

# DECAF: MEG-based Multimodal Database for Decoding Affective Physiological Responses

Mojtaba Khomami Abadi, *Student Member, IEEE*, Ramanathan Subramanian, *Member, IEEE*, Seyed Mostafa Kia, *Student Member, IEEE*, Paolo Avesani, *Member, IEEE*, Ioannis Patras, *Senior Member, IEEE*, Nicu Sebe, *Senior Member, IEEE*

**Abstract**—In this work, we present **DECAF**— a multimodal dataset for **decoding** user physiological responses to **affective** multimedia content. Different from datasets such as DEAP [15] and MAHNOB-HCI [31], DECAF contains (1) Brain signals acquired using the Magnetoencephalogram (MEG) sensor, which requires little physical contact with the user's scalp and consequently facilitates naturalistic affective response, and (2) Explicit and implicit emotional responses of 30 participants to 40 one-minute music video segments used in [15] and 36 movie clips, thereby enabling comparisons between the EEG vs MEG modalities as well as movie vs music stimuli for affect recognition. In addition to MEG data, DECAF comprises synchronously recorded near-infra-red (NIR) facial videos, horizontal Electrooculogram (hEOG), Electrocardiogram (ECG), and trapezius-Electromyogram (tEMG) peripheral physiological responses. To demonstrate DECAF's utility, we present (i) a detailed analysis of the correlations between participants' self-assessments and their physiological responses and (ii) single-trial classification results for *valence*, *arousal* and *dominance*, with performance evaluation against existing datasets. DECAF also contains *time-continuous* emotion annotations for movie clips from seven users, which we use to demonstrate dynamic emotion prediction.

**Index Terms**—Emotion recognition, User physiological responses, MEG, Single-trial classification, Affective computing

## 1 INTRODUCTION

**A**FFECT recognition is a necessity in human-computer interaction. Users' demands can be implicitly inferred from their emotional state, and systems effectively responding to emotional inputs/feedback can greatly enhance user experience. However, affect recognition is difficult as humans emotions manifest both explicitly in the form of affective intonations and facial expressions, and subtly through physiological responses originating from the central and peripheral nervous system. Given that the majority of multimedia content is created with the objective of eliciting emotional reactions from viewers, representing, measuring and predicting emotion in multimedia content adds significant value to multimedia systems [1]. Approaches to predict affect from multimedia can be categorized as (i) *content-*

*centric* [10], [32], using primitive audio-visual features which cannot adequately characterize the emotion perceived by the viewer, or (ii) *user-centric*, employing facial expressions [28] and speech intonations [26], which denote a conscious and circumstantial manifestation of the emotion, or peripheral physiological responses [21], which capture only a limited aspect of human emotion.

Recently, cognition-based approaches employing imaging modalities such as fMRI and EEG to map brain signals with the induced affect [11], [15], [31] have gained in popularity, and brain signals encode emotional information complementary to multimedia and peripheral physiological signals, thereby enhancing efficacy of user-centric affect recognition. However, acquisition of high-fidelity brain signals is difficult and typically requires the use of specialized lab equipment and dozens of electrodes positioned on the scalp, which impedes naturalistic user response. Magnetoencephalogram (MEG) is a non-invasive technology for capturing functional brain activity, which requires little physical contact between the user and the sensing coil (Fig. 2), and therefore allows for (1) recording meaningful user responses, with little psychological stress and (2) compiling affective responses over long time periods. Also, MEG responses can be recorded with higher spatial resolution as compared to EEG.

In this paper, we present **DECAF**— a MEG-based multimodal database for **decoding** affective user responses. Benefiting from facile data acquisition, DE-

- Mojtaba Khomami Abadi is with the Dept. of Information Engineering and Computer Science, University of Trento, 38123 Trento and Semantic Knowledge an Innovation lab (SKIL), Telecom Italia.
- Ramanathan Subramanian is with the Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore. (Email: subramanian.r@adsc.com.sg)
- Seyed Mostafa Kia and Paolo Avesani are with the NeuroInformatics Lab, Fondazione Bruno Kessler, 38122 Trento, Italy. (Email: moskia@fbk.eu, avesani@fbk.eu)
- Ioannis Patras is with the School of Computer Science and Electronic Engg., Queen Mary University of London (QMUL). (Email: I.Patras@eecs.qmul.ac.uk)
- Nicu Sebe is with the Dept. of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy. (Email: khomami-abadi@disi.unitn.it, sebe@disi.unitn.it)

CAF comprises affective responses of 30 subjects to 36 movie clips (of length  $\mu=80$  sec,  $\sigma=20$ ) and 40 1-minute music video segments (used in [15]), making it one of the largest emotional databases available<sup>1</sup> to the affective computing community. In addition to MEG signals, DECAF contains synchronously recorded near-infra-red (NIR) facial videos, and horizontal Electrooculogram (hEOG), Electrocardiogram (ECG), and trapezius-Electromyogram (tEMG) peripheral physiological responses<sup>2</sup>. A major limitation of affective computing works [15], [21], [31] that DECAF seeks to address is the lack of benchmarking with respect to stimuli and sensing modalities. DECAF facilitates comparisons between (1) MEG vs EEG modalities for affect sensing via their performance on the DEAP database [15], and (2) music-video vs movie clips concerning their suitability for emotion elicitation.

We present analyses concerning (i) participants' self-assessment ratings for *arousal* and *valence* for music and movie stimuli, (ii) correlations between user ratings (explicit feedback) and implicitly observed MEG responses, and (iii) single-trial classification of *valence*, *arousal* and *dominance* from MEG, peripheral responses, facial activity, content-based audio visual features and fusion of these modalities. Finally, *time-continuous* emotion annotations which are useful for dynamic emotion analysis, were compiled from seven experts for the movie clips— as an application, we show dynamic emotion prediction on time-contiguous snippets from the movie clips with a model trained using these annotations and audio-visual/MEG features.

The paper is organized as follows: Section 2 overviews related work. Methodology adopted for movie clip selection is described in Section 3, while experimental protocol is detailed in Section 4. Analysis of users' self assessments is presented in Section 5, while features extracted for affect recognition are described in Section 6. Correlations between self-assessments and physiological responses along with single-trial classification results are presented in Sections 7 and 8. Dynamic emotion estimation is detailed in Section 9, and conclusions are stated in Section 10.

## 2 RELATED WORK

Creating a stimulus database for eliciting emotions is crucial towards understanding how affect is expressed in controlled lab conditions. The *actual* emotion induced upon perceiving a stimulus designed to elicit an *intended* emotion is influenced by a number of psychological and contextual factors, and can therefore be highly subjective. Consequently, ensuring that the *actual* affective response is in agreement with

the *intended* response is non-trivial, and is typically achieved in practice as follows: (1) Many affective studies assume that the entire gamut of human emotions can be represented on the valence-arousal-dominance<sup>3</sup> (VAD) space as proposed by Bradley [5], and (2) To largely ensure that the elicited and intended emotions are consistent, presentation stimuli are carefully selected based on literature, or based on 'ground truth' VA ratings acquired from a large population that evaluates them prior to the actual study.

Gross and Levenson's seminal work on affective database creation [9] evaluates the responses of 494 subjects to 250 movie clips for identifying 16 movie clips capable of evoking eight target emotions. Content-based affect recognition works [10], [32] also perform emotion analysis on movie clips/scenes. User-centric emotion recognition works have employed a variety of stimuli to elicit emotions— Joho *et al.* [12] use a combination of movie and documentary clips to evoke facial activity, which is then used for highlights detection. Use of physiological responses for recognizing affect, pioneered by Sinha and Parsons [29] to distinguish between neutral and negative imagery, has gained popularity recently. Lisetti and Nasoz [21] use movie clips and mathematical equations to evoke emotions, which are decoded from users' skin conductance, heart rate, temperature, EMG and heat flow responses. Kim and André [14] use audio music clips to induce emotions, recognized through heart rate, EMG, skin conductivity and respiration changes.

Among cognition-based approaches, the DEAP dataset [15] is compiled to develop a user-adaptive music recommender system. It contains EEG, galvanic skin response (GSR), blood volume pressure, respiration rate, skin temperature and EOG patterns of 32 viewers watching 40 one-minute music video excerpts. The MAHNOB-HCI database [31] is compiled to model emotional responses of users viewing multimedia stimuli. It contains face and upper-body video, audio, physiological and eye-gaze signals of 27 participants watching 20 emotional movie/online clips in one experiment, and 28 images and 14 short videos in another. Analyses on the DEAP and MAHNOB-HCI datasets confirm that EEG effectively encodes emotional information, especially arousal.

Examination of related works reveals that user-centered affect recognition has been achieved with diverse stimuli, reflecting the fact that human affect sensing is multimodal. However, indigenous stimuli and signals employed by each of these works provides little clarity on (1) which stimulus most effectively

3. *Valence* indicates emotion type (*pleasant* or *unpleasant*), while *arousal* denotes the intensity of emotion (*exciting* or *boring*). *Dominance* measures the extent of control on viewing a stimulus (feeling *empowered* or *helpless*) [15]. We mainly use the VA-based affect representation, shown to account for most emotional responses by Greenwald *et al.* [8] in this work.

