SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild

Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic

Abstract—Natural human-computer interaction and audio-visual human behaviour sensing systems, which would achieve robust performance in-the-wild are more needed than ever as digital devices are becoming indispensable part of our life more and more. Accurately annotated real-world data are the crux in devising such systems. However, existing databases usually consider controlled settings, low demographic variability, and a single task. In this paper, we introduce the SEWA database of more than 2000 minutes of audio-visual data of 398 people coming from six cultures, 50% female, and uniformly spanning the age range of 18 to 65 years old. Subjects were recorded in two different contexts: while watching adverts and while discussing adverts in a video chat. The database includes rich annotations of the recordings in terms of facial landmarks, facial action units (FAU), various vocalisations, mirroring, and continuously valued valence, arousal, liking, agreement, and prototypic examples of (dis)liking. This database aims to be an extremely valuable resource for researchers in affective computing and automatic human sensing and is expected to push forward the research in human behaviour analysis, including cultural studies. Along with the database, we provide extensive baseline experiments for automatic FAU detection and automatic valence, arousal and (dis)liking intensity estimation.

Index Terms—SEWA, Affect Analysis In-the-wild, emotion recognition, regression.

1 Introduction

Artificial Intelligence (AI) technologies are enabling the development of intelligent systems that are human-affect-aware and trustworthy, meaning that they can automatically detect and intelligently respond to users affective states [1]. Affecting computing requires methods that can robustly and accurately analyse human facial, vocal as well as verbal behaviour and interactions in-the-wild, that is, from data captured by omnipresent audio-visual sensors in digital devices in almost arbitrary recording conditions including semi-dark, dark and noisy rooms.

Albeit notable progress has been made so far in machine analysis of human emotion and sentiment, there are still important challenges that need to be addressed in order to deploy and integrate affect-aware systems into everyday interfaces and real-world contexts. Concretely [2], [3]:

- The vast majority of available datasets suitable for audio-visual emotion and sentiment research (cf. Section 2.4 for a brief overview) have been collected in laboratory or controlled conditions, with controlled noise level and reverberation, often limited verbal content, illumination and calibrated cameras. Clearly, such conditions are not present in real-world applications and tools trained on such data usually do not generalise well to behavioural recordings made in-the-wild.
- Many of the available datasets contain examples of induced behaviour as opposed to spontaneous behaviour (occurring in real-life settings of users natural environment like their home). As explained in various studies e.g., [4], [5], spontaneous facial movements are smooth and ballistic, and are more typical of the subcortical system (not associated with cortex and displayed unconsciously). On the other hand, induced facial expressions may be planned and socially modified to a certain extent (i.e., associated with cortex and produced consciously), with entirely different dynamical characteristics than fully spontaneous facial expressions. Consequently, the dynamics of the behaviour (timing, velocity, frequency, temporal inter-dependencies between gestures) crucially affect facial and vocal behaviour interpretation, and currently they are not taken into account.
- Observed behaviours may be influenced by those of an interlocutor and thus require analysis of both interactants, especially to measure such critically important patterns as mimicry, rapport or sentiment. However, existing approaches typically perform analysis of a single individual, and webcam-mediated face-to-face human computer interaction (FF-HCI) is not addressed as a problem of simultaneous analysis of both interacting parties.
- Available audio-visual databases are typically culture specific, e.g., the VAM faces database [6] consists of 20 German speakers, the SEMAINE database consist of UK subjects [7], the RECOLA database consists of French speaking participants, the CONFER dataset contains only Greeks [8]. Hence, there is no database that would enable a large scale study on the effect of culture on expression recognition and communication of emotions and sentiment.
- Most of the existing databases are only annotated in terms of certain behaviour and affect dimensions, for instance, the SEMAINE database contains continuous annotations of valence and arousal, the CONFER dataset is annotated only in terms of conflict intensity etc. Moreover, is no database annotated in terms of multiple behavioural cues: facial action units (FAUs),

affect dimensions and social signals.

In this paper, we aim to address the above mentioned challenges and limitations of existing datasets by introducing the SEWA database (SEWA DB), in Section 3, which is an audio-visual, multilingual dataset of richly annotated facial, vocal, and verbal behaviour recordings made in-the-wild. The SEWA DB extends and contrasts considerably the available audio-visual datasets for affect research by providing the following key features:

- The SEWA DB consists of audio-visual recordings of spontaneous behaviour of volunteers, captured in completely unconstrained, real-world, environments using standard web-cameras and microphones.
- It contains episodes of unconstrained interactions of subjects of different age, gender, and cultural backgrounds. In particular, 6 groups of volunteers with around 66 subjects per group (50% females, uniformly divided over 5 age groups, 20+, 30+, 40+, 50+, 60+) from six different cultural backgrounds, namely British, German, Hungarian, Greek, Serbian, and Chinese were recoded. This makes the SEWA DB the first publicly available benchmark dataset for affect analysis in the wild across age and cultures.
- Audio-visual recordings in the SEWA DB are richly annotated in terms of FAUs, facial landmarks, vocal and verbal cues as well as continuously valued emotion dimensions such as valence, arousal, liking and social signals including agreement and mimicry. This unique feature will allow for the first time to study different aspects of human affect simultaneously, investigate how observed behaviours are influenced in dyadic interactions, and exploit behaviour dynamics in affect modelling and analysis. Furthermore, the breadth of the annotations will allow to exploit interdependencies between age, gender, word and language usage, affect and behaviour, hence enabling robust and context-sensitive interpretation of speech and non-verbal behaviour.

For benchmarking and comparison purposes, we provide exhaustive baseline experimental results FAUs detection and valence, arousal and liking/disliking estimation are provided in Section 4.

The SEWA database is available online at http://db.sewaproject.eu/ and will not only be an extremely valuable resource for researchers in affective computing and automatic human sensing but it may also push forward the endeavour in human behaviour analysis, especially when it comes to cross-cultural studies.

2 State-of-the-art in audio-visual emotion databases

The standard approach in automatic emotion recognition relies on machine learning models trained on a collection of recordings, annotated in terms of different categories or dimensions of affect. As a consequence, the quality of the trained models, and especially their generalisation ability on new data acquired in different conditions, strongly depends on a myriad of factors that shape the construction of the emotional dataset itself. In this section, we discuss three of those main factors, namely elicitation methods, models of emotion representation, and data annotation techniques. In section 2.4, we provide an overview of existing corpora.

2.1 Elicitation methods

One of the factors that has a significant impact on the models of emotion is the type of elicitation methods used to collect affective data. In order to record expressions of affect, one needs to consider a suitable context in which those expressions will be observed. Three main types of context have been used so far to collect such data: (i) *posed behaviour* – emotion is portrayed by a person upon request, e. g., [9], [10], (ii) *induced behaviour* – a controlled setting is designed to elicit a reaction to a given affective stimulus, e. g., watching audio-visual clips or interacting with a manipulated system [11], [12], and (iii) *spontaneous behaviour* – natural interactions between individuals, or between a human and a machine, are collected in a given context, e. g., chatting with a sensitive artificial listener [7], or resolving a task in collaboration [13].

2.1.1 *Posed behaviour*

Interaction scenarios based on a posed behaviour present the advantage to know in advance the expressed emotion, since the portrayals are acted. Targeted emotions usually include the six "basic emotions“ [14], and for which evidence for some universality over various cultures has been shown [15]. Acted scenarios further provide a fine-grained control of the material used to collect data, e. g., phonetic complexity of the spoken utterances can be balanced for vocal analysis [10], as well as illumination or pose variations for facial analysis [16]. In order to facilitate the portrayal of emotion, scripted scenarios can be exploited, with eventually the help of a professional director, who can interact with the actor, thus providing a more natural context [10], [17].

The automatic analysis of acted expressions of six basic emotions is now considered a solved problem with high accuracy performance reported in the literature [18], [19], [20]. Acted data can be of great interest when one wants to focus on specific details of emotional expressions. For instance, this can be very helpful for building rule-based prediction systems [21], [22], or when the targeted population presents major difficulties in handling complex display of emotion, such as in the autism spectrum conditions [23], [24]. On the other hand, acted data cannot be used for training when one wants to predict natural display of emotion. Spontaneous expressions are much more subtle in comparison with acted portrayals. As a consequence, they are also much more challenging to recognise [25].

2.1.2 *Induced behaviour*

In order to collect naturalistic expressions of emotion, one can induce affect by using either passive or active methods. Passive methods consist in (dis)playing a set of standardised stimuli to subjects whose reactions (vocal, facial, and physiological) are recorded. Stimuli can be either static, e. g., the International Affective Picture Systems (IAPS) [26], or dynamic, e. g., audio clips [27], video clips [28] or excerpts from movies [29]. They can also be incorporated into a human-computer interaction system, in order to provoke affective reaction from the user. For instance, system malfunctions or unexpected events can be generated automatically, or by a Wizard-of-Oz, in order to induce emotion [30]. Induced behaviours are also of interest for emotionally driven marketing research, e.g., the efficiency of an audiovisual advertisement can be measured automatically through the affective reactions of the audience, instead of self-reported questionnaires.

2.1.3 Spontaneous behaviour

The most appealing approach for capturing a wide range of fully naturalistic displays of emotions consists in recording spontaneous human interactions. Ecologically valid situations, i. e., observing humans in their natural environments, would be ideal as it ensures unobtrusiveness and thus guarantees the observation of fully natural behaviours. This step out of the laboratory has not been accomplished until now for the collection of affective data produced during human interactions. The SEWA database presented in this paper is the very first such collection of interactive human behaviour, annotated in terms of displayed affective dimensions, recorded in-the-wild. Until now, various characters presenting different personality traits (e. g., joyful, depressed, introvert, etc), and simulating an artificial sensitive listener, have served as human interlocutors [7].

2.2 Representation of emotion

Emotion is a subjective feeling, and a complex internal phenomenon. Hence, using a simplistic classification-based model relying on few emotion categories provides only a very limited description of the phenomenon. Additionally, whether or not a certain expression stems from one emotion or the other (e. g., sadness versus boredom) could be matter of subjective interpretation.

In the pursuit of a finer model, several continuous-valued, multidimensional models have been proposed for more precise emotion description. Arguably, the most popular model employed by the affective computing research community is the two dimensional model describing the degree of activation (arousal) and pleasantness (valence) of displayed affect expressions as a point in the Cartesian plane. A well known problem with this approach is the dynamically varying time-delay between an expression and the annotation due to reaction lag of the annotators, which also varies among different annotators, over the sessions, and even during every session [31]. Yet methods have been proposed for spatio-temporal alignment of annotations [31], [32] to remedy this problem and come up with reliable ground truth to be used for training dimensional affect regressors.

