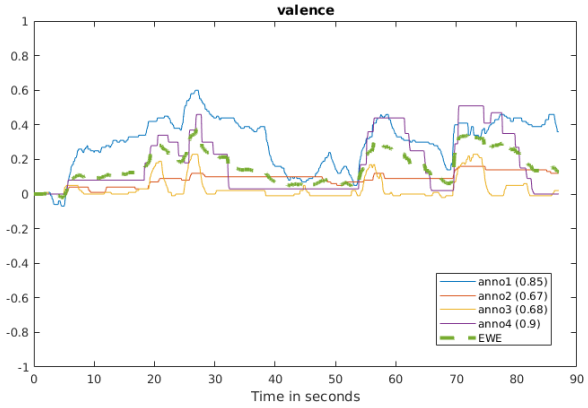

Figure 8: Annotations of valence from 4 annotators and the corresponding gold standard in dashed green line. In brackets it is shown the evaluator score, namely the inter-rater agreement.

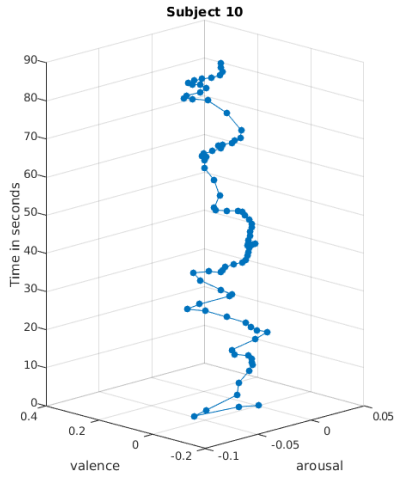

Figure 9: Visualization of gold standard annotation in terms of valence, arousal and time, binned with 2 seconds overlapped moving windows.

valence. The CCC reflects such trend, giving 0.093 for the arousal and 0.509 for the valence.

5 CONCLUSIONS AND PERSPECTIVES

Two are the main contributions of this study: i) AMHUSE, a multimodal dataset of induced amusement emotional states, and ii) DANTE, a novel web-based annotation tool for valence and arousal continuous values. Both are made available to the research community. Recordings include multimodal data (video, depth sequence, facial landmarks and AUs, EDA, BVP and temperature) gathered from 36 participants who watched and rated their emotional response to 4 stimuli (videos). The subjects' reactions were also annotated by 4 annotators on all recorded sequences in terms of valence and arousal. Significant correlation were found between the participant self-reports.

Note that, since there is a gender imbalance (3:1 males to females) both in terms of participants and annotators this could be a possible issue to take into account when using the dataset. fMRI studies suggest that women show greater correlates to humour appreciation than men; however, Chan's study [7] more subtly identified gender differences in amusement that were specific to particular types of jokes. In this terms, gender differences should be averaged, because of the different kinds of humour stimuli. More controversial are physiological measures: men and women respond electrodermally similarly to pleasant stimuli, except for erotic, where men show larger SCRs than women [4].

Amusement is important from a theoretical standpoint. Its distinction with respect to joy [21], the subtle autonomic behaviour with respect to that of negative emotions, the involvement of key expressive components such as smiling [33], all make its analysis and modelling a challenging task. For instance, smiles themselves can be either simple or complicated things [33]. They can be triggered by positive emotion, positive social motives, but also exploited to communicate and maintain social status. Further, coping with such challenge, might require to go beyond the current paradigm of the pattern recognition "pipeline" [47] and engage with embodied and simulation-based accounts [18, 33, 52]. Clearly, in such case, the availability of multimodal data involving physiological signals is mandatory. Eventually, positive emotions serve important social functions [17] and there is evidence that human-computer interaction is natural and social too [42]. We surmise that such emotions and amusement to a great extent, are likely to play a key role for a number of human-computer applications.

ACKNOWLEDGMENTS

This research was carried out as part of the project "Interpreting emotions: a computational tool integrating facial expressions and biosignals based shape analysis and bayesian networks", supported by the Italian Government, managed by MIUR, financed by the *Future in Research* Fund.

REFERENCES

[1] Ralph Adolphs. 2002. Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews* 1, 1 (2002), 21–62.
[2] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *11th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, Vol. 6. IEEE, 1–6.
[3] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. 2015. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern recognition letters* 66 (2015), 52–61.
[4] Margaret M Bradley, Maurizio Codispoti, Dean Sabatinelli, and Peter J Lang. 2001. Emotion and motivation II: sex differences in picture processing. *Emotion* 1, 3 (2001), 300.
[5] Joost Broekens and Willem-Paul Brinkman. 2013. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies* 71, 6 (2013), 641–667.
[6] R.A. Calvo and S. D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1, 1 (2010), 18–37.
[7] Yu-Chen Chan. 2016. Neural Correlates of Sex/Gender Differences in Humor Processing for Different Joke Types. *Frontiers in Psychology* 7 (2016).
[8] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.
[9] L. Devillers, S. Rosset, G. D. Duplessis, M. A. Sehili, L. BÃlchade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. TÃijrker, M. Sezgin, K. E. Haddad, S. Dupont, D. Luzzati, Y. Esteve, E. Gilmartin, and N. Campbell. 2015. Multimodal

data collection of human-robot humorous interactions in the Joker project. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 348–354.