1. <http://disi.unitn.it/~mhug/DECAF.html>

2. DECAF represents a significant extension of the dataset reported in [2], which only contains MEG and peripheral physiological responses of 18 subjects.

elicits consistent emotional responses across users, in order to maximize our understanding of affect perception and expression?, and (2) which modality best characterizes user emotional responses?— answers to these questions can increase efficacy of affect recognition approaches. DECAF is compiled with the aim of evaluating both stimuli and sensing modalities for user-centered affect recognition.

### 3 STIMULI SELECTION

One of our objectives was to compile a large database of affective movie stimuli (comparable in size to DEAP [15]) and user responses for the same. This section describes how the 36 movie clips compiled to this end were selected. Based on previous studies that have identified movie clips suited to evoke various target emotions [3], [9], we initially compiled 58 Hollywood movie segments. These clips were shown to 42 volunteers, who self-assessed their emotional state on viewing each video to provide: valence level (very negative to very positive), arousal level (very calm to very excited), and the most appropriate tag that describes the elicited emotion (Table 1).

These annotations were processed to arrive at the final set of 36 clips as follows:

(1) To ensure that the annotations are comparable, we transformed all V and A annotations using the  $z$ -score normalization.

(2) To better estimate the affective perception of annotators, we discarded the outliers from the pool of annotators for each video clip as follows: Along the V,A dimensions, we thresholded the annotations at zero to associate *high* ( $H_i$ ) and *low* ( $L_i$ ) video sets to each annotator ( $i = 1 \dots 42$ ). We then computed Jaccard distances  $D_H, D_L$  ( $42 \times 42$  matrices) between each pair of annotators  $i, j$  for the *high*, *low* sets, e.g.,  $D_H(i, j) = 1 - \frac{|H_i \cap H_j|}{|H_i \cup H_j|}$ , where  $|\cdot|$  denotes set cardinality, and cumulative distance for each annotator from peers as the sum of each row. Finally, we derived Median Absolute Deviation of the cumulative distance distribution, and those annotators more than 2.5 deviations away from the median were considered outliers as per [19]. In all, 5 and 2 outlier annotators were respectively removed for the V and A dimensions.

(3) Similar to [15], we computed  $\mu/\sigma$  from the inlier V,A ratings for each movie clip as plotted in Fig. 1, and chose 36 clips such that (a) their ratings were close to the corners of each quadrant, (b) they were uniformly distributed over the valence-arousal plane, and (c) only one clip per movie was chosen from each quadrant to avoid priming effects. Table 1 contains descriptions of the selected movie clips, while Fig. 1 presents the distribution of  $\mu/\sigma$  ratings for the original 58 clips and highlights the 36 selected clips. The mean A, V ratings listed in Table 1 are considered as *ground truth* annotations in our work. The chosen movie clips were 51.1–128.2 sec long ( $\mu = 80, \sigma = 20$ ) and were

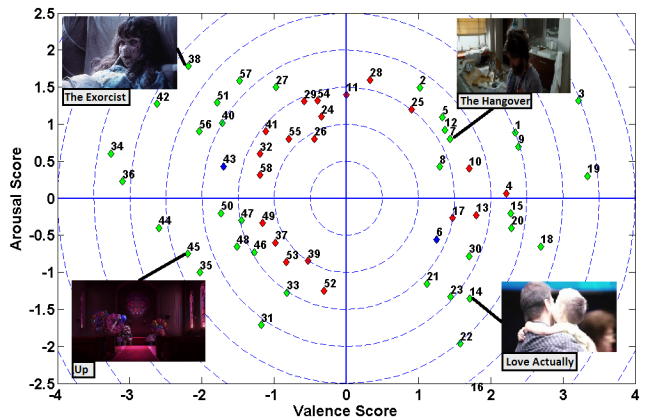


Fig. 1. Distribution of videos'  $\mu/\sigma$  ratings in the VA plane. The 36 selected videos are highlighted in green, while two introductory videos are highlighted in blue.

associated with diverse emotional tags. For benchmarking affective stimuli, we also recorded emotional responses to 40 one-minute music video used in the DEAP study [15].

### 4 EXPERIMENT SETUP

In this section, we present a brief description of (a) MEG, peripheral physiological and facial signals recorded in the study before detailing the (b) experimental set-up and protocol.

#### 4.1 MEG, peripheral physiological signals, and NIR facial videos

To collect users' implicit affective responses, we recorded (i) Magnetoencephalogram (MEG), (ii) horizontal Electrooculogram (hEOG), (iii) Electrocardiogram (ECG), (iv) Trapezius Electromyogram (tEMG) and (v) Near Infra-red (NIR) facial video signals that are described below:

**MEG:** MEG technology enables non-invasive recording of brain activity and is based on SQUIDS (Super-conducting Quantum Interference Devices), which enables recording of very low magnetic fields. Magnetic fields produced by the human brain are of the order of femtoTesla (fT) and since sensors are really sensitive to noise, the MEG equipment is located in a magnetically shielded room insulated from other electrical/metallic installations. A multiple coils configuration enables measurement of magnetic fields induced by tangential currents, and thus, brain activity in the sulci of the cortex can be recorded. We used the *ELEKTA Neuromag* device which outputs 306 channels (corresponding to 102 magnetometers and 204 gradiometers, as in Fig. 5) with a sampling frequency of 1 KHz.

Unlike in EEG, MEG sensors do not touch the subject's head and the participant can potentially make head movements during the recordings. However, due to high spatial resolution, even small head movements will cause a sensor to sense another part

TABLE 1

Description of movie clips selected for the DECAF study with their duration in seconds ( $L$ ), most frequently reported emotion tag and statistics derived from 42 annotators. Introductory videos are marked with \*\*.

Emotion	ID	Source Movie	L	Valence		Arousal		Scene Description
				$\mu$	$\sigma$	$\mu$	$\sigma$	
Amusing	01	Ace-Ventura: Pet Detective	102.1	1.22	0.53	1.03	1.00	Ace Ventura successfully hides his pets from the landlord
	02	The Gods Must be Crazy II	67.1	1.56	0.50	1.20	0.96	A couple stranded in the desert steal ostrich eggs for food
	04	Airplane	85.2	0.99	0.83	1.15	0.88	Woman and co-passengers react as pilot struggles to control aircraft
	05	When Harry Met Sally	100.2	1.05	0.61	1.08	1.02	Sally shows Harry how women fake orgasms at a restaurant
	**	Modern Times	106.4	0.87	0.69	-0.35	0.86	Bewildered factory worker in an assembly line
Funny	03	Liar Liar	55.1	0.95	0.65	0.56	0.96	Prosecution and defense discuss a divorce case in court
	06	The Gods Must be Crazy	52.1	1.26	0.56	0.81	1.15	Man tries to get past an unmanned gate on a brakeless jeep
	07	The Hangover	90.2	0.95	0.70	0.85	1.06	Group of friends on the morning after a drunken night
	09	Hot Shots	70.1	0.98	0.66	0.81	0.90	A hilarious fight sequence
Happy	08	Up	67.1	1.42	0.43	0.35	1.18	Carl- a shy, quiet boy meets the energetic Elle
	10	August Rush	90.1	0.76	0.68	-1.17	1.02	A son meets his lost mother while performing at a concert
	11	Truman Show	60.1	0.90	0.50	-1.98	0.69	Truman and his lover go to the beach for a romantic evening
	12	Wall-E	90.2	1.41	0.53	-0.82	0.91	Wall-E and Eve spend a romantic night together
	13	Love Actually	51.1	1.03	0.70	-1.38	0.80	Narrative purporting that 'Love is everywhere'
	14	Remember the Titans	52.1	0.79	0.58	-0.99	0.82	Titans win the football game
	16	Life is Beautiful	58.1	1.10	0.42	-0.16	0.79	Funny Guido arrives at a school posing as an education officer
	17	Slumdog Millionaire	80.1	0.94	0.35	-0.34	0.85	Latika and Jamal unite at the railway station
18	House of Flying Daggers	77.2	0.84	0.56	-1.79	0.88	Young warrior meets with his love with a bouquet	
Exciting	15	Legally Blonde	51.1	0.64	0.37	-0.62	0.80	Elle realizes that she has been admitted to Harvard Law School
	33	The Untouchables	117.2	-0.70	0.60	1.05	0.70	Shoot-out at a railway station
Angry	19	Gandhi	108.1	-0.50	0.67	-1.00	0.92	Indian attorney gets thrown out of a first-class train compartment
	21	Lagaan	86.1	-0.98	0.49	-0.69	0.71	Indian man is helpless as a British officer threatens to shoot him
	23	My Bodyguard	68.1	-0.81	0.59	-1.35	0.79	Group of thugs provoke a teenager
	35	Crash	90.2	-1.56	0.45	0.45	0.95	A cop molests a lady in public
Disgusting	28	Exorcist	88.1	-1.52	0.64	1.71	0.90	An exorcist inquires a possessed girl
	34	Pink Flamingos	60.2	-1.95	0.61	0.18	0.83	A lady licks and eats dog faeces
Fear	30	The Shining	78.1	-0.85	0.49	1.01	0.95	Kid enters hotel room searching for his mom
	36	Black Swan	62.2	-1.07	0.35	1.00	0.73	A lady notices paranormal activity around her
	**	Psycho	76.2	-1.23	0.73	0.44	1.01	Lady gets killed by intruder in her bath tub
Sad	20	My girl	60.1	-0.85	0.62	-0.82	1.06	Young girl cries at her friend's funeral
	22	Bambi	90.1	-0.95	0.37	-0.43	1.07	Fawn Bambi's mother gets killed by a deer hunter
	24	Up	89.1	-0.99	0.45	-0.97	0.76	Old Carl loses his bedridden wife
	25	Life is Beautiful	112.1	-0.62	0.41	-0.16	0.81	Guido is caught, and shot to death by a Nazi soldier
	26	Remember the Titans	79.1	-0.84	0.53	-0.55	0.87	Key Titans player is paralyzed in a car accident
	27	Titanic	71.1	-0.98	0.57	-0.30	0.99	Rescuers arrive to find only frozen corpses in the sea
	31	Prestige	128.2	-1.24	0.73	1.20	0.88	Lady accidentally dies during magician's act
Shock	29	Mulholland Drive	87.1	-1.13	0.55	0.82	0.97	Man shocked by suddenly appearing frightening figure
	32	Alien	109.1	-0.99	0.71	1.22	0.76	Man is taken by an alien lurking in his room