2.3 Data Annotation and generation of the Gold Standard

Depending on the choice of the model for emotion representation, several research groups have developed their own annotation tools. Some of these tools have been now made available to researchers across the world, some even with open source licenses. Popular annotation tools in use today are ANVIL [33], ATLAS [34], Ikannotate [35], EmoWheel (Geneva emotion wheel) [36], FEELtrace [37], Gtrace [1], ANNEMO [38], and the frame by frame Valence/Arousal Online Annotation Tool [39].

Several methods exist for creating a unified view of the perceived emotion from a set of annotations, generally referred as the 'Gold Standard' to differentiate with 'ground-truth', which is avoided for affective computing as there does not exist a truth on a subjective feeling such as emotion. The basic principle is to use consensus among the evaluators to come up with a common, best representative annotation by using different metrics such as the correlation coefficients, dynamical time warping (DTW) distance [32], [40], average of the data post standardization or normalization, or assigning individual annotations certain weight percentages (e. g., evaluator weighted estimator (EWE)).

2.4 The Existing Corpora

Here, we focus here on databases containing dyadic interaction recordings annotated in terms of displayed affective reactions. For an overview of databases containing recordings of non-interactive subjects, the reader is referred to recent survey papers (e.g. [19], [41]) and recent database papers (e.g. [39]). Most of the databases of dyadic interaction recordings annotated in terms of displayed affective reactions contain recordings made in controlled settings, concern a constrained dyadic task, have low demographic variability, and are made in primarily one language –that has mostly been English so far. Predominance of one language in the corpora limits usability of the database for cross-lingual, cross-cultural study of emotion recognition. Table 1 presents a summarized overview of the surveyed databases containing dyadic interactions.

2.4.1 Elicitation using Conversational Context

The *Geneva Airport Lost Luggage Study* database [42] is amongst the very few databases featuring cultural diversity in terms of its subjects. 112 passengers –that were required to claim their lost baggage at the airline's baggage retrieval office, were recorded surreptitiously during their interaction with the airline agents processing their claims. Because of this unobtrusive recording paradigm, the dataset features truly natural emotional responses, and can be claimed to be free from Labov's paradox [49]. The gender split of the data is adequately balanced with 59.8 % male, and 40.2 % female subjects. The linguistic/cultural distribution however is quite unbalanced. In addition, individual languages of the participants is not reported, only 'language groups' are given [42]. No continuous annotations are available, only subjects overall feeling for the whole episode are provided. Specifically, the subjects self-reported their emotional states before and after the interaction in a 7-point scale for 5 emotion categories; namely 'angry/irritated', 'resigned/sad', 'indifferent', 'worried/stressed', and 'in good humour'.

The *Cardiff Conversation Database (CCDb)* [46] and *4D Cardiff Conversation Database (4D CCDb)* [48] databases follow another interesting recording paradigm where the subjects freely discuss topics of their own interest and lead the

[1] Successor to FEELtrace, https://sites.google.com/site/roddycowie/work-resources

TABLE 1: Summary of corpus available for estimation of arousal and valence from audiovisual data, featuring unscripted interactive discourse. Information that is not available in the citation is indicated as 'NI', short for 'No Information'.

Dataset	#Subjects				# audio--visuals	Annotation	Duration	Language(s)	Elicitation	Year
	Total	M	F	Age range						
GENEVA [42]	112	55	45	20-60+	112	NI	NI	French English German/ northern Europe Asia Other	conversation	1997
SMARTKOM [43]	224	NI	NI	16-45+	466	NI	17 hours	German	HMI	2002
VAM-faces [6]	20	NI	NI	16-69 70%<35	1421	Linkert-like scale (5 points from -1 to 1), 7–8 raters	12 hours	German	Talk-show	2008
SEMAINE [7]	150	57	93	NI	959	continuous, Feel-trace, 7 raters	80 hours	English	HMI	2012
Belfast naturalistic [2]	125	31	94	NI	298	Continuous, Feeltrace, 6–258 raters	86 minutes	English	Talk-show	2010
AVEC'13 [44]	292	NI	NI	18-63	340	Continuous, Feeltrace, 1 rater	240 hours	German	HMI	2013
Belfast induced 1 [45]	114	70	44	NI	570	Continuous, Feeltrace, 6–258 raters	237 minutes	English	TV/interviews	2012
Belfast induced 2 [45]	82	37	45	NI	650	Valence only, Continuous, Feeltrace, 1 rater	458 minutes		laboratory based tests	
CCDb [46]	16	12	4	25-56	30	NI	300 minutes	English	conversations	2013
RECOLA [13]	46	19	27	NI	46	Continuous, Feeltrace, 7 raters	230 minutes	French	online conversation	2013
AVEC'14 [47]	84	NI	NI	18-63	300	Continuous, Feeltrace, 3+ raters	240 hours	German	HMI	2014
MAHNOB Mimicry [9]	60	31	29	NI	54	Continuous, Feeltrace, $\simeq$ 5 raters	11 hours	English	dyadic conversations	2015
4D CCDb [48]	4	2	2	20-50	34	NI	17 minutes	English	conversation	2015
SEWA (this work)	398	201	197	18-60+	1990	Continuous, Feeltrace, 5 raters	44 hours	Chinese English German Greek Hungarian Serbian	Watching videos – Dyadic conversations	2017

conversations themselves. They are not given any specific task, nor a topic for conversation, nor any specific audiovisual stimuli to elicit emotions. Both the databases contain conversations only in English, each containing 30 and 6 conversations respectively. The gender split is quite skewed with 12 male and only 4 female subjects. The dataset is annotated in terms of Frontchannel (main speaker periods), Backchannel (qualified utterances and expressions), agreement/disagreement episodes, smiles, laughter, negative and positive surprises, thinking phases, confusion, head motions, made using ELAN [50] framework.

The *MAHNOB Mimicry* [9] dataset features dyadic conversations where subjects engage in socio-political discussions, or negotiate a tenancy agreement. The subjects span range of nationalities including Spanish, French, Greek, English, Dutch, Portuguese,and Romanian. All conversations were recorded in English. Subjects are 18 to 34 years old, with 4.8 years of standard deviation. Continuous annotations were obtained using FeelTrace by approximately 5 raters.

The *Conflict Escalation Resolution (CONFER)* [8] dataset is constituted of 120 video clips of interactions between 54 subjects from Greek televised political debates. The data was annotated by 10 experts in terms of continuous conflict intensity.

2.4.2 Elicitation using Human-Machine Interfaces

The *Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression Dataset (SEMAINE)* [7] presents richly annotated recordings (7 basic emotion states, 6 types of epistemic states, transcripts, laughs, head movement and FACS) of interactions in laboratory conditions between a human and a machine-like agent in three different Sensitive Active Listener (SAL) scenarios. It features 150 participants, most of which come from Caucasian background and 38 % are male. The language of communication is predominantly English.

The *SMARTKOM* dataset [43] features subjects interacting in laboratory conditions, in German with a pretense/WOZ multimodal dialogue system that supposedly allows the user to interact almost naturally with a computer. The recording sessions were first split into subject-state episodes by the labellers marking start and end of each perceived episode. The segments were then labelled with the following 7 categories: 'joy/gratification', 'anger/irritation', 'helplessness', 'pondering/reflecting', 'surprise', 'neutral' and 'unidentifiable episodes'. Gender split is 20 male and 25 female speakers.

2.4.3 Elicitation through Tasks

The RECOLA [13] dataset contains multimodal recordings of French students performing a collaborative task. The participants discuss and rank 15 items in the order of their significance for their survival in a remote and hostile region in cold winter. The subjects are from different parts of Switzerland, and thus have different cultural backgrounds (33 French, 8 Italian, 4 German, 1 Portuguese). The database however features French language alone, and mean age of the subjects is 22 years with only 3 years of standard deviation. Continuous levels of valence and arousal were annotated by 7 raters using FeelTrace.

To collect *Belfast Induced Natural Emotion Database* [45], English speaking participants were asked to perform select set of tasks specifically designed to induce mild to moderately strong emotionally coloured responses (e.g. reaching into a box that sets off a very loud alarm). Mean age of subjects is 24 years with 6 years of deviation. Continuous values of valence and arousal were obtained for each clip by 6 to 258 raters using FeelTrace.

The corpus for Audio-Visual Emotion recognition Challenges in 2013 and 2014, namely *AVEC'13* [44] and *AVEC'14* [47], used a subset of audio-visual depressive language corpus (AViD-Corpus) which consists of recordings of subjects performing human-computer interaction tasks, labelled by 23 annotators continuous for arousal and valence estimates. The mean age of subjects is 31 years, with 6 years standard deviation.

2.4.4 Corpus collected by segmenting existing recordings

Belfast Naturalistic Database [2] contains 10 to 60 seconds–long audiovisuals taken from English television chat shows, current affairs programmes and interviews. It features 125 subjects, of which 31 are male, and 94 are females. Out of 298 clips, 100 videos totalling 86 minutes in duration have been labelled with continuous-valued emotion labels for activation and evaluation dimensions, with additionally 16 basic classifying emotion labels.

Similarly, *Vera am Mittag (VAM) database* [6] contains 12 hours of recordings of the German TV talk-show Vera am Mittag (Vera at noon) with continuous-valued emotion labels for arousal, valence, and dominance. It contains 20 participants with age ranging from 16 to 69.

The *AFEW-VA database* [39] is a visual dataset containing 600 challenging video clips extracted from feature films and annotated per-frame in term of levels of valence and arousal, as well as 68 facial landmarks. It contains 240 subjects, 50% female, with age ranging from 8 to 76.

3 SEWA DATABASE

The main aim of the SEWA DB is to provide enough suitable data of labelled examples to facilitate the development of robust tools for automatic machine understanding of human behaviour.

In particular, we recorded 6 groups of volunteers (around 66 persons per group) from six different cultural backgrounds: British, German, Hungarian, Greek, Serbian, and Chinese. The volunteers in each group have a broad distribution in gender and age. Specifically, there are at least one native speakers each age group (18-29, 30-39, 40-49, 50-59, and 60+) for each culture. The resulting database contains a total of 199 sessions of experiment recordings: 1600 minutes of audio-visual data of people's reaction to adverts from 398 individuals, and 1057 minutes of recorded computer-mediated face-to-face interactions between pairs of subjects.

The SEWA database includes annotations of the recordings in terms of facial landmarks, facial action units, various vocalisations, verbal cues, mirroring, and rapport, continuously valued valence, arousal, liking, and prototypic examples (templates) of (dis)liking and sentiment. The data has been annotated in an iterative fashion, starting with a sufficient amount of examples to be annotated fully in a semi-automated manner.

3.1 Data collection

To create the SEWA dataset, a data collection experiment has been conducted. In this experiment, participants were divided into pairs based on their cultural background, age and gender. During initial sign-up, participants were asked to complete a questionnaire of demographic measures including gender, age, cultural background, education, personality traits, and familiarity with the other person in the pair. To promote natural interactions, participants within each pair were required to know each other personally in advance of the experiment. Each pair of the participants then took part in two parts of the experiment, resulting in two sets of recordings.