[10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2106–2112.

[11] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19, 3 (July 2012), 34–41.

[12] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 43.

[13] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. 2000. A new emotion database: considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

[14] P. Ekman and E. L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

[15] Anna Esposito and Maria Teresa Riviello. 2010. *The New Italian Audio and Video Emotional Database*. Springer Berlin Heidelberg, Berlin, Heidelberg, 406–422.

[16] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological science* 18, 12 (2007), 1050–1057.

[17] Barbara L Fredrickson. 2003. The value of positive emotions. *American scientist* 91, 4 (2003), 330–335.

[18] Alvin I Goldman and Chandra Sekhar Sripada. 2005. Simulationist models of face-based emotion recognition. *Cognition* 94, 3 (2005), 193–213.

[19] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 381–385.

[20] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[21] David R Herring, Mary H Burleson, Nicole A Roberts, and Michael J Devine. 2011. Coherent with laughter: Subjective experience, behavior, and physiological responses during amusement and joy. *International Journal of Psychophysiology* 79, 2 (2011), 211–218.

[22] Christian M Jones, Laura Scholes, Daniel Johnson, Mary Katsikitis, and Michelle C Carras. 2014. Gaming well: links between videogames and flourishing mental health. *Frontiers in psychology* 5 (2014).

[23] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 46–53.

[24] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW for Valence and Arousal estimation In-The-Wild. *Image and Vision Computing* (2017), –.

[25] Peter Kuppens, Zita Oravecz, and Francis Tuerlinckx. 2010. Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology* 99, 6 (2010), 1042.

[26] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (1993), 261–273.

[27] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[28] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 94–101.

[29] Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology*. the MIT Press. 216–217 pages.

[30] Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–8.

[31] Arturo Nakasone, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion Recognition from Electromyography and Skin Conductance. *The 5th International Workshop on Biosignal Interpretation* (2005), 219–222.

[32] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[33] Paula M Niedenthal, Martial Mermillod, Marcus Maringer, and Ursula Hess. 2010. The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression. *Behavioral and brain sciences* 33, 06 (2010), 417–433.

[34] Martha C Nussbaum. 2003. *Upheavals of thought: The intelligence of emotions*. Cambridge University Press, Cambridge, UK.

[35] Zita Oravecz, Francis Tuerlinckx, and Joachim Vandekerckhove. 2011. A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological methods* 16, 4 (2011), 468.

[36] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *International Conference on Multimedia and Expo, 2005. ICME 2005. IEEE*. IEEE, 5–pp.

[37] L. Pessoa. 2008. On the relationship between emotion and cognition. *Nature Reviews Neuroscience* 9, 2 (2008), 148–158.

[38] R. W. Picard. 2000. *Affective computing*. MIT press, Cambridge, MA.

[39] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Isntelligence* 23, 10 (2001), 1175–1191.

[40] Robert Plutchik. 1956. The psychophysiology of skin temperature: A critical review. *Journal of General Psychology* 55, June (1956), 249–268.

[41] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[42] Byron Reeves and Clifford Nass. 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. (1996).

[43] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–8.

[44] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–8.

[45] S. Rukavina, S. Gruss, S. Walter, H. Hoffmann, and H. C. Traue. 2015. OPEN EmoRec II – A Multimodal Corpus of Human-Computer Interaction. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 9, 5 (2015), 1181–1187.

[46] J. A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[47] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 6 (2015), 1113–1133.

[48] N Selvaraj, a Jaryal, J Santhosh, K K Deepak, and S Anand. 2008. Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *Journal of medical engineering & technology* 32, 6 (2008), 479–484.

[49] Michelle N Shiota, Samantha L Neufeld, Wan H Yeung, Stephanie E Moser, and Elaine F Perea. 2011. Feeling good: autonomic nervous system responding in five positive emotions. *Emotion* 11, 6 (2011), 1368.

[50] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55.

[51] Virginia E Sturm, Jennifer S Yokoyama, Janet A Eckart, Jessica Zakrzewski, Howard J Rosen, Bruce L Miller, William W Seeley, and Robert W Levenson. 2015. Damage to left frontal regulatory circuits produces greater positive emotional reactivity in frontotemporal dementia. *Cortex* 64 (2015), 55–67.

[52] J. Vitale, M-A. Williams, B. Johnston, and G. Boccignone. 2014. Affective facial expression processing via simulation: A probabilistic model. *Biologically Inspired Cognitive Architectures Journal* 10 (2014), 30–41.

[53] Adrienne Wood, Magdalena Rychlowska, Sebastian Korb, and Paula Niedenthal. 2016. Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends in cognitive sciences* 20, 3 (2016), 227–240.