of the brain and induce changes in the MEG signal. Therefore, we asked subjects to not move their head during the recordings. To compensate for inadvertent head movements, before each recording, we attached five Head Position Indicator (HPI) coils to accurately determine the subject's head pose. Two HPI coils were attached behind the ears without being in the hair, while three coils were interspersed on the forehead. Prior to the experiment, we also recorded the subject's skull shape by sampling the 3D positions of 210 points uniformly distributed around the skull.<sup>4</sup>

**ECG:** ECG is well known for its relevance in emotion recognition [14], [15], [31]. ECG signals were recorded using three sensors attached to the participant. Two electrodes were placed on the wrist, and a reference was placed on a bony part of the arm (ulna bone). This setup allows for precise detection of heart beats, and subsequently, accurate computation of heart rate (HR) and heart rate variability (HRV).

**hEOG:** Electrooculography denotes the measurement of eye movements, fixations and blinks. In this study, we used hEOG which reflects the horizontal eye movement of users by placing two electrodes on the left and right side of the user's face close to the eyes. Zygomatic muscle activities produce high

frequency components in the bipolar EOG signal, and hence the EOG signal also captures facial activation information.

**tEMG:** Different people exhibit varying muscle movements while experiencing emotions. However, some movements are involuntary— *e.g.*, nervous twitches produced when anxious, nervous or excitable. Trapezius EMG is shown to effectively correlate with users' stress level in [33]. We placed the EMG bipolar electrodes above the trapezius muscle to measure the mental stress of users as in [14], [15]. The ECG reference electrode also served as reference for hEOG and tEMG.

**NIR Facial Videos:** As the MEG equipment needs to be electrically shielded, traditional video cameras could not be used for recording facial activity, and we therefore used a near infra-red camera for the same. Facial videos were recorded as *avi* files at 20 fps.

The ELEKTA Neuromag device accurately synchronizes MEG signals with the peripheral physiology signals. Synchronization of the NIR videos was handled by recording the sound output of the stimulus presentation PC with the user's facial videos, and using this information to determine stimulus beginning/end.

## 4.2 Experimental set-up

**Materials:** All MEG recordings were performed in a shielded room with controlled illumination. Due to sensitivity of the MEG equipment, all other devices used for data acquisition were placed in an

4. While DECAF contains HPI information, HPI-based MEG signal compensation will be attempted in future work. Since head-movement can induce noise in the MEG data, HPI MEG compensation can be useful for discarding noise and improving signal-to-noise ratio.

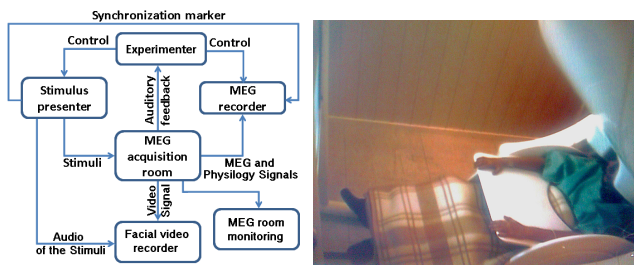


Fig. 2. (Left) Illustration of the experimental set-up. (Right) A subject performing the experiment– the stimulus is presented on the screen to the left, while the subject is seated under the MEG equipment on the right.

adjacent room, and were controlled by the experimenter. Three PCs were used, one for stimulus presentation, and two others for recording NIR videos and MEG, physiology data as seen in Fig. 2. The stimulus presentation protocol was developed using MATLAB’s Psychtoolbox (<http://psychtoolbox.org/>) and the ASF framework [27]. Synchronization markers were sent from the stimulus presenter PC to the MEG recorder for marking the beginning and end of each stimulus. All stimuli were shown at  $1024 \times 768$  pixel resolution and a screen refresh rate of 60 Hz, and this display was projected onto a screen placed about a meter before the subject inside the MEG acquisition room (Fig. 2). All music/movie clips were played at 20 frames/second, upon normalizing the audio volume to have a maximum power amplitude of 1. Participants were provided with a microphone to report their emotional state and communicate with the experimenters.

**Protocol:** 30 university graduate students (16 male, age range  $27.3 \pm 4.3$ ) participated in the experiments. Data acquisition for each participant was spread over two sessions– movie clips were presented in one session, and music videos in the other. The presentation order of the music and movie clips was counterbalanced across subjects. During each session, music/movie clips were shown in random order, such that two clips with similar valence, arousal characteristics did not follow one another. To avoid fatigue, each recording session was split into two halves (20 music/18 movie clips shown in each half) and lasted one hour. We recorded the resting state brain activity for five minutes at the beginning of each session, and for one minute at the end or before/after breaks.

**Subject Preparation:** To ensure the absence of metallic objects near the MEG equipment, prior to each recording session, participants had to change their clothing and footwear– those wearing glasses were given suitable metal-free replacements. First, participants were briefed about the experiment and asked to provide written informed consent. HPI coils were placed on their head and their head shapes

and coil positions were registered as explained in section 4.1. Once inside the MEG room, electrodes of physiological sensors were attached to participants, and by checking the impedance level of the electrodes from the MEG recorder, we made sure that they were comfortable and were positioned correctly under the MEG sensor. Participants were provided with a desk pad, pillows and blanket to relax during the experiment. We then recorded five minutes resting state brain activity while the subject was fixating on a cross at the middle of the screen. Then, two practice trials (with the videos highlighted in blue in Fig 1, and denoted using \*\* in Table 1) were conducted to familiarize subjects with the protocol.

Each acquisition session involved a series of trials. During each trial, a fixation cross was first shown for four seconds to prepare the viewer and to gauge his/her rest-state response. Upon stimulus presentation, the subject conveyed the emotion elicited in him/her to the experimenter through the microphone. Ratings were acquired for (i) Arousal (‘How intense is your emotional feeling on watching the clip?’) on a scale of 0 (very calm) to 4 (very excited), (ii) Valence (‘How do you feel after watching this clip?’) on a scale of -2 (very unpleasant) to 2 (very pleasant), and (iii) Dominance on a scale of 0 (feeling empowered) to 4 (helpless). A maximum of 15 seconds was available to the participant to convey each rating. All in all, the whole experiment (spread over two sessions) including preparation time took about three hours per subject, who was paid a participation fee of €40.

## 5 RATING ANALYSIS

### 5.1 Self-assessments: Music vs movie clips

As mentioned earlier, one objective behind compiling the DECAF database was to examine the effectiveness of different stimuli in eliciting similar emotional responses across subjects. In this section, we compare the self-assessment (or explicit) valence-arousal ratings for music and movie clips provided by the DECAF participants. Since self-reports are a conscious reflection of the user’s emotional state upon viewing the stimulus, one can expect any differences between the ratings for music and movie clips to also impact affect recognition from physiological responses.

Fig. 4 presents distributions of the AV ratings provided by the 30 DECAF participants for movie and music clips. The blue, magenta, black and red colors respectively denote high arousal-high valence (HAHV), low arousal-high valence (LAHV), low arousal-low valence (LALV) and high arousal-low valence (HALV) stimuli as per the ground-truth ratings derived from Table 1 for movie clips and [15] for music videos. A U-shape, attributed to the difficulty in evoking low arousal but strong valence responses [15], [17], is observed for both movie and music clips. The ‘U’ bend is particularly pronounced in the case of



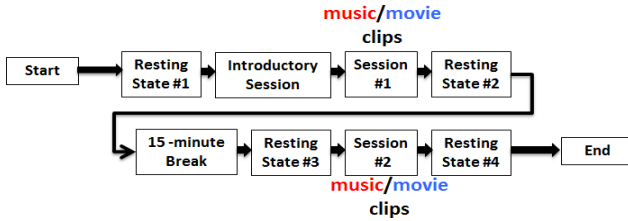


Fig. 3. Timeline for experimental protocol.

music clips, implying that a number of stimuli were perceived to be close-to-neutral in valence, and there is considerable overlap among the four quadrants. For movie clips, perfect agreement with the ground-truth is noted for valence, but cluster overlap is observed along the arousal dimension.