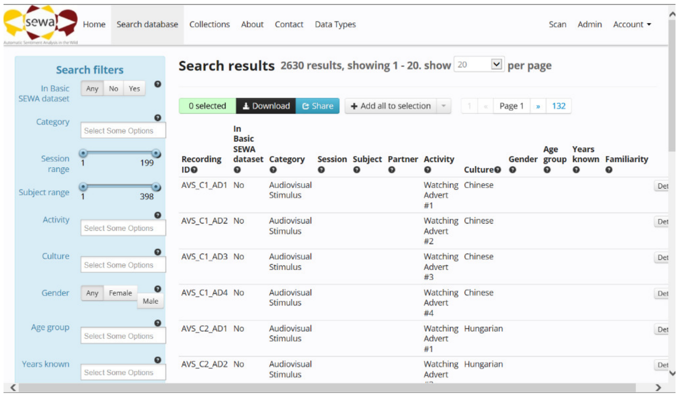
The SEWA data collection experiment was conducted using a website specifically built for this task (shown in Figure 3.4). The website (http://videochat.sewaproject.eu) utilises WebRTC/OpenTok to facilitate the playing of adverts, video-chat, and synchronized audio/video recording using the microphone and webcam on the participants own computer. This setup allowed the participants to be recorded in truly unconstrained in-the-wild environments with various lighting conditions, poses, background noise levels, and sensor qualities.

Experimental Setup Part 1: Each participant was asked to watch 4 adverts, each of being around 60 seconds long. These adverts had been chosen to elicit mental states including amusement, empathy, liking and boredom. For consistent understanding of the advertisement content across cultures, the advertisements chosen have no dialogues, but are driven primarily by the visuals and the accompanying music. The four advertisements [3] in order are:

1) A violent advertisement of the *National Domestic Violence Hotline* –eliciting disgust, distress, and yet the liking for the effectiveness of the advertisement,

[2] http://sspnet.eu/2010/02/belfast-naturalistic/

[3] These videos, including all subtitled versions prepared for subjects of different cultural backgrounds, are included in our dataset for referencing purposes.



Search filters

In Basic SEWA dataset: Any | No | Yes

Category: Select Some Options

Session range: 1 – 199

Subject range: 1 – 398

Activity: Select Some Options

Culture: Select Some Options

Gender: Any | Female | Male

Age group: Select Some Options

Years known: Select Some Options

Search results 2630 results, showing 1 - 20. show 20 per page

Recording ID	In Basic SEWA dataset	Category	Session	Subject	Partner	Activity	Culture	Gender	Age group	Years known	Familiarity
AVS_C1_AD1	No	Audiovisual Stimulus				Watching Advert #1	Chinese				
AVS_C1_AD2	No	Audiovisual Stimulus				Watching Advert #2	Chinese				
AVS_C1_AD3	No	Audiovisual Stimulus				Watching Advert #3	Chinese				
AVS_C1_AD4	No	Audiovisual Stimulus				Watching Advert #4	Chinese				
AVS_C2_AD1	No	Audiovisual Stimulus				Watching Advert #1	Hungarian				
AVS_C2_AD2	No	Audiovisual Stimulus				Watching Advert	Hungarian				

Fig. 1: Front page of the online SEWA database and the search filters.

2) A self-deprecating, witty advertisement of the *Smart Fortwo* car –eliciting pleasure and liking for the advertisement,
3) A bizarre, abstract advertisement of the *Jean Paul Gaultier Le Male Terrible* perfume with a highly blurred product emphasis, with illegible product name in the visuals –eliciting confusion and a strong disliking for the advertisement,
4) An advertisement of a touch-activated *Grohe* faucet presenting use cases for the newly introduced sensor-based activation feature, –eliciting interest and liking for the product, and boredom for the advertisement overall.

After watching the advert, the participant was also asked to fill-in a questionnaire to self-report his/her emotional state and sentiment toward the advert.

Experimental Setup Part 2: After watching the 4th advert, the two participants were asked to discuss the advert they had just watched by using the video-chat function provided by the SEWA data collection website. On average, the recorded conversation was 3 minutes long. The discussion was intended to elicit further reactions and opinions about the advert and the advertised product, such as whether the advertised is to be purchased, whether it is to be recommended to others, what are the best parts of the advert, whether the advert is appropriate, how it can be enhanced, etc.. After the discussion, each participant was asked to fill-in a questionnaire to self-report his/her emotional state and sentiment toward the discussion.

3.2 The Data Statistics and subject demographics

During the SEWA experiment, 198 recording sessions have been successful, with a total of 398 subjects being recorded. The subjects were coming from 6 different cultural backgrounds: British, German, Hungarian, Serbian, Greek, and Chinese. 201 of the participants are male, 197 are female, resulting in a gender ratio (male / female) of 1.020. Furthermore, the participants are categorized into 5 age groups: 18 29, 30 39, 40 49, 50 59 and 60+, with the 18 29 group being most numerous. The detailed participant demographics are shown in Table 2.

TABLE 2: SEWA Demographics

		Cultures					
		Chinese	English	German	Greek	Hungarian	Serbian
Gender	Male	36	33	39	34	26	33
	Female	34	33	25	22	44	39
Interactions	F–F	22	20	16	8	30	16
	M–F	24	26	18	26	28	46
	M–M	24	20	30	22	12	10
Age	18–29	44	34	41	18	44	22
	30–39	16	12	13	29	9	15
	40–49	4	6	1	1	5	8
	50–59	6	8	5	8	5	14
	60+	0	6	4	0	7	13
Total		70	66	64	56	70	72

A total of 1990 audio-visual recording clips (5 clips per subject: 4 recorded during the advert-watching part and 1 recorded during the video-chat part) were collected during the experiment, comprising of 1600 minutes of audio-visual data of people's reaction to adverts and 1057 minutes of video-chat recordings. Due to the wide spread of the participants computers hardware capacity, the quality of the video and audio recordings is not constant. Specifically, the spatial resolution of the video recordings ranges from 320x240 to 640x360 pixels and the frame rate is between 20 and 30 fps. The audio recordings sample rate is either 44.1 or 48 kHz.

3.3 Data annotation

The SEWA database contains annotations for facial landmarks, (pre-computed) acoustic low-level descriptors (LLDs) [51] [52], hand gestures, head gestures, facial action units, verbal and vocal cues, continuously-valued valence, arousal and liking / disliking (toward the advertisement), template behaviours, episodes of agreement / disagreement, and mimicry episodes.

Due to the large amount of raw data acquired from the experiment, the annotation process has been conducted iteratively, starting with sufficient amount of examples to be annotated in a semi-automated manner and used to train various feature extraction algorithms developed in SEWA. Specifically, 538 short (10-30s) video-chat recording segments were manually selected to form the fully-annotated **basic SEWA dataset**. These segments were selected based on the subjects to the subjects emotional state of low / high valance, low / high arousal, and liking / disliking. All 6 cultures were evenly represented in the basic SEWA dataset, with approximately 90 segments selected from each culture based on the consensus of at least 3 annotators from the same culture.

3.3.1 Facial landmarks

Facial landmarks were annotated for all segments included in the basic SEWA dataset using a 49-point mark-up, as described in [53]. Manual annotation of facial landmarks is highly labour intensive. Based on previous experience [53], we know that trained annotators can only achieve a sustained annotation speed of 30 frames per hour. Since the basic SEWA dataset contains a total of 369974 video frames, it would be impractical to annotate all of them manually (which would require more than 12000 hours of work). Therefore, the annotation was performed semi-automatically.

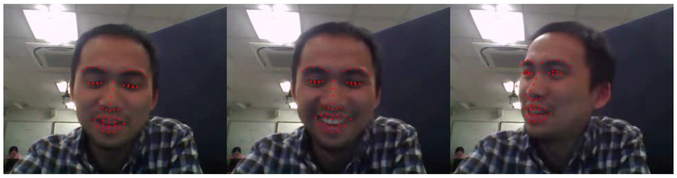


Fig. 2: Example of facial landmark annotation. The 49 facial landmarks were annotated for all segments included in the basic SEWA dataset.

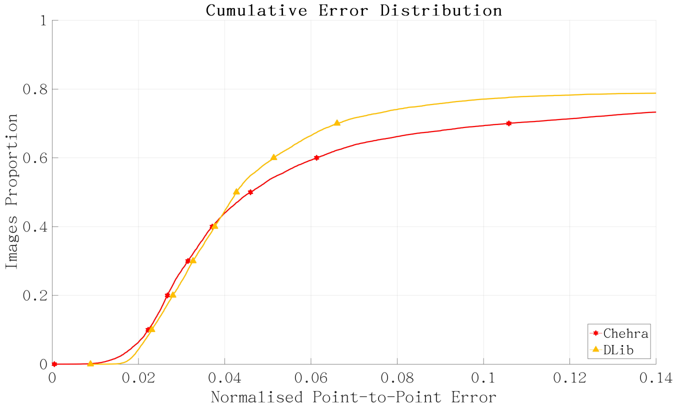


Fig. 3: CED curves of the Chehra tracker [54] and the Dlib tracker [57] on the manually corrected frames.

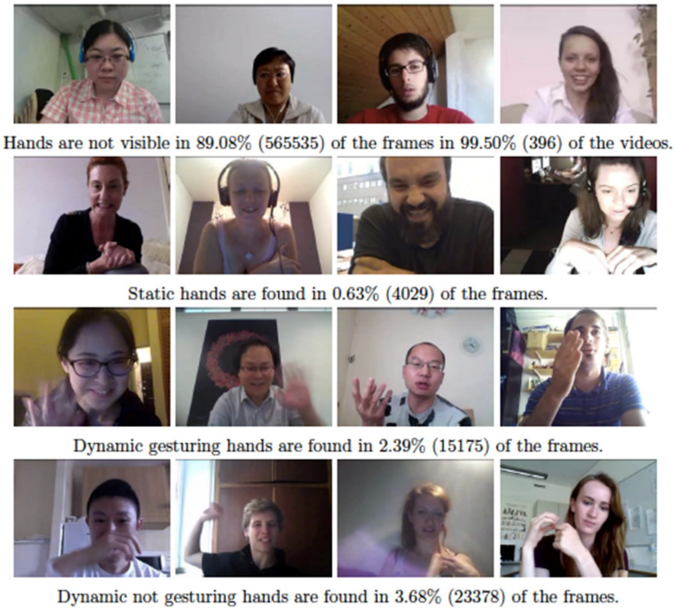
During the annotation process, we first applied the Chehra facial landmark tracker [54] [55] on all video segments. Using a discriminative model trained by a cascade of regressors, the tracker can construct personalised model by incremental updating of the generic model. More than 95.1% of the tracking results (in 351875 frames) produced by Chehra are accurate and require no further correction. For the remaining 18099 frames, manual annotation was performed in a similar way as in preparation of the 300VW dataset [53] [56]. Specifically, we manually annotated 1 in every 8 frames and used the results to train a set of person-specific trackers. These person-specific trackers were applied to the rest of the frames to obtain the annotations. Finally, a visual inspection was performed on the annotations and those deemed unsatisfactory were further corrected. An example of the facial landmark annotation obtained from this process is show in Figure 2.