We performed two-sample  $t$ -tests to check if the arousal characteristics of movie/music stimuli influenced their valence ratings— these tests revealed that valence ratings differed very significantly for HA music ( $t(18) = 9.4208, p < 0.000001$ ), HA movie ( $t(16) = 13.5167, p < 0.000001$ ) clips and LA movie clips ( $t(16) = 11.586, p < 0.000001$ ), but somewhat less significantly for LA music clips ( $t(18) = 5.6999, p < 0.00005$ ). Conversely, similar significance levels were observed while comparing arousal ratings for HV music ( $t(18) = 4.2467, p < 0.0005$ ) and movie ( $t(16) = 4.2988, p < 0.0005$ ), as well as LV music ( $t(18) = -4.8256, p < 0.005$ ) and movie ( $t(16) = -3.3194, p < 0.005$ ) stimuli. Overall, the valence-arousal distinction was slightly better for movie vis-à-vis music clips.

To evaluate how consistently emotional responses were elicited across subjects, we measured agreement between the ground-truth and participant ratings using the Cohen’s Kappa measure assuming that ground-truth AV labels were provided by an ‘ideal’ annotator. To this end, we assigned high/low V, A labels to the stimuli based on each user’s median ratings, and computed  $\kappa$  between the ground-truth and user judgements. The mean  $\kappa$  over all subjects for music-valence, movie-valence, music-arousal and movie-arousal were found to be  $0.50 \pm 0.17$ ,  $0.67 \pm 0.24$ ,  $0.14 \pm 0.17$  and  $0.19 \pm 0.17$  respectively. Agreement with the ground-truth was higher for movie stimuli, implying that movie stimuli evoked intended emotions more consistently across users. Also, agreement was considerably higher for valence, indicating stronger differences in arousal perception across subjects.

## 6 DATA ANALYSIS

This section describes the procedure for data preprocessing and feature extraction from (i) MEG signals, (ii) physiology signals, (iii) face videos and (iv) multimedia signals. All the cut-off frequencies and smoothing parameters employed were adopted from [14], [15], [31]. For both MEG and peripheral physiological modalities, we computed (1) time-continuous features

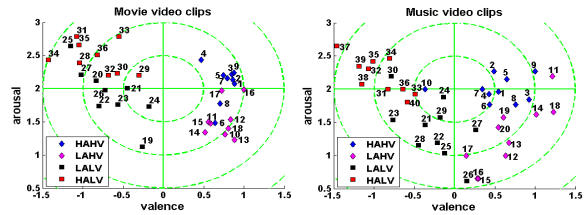


Fig. 4. Mean AV ratings for movie (left) and music clips (right) derived from DECAF participants.

for dynamic emotion analysis and (ii) statistical measures<sup>5</sup> computed over the time-continuous features, considering only the final 50 seconds.

### 6.1 MEG preprocessing and feature extraction

MEG preprocessing involved three main steps, (i) Trial segmentation, (ii) Spectral filtering and (iii) Channel correction, that were handled using the MATLAB Fieldtrip toolbox [25]. Since magnetometer outputs are prone to environmental and physiological noise, we only used the gradiometer outputs for our analysis.

**Trial Segmentation:** Participant responses corresponding to each trial were extracted by segmenting the MEG signal from 4 seconds prior to stimulus presentation (pre-stimulus) to the end of stimulus. Per subject, there were 36 and 40 trials for the movie clips and music videos respectively.

**Frequency domain filtering:** Upon downsampling the MEG signal to 300 Hz, low-pass and high-pass filtering with cut-off frequencies of 95 Hz and 1 Hz respectively were performed. The high-pass filter removes low frequency ambient noise in the signal (e.g., generated by moving vehicles). Conversely, the low-pass filter removes high frequency artifacts generated by muscle activities (between 110-150 Hz).

**Channel correction:** Dead and bad channels were removed from the MEG data. Dead channels output zero values, while bad channels are outliers with respect to metrics such as signal variance and signal amplitude  $z$ -score over time. To preserve the dimensional consistency of MEG data over all trials and subjects, removed channels were replaced with interpolations from neighboring channels.

**Time-Frequency analysis (TFA):** The spectral power in certain frequency bands has been found to contain valuable information for affect recognition in a number of EEG studies. The multitaper and wavelet transforms are typically used in order to achieve better control over frequency smoothing, and high frequency smoothing has been found to be beneficial when dealing with brain signals above 30 Hz [23]. Therefore, we used variable-width wavelets to transform the preprocessed MEG signal to the time-frequency domain for spectral power analysis.

5. mean ( $\mu$ ), standard deviation ( $\sigma$ ), skewness, kurtosis, percentage of values above  $\mu + \sigma$ , and percentage of values below  $\mu - \sigma$

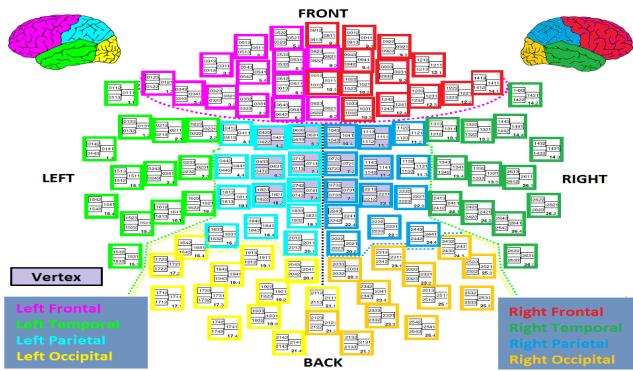


Fig. 5. Elekta Neuromag MEG channel positions. Channels corresponding to different lobes are color-coded (figure adapted from [www.megwiki.org](http://www.megwiki.org), best viewed under zoom).

**MEG-TFA Features:** We used a time-step of 1 second for temporal processing of the MEG signal from each trial, and a frequency step of 1 Hz to scan through a frequency range of 1-45 Hz. We linearly varied the wavelet width with frequency, increasing from 4 for lower frequencies to 8 for higher frequencies. Upon applying a wavelet transform on the MEG data, we performed the following steps: (a) We used a standard Fieldtrip function for combining the spectral power of each planar gradiometer pair to obtain 102 combined-gradiometer (GRAD) responses. (b) In order to better elucidate the MEG response dynamics following stimulus presentation for each subject, individual trial power was divided by a *baseline* power, obtained as the mean over two seconds pre-stimulus from all trials. (c) To increase dynamic range of the spectral power, the time-frequency output was logarithm transformed.

**Channel Grouping:** On computing the MEG spectral power over 102 GRAD pairs, in order to reduce data dimensionality while preserving spatial information, the 102 channels were divided into nine groups according to functionality of different brain regions namely: Vertex, left temporal, right temporal, left parietal, right parietal, left occipital, right occipital, left frontal and right frontal (Fig. 5). The sensors in each group encode different brain functionalities that may directly or indirectly relate to emotions, and we show that this grouping is beneficial for affect recognition in Sec. 8. Per subject and movie/music clip, time-frequency analysis outputs nine (one per group) 3D matrices with the following dimensions:  $K \times \text{clip length time points} \times 45 \text{ frequencies}$ , where  $K$  denotes the number of GRAD channels per group.

**DCT features:** The Discrete Cosine Transform (DCT) is often used in signal, image and speech compression applications due to its strong energy compaction ability. Also, the DCT feature space has been shown to efficiently compress spatio-temporal patterns of MEG data without impacting model pre-

cision [13]. We employed DCT to compress the MEG-TFA output on a per-second basis, as well as for single-trial classification. Per second, from each of the 9 lobes we extracted 60 DCT coefficients (4 along spatial and 15 along spectral respectively), and concatenated them to extract 540 DCT features. For single-trial classification, from each brain lobe, we used the first  $n = 2$  DCT coefficients from the spatial, temporal and spectral dimensions to obtain a total of  $9 \times 8 = 72$  features. We observed that classification results did not improve with  $n > 2$  DCT coefficients per dimension—this could be attributed to the fact that our model training involves much fewer examples as compared to the feature dimensionality.

## 6.2 Peripheral physiological feature extraction

### 6.2.1 hEOG features

The horizontal EOG signal has information about eye movements, point-of-gaze and eye blinks. Muscular facial activities and eye blinks appear as high frequency components in the EOG signal. Eye movements, blinks and facial muscular activities have been found to be highly correlated with emotional responses [15], [31].

**Eye movements:** To extract eye movement information, we low-pass filtered the signal with 5 Hz cut off, and then used wavelet transform to extract power spectral density (PSD) in 0-2 Hz range with a frequency resolution of 0.2 Hz, and temporal resolution of 50ms. Then for each second, we averaged the PSD values over frequency ranges of  $\{[0, 0.1), [0.1, 0.2), [0.2, 0.3), [0.3, 0.4), [0.4, 0.6), [0.6, 1.0), [1.0, 1.5), [1.5, 2)\}$ . Therefore, we obtained 8 features per second to describe eye movements.

**Facial muscle activity:** Facial muscular activities mainly relate to the movement of zygomatic major muscles, which occurs when a subject exhibits a smile, frown or other facial expressions. We limited the signal to 105-145 Hz, and then used wavelet transform to extract PSD with a frequency resolution of 1 Hz and temporal resolution of 500 ms.

Then for each second, we averaged the PSD values over  $\{[105, 115), [115, 130), [130, 145)\}$  frequency ranges. Since there are many muscles controlling facial activities, we used the three bands to obtain fine-grained information regarding muscular activities. Therefore per second, we obtained three values to represent zygomatic activities. Overall, from hEOG, we obtained 11 vectors of clip-length duration.

### 6.2.2 ECG features

From the ECG signal, we extracted information from both the original signal and its PSD.

**Heart beats:** We detected heart beats through R-peak detection in the ECG signal. Upon removal of low frequency components, R-peaks were detected as

the amplitude peaks. We then computed inter-beat-intervals (IBI), heart rate (HR) and heart rate variability (HRV) as the derivative of HR. Upon smoothing HR with a Kaiser window of temporal width 10 sec, and shape parameter  $\beta = \frac{1}{6}$ , we computed two features (smoothed HR and HRV) per second from which, statistical measures over IBI, smoothed HR, and HRV during the final 50 seconds of each trial were derived for affect recognition.

**Power spectral density:** ECG was recorded at 1 KHz sampling rate, and we used a wavelet transform over the ECG signal to extract the PSD in the frequency range of 0-5 Hz. Then, the mean PSD magnitudes over the frequency intervals  $\{(0, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6], (0.6, 1], (1, 1.5], (1.5, 2], (2, 2.5], (2.5, 5.0)\}$  were used as features– this gave us 11 values per second.

For single-trial classification alone, additional low-frequency information characterizing emotions was extracted as in [15]. We downsampled the ECG signal from 1 KHz to 256 Hz, and removed the low frequency drift. Then, we estimated the signal PSD using Welch’s method with a window length of  $15 \times sr$  and the overlap of  $10 \times sr$ , where  $sr$  denotes signal sampling rate. We used the mean PSD over  $\{[0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4]\}$  bands, and the logarithm PSD obtained for the sub-bands obtained on dividing  $[0, 2.4]$  into 10 equal intervals to obtain 14 more ECG PSD features.

### 6.2.3 Trapezius EMG

EMG effectively captures the mental stress of users [30]. As bipolar EMG electrodes are placed above the trapezius muscle, heart-related artifacts are observed in the signal and the EMG signal consists of two components: (1) Heart activities such as heart beats can be mainly inferred from the 0-45 Hz range, and (2) Trapezius EMG can be obtained from the  $\{[55, 95], [105, 145]\}$  range.

**Heart activities:** We low-passed the signal to within 45 Hz, and used wavelet transform to extract the PSD map with frequency and temporal resolution of 0.2 Hz and 50 ms respectively. Per second and trial, we computed the mean PSD over the following frequency bands:  $\{[0, 0.5], [0.5, 1.5], [1.5, 2.5], [2.5, 3.5], [3.5, 5.0], [5.0, 10], [10, 15], [15, 25], [25, 45]\}$ , to describe heart activities when the ECG signal was unavailable.

**Muscle activities:** We band-passed the EMG signal between 55-145 Hz and employed wavelet transform to extract the PSD map with frequency resolution of 1 Hz, and temporal resolution of 500 ms. Per each second and trial, we computed two values corresponding to mean PSD over the  $\{[55, 95], [105, 145]\}$  frequency bands to characterize trapezius muscle activities, and aforementioned statistical measures over the final 50 seconds were used for affect recognition.



Fig. 6. Participant’s facial video before (left) and after (middle) histogram equalization. Tracking 3D grid is shown on the right.

TABLE 2

Extracted audio-visual features from each movie clip (feature dimension listed in parenthesis).

Audio features	Description
<b>MFCC features (39)</b>	MFCC coefficients [20], Derivative of MFCC, MFCC Autocorrelation (AMFCC)
<b>Energy (1) and Pitch (1)</b>	Average energy of audio signal [20] and first pitch frequency
<b>Formants (4)</b>	Formants up to 4400Hz
<b>Time frequency (8)</b>	mean and std of: MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio [20]
<b>Zero crossing rate (1)</b>	Average zero crossing rate of audio signal [20]
<b>Silence ratio (2)</b>	Mean and std of proportion of silence in a time window [6], [20]
Video features	Description
<b>Brightness (6)</b>	Mean of: Lighting key, shadow proportion, visual details, grayness, median of Lightness for frames, mean of median saturation for frames
<b>Color Features (41)</b>	Color variance, 20-bin histograms for hue and lightness in HSV space
<b>VisualExcitement (1)</b>	Features as defined in [32]
<b>Motion (1)</b>	Mean inter-frame motion [22]

### 6.3 Facial Expression Analysis

We used histogram equalization to enhance contrast in the recorded NIR facial videos, and then employed the facial tracker described in [28] to track 12 facial landmarks (Figure 6). Statistical measures over the activation of these landmarks in the final 50 seconds of each trial were used for classification.

### 6.4 Multimedia features

We computed low-level audio visual features from the movie and music clips as described in [15] for comparing different modalities, and identifying the salient emotional information sources– extracted features are listed in Table 2. All in all, 49 video features and 56 audio features were extracted. For single-trial classification, we computed statistics over 1-second segments, while using statistics from features computed at the frame level for fine-grained, per-second emotion estimation described in Sec. 9.

## 7 MEG CORRELATES WITH USER RATINGS

We now present correlations observed between users’ self-assessments and their MEG responses. In order to directly compare our results with [15], we performed MEG feature extraction identical to [15] briefly described as follows. Following artefact rejection, we downsampled the MEG signal to 256Hz and then band-limited the same to within 1-48 Hz. Upon combining gradiometer outputs, the spectral power between 3 and 47 Hz over the last 30 seconds of each clip was extracted using Welch’s method with a window



size of 256 samples. Mean power over the  $\theta$  ([3-8] Hz),  $\alpha$  ([8-14] Hz),  $\beta$  ([14-30] Hz) and  $\gamma$  ([30-45] Hz) for each of 102 MEG sensors were correlated with the users' self-assessments.

We computed Spearman correlations between the above MEG-PSD outputs and participants' self ratings. Following [15], per subject, trial, emotion dimension and frequency band, correlations were computed over the 102 combined GRAD outputs. Upon computing correlations for each subject, and assuming independence [18],  $p$ -values obtained for each subject and condition were fused over all users using Fisher's method. Different from [15], we also accounted for multiple comparisons by controlling false discovery rate (FDR) using the procedure proposed in [4], and the observed significant correlations are highlighted in Fig. 7 ( $p < 0.05, 0.01, \text{ and } 0.001$  are respectively denoted in cyan, magenta, and red).

**Observations:** Observations similar to [15] can also be noted from Fig. 7. Thanks to the higher spatial resolution of MEG, a greater number of significant correlates and a wider range of correlations ([-0.15,0.25] with MEG vs [-0.1,0.1] with EEG) are observed with MEG signals as compared to EEG. For both movie and music stimuli, we observe a negative correlation between  $\alpha, \beta$  and  $\gamma$  powers and the arousal level over the vertex, the parietal and occipital lobes, which is consistent with the findings in [15]. Over the temporal and occipital lobes, we observe a positive correlation between the  $\theta, \beta$  and  $\gamma$  powers and the valence level. Note that the occipital and temporal lobes encode low-level audio-visual information which are responsible for inducing emotions [32]. The possibility of facial muscle activities, which are also prominent at high frequencies, influencing the observed correlations between valence/arousal ratings and MEG responses is minimal as facial activities are likely to occur in response to both negative and positive valence stimuli (e.g., funny and disgust). Finally, a few significant negative correlates in the parietal lobe, and few positive correlates in the occipital lobe are observed between dominance ratings and the MEG  $\beta, \gamma$  powers.

**Movie vs music:** As evident from Fig. 7, larger and more significant correlations are observed for movie clips as compared to music video clips, which suggests that emotions are more strongly and consistently evoked by movie stimuli. In particular, no correlations with  $p < 0.001$  are observed for music videos for the arousal and dominance dimensions. However, a larger number of correlations are observed over all frequency bands for arousal with music clips. We mention here that some of the detectable correlates for movie stimuli may have arisen from extraneous factors— e.g., correlates between  $\theta, \alpha$  powers and valence ratings may be attributed to eye movements/blinks. Likewise, positive correlation between  $\gamma$  power and dominance over the occipital lobes could

TABLE 3

Mean binary classification performance for music-video clips with the schema described in [15]. F1-scores of distributions significantly over 0.5 are highlighted (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ). NR denotes 'not reported'.

	Music (SS)					
	Arousal		Valence		Dominance	
	Acc	F1	Acc	F1	Acc	F1
EEG [15]	0.62	0.58**	0.58	0.56**	NR	NR
Max Baseline [15]	0.64	0.50	0.59	0.50	NR	NR
MEG	0.62	0.58***	0.59	0.55*	0.62	0.53*
Max Baseline	0.52	0.50	0.54	0.50	0.66	0.50

be explained by low-level visual cues [24], while the similar but weaker correlate observed for arousal could be owing to the strong positive correlation between arousal and dominance ratings ( $0.57 \pm 0.24$ ) across participants. Further examination to more accurately identify the information source responsible for the above correlations would involve (1) HPI-based MEG signal compensation, (ii) Independent component analysis, and (iii) Brain source localization using MR brain scans, which is left to future work.

## 8 EXPERIMENTAL RESULTS

We now present comparisons between MEG vs EEG, and movie vs music clips based on single-trial classification results.