3.3.2 *Hand Gesture*

Hand gestures were annotated for all video-chat recordings in 5 frame steps. Five types of hand gestures were labelled: hand not visible (89.08%), hand touching head (3.32%), hand in static position (0.63%), display of hand gestures (2.39%), and other hand movements (3.68%). Some examples of the labelled frames are shown in Figure 4.

3.3.3 *Head Gesture*

Head gestures were annotated in terms of nod and shake for all segments in the basic SEWA dataset. The annotation was performed manually on a frame-by-frame basis. To be able to provide good training examples for the head nod/shake detector, we emphasised specifically on high precision during the annotation process. Specifically, only un-ambiguous displays of head nod / shake were labelled. In the end, a total of 282 head nod sequences and 122 head shake sequences were identified. Examples of the labelled head nod / shake sequences are shown in Figure 5



Hands are not visible in 89.08% (565535) of the frames in 99.50% (396) of the videos.

Static hands are found in 0.63% (4029) of the frames.

Dynamic gesturing hands are found in 2.39% (15175) of the frames.

Dynamic not gesturing hands are found in 3.68% (23378) of the frames.

Fig. 4: Examples of hand gesture annotation.

Fig. 5: Examples of head nod (top row) and head shake (bottom row) sequences.

3.3.4 *Transcript*

We provide the audio transcript of all video-chat recordings. In addition to the verbal content, the transcript also contains labels of certain non-verbal cues, such as sighing, coughing, laughter, etc. Utterances were transcribed lexically, including markers for non-linguistic vocalizations including laughs, "back-channel" expressions of consent and hesitation, rapid audible nasal exhalation (like in smirk and snigger), and audible oral exhalations. To minimise the efforts for transcription, semi-automatic methods such as active learning were in an iterative fashion, starting with existing modules for automatic speech recognition and spotting of non- linguistic vocalisations [58] [59]. Since lexical transcription does not require special training, crowdsourcing was used.

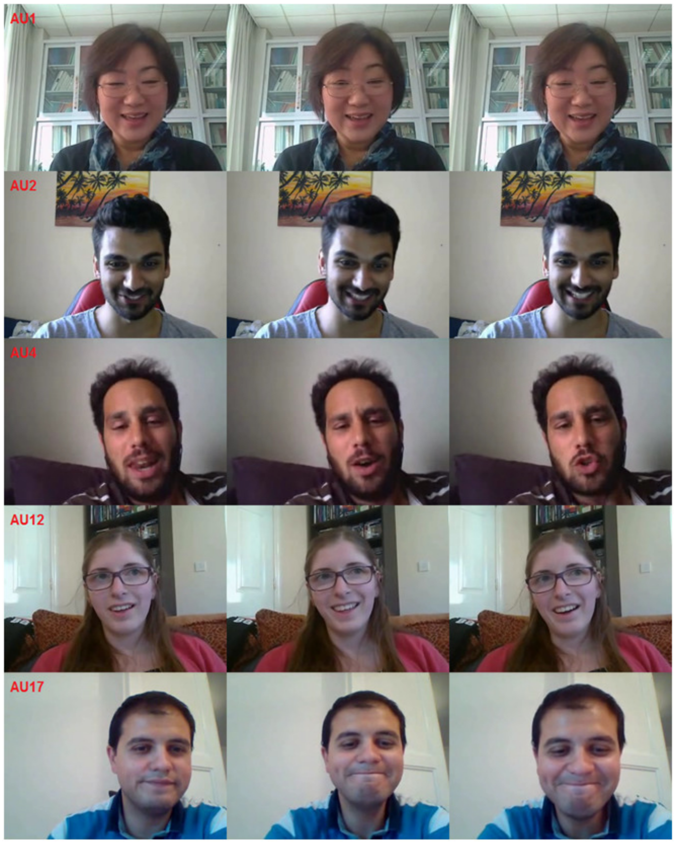


Fig. 6: Examples of AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (Brow lowerer), AU12 (Lip corner puller) and AU17 (Chin raiser)

3.3.5 Facial Action Units annotation

Manual annotation of Action Units (AUs) has to be performed by trained experts which is expansive and time consuming. Especially due to the size of the SEWA database, such manual annotation is prohibitive. Therefore, we focus on accurate semi-automatic annotation of five AUs (1,2,4,12 and 17, depicted in Fig. 6). We selected these AUs as they are occurring most in naturalistic settings, and are important for high-level reasoning about sentiment.

We leverage the state-of-the-art method of [60] for detection of AUs. Specifically, we employ publicly available datasets (DISFA [61] and FERA2015 [62]), annotated in terms of AU intensity, to train our models for sequence modelling. This allows us to automatically obtain segments where target AUs are active (intensity $\neq 0$) and non-active (intensity=0). Although this allowed us to narrow down the possible number of AU activations in target videos, the annotation process could not be fully automated. This is mainly because of a high number of false positives that can occur in such obtained automated annotations of AUs due to the training of target models being performed on videos from different datasets, which can in some instances differ significantly in lighting, head-pose and other conditions, from the SEWA videos. Once the automatic annotation is performed, several annotators have manually inspected the obtained active segments of target AUs, defining the starting and end frame of the AU within the segments classified as active by the model.

We used this annotations to train the AU detector of 5 facial action units (AU) from the basic SEWA dataset: inner eyebrow raiser (AU1, 109 examples), outer eyebrow raiser (AU2, 79 examples), eyebrow lowerer (AU4, 94 examples), lip corner puller (AU12, 104 examples), and chin raiser (AU17, 61 examples). Similarity, the AU examples were again identified in a semi-automatic manner. Specifically, we first applied automatic AU detectors to the video segments and manually removed all false-positives from the detection results. Consequently, the AU annotation is not exhaustive, meaning that some AU activations may be missed.

Using this semi-automatic approach, we annotated 500 sequences (150 frames each) containing at least one of the 5 target AUs. We will refer to this dataset as SEWA AU DATASET. For the baseline experiments, we split this dataset in subject independent training, development and test sets. The size of each dataset and for each AU are shown in table 3.

The proposed method to semi-automatic AU detection has been implemented into a standalone module (VSL-AU detector) in C++/Matlab. This detector is then further integrated into the SEWA back-end emotion recognition server using the HCI2 Framework [63].

TABLE 3: Number of frames with active AUs in training, test and validation set.

AU	**TR**	**TE**	**VA**
1	5180	4340	5740
2	4060	3220	3920
4	4620	4340	4200
12	5600	4620	4340
17	3500	3080	2940
Total	22960	19600	21140

3.3.6 Valence, Arousal, and Liking/Disliking annotation

Continuously-valued valance, arousal and liking / disliking (toward the advertisement) were annotated for all segments in the basic SEWA dataset. In order to identify the subtle changes in the subjects emotional state, annotators were always hired from the same cultural background of the recorded subjects. In addition, to reduce the effect of the annotator bias, 5 annotators were recruited for each culture. The annotation was performed using a custom-built tool, which played the recordings while asked the annotators to push / pull a joystick in real-time to indicate the subject's level of valence, arousal, or liking/disliking. The joystick's pitch value was then sampled at approximately 66 Hz and saved as the annotation. To avoid cognitive overload on the annotators, the three dimensions (valence, arousal and liking / disliking) were annotated separately in three passes. Furthermore, for each dimension, the segments were annotated three times, first based on audio data only, then based on video data only and finally based on audio-visual data. An example of the continuous annotations obtained with this process is illustrated in Fig. 7. Only the segments where the subject on camera was speaking and his chat partner silent were considered. Of the segments satisfying this condition, 90 were selected for annotation so as to contain 15 segments for each of the following criteria: (a) high arousal, (b) low arousal, (c) positive valence, (d) negative valence,

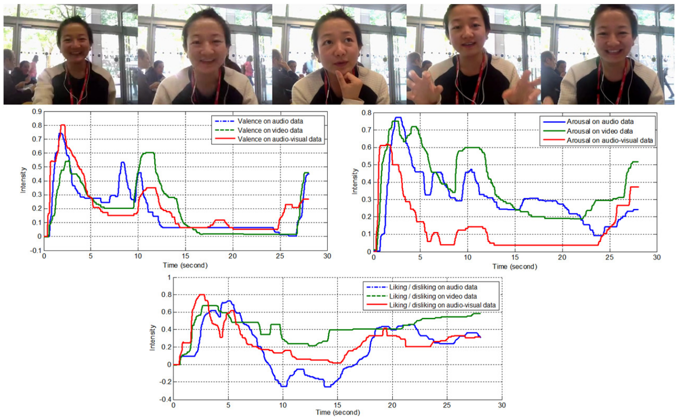


Fig. 7: An example of the continuously valued annotation results on valence, arousal and liking and disliking.

TABLE 4: Behaviour templates identified in the basic SEWA dataset.

	British	German	Hungarian	Serbian	Greek	Chinese
Low Valence	2	4	2	6	2	3
High Valence	2	4	2	5	2	4
Low Arousal	2	3	2	2	2	2
High Arousal	2	3	2	6	2	4
Liking	2	4	2	6	2	5

(e) presence of liking and (f) presence of disliking. The latter two 30 segments were continuously annotated for liking and disliking while all 90 segments were fully annotated for continuous valence and arousal.

These continuous annotations obtained are further combined into one single ground-truth employing Canonical Time Warping for each sequence to construct a subspace were the annotations of all raters are maximally correlated with each other and with the corresponding audio-visual features. The ground-truth annotation is then derived from the correlated subspace. More precisely, it is obtained by keeping only the coefficient corresponding to the first component each annotation. This is additionally normalised in the continuous range $[0, 1]$.

3.3.7 Behaviour Templates

Moreover, we identified behaviour templates –that is prototypical behaviours– for each culture when the subjects are in the emotional state of low / high valence, low / high arousal or showing liking / disliking toward the advertisement. For each category, at least two examples were identified. Table 4 shows the exact distribution of the templates found in the basic SEWA dataset. These templates can be used to train and test the behaviour similarity detector. Figure. 8 illustrates some examples of these behaviour templates.

In addition to continuous values such as valence and arousal, we extracted a number of episodes from the video-chat recordings in which the pair of subjects were in low, mid or high level of agreement / disagreement with each other and annotated the level of agreement/disagreement. The selections were based on the consensus of at least 3 annotators from the same culture of the recorded subjects. The exact numbers of agreement / disagreement episodes are shown in Table 5. Two examples of the agreement / disagreement episodes are shown in Figure 9.