### 8.1 Single-trial Classification: MEG versus EEG

In order to evaluate our MEG-based approach against the EEG framework described in [15], we attempted single-trial binary (*high/low*) classification of valence and arousal employing (i) labels derived from subject-wise self-reports and (ii) extracting MEG features in a manner identical to [15]. Employing the Naive-Bayes classifier and subject-specific models, only the top 10% discriminative features based on Fisher feature selection criteria were used in each loop of a leave-one-trial-out cross-validation scheme. Very comparable results with EEG and MEG obtained with this procedure (Table 3) suggest that the affect encoding power of EEG and MEG are comparable. However, the increased spatial resolution of MEG allows for fine-grained affective analysis, which enables similar or superior recognition performance on music and movie clips using the features extracted in Sec. 6 as described later.

While the fairest comparison between EEG and MEG would entail simultaneous recording of the two modalities for identical subjects and stimuli, such a study may be impossible to implement in practice. We have compared emotion recognition performance based on the results observed on two random subject populations that are comparable in size, and this is the second best possible way of performing a comparison

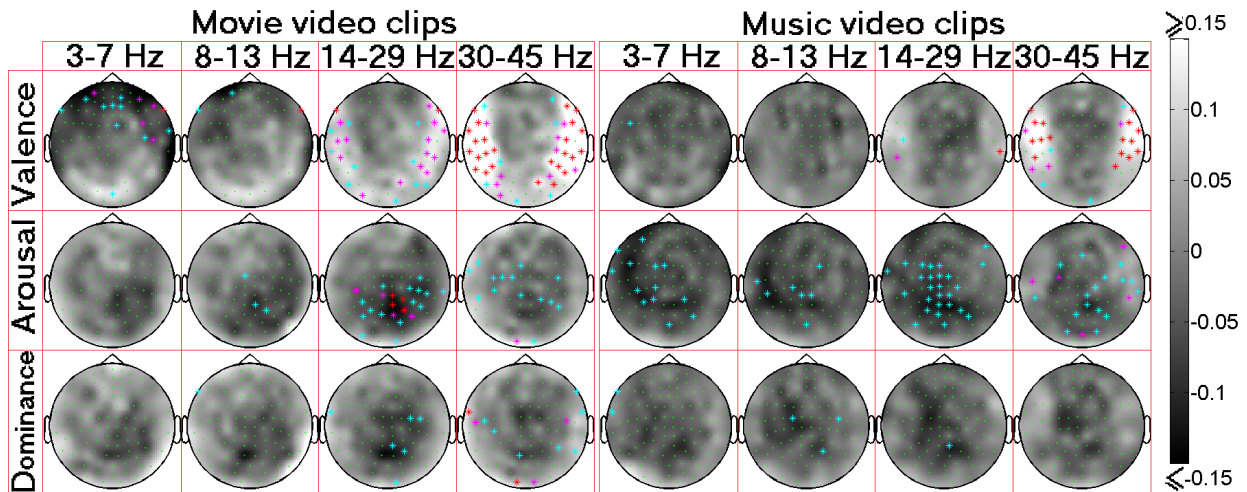


Fig. 7. Spearman correlation analysis between the MEG responses and participants’ self-assessments. Correlation over each channel (in green) is denoted by the gray level, and significant ( $p < 0.05$ ,  $p < 0.01$ , and  $p < 0.001$ ) correlations are highlighted with \* marks (in cyan, magenta, and red).

in our view. Designing better approaches for comparing the efficacy of different modalities for user-centric emotion recognition is a research problem requiring further investigation.

## 8.2 Classification procedure and results

On a per-user basis, we attempted to recognize the emotional *valence* (V), *arousal* (A) and *dominance* (D) of a test music/movie clip as *high/low* based on the MEG and peripheral physiological responses. Given the large subjectivity in user responses for music videos in [15], subject-specific labels were used for each stimulus. However, as (i) many significant correlates observed between ratings and MEG responses of the user population, and (ii) the stimulus label should reflect the perception of the population instead of individuals, we repeated the classifications with both population-based (denoted as PB in Table 4) and subject-based (SB in Table 4) labels.

Under PB labeling, each stimulus was assigned a *high/low* (V/A/D) label based on whether its rating was higher or lower than the mean rating provided by the participant population for the stimulus set. Likewise, the SB label for each stimulus denoted whether its rating was higher/lower than the mean subject rating. The proportion/distribution of positive and negative classes for movie and music V,A,D under PB/SB tagging is presented in Table 4. For SB labeling, the mean and standard deviation of the positive class distribution are specified. Under PB labeling, the proportion of positive and negative classes is most imbalanced for music and movie arousal, whereas the most balanced distributions under SB labeling are observed for movie valence and music arousal. Given the unbalanced positive and negative classes, we use F1-scores as the primary measure to compare classifi-

cation performance with different stimulus types and information modalities.

We used a linear SVM classifier for our experiments and the mean accuracy and F1-scores obtained over the 30 participants using leave-one-trial-out cross-validation are tabulated in Table 4. The optimal SVM slack parameter was tuned by considering values in  $[10^{-4}, 10^4]$  using an inner leave-one-out cross-validation loop. As baselines, we present the F1-scores of (i) a random classifier, (ii) majority-based voting<sup>6</sup> and (iii) voting based on training class distribution—note that the maximum baseline F1-score is 0.50. Instances where the F1-score distribution across subjects is significantly higher than 0.5 as determined by a paired *t*-test are highlighted in Table 4.

To demonstrate how the higher spatial resolution of MEG benefits affect recognition, we present results achieved with features extracted exclusively from each brain lobe, and also the concatenation of features from all lobes (MEG Early Fusion or MEF). In addition, we present accuracies and F1-scores achieved using (i) the combination of hEOG, ECG and tEMG responses (peripheral physiology or PP), (ii) facial expressions (FE), (iii) multimedia features (MM), and (iv) late fusion of the decisions from the the MEF, PP, FE and MM classifiers following the methodology proposed in [16]. If  $\{p_i\}_{i=1}^4$  denote the posterior probabilities output by the four classifiers and  $t_i = \alpha_i F_i / \sum_{i=1}^4 \alpha_i F_i$ , where  $\alpha_i$ 's denote fusion weights and  $F_i$  denotes F1-score of the  $i^{th}$  classifier on training data, the optimal weights  $\{\alpha_i^*\}$  are chosen as those maximizing F1-score on the training set using an inner cross-validation loop. Posterior probability of the test sample is computed as  $\sum \alpha_i^* p_i t_i$ , which is then used

6. With leave-one-out classification on a balanced class distribution (Table 4), majority-based voting would yield 0% accuracy as the test-label class is in minority in the training set.

to assign the test label.

### 8.3 Discussion of classification results

In Table 4, the obtained F1-scores clearly demonstrate that the increased spatial resolution of MEG benefits affect analysis and recognition. For all conditions, the classification performance obtained with MEG features from at least one of the nine brain lobes is similar to or better than the performance achieved with MEF, where features of all the brain lobes are pooled together. This result is unsurprising as the various brain lobes are known to encode different types of emotional information, as also suggested by the correlation analysis in Sec. 7. Under PB stimulus labeling, the best F1-scores for movie and music arousal are obtained for the right temporal lobe, while the left and right temporal lobes respectively are found to encode optimal information for decoding the valence of movie and music stimuli. Best performance for dominance is obtained with right-frontal lobe features for movies, and left parietal for music.

Another salient observation is that despite the subjectivity in emotion perception and expression, reliable and above-chance emotion recognition is achieved upon associating the physiological responses of each user with stimulus labels assigned by the population. For movie clips in particular, much better classification performance achieved under PB labeling as compared to SB labeling. In practice, emotion (or genre) tags to movies or music videos are attached based on the perception of the *general audience*, and not on the basis of individual perception. Likewise, for the purpose of affect recognition and emotion elicitation, it would be desirable to work with control stimuli consistently capable of evoking the target emotion from target users. Movie clips (and corresponding user responses) compiled as part of DECAF are an important contribution in this respect.

The obtained results also point to the complementarity of different signals in encoding emotions. Consistent with the findings in [15], MEG signals are seen to effectively encode arousal and dominance, while peripheral physiology signals efficiently encode valence. Facial expressions are also seen to best encode valence, while audio-visual features achieve best arousal recognition for music clips with PB labels. This complementarity was also evident when finding the best two and three information modalities for recognizing valence and arousal under PB labeling— considering feature pairs, MEG and peripheral physiological features produced the best arousal recognition for movie clips ( $F1=0.66^{***}$ ), while peripheral and audio-visual features best recognized valence from music clips ( $F1=0.83^{***}$ ). Facial activities and multimedia content provided best recognition of valence from movies ( $F1=0.78^{***}$ ) and arousal from music clips

( $F1=0.87^{***}$ ). Considering triplets, the combination of MEF, PP and MM consistently produced the best F1-scores for movie-arousal ( $F1=0.71^{***}$ ), movie-valence ( $F1=0.81^{***}$ ), music-arousal ( $F1=0.87^{***}$ ), music-valence ( $F1=0.85^{***}$ ). F1-scores obtained by fusing the outputs of all modalities are slightly lower than those obtained from combinations of feature triplets, suggesting that feature selection may be necessary for optimal fusion results.

Finally, comparing the emotion recognition performance with music and movie clips, superior F1-scores achieved using MEG features for population-rated movie clips again confirms that they serve as better control stimuli for affect recognition studies. For music stimuli, relatively higher recognition is achieved with subject-specific labels, and the best performance with PB labels is achieved for arousal using multimedia features.

## 9 CONTINUOUS EMOTION ESTIMATION

DECAF also contains time-continuous arousal (A) and valence (V) annotations for the 36 movie clips acquired from seven experts, who were very familiar with the movie clips, but were not part of the MEG study. While the user ratings acquired in Sec. 4 are useful for recognizing the *general* stimulus emotion, dynamic AV ratings are used for estimating the *emotional highlight* in a given clip. We show how these annotations were utilized to predict A,V levels of time-contiguous snippets using (i) multimedia audio-visual (MM), and (ii) MEG features.

**Experiments and Results:** We asked seven experts to provide per-second V,A ratings for 36 movie clips listed in Table 1 using the G-Trace software [7]. The experts, who could familiarize themselves with scene dynamics by viewing the movie clips as many times as they wanted to prior to rating them, were required to annotate the *target emotion* meant to be evoked in the viewer (in terms of V, A levels) for each second of the video. Upon rescaling the annotations using  $z$ -score normalization, Kendall's coefficient of concordance ( $W$ ) was used to measure the dynamic inter-annotator agreement— overall  $W$  was found to be  $0.47\pm 0.27$  for arousal, and  $0.64\pm 0.18$  for valence, signifying good agreement. Re-computing  $W$  over the *first* and *second* half of the clips, we observed  $W$  to be  $0.35\pm 0.25$ ,  $0.43\pm 0.28$  and  $0.58\pm 0.24$ ,  $0.54\pm 0.23$  for A, V respectively, implying that expert assessments were more consistent for the emotionally salient second halves of the clip (all clips began with a neutral segment). Finally, the median annotation was used as the *gold standard* dynamic rating for each clip. Dynamic V, A ratings are illustrated in Fig. 8. We then attempted prediction of dynamic V, A levels in time-contiguous snippets derived from the movie clips using (i) audio-visual and (ii) MEG features. Per-second features extracted in Sec. 6 were used to this end.

TABLE 4

**Single trial classification for music and movie clips**– (Upper) classification results using MEG information from each of the brain lobes. (Middle) Unimodal and multimodal classification results. (Bottom) Baseline comparisons along with the distribution of positive samples are tabulated. Mean F1 scores derived from a distribution significantly above chance level (0.50) are highlighted (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ). PB, SB respectively denote use of population and subject-based labels in the classification framework.

		Movie (PB)			Music (PB)			Movie (SB)			Music (SB)		
		A	V	D	A	V	D	A	V	D	A	V	D
Vertex	Acc	0.59	0.57	0.57	0.51	0.51	0.52	0.55	0.55	0.51	0.53	0.50	0.53
	F1	0.58***	0.57***	0.57***	0.51	0.51	0.51	0.54	0.53	0.48	0.52	0.49	0.49
Left Temporal	Acc	0.60	0.60	0.58	0.51	0.51	0.52	0.59	0.58	0.51	0.54	0.50	0.54
	F1	0.60***	0.60***	0.58***	0.51	0.51	0.51	0.59***	0.57**	0.49	0.52	0.49	0.51
Right Temporal	Acc	0.62	0.56	0.57	0.55	0.53	0.53	0.59	0.55	0.54	0.60	0.54	0.54
	F1	0.62***	0.55**	0.57***	0.55*	0.53*	0.53*	0.58**	0.53	0.51	0.58***	0.53	0.51
Left Parietal	Acc	0.60	0.56	0.57	0.52	0.52	0.55	0.55	0.56	0.53	0.53	0.48	0.52
	F1	0.60***	0.55**	0.57***	0.52	0.51	0.54*	0.54*	0.54*	0.49	0.52	0.47	0.49
Right Parietal	Acc	0.58	0.57	0.57	0.51	0.51	0.52	0.55	0.55	0.58	0.51	0.53	0.54
	F1	0.57**	0.57***	0.56***	0.50	0.50	0.52	0.53	0.53	0.55**	0.50	0.52	0.51
Left Occipital	Acc	0.58	0.59	0.57	0.51	0.50	0.52	0.53	0.56	0.54	0.55	0.48	0.53
	F1	0.57**	0.58***	0.56**	0.51	0.50	0.52	0.51	0.54*	0.50	0.54*	0.47	0.50
Right Occipital	Acc	0.60	0.56	0.56	0.50	0.53	0.50	0.57	0.54	0.55	0.54	0.53	0.53
	F1	0.60***	0.55**	0.56*	0.50	0.53	0.50	0.56**	0.53	0.52	0.53	0.51	0.49
Left Frontal	Acc	0.59	0.56	0.57	0.55	0.51	0.51	0.56	0.56	0.53	0.57	0.55	0.60
	F1	0.58***	0.56**	0.57***	0.54*	0.50	0.51	0.55**	0.55**	0.50	0.55**	0.54*	0.56**
Right Frontal	Acc	0.55	0.59	0.61	0.50	0.52	0.50	0.51	0.54	0.53	0.54	0.52	0.53
	F1	0.55***	0.59***	0.61***	0.49	0.52	0.49	0.50	0.53	0.49	0.53	0.51	0.49
MEG Early Fusion	Acc	0.60	0.61	0.59	0.53	0.53	0.54	0.55	0.58	0.55	0.58	0.56	0.55
	F1	0.60***	0.61**	0.59***	0.52	0.53	0.54*	0.54*	0.58***	0.53	0.55**	0.55**	0.53*
Peripheral Physiology	Acc	0.55	0.60	0.50	0.55	0.59	0.56	0.56	0.60	0.56	0.57	0.55	0.57
	F1	0.54*	0.59***	0.50	0.54*	0.59***	0.55**	0.55**	0.59***	0.54*	0.56**	0.54*	0.54**
Facial Expressions	Acc	0.58	0.64	0.53	0.60	0.61	0.53	0.56	0.61	0.55	0.58	0.60	0.55
	F1	0.57**	0.64***	0.53	0.59**	0.60***	0.53	0.54**	0.61***	0.54	0.56**	0.58***	0.52
Multimedia Content	Acc	0.58	0.64	0.33	0.85	0.73	0.57	0.52	0.61	0.53	0.62	0.68	0.58
	F1	0.57	0.64	0.33	0.85	0.72	0.57	0.51	0.60***	0.52	0.61***	0.67***	0.55*
Late Fusion	Acc	0.70	0.79	0.66	0.85	0.82	0.66	0.66	0.73	0.72	0.73	0.76	0.74
	F1	0.68***	0.77***	0.64***	0.84***	0.81***	0.65***	0.62***	0.71***	0.66***	0.70***	0.73***	0.67***
Random	Acc	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	F1	0.49	0.50	0.50	0.49	0.50	0.50	0.49	0.49	0.48	0.49	0.49	0.48
Majority	Acc	0.58	0.00	0.53	0.57	0.53	0.00	0.57	0.53	0.60	0.52	0.54	0.66
	F1	0.37	0.00	0.35	0.37	0.34	0.00	0.37	0.33	0.36	0.32	0.34	0.39
Class-ratio	Acc	0.51	0.50	0.50	0.51	0.50	0.50	0.54	0.52	0.56	0.52	0.53	0.57
	F1	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
+ve Class proportion	Mean	58.3%	50.0%	52.8%	57.5%	52.5%	50.0%	48.4%	49.3%	41.9%	49.3%	46.3%	45.6%
	STD	-	-	-	-	-	-	13.6%	9.5%	14.9%	10.7%	10.9%	19.0%

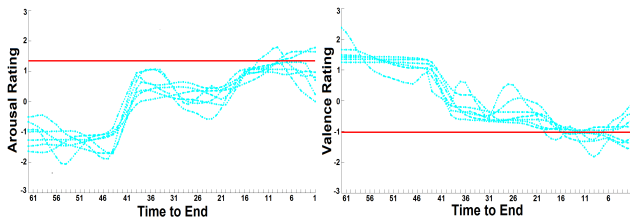


Fig. 8. Time-continuous A (left), V (right) ratings for Clip 36 in Table 1 from seven experts are plotted in cyan. Both continuous and static ratings (red) are  $z$ -score normalized and are in the range  $[-3, 3]$ .

Apart from Lasso sparse regression, we also employed Multi-task learning (MTL) based regressors– given a set of  $T$  related tasks (movie clips related in terms of V, A in this case), MTL seeks to *jointly* learn a set of weights  $W = \{W_t\}_{t=1}^T$ , where  $W_t$  models task  $t$ . MTL enables simultaneous learning of similarities as well as differences among tasks, leading to a more efficient model than learning each task independently. In this work, we employed three MTL variants from the MALSAR library [34]– multi-task Lasso, Dirty MTL

where the weight matrix  $W = P + Q$ , with  $P$  and  $Q$  denoting group-common and task-specific components, and sparse graph-regularized MTL (or SR MTL), where *a priori* knowledge on task-relatedness is incorporated in the learning process so that weight similarity is only enforced among related tasks.

A, V weights for the 36 movie clips learned from audio-visual (MM) features (concatenation of audio and video features) through the Dirty and SR MTL approaches are presented in Fig. 9. *A-priori* knowledge available in the form of ground truth labels (Table 1) were used to group related stimuli and input to the SR MTL algorithm. SR MTL weights learnt for high and low arousal clips are shown in the top row, while the bottom row presents weights learned for high and low valence clips. MFCCs are found to be the most salient audio features, while color and brightness video features are the best predictors for both valence and arousal. Concerning SR MTL outputs, visual excitement features are found to be characteristic of high arousal clips, while inter-frame motion is indicative of high-valence clips.