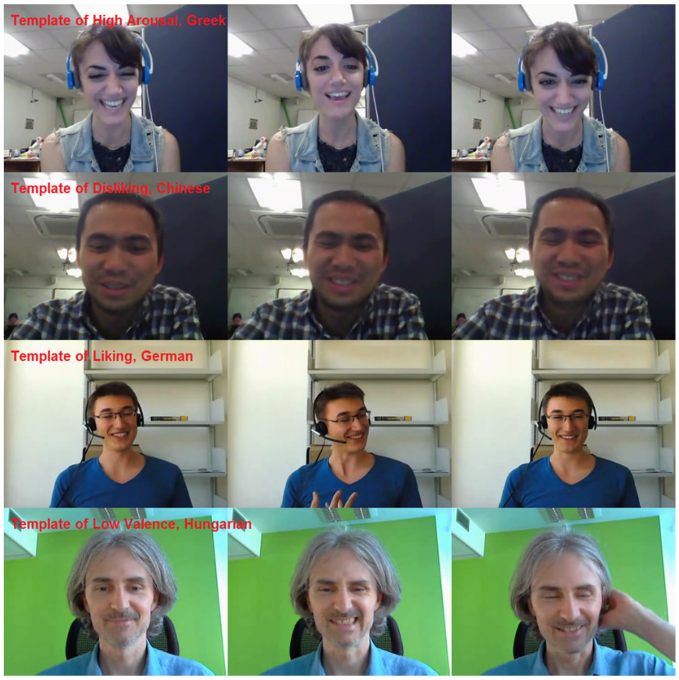

Template of High Arousal, Greek
Template of Disliking, Chinese
Template of Liking, German
Template of Low Valence, Hungarian

Fig. 8: Examples of behaviour templates identified from the basic SEWA dataset.

TABLE 5: Agreement / disagreement episodes identified in the video-chat recordings.

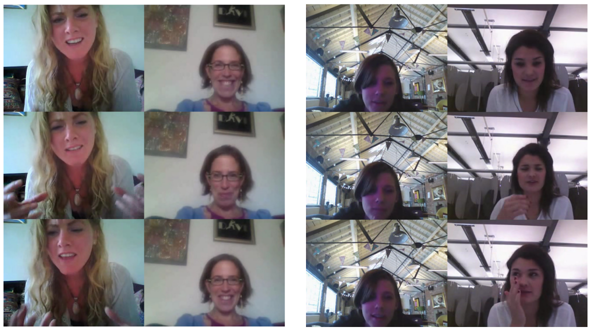
	British	German	Hungarian	Serbian	Greek	Chinese
Strong Agreement	12	7	7	7	5	5
Moderate Agreement	26	7	6	7	5	6
Weak Agreement	29	7	6	7	5	6
Weak Disagreement	7	6	5	4	5	4
Moderate Disagreement	3	9	5	6	5	5
Strong Disagreement	3	6	5	4	5	3

3.3.8 Mimicry Episodes

Lastly, 197 mimicry episodes (48 British, 31 German, 39 Hungarian, 20 Serbian, 41 Greek and 17 Chinese), in which one subject mimicked the facial expression and / or head gesture of the other subject, were identified from the video-chat recordings. Two examples of the identified mimicry episodes are shown in Figure 10.

3.4 Database availability

The SEWA database is available online at: http://db.sewaproject.eu/. The web-portal provides a comprehensive search filter (shown in Fig.) allowing users to search for specific recordings based on various criteria, such as demographic data (gender, age, cultural background, etc.), availability of certain types of annotation, and so on. This will facilitate investigations during and beyond the project



(a) Strong agreement (b) Strong disagreement

Fig. 9: Examples of agreement and disagreement episodes.

(a) Chinese culture (b) Hungarian culture

Fig. 10: Examples of the mimicry episodes.

in the field of machine analysis of facial behaviour as well as in other research fields.

The SEWA database is made available to researchers for academic-use only. To comply with clauses stated in the Informed Consent signed by the recorded participants, all non-academic/commercial uses of the data are prohibited. Only researchers who signed the EULA will be granted access to the database. In order to ensure secure transfer of data from the database to an authorised users PC, the data are protected by SSL (Secure Sockets Layer) with an encryption key. If at any point, the administrators of the SEWA database and/or SEWA researchers have a reasonable doubt that an authorised user does not act in accordance to the signed EULA, they will be declined the access to the database.

4 Baseline experiments

In this section, we introduce the experimental setting and results for action unit detection, as well as valence, arousal and liking/disliking estimation.

4.0.1 *Methods*

We performed experiments with three widely used and established methods:

Support Vector Machine for Regression (SVR): We used a Support Vector Machine for regression, a common approach for affect estimation that has been widely used as a baseline for valence and arousal estimation [64], [65], [66] and to produce state-of-the-art results [67], [68]. In this paper, we use the Scikit-Learn implementation [69] and a linear kernel.

Tree-based The Random Forest Regressor (*RF*) was used to produce the the second set of baseline results. It has been shown to produce state-of-the-art results on a wide range of problems [70] and especially for continuous emotion recognition [71]. The scikit-learn [69] implementation of random forests was used.

Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) : We utilize LSTM-RNN as our third baseline method, owing to their popularity and their ability to learn long-range contextual information for sequential patterns [72], [73] and their successful application for continuous emotion recognition [74], [75]. To implement the LSTM-RNN models, we utilized the CURRENNT [76] toolkit.

4.1 Feature extraction

In this section, we describe in detail the feature extraction procedure. For video features, we used appearance-based and geometric-based features. For audio features, low level descriptors (LLDs) were used.

4.1.1 *Appearance features*

To model appearance, we used dense SIFT [77], which are much more robust than raw pixels. After facial landmarks have been detected, images are normalised in term of similarity transformation (translation, scaling and rotation). Dense SIFT features with 8-bins are then extracted from patches of size 11x11 around each of the facial landmarks. The resulting descriptors therefore encode both geometric and appearance features. We reduced the dimensionality of these feature vectors by applying Principle Component Analysis (PCA). In particular, we kept the 300 first component with the highest associated eigen-values to obtain a lower-dimensionality subspace on which we then project the appearance vectors to obtain a compact but informative facial representation.

4.1.2 *Geometric features*

We also use geometric information directly obtained from the detected facial landmarks (shape features). After variations due to translation, scaling and in-plane rotation have been removed, the feature vector is then represented by the coordinates $[x_k, y_k]$ for $k \in \{1, ..., 49\}$ of the facial landmarks, stacked into a vector $(x_1, y_1, \cdots, x_{49}, y_{49}) \in (R)^{98}$.

In particular, our shape normalization follows the approach [78], [79], [80] and leverages a linear shape model built from images annotated with $u = 68$ fiducial points. The annotated shapes are first normalized using Procrustes Analysis to remove variations due to similarity transformations (that is translation, rotation and scaling). From these we then obtain the aligned mean shape $\mathbf{s}_0$. To model similarity variations, we then explicitly construct 4 bases from $\mathbf{s}_0$ compactly represented as columns of $\mathbf{Q}^{2u \times 4}$. Given a shape $\mathbf{s}_t \in \mathcal{R}^{2u \times 1}$ a shape feature vector detected at frame t, the similarity normalized features is then given by $\mathbf{s}_{\text{sim}} = \mathbf{s}_y - \mathbf{Q}\mathbf{Q}^T(\mathbf{s}_y - \mathbf{s}_0)$.

4.1.3 Audio features

We used the established [81] as our audio feature sets. For each audio recording, we capture the acoustic LLDs with the openSMILE toolkit [82] at a step size of $10\,ms$. Specifically, we extract frame-wise LLDs based on two different sets, namely, the *Interspeech 2013 Computational Para-linguistics Challenge (*ComParE*) set* and the *Geneva Minimalistic Acoustic Parameter Set (*GeMAPS*)*, detailed descriptions of which will be given, respectively.

(ComParE) consists of 6 373 acoustic features [51], [83], [84]. It contains 65 LLDs, covering spectral, cepstral, prosodic and voice quality information, which are summarised in Table 6. From these LLDs extracted from each frame ($20\,ms - 60\,ms$) of the audio signal, the first order derivatives (deltas) are computed and then functionals, such as, e. g., moments and percentiles, are applied to each frame-level LLD and its delta coefficient over the whole audio signal, to form the ComParE feature set. In the feature sets provided with the SEWA database, however, deltas and functionals are not applied to enable the user to perform time-continuous emotion recognition.

TABLE 6: INTERSPEECH 2013 Computational Paralinguistics Challenge feature set. Overview of 65 acoustic low-level descriptors (LLDs)

4 energy related LLD	**Group**
Loudness	Prosodic
Modulation loudness	Prosodic
RMS energy, zero-crossing rate	Prosodic
55 spectral related LLD	**Group**
RASTA auditory bands 1-26	Spectral
MFCC 1-14	Cepstral
Spectral energy 250-650 Hz, 1-4 KHz	Spectral
Spectral roll-off Pt. .25, .50, .75, .90	Spectral
Spectral flux, entropy, variance	Spectral
Spectral skewness and kurtosis	Spectral
Spectral slope	Spectral
Spectral harmonicity	Spectral
Spectral sharpness (auditory)	Spectral
Spectral centroid (linear)	Spectral
6 voicing related LLD	**Group**
F_0 via SHS	Prosodic
Probability of voicing	Voice quality
Jitter (local and delta)	Voice quality
Shimmer	Voice quality
Log harmonics-to-noise ratio	Voice quality

The second acoustic feature set provided is based on GeMAPS [52], a minimalistic expert-knowledge based feature set for the acoustic analysis of speaker states and traits. Compared with large-scale sets, such as ComParE, its main aim is to reduce the risk of over-fitting in the training phase. GeMAPS contains a compact set of 18 LLDs, covering spectral, prosodic and voice quality information, cf. Table 7. The LLDs were selected with respect to their capability to describe affective physiological changes in voice production.

For the baseline experiments, described next, the ComParE LLDs were summarized over a block of 6 seconds computing the *mean* and the *standard deviation* of each LLD resulting in a feature vector of dimension 130. This is done as a single LLD frame does not convey meaningful information about the affective state of a speaker. Using ComParE only is justified by the fact that the LLDs in the eGeMAPS set are mostly redundant and the results achieved are not superior on average [52].

TABLE 7: Geneva Minimalistic Acoustic Parameter Set. Overview of 18 acoustic low-level descriptors (LLDs)

6 frequency related LLD	**Group**
Pitch	Prosodic
Jitter	Voice quality
Formant 1, 2, 3 frequency	Voice quality
Formant 1 bandwidth	Voice quality
3 energy related LLD	**Group**
Shimmer	Voice quality
Loudness	Prosodic
Harmonics-to-Noise ratio	Voice quality
9 spectral related LLD	**Group**
α ratio	Spectral
Hammarberg Index	Spectral
Spectral slope 0-500 Hz and 500-1500Hz	Spectral
Formant 1, 2, 3 relative energy	Voice quality
Harmonic difference H1-H2, H1-A3	Voice quality

4.2 Experimental setting

Extensive baseline experiments were conducted in two different settings:

Multi-culture, person independent experiment –coined *multi*–: This is the generic context in which we perform experiments on all cultures mixed (i. e., training and testing on all cultures) but in a person-independent way.

Culture independent –coined $C1, \cdots, C6$: The goal of this setting is to test performance in a culture-specific manner (English, German, Hungarian, etc.). In this case, for each culture, the data of that culture was divided into person independent training, validation and testing sets.

In both cases, we ensured that the split of the data was person-independent by manually dividing the data into subject-independent training, development, and test partitions with a 3:1:1 ratio. All partitions were balanced with respect to age, gender and the criteria after that the segments have been selected, to make sure that we do not end up, e. g. in too many segments of liking in the training partition and in only segments of disliking in the test partition.

For each experiment, we optimised the model parameters by performing a grid-search on the development set to find the best setting of the regularization parameter C for the SVR and number of trees n for the Random Forest, and report results on the testing set.

4.2.1 Performance measure

The problem of AU detection is a classification one while that of valence, arousal and liking/disliking level estimation is a regression one, mandating different error measures. Given a ground-truth and a prediction, for Action Units, we measure performance with the F_1 score. The f_1 score is defined as:

$$f_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

This score is widely used for AU-detection and classification of facial expressions of emotions [41], [60] because of its robustness to the imbalance in positive and negative samples, which is very common in the case of AUs.

For valence, arousal and liking/disliking, performance is measured using the Pearson product-moment correlation coefficient (*CORR*), which is the standard measures used for measuring valence and arousal estimation accuracy [19]. We also report the Concordance Correlation Coefficient (*CCC*), recently used in the last AVEC competitions [66], [85].

The correlation coefficient (CORR) is defined as follows. Let θ be a series of n ground-truth labels, $n \in \mathcal{N}$ and $\hat{\theta}$ a series of n corresponding prediction labels.

$$\text{CORR}(\hat{\theta}, \theta) = \frac{\text{COV}(\hat{\theta}, \theta)}{\sigma_{\hat{\theta}} \sigma_{\theta}} = \frac{E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}} \sigma_{\theta}} \tag{2}$$

Finally, the concordance correlation coefficient (CCC) is defined as:

$$\text{CCC}(\hat{\theta}, \theta) = \frac{2 \times \text{COV}(\hat{\theta}, \theta)}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2}, = \frac{2E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2}, \tag{3}$$

4.3 Experimental results

Here we present the experimental results for action unit detection and valence, arousal and liking/disliking estimation.

4.3.1 Action unit detection

We used the SVM and the Random Forest for AU detection using geometric and appearance features and feature fusion as described in section 4.1. The results in terms of F1-score for per-frame detection are shown in Table 8 on the test set, and in Table 9 on the development set. The tables show that the AU detector perform well, but clearly not good enough for AU detection in a fully automatic manner. AU 12 has the highest F1-score (0.618) with feature fusion and SVM classifier. These results demonstrate again that it is important to use both types of features, texture and appearance, to achieve superior results. In particular, and in line with previous research, the average results achieved by landmarks are higher than those by texture features which confirms the representative power of geometric features. In comparison to the baseline results with those in the FERA2015 [62] database, our results obtained here are lower on the overlapping AUs (10 and 17). This is mainly because the SEWA dataset contains facial expressions recorded in different contexts and in the wild, while the FERA2015 recordings are made in an controlled environment or laboratory with controlled noise level, illumination and calibrated cameras.

4.4 Estimation of valence, arousal and liking/disliking

The setting of our baseline experiment allows us to investigate the effect of audio, video and the fusion of both on the results. In addition, we are able to separate the effect of culture on the results. As annotations were performed *separately* but by the *same annotators* on the audio, video and audio-video feeds respectively, we are also able to infer the human-level-performance of recognizing the levels of valence and arousal displayed by a subject given each type of information. Results are reported in term of CORR in Table. Table.11 and in term of CCC in Table.10 .

TABLE 8: F1-score for AU detection on the test partition

	Landmarks		SIFT		Fusion	
AU	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**
1	0.401	0.285	0.512	0.265	0.514	0.272
2	0.323	0.415	0.300	0.211	0.293	0.275
4	0.409	0.123	0.345	0.265	0.345	0.183
12	0.513	0.492	0.518	0.321	0.613	0.421
17	0.361	0.068	0.303	0.177	0.302	0.247
av	0.385	0.251	0.378	0.226	0.407	0.271

TABLE 9: F1-score for AU detection on the development partition

	Landmarks		SIFT		Fusion	
AU	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**
1	0.345	0.198	0.477	0.161	0.479	0.301
2	0.583	0.470	0.406	0.276	0.404	0.271
4	0.405	0.255	0.461	0.290	0.460	0.289
12	0.533	0.421	0.588	0.413	0.618	0.431
17	0.419	0.282	0.271	0.297	0.271	0.246
av	0.432	0.296	0.417	0.261	0.429	0.290

Results show that, for valence, we obtained better results on annotation obtained using exclusively video. These results are slightly lower when using labels obtained by annotating audio-video, while the worst results were obtained on the labels collected from the audio feed only. This is in line with the recent finding by psychologists that valence is much better estimated from video imagery than from audio only, while arousal is much better predicted from audio than from video [86], [87]. As expected, using a fusion of audio-video features increases the results, while audio features are the least helpful, supporting the theory that the face and its deformation is the main medium of communication between humans when it comes to emotions.

On a model level, we observe that performances of different regression models vary from each other and that, overall, SVM performs better than RF which in turn outperforms LSTM. However, perhaps surprisingly, in the experiments on audio features, while Support Vector Machine for Regression (SVR) and Random Forest (RF) are expected to perform well for arousal and valence prediction, Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) noticeably outperform them for liking/disliking prediction. For example, while the average CCC of liking prediction based on audio features and A+V annotations is 0.194 and 0.087 by SVR and RF, respectively, a CCC of 0.254 is achieved with LSTM.

Still when it comes to audio features, in most cases, arousal is better predicted than valence, which conforms repeated findings in the literature [66], [67], [68]. Conversely, when using video features, valence seems to be more accurately predicted than arousal. These observation would confirm that acoustic features are more informative for arousal while valence can result in more subtle facial expressions requiring geometric and appearance features to be predicted accurately.

However, for liking or disliking, there is no such notice-

			Valence			Arousal			Liking/Disliking		
Method	Feature	Annotation:	A	V	AV	A	V	AV	A	V	AV
SVM	A	C1	0.132	0.170	−0.070	0.288	0.282	0.193	0.301	0.008	−0.072
		C2	0.013	0.396	0.193	0.611	−0.002	0.083	0.104	0.178	0.224
		C3	0.251	0.087	0.116	0.619	0.545	0.694	−0.119	−0.100	0.002
		C4	−0.069	0.149	−0.002	0.101	−0.222	0.043	0.496	0.132	0.640
		C5	−0.106	0.398	0.109	0.067	0.243	0.352	0.279	0.141	0.046
		C6	0.027	0.209	0.230	0.488	0.494	0.402	−0.253	0.186	0.287
		multi	0.123	0.297	0.196	0.427	0.351	0.263	0.144	0.228	0.229
		avr.	0.053	0.244	0.110	0.372	0.242	0.290	0.136	0.110	0.194
	V	C1	0.193	0.427	0.294	0.097	0.228	0.228	0.163	0.299	0.342
		C2	0.238	0.154	0.203	0.174	0.457	0.337	0.080	−0.060	0.188
		C3	−0.081	0.278	0.215	0.055	0.386	0.193	0.101	0.417	0.027
		C4	0.005	0.301	0.376	0.244	0.236	0.376	0.231	0.165	0.390
		C5	0.155	0.495	0.252	0.019	0.271	0.150	0.494	0.170	0.406
		C6	0.038	0.264	0.171	0.099	0.268	0.297	0.004	0.325	−0.036
		multi	0.195	0.312	0.194	0.249	0.202	0.172	0.117	0.154	0.048
		avr.	0.106	0.319	0.244	0.134	0.293	0.250	0.170	0.210	0.195
	AV	C1	0.268	0.445	0.305	0.116	0.224	0.255	0.188	0.436	0.268
		C2	0.215	0.184	0.264	0.166	0.501	0.405	0.076	−0.096	0.231
		C3	0.057	0.359	0.284	0.100	0.448	0.296	0.195	0.374	0.112
		C4	0.063	0.282	0.320	0.198	0.183	0.179	0.055	0.254	0.341
		C5	0.188	0.468	0.236	0.095	0.261	0.217	0.452	0.165	0.406
		C6	−0.020	0.296	0.212	0.035	0.229	0.190	0.242	0.294	−0.050
		multi	0.171	0.326	0.199	0.267	0.164	0.175	0.103	0.148	0.054
		avr.	0.135	0.337	0.260	0.140	0.287	0.245	0.187	0.225	0.195
RF	A	C1	0.203	0.116	0.018	0.341	0.169	0.235	0.264	−0.070	0.004
		C2	0.060	0.117	−0.037	−0.023	0.433	−0.141	−0.029	0.184	0.007
		C3	0.165	0.054	0.105	0.665	0.388	0.531	−0.191	−0.066	0.069
		C4	−0.130	−0.175	−0.017	0.067	0.054	0.090	0.303	−0.016	0.182
		C5	0.002	0.154	0.206	0.007	0.217	0.062	0.055	0.111	0.203
		C6	0.115	0.238	0.155	0.560	0.370	0.363	0.002	0.106	0.088
		multi	0.104	0.156	0.078	0.294	0.288	0.223	0.078	0.121	0.059
		avr.	0.074	0.094	0.073	0.273	0.274	0.195	0.069	0.053	0.087
	V	C1	0.002	0.366	0.151	0.104	0.123	0.166	0.324	0.109	0.319
		C2	0.115	−0.007	0.099	0.127	0.511	0.204	0.156	−0.042	−0.070
		C3	0.123	0.238	0.177	0.129	0.104	0.255	−0.078	0.021	0.158
		C4	0.047	0.387	0.324	0.050	0.234	0.193	0.185	0.027	0.304
		C5	0.037	0.259	0.129	−0.016	0.265	0.134	0.114	−0.093	−0.035
		C6	0.162	0.358	0.396	0.077	0.174	0.139	0.078	0.247	0.075
		multi	0.034	0.207	0.192	0.047	0.123	0.127	−0.023	0.062	0.077
		avr.	0.074	0.258	0.210	0.074	0.219	0.174	0.108	0.047	0.118
	AV	C1	0.135	0.358	0.091	0.027	0.225	0.056	0.347	0.293	0.376
		C2	0.064	0.126	0.133	0.145	0.355	0.158	−0.064	0.136	−0.081
		C3	0.064	0.193	0.226	0.056	0.120	0.115	−0.029	0.177	0.083
		C4	0.058	0.374	0.256	0.084	0.152	0.162	0.130	0.045	0.347
		C5	0.043	0.262	0.119	0.052	0.241	0.079	0.109	−0.133	−0.121
		C6	0.103	0.360	0.335	0.028	0.173	0.089	0.078	0.272	0.032
		multi	0.023	0.220	0.137	0.043	0.143	0.125	−0.065	0.113	0.071
		avr.	0.070	0.270	0.185	0.062	0.201	0.112	0.072	0.129	0.101
LSTM	A	C1	0.165	0.135	0.120	0.347	0.099	0.321	0.415	0.206	0.269
		C2	0.317	0.188	0.216	0.212	0.072	0.253	0.222	0.146	0.164
		C3	0.249	0.152	0.226	0.669	0.298	0.540	0.108	0.118	0.404
		C4	0.132	0.251	0.300	0.120	0.110	0.210	0.173	0.149	0.295
		C5	0.121	0.381	0.279	0.116	0.115	0.239	0.349	0.324	0.407
		C6	0.238	0.326	0.304	0.500	0.532	0.616	0.219	0.141	0.087
		multi	0.118	0.212	0.082	0.346	0.296	0.234	0.215	0.206	0.151
		avr.	0.192	0.235	0.218	0.330	0.218	0.344	0.243	0.184	0.254
	V	C1	0.076	0.112	0.074	0.106	0.180	0.112	0.197	0.146	0.243
		C2	0.167	0.084	0.151	0.221	0.106	0.125	0.248	0.052	0.144
		C3	0.116	0.168	0.190	0.171	0.187	0.240	0.055	0.252	0.136
		C4	0.109	0.010	0.081	0.083	0.120	0.192	0.169	0.096	−0.022
		C5	0.212	0.325	0.171	0.225	0.243	0.254	0.251	0.256	0.259
		C6	0.295	0.245	0.158	0.199	0.137	0.204	0.107	0.129	0.241
		multi	0.091	0.281	0.153	0.115	0.115	0.119	0.164	0.076	0.086
		avr.	0.152	0.175	0.140	0.160	0.155	0.178	0.170	0.144	0.155
	AV	C1	0.236	0.238	0.231	0.244	0.238	0.195	0.181	0.114	0.172
		C2	−0.008	0.128	0.067	0.093	0.172	0.187	0.003	0.015	−0.000
		C3	0.202	0.160	0.232	0.168	0.128	0.128	−0.036	0.120	0.270
		C4	0.107	0.056	0.138	0.122	0.141	0.186	0.095	0.026	0.055
		C5	0.101	0.241	0.243	0.111	0.105	0.251	0.153	0.179	0.201
		C6	0.093	0.140	0.254	0.178	0.221	0.287	0.234	0.064	0.120
		multi	0.119	0.228	0.195	0.167	0.112	0.162	0.095	0.064	0.065
		avr.	0.121	0.170	0.194	0.155	0.160	0.199	0.103	0.083	0.126

TABLE 10: Results in term of CCC for valence, arousal and liking/disliking

			Valence			Arousal			Liking/Disliking		
Method	Feature	Annotation:	A	V	AV	A	V	AV	A	V	AV
SVM	A	C1	0.139	0.177	−0.078	0.336	0.305	0.218	0.420	−0.012	−0.012
		C2	−0.178	0.464	−0.161	0.628	−0.002	0.035	0.049	0.217	−0.205
		C3	0.401	0.089	0.139	0.671	0.523	0.688	0.010	0.306	0.167
		C4	−0.071	−0.168	−0.004	0.299	−0.085	0.199	0.625	0.186	0.641
		C5	−0.118	0.557	0.051	0.367	0.433	0.396	0.351	0.547	0.198
		C6	−0.023	0.262	0.330	0.695	0.536	0.533	0.003	0.151	0.103
		multi	0.165	0.332	0.251	0.370	0.362	0.259	0.171	0.247	0.280
		avr.	0.045	0.245	0.075	0.481	0.296	0.333	0.233	0.235	0.167
	V	C1	0.197	0.433	0.347	0.097	0.272	0.266	0.201	0.333	0.397
		C2	0.231	0.189	0.222	0.235	0.460	0.444	0.051	−0.069	0.179
		C3	−0.059	0.334	0.229	−0.008	0.443	0.229	0.117	0.542	0.034
		C4	0.182	0.403	0.463	0.338	0.349	0.254	0.140	0.236	0.432
		C5	0.167	0.470	0.301	0.022	0.289	0.220	0.570	0.208	0.408
		C6	0.173	0.389	0.319	0.102	0.348	0.235	0.001	0.431	0.027
		multi	0.171	0.321	0.266	0.281	0.182	0.209	0.132	0.134	0.024
		avr.	0.152	0.363	0.307	0.152	0.335	0.265	0.173	0.259	0.214
	AV	C1	0.266	0.463	0.325	0.118	0.308	0.318	0.232	0.430	0.477
		C2	0.247	0.109	0.239	0.222	0.423	0.351	0.081	−0.107	0.160
		C3	−0.137	0.407	0.317	0.130	0.437	0.273	0.229	0.467	0.122
		C4	−0.064	0.392	0.434	0.327	0.355	0.325	0.206	0.319	0.383
		C5	0.186	0.489	0.370	0.097	0.258	0.290	0.500	0.185	0.432
		C6	0.062	0.270	0.120	0.032	0.280	0.231	0.102	0.411	0.050
		multi	0.203	0.333	0.268	0.282	0.171	0.169	0.192	0.157	0.057
		avr.	0.109	0.352	0.296	0.173	0.319	0.280	0.220	0.266	0.240
RF	A	C1	0.289	0.162	0.025	0.387	0.242	0.334	0.519	−0.107	0.008
		C2	0.156	0.134	−0.112	−0.044	0.562	−0.384	−0.060	0.385	0.013
		C3	0.328	0.079	0.166	0.772	0.492	0.647	−0.424	−0.214	0.102
		C4	−0.180	−0.324	−0.087	0.135	0.123	0.180	0.523	−0.022	0.454
		C5	0.006	0.502	0.378	0.030	0.472	0.139	0.199	0.375	0.662
		C6	0.171	0.341	0.287	0.682	0.556	0.573	0.003	0.156	0.154
		multi	0.227	0.268	0.140	0.399	0.411	0.312	0.168	0.198	0.116
		avr.	0.142	0.166	0.114	0.337	0.408	0.257	0.133	0.110	0.216
	V	C1	0.139	0.392	0.266	0.144	0.138	0.094	0.346	0.090	0.526
		C2	0.128	0.049	0.158	0.241	0.512	0.268	0.259	−0.093	−0.054
		C3	0.205	0.301	0.278	0.178	0.136	0.294	0.050	−0.124	0.255
		C4	0.127	0.393	0.358	0.024	0.315	0.148	0.415	−0.137	0.393
		C5	0.110	0.336	0.239	−0.075	0.261	0.145	0.142	−0.237	0.033
		C6	0.104	0.375	0.455	0.054	0.219	0.213	0.240	0.316	0.081
		multi	0.081	0.268	0.227	0.077	0.140	0.181	0.056	0.155	0.050
		avr.	0.128	0.302	0.283	0.092	0.246	0.192	0.215	−0.004	0.183
	AV	C1	0.226	0.457	0.090	0.059	0.251	0.173	0.261	0.356	0.505
		C2	0.107	0.188	0.212	0.211	0.460	0.222	−0.111	0.079	0.062
		C3	0.098	0.303	0.233	0.175	0.157	0.191	−0.124	0.227	0.055
		C4	0.042	0.341	0.353	0.137	0.219	0.183	0.239	0.053	0.475
		C5	0.162	0.318	0.177	−0.029	0.399	0.218	0.121	0.123	−0.182
		C6	0.110	0.416	0.380	0.102	0.223	0.149	0.144	0.343	0.128
		multi	0.028	0.243	0.201	0.097	0.164	0.197	0.040	0.147	0.113
		avr.	0.110	0.324	0.235	0.107	0.268	0.190	0.081	0.190	0.165
LSTM	A	C1	0.197	0.144	0.125	0.390	0.115	0.346	0.518	0.221	0.296
		C2	0.331	0.377	0.348	0.345	0.115	0.462	0.453	0.347	0.385
		C3	0.306	0.159	0.258	0.680	0.304	0.569	0.299	0.190	0.560
		C4	0.136	0.318	0.354	0.162	0.135	0.276	0.220	0.206	0.350
		C5	0.142	0.525	0.288	0.227	0.158	0.282	0.463	0.492	0.501
		C6	0.328	0.365	0.307	0.649	0.661	0.672	0.303	0.158	0.123
		multi	0.154	0.221	0.101	0.400	0.346	0.285	0.243	0.225	0.200
		avr.	0.228	0.301	0.254	0.408	0.262	0.413	0.357	0.263	0.345
	V	C1	0.105	0.137	0.094	0.138	0.236	0.139	0.238	0.153	0.260
		C2	0.403	0.119	0.274	0.259	0.115	0.162	0.389	0.107	0.467
		C3	0.143	0.188	0.227	0.208	0.190	0.308	0.097	0.258	0.218
		C4	0.155	0.013	0.103	0.125	0.185	0.271	0.228	0.135	−0.035
		C5	0.235	0.384	0.188	0.237	0.269	0.383	0.254	0.485	0.261
		C6	0.338	0.310	0.195	0.321	0.150	0.275	0.139	0.145	0.285
		multi	0.135	0.322	0.187	0.140	0.148	0.173	0.186	0.110	0.105
		avr.	0.216	0.210	0.181	0.204	0.185	0.244	0.219	0.199	0.223
	AV	C1	0.329	0.293	0.286	0.250	0.330	0.239	0.195	0.132	0.178
		C2	−0.014	0.134	0.158	0.104	0.193	0.251	0.007	0.036	−0.001
		C3	0.207	0.218	0.266	0.200	0.136	0.147	−0.038	0.129	0.272
		C4	0.144	0.070	0.189	0.158	0.175	0.223	0.113	0.029	0.090
		C5	0.157	0.286	0.273	0.138	0.114	0.303	0.205	0.277	0.266
		C6	0.093	0.202	0.254	0.268	0.257	0.352	0.257	0.064	0.132
		multi	0.150	0.252	0.232	0.187	0.135	0.202	0.127	0.079	0.099
		avr.	0.152	0.208	0.237	0.187	0.191	0.245	0.124	0.106	0.148

TABLE 11: Results in term of PCC for valence, arousal and liking/disliking

able tendency. In most cases, as can be seen from the table, the performance for liking or disliking is lower than for arousal and valence. This could be mainly because the prediction of liking and disliking is more content-related and could not obtain sufficient useful information via acoustic cues only, lacking linguistic cues.

Moreover, regarding the three different types of annotations, we also note that, in most cases the best performance in terms of CCC was obtained by audio-based annotations for arousal and by video-based ones for valence, respectively, while no obvious performance improvement was seen when the combination of audio and video was provided during annotations. However, for liking or disliking, in many cases the best results of prediction of liking were achieved when the audio/video-based annotations were utilised. This may be because prediction of liking or disliking is a quite complex problem which is difficult to address with limited data. It could be improved when more data with information of multiple modalities is given.

Using video features, culture 5 (Hungarian) is best predicted (with a CCC of 0.495) for valence using SVR. Interestingly, this same culture is also best predicted for valence using audio features (CCC 0.398), again with SVR based on audio features but video-based annotation. For arousal and using audio features only, regarding the six different cultures, the performance in term of CCC (0.694) is obtained for culture 3 (German) with SVR on audio/video-based annotations. In contrast, using a fusion of audio *and* video features, best results are obtained for arousal on culture 2 with a CCC of 0.501.

It is interesting to notice that, considering experiments of SVR, predictions of valence with video-based annotations outperforms that with audio-based annotations for all cultures except for culture 3 (German). Similarly, predictions of arousal with audio-based labels outperforms that with video-based labels for all cultures except for culture 5 (Hungarian) and 6 (Serbian). Such a contrast could be mainly due to the close connection between the two dimensions in spontaneous conversation. Therefore, it might be good to predict them together, i. e., conducting multi-task learning to take advantage of the interconnections between the two different aspects.

5 Conclusion

We introduced the SEWA database (SEWA DB), a multilingual dataset of annotated facial, vocal and verbal behaviour recordings made in-the-wild. In addition to providing training data for the technologies developed during the SEWA projects, the SEWA DB has also been made publicly available to the research community, representing a benchmark for efforts in automatic analysis of audio-visual behaviour in the wild. The SEWA DB contains the recordings of 204 experiment sessions, covering 408 subjects recruited from 6 different cultural backgrounds: British, German, Hungarian, Greek, Serbian, and Chinese. The database includes a total of 1525 minutes of audio-visual recordings of the subjects reaction to the 4 advertisement stimuli and 568 minutes of video-chat recordings of the subjects discussing the advertisement. In addition to the raw audio and video data, the SEWA DB also contains a wide range of annotations including: low-level audio descriptor (LLD) features, facial landmark locations, hand-gesture, head gesture, facial action units, audio transcript, continuously-valued valence, arousal and liking / disliking (toward the advertisement), template behaviours, agreement / disagreement episodes, and mimicry episodes.

We provide exhaustive baseline experiments to assess Action Unit detection and valence, arousal and liking/disliking prediction, which is both helpful in advancing the field of affect estimation and will help advance the state-of-the art by providing a comparison benchmark.

We believe this large corpus will be helpful to the community, both in the psychological field in helping test hypothesis and in the computer science field to advance the state of automatic sentiment analysis in the wild.

Acknowledgments

This work was funded by the European Community Horizon 2020 [H2020/2014- 2020] under grant agreement no. 645094 (SEWA).

References

[1] S. Brave and C. Nass, "The human-computer interaction handbook," J. A. Jacko and A. Sears, Eds. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2003, ch. Emotion in Human-computer Interaction, pp. 81–96.

[2] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[3] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 43, 2015.

[4] W. E. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions," *Psychological bulletin*, vol. 95, no. 1, pp. 52–77, 01 1984.

[5] Z. Ambadar, J. F. Cohn, and L. I. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *Journal of nonverbal behavior*, vol. 33, no. 1, p. 1734, March 2009. [Online]. Available: http://europepmc.org/articles/PMC2701206

[6] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *ICME*. IEEE, 2008, pp. 865–868.

[7] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[8] C. Georgakis, Y. Panagakis, S. Zafeiriou, and M. Pantic, "The conflict escalation resolution (confer) database," *Image and Vision Computing*, 2017.

[9] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic, "The mahnob mimicry database: A database of naturalistic human interactions," *Pattern recognition letters*, vol. 66, pp. 52–61, 2015.

[10] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception," *Emotion*, vol. 12, no. 2, pp. 1161–1179, 2012.

[11] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijohlt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[13] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "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, 2013, pp. 1–8.

[14] P. Ekman and D. Cordaro, "What is meant by calling emotions basic," *Emotion Review*, vol. 3, no. 4, pp. 364–370, 2011.

[15] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 28, no. 4, pp. 384–392, 1993.

[16] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[18] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[19] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[20] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[21] M. Pantic and L. J. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

[22] D. Bone, C.-C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A Rule-Based framework with knowledge-inspired vocal features," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 201–213, 2014.

[23] S. B.-C. Ofer Golan, Yana Sinai-Gavrilov, "The Cambridge mindreading face-voice battery for children (CAM-C): Complex emotion recognition in children with and without autism spectrum conditions," *Molecular Autism*, vol. 22, no. 6, 2015.

[24] E. Marchi, Y. Zhang, F. Eyben, F. Ringeval, and B. Schuller, "Autism and Speech, Language, and Emotion - a Survey," in *Evaluating the Role of Speech Technology in Medical Case Management*, H. Patil and M. Kulshreshtha, Eds. Berlin: De Gruyter, 2015.

[25] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the 1st Challenge," *Speech Communication, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Process.*, vol. 53, no. 9/10, pp. 1062–1087, Nov./Dec. 2011.

[26] P. Lang and M. M. Bradley, "The international affective picture system (iaps) in the study of emotion and attention," *Handbook of emotion elicitation and assessment*, vol. 29, 2007.

[27] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[28] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing Journal*, vol. 31, no. 2, pp. 186–202, February 2013.

[29] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[30] S. Walter, J. Kim, D. Hrabal, S. C. Crawcour, H. Kessler, and H. C. Traue, "Transsituational individual-specific biopsychological classification of emotions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 4, pp. 988–995, 2013.

[31] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic cca for analysis of affective behaviour and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[32] F. Zhou and F. De la Torre, "Generalized canonical time warping," *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 38, no. 2, pp. 279–294, 2016.

[33] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," in *Proc. 7th European Conference on Speech Communication and Technology*, 2001.

[34] S. Meudt, L. Bigalke, and F. Schwenker, "Atlas–an annotation tool for hci data utilizing machine learning methods," *Proc. of APD*, vol. 12, pp. 5347–5352, 2012.

[35] R. Böck, I. Siegert, M. Haase, J. Lange, and A. Wendemuth, "ikannotate–a tool for labelling, transcription, and annotation of emotionally coloured speech," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 25–34.

[36] S. Scherer, I. Siegert, L. Bigalke, and S. Meudt, "Developing an expressive speech labelling tool incorporating the temporal characteristics of emotion." in *LREC*, 2010.

[37] 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 *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[38] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), FG*, Shanghai, China, 2013, 8 pages.

[39] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23 – 36, 2017, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

[40] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence, Special Issue in Multimodal Pose Estimation and Behaviour Analysis*, 2016.

[41] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[42] K. R. Scherer and G. Ceschi, "Lost luggage: a field study of emotion–antecedent appraisal," *Motivation and emotion*, vol. 21, no. 3, pp. 211–235, 1997.

[43] F. Schiel, S. Steininger, and U. Türk, "The SmartKom Multimodal Corpus at BAS." in *LREC*, 2002.

[44] 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*, ACM. Barcelona, Spain: ACM, Oct. 2013, pp. 3–10.

[45] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[46] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (ccdb): A database of natural dyadic conversations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 277–282.

[47] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014 – The Three Dimensional Affect and Depression Challenge," in *Proc. 4th ACM international workshop on Audio/Visual Emotion Challenge*, ACM. Orlando, FL: ACM, Nov. 2014.

[48] J. Vandeventer, A. J. Aubrey, P. L. Rosin, and D. Marshall, "4D Cardiff Conversation Database (4D CCDb): A 4D database of natural, dyadic conversations," in *Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, 2015.

[49] W. Labov, *Sociolinguistic patterns*. University of Pennsylvania Press, 1972, no. 4.

[50] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, vol. 2006, 2006, p. 5th.

[51] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[52] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[53] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1003–1011.

[54] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," pp. 1859–1866, 2014.

[55] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, and M. Pantic, "From pixels to response maps: Discriminative image filtering for face alignment in the wild," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 6, pp. 1312–1320, 2015.

[56] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 1–9.

[57] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[58] M. Wöllmer, E. Marchi, S. Squartini, and B. Schuller, "Robust multi-stream keyword and non-linguistic vocalization detection for computationally intelligent virtual agents," in *International Symposium on Neural Networks*. Springer, 2011, pp. 496–505.

[59] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The munich 2011 chime challenge contribution: Nmf-blstm speech enhancement and recognition for reverberated multisource environments."

[60] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'15)*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[61] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[62] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–8.

[63] J. Shen and M. Pantic, "Hci$\hat{2}$ framework: A software framework for multimodal human-computer interaction," *Transactions on Cybernetics*, vol. 43, no. 6, pp. 1593–1606, 2013.

[64] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," *Proc. 14th Int'l Conf. Multimodal Interaction Workshops*, pp. 449–456, 2012.

[65] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pp. 3–10, 2014.

[66] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 3–10.

[67] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Automatic Face and Gesture Recognition Workshops)*. IEEE, 2013, pp. 1–8.

[68] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[70] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*. New York, NY, USA: ACM, 2008, pp. 96–103.

[71] H. Chen, J. Li, F. Zhang, Y. Li, and H. Wang, "3d model-based continuous emotion recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1836–1845.

[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[73] A. Graves, *Supervised sequence labelling with recurrent neural networks*. Berlin/Heidelberg, Germany: Springer, 2012, vol. 385.

[74] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 3593–3597.

[75] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, 2016.

[76] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: the Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *J. Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[77] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[78] J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Fast and exact bidirectional fitting of active appearance models," in *Proceedings of the IEEE Intl Conf. on Image Processing (ICIP15)*, Quebec City, QC, Canada, September 2015, pp. 1135–1139.

[79] ——, "Fast newton active appearance models," in *Proceedings of the IEEE Intl Conf. on Image Processing (ICIP14)*, Paris, France, October 2014, pp. 1420–1424.

[80] ——, "Fast and exact newton and bidirectional fitting of active appearance models," *IEEE Transactions on Image Processing (TIP), accepted for publication*, 2016.

[81] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Springer, 2016.

[82] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM International Conference on Multimedia (ACM MM)*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[83] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load." in *Proc. of INTERSPEECH*. Singapore, Singapore: ISCA, 2014, pp. 427–431.

[84] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson & eating condition," in *Proc. of INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 478–482.

[85] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*. Brisbane, Australia: ACM, 2015, pp. 3–8.

[86] J. Russell, B. J.A, and J. Fernandez-Dols, "Facial and vocal expressions of emotions," *Annu. Rev. Psychol.*, vol. 54, pp. 329–349, 2003.

[87] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3d emotion space concept," in *Robust Speech*, M. Grimm and K. Kroschel, Eds. Rijeka: IntechOpen, 2007, ch. 16.