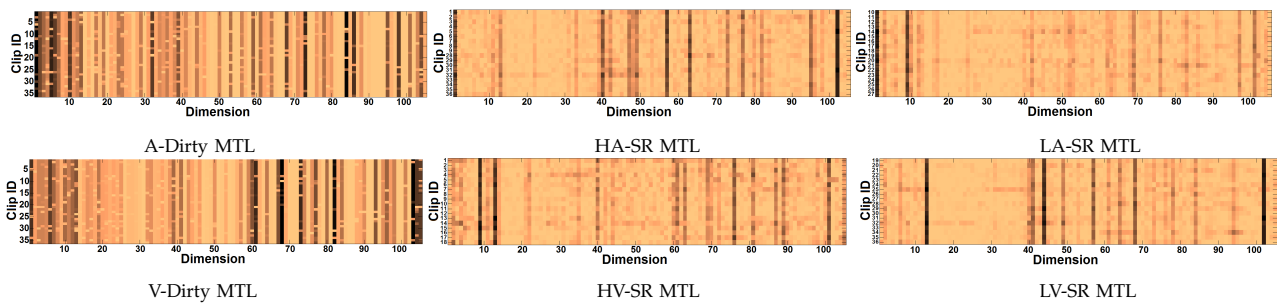


Fig. 9. Learned weights for arousal (top) and valence (bottom) for the movie clips with Dirty MTL and SR MTL. Audio-visual features over the entire clip length were used for model training. Larger weights are denoted using darker shades. MM features (106 in total) are arranged in the order specified in Sec. 6. Best viewed under zoom.

TABLE 5

Valence/Arousal prediction with multimedia (MM) and MEG features. RMSE mean, standard deviation over four runs are reported. Range of V, A levels is [-3, 3]. Best model is shown in bold.

		First		Second		
		5 s	15 s	5 s	15 s	
Val	MM	Lasso	1.98±1.25	3.07±1.48	1.68±0.18	2.81±0.97
		MT-Lasso	<b>1.00±0.05</b>	1.66±0.54	<b>1.18±0.14</b>	<b>2.03±0.71</b>
		Dirty MTL	1.11±0.06	1.79±0.55	1.27±0.16	2.10±0.69
		SR MTL	1.09±0.09	<b>1.55±0.39</b>	1.89±0.13	2.80±0.74
	MEG	Lasso	1.30±0.09	1.87±0.46	2.03±0.25	2.93±0.78
		MT-Lasso	1.32±0.09	1.98±0.54	1.54±0.21	2.47±0.81
Arl	MM	Dirty MTL	1.42±0.10	2.44±0.82	<b>1.51±0.19</b>	<b>2.44±0.82</b>
		SR MTL	<b>1.09±0.05</b>	<b>1.58±0.41</b>	2.07±0.17	2.84±0.69
		Lasso	1.54±0.47	2.11±0.77	2.18±0.58	3.28±2.17
		MT-Lasso	<b>0.91±0.11</b>	1.47±0.47	<b>1.10±0.08</b>	1.89±0.66
	MEG	Dirty MTL	1.07±0.09	1.62±0.46	1.23±0.08	<b>1.97±0.61</b>
		SR MTL	1.01±0.07	<b>1.42±0.35</b>	1.86±0.13	2.48±0.53
Arl	MM	Lasso	1.11±0.08	1.65±0.45	1.75±0.06	2.53±0.66
		MT-Lasso	1.12±0.09	1.71±0.51	1.41±0.11	2.27±0.73
		Dirty MTL	1.19±0.11	1.84±0.56	<b>1.38±0.11</b>	<b>2.25±0.75</b>
		SR MTL	<b>0.99±0.08</b>	<b>1.42±0.36</b>	1.73±0.06	2.44±0.60
	MEG	Lasso	1.11±0.08	1.65±0.45	1.75±0.06	2.53±0.66
		MT-Lasso	1.12±0.09	1.71±0.51	1.41±0.11	2.27±0.73

Finally, dynamic V, A level prediction performance using MM and MEG features (average MEG response of the 30 DECAF participants was used here) on 5 and 15 second snippets randomly extracted from the first and second half from each of the movie clips is presented in Table 5— remainder of the movie clips was used for model training. The root mean square error (RMSE) measure is used for comparison—evidently, larger prediction errors are noted for snippets from the second half, and for 15-sec segments. MTL considerably outperforms Lasso regression, implying that jointly learning from features of multiple movie clips is beneficial as compared to clip-wise learning, while slightly better prediction performance is achieved with MM features considering the best model for each condition.

## 10 CONCLUSION

The DECAF database compiled with the aim of evaluating user-centered affect recognition with (i) MEG vs EEG sensing, and (ii) movie vs music clips, is presented in this paper. The increased spatial resolution of MEG enables fine-grained analysis of cognitive responses over brain lobes in turn aiding affect recognition, while coherence between explicit ratings and implicit responses is greater across users for

movie clips, suggesting that they are better control stimuli for affect recognition studies. While classification results for valence, arousal and dominance are presented with the aim of comparing with [15], dominance may be hard to qualify in a movie-watching context even if it has been found to be relevant with regard to musical compositions. This study was limited to sensor-space analyses of MEG responses—source-space analysis was not performed, and is left to future work. Finally, dynamic emotion prediction with time-continuous emotion annotations available as part of DECAF is demonstrated, and simultaneously learning from multimedia/MEG features from all clips is found to be more beneficial than learning one model per clip. Unlike EEG, MEG is a relatively new technology, and with improvements in techniques such as HPI-based MEG signal compensation, we believe that much higher recognition performance than that achieved in this introductory work is possible.

## REFERENCES

- [1] Emotion and multimedia content. In B. Furht, editor, *Encyclopedia of Multimedia*. Springer, 2006.
- [2] M. K. Abadi, S. M. Kia, R. Subramanian, P. Avesani, and N. Sebe. User-centric Affective Video Tagging from MEG and Peripheral Physiological Responses. In *Affective Computing and Intelligent Interaction*, pages 582–587, 2013.
- [3] E. E. Bartolini. Eliciting emotion with film: Development of a stimulus set. Master’s thesis, Wesleyan University, 2001.
- [4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- [5] M. Bradley. Emotional memory: a dimensional analysis. In *Emotions: Essays on Emotion Theory*. Lawrence Erlbaum, Hillsdale, NJ, 1994.
- [6] L. Chen, S. Gunduz, and M. T. Ozsu. Mixed type audio classification with support vector machine. In *IEEE International Conference on Multimedia and Expo*, pages 781–784, 2006.
- [7] R. Cowie, G. McKeown, and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions (IJSE)*, 3(1):1–17, 2012.
- [8] M. K. Greenwald, E. W. Cook, and P. J. Lang. Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology*, 3:51–64, 1989.
- [9] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.
- [10] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

- [11] U. Hasson, R. Malach, and D. J. Heeger. Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1):40 – 48, 2010.
- [12] H. Joho, J. Staiano, N. Sebe, and J. M. Jose. Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications*, 51(2):505–523, 2011.
- [13] S. M. Kia, E. Olivetti, and P. Avesani. Discrete cosine transform for MEG signal decoding. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 132–135, 2013.
- [14] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.
- [15] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
- [16] S. Koelstra and I. Patras. Fusion of facial expressions and eeg for implicit affective tagging. *Image and Vision Computing*, 31(2):164–174, 2013.
- [17] P. Lang, M. Bradley, and B. Cuthbert. IAPS: Affective ratings of pictures and instruction manual. Technical report, University of Florida, 2008.
- [18] N. A. Lazar, B. Luna, J. A. Sweeney, and W. F. Eddy. Combining brains: a survey of methods for statistical pooling of information. *Neuroimage*, 16(2):538–550, 2002.
- [19] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.
- [20] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.
- [21] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Adv. Sig. Proc.*, 2004(11):1672–1687, 2004.
- [22] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Int'l Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
- [23] P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophysical Journal*, 76(2):691 – 708, 1999.
- [24] E. Mower, M. J. Mataric, and S. Narayanan. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Transactions on Multimedia*, 11(5):843–855, 2009.
- [25] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011.
- [26] P.-Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Interaction*, 59(1-2):157–183, 2003.
- [27] J. Schwarzbach. A simple framework (asf) for behavioral and neuroimaging experiments based on the psychophysics toolbox for matlab. *Behavior Research Methods*, pages 1–8, 2011.
- [28] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang. Authentic Facial Expression Analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.
- [29] R. Sinha and O. A. Parsons. Multivariate response patterning of fear and anger. *Cognition and Emotion*, 10(2):173–198, 1996.
- [30] D. M. Sloan. Emotion regulation in action: emotional reactivity in experiential avoidance. *Behaviour Research and Therapy*, 42(11):1257–1270, 2004.
- [31] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.
- [32] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.
- [33] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens. Trapezius muscle emg as predictor of mental stress. In *Wireless Health 2010*, pages 155–163, 2010.
- [34] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-task Learning via Structural Regularization*. Arizona State University, 2011.



**Mojtaba Khomami Abadi** received his B.Sc. in Computer Science from the College of Science, University of Tehran, Iran, in 2010. He is now a PhD candidate at the Department of Information Engineering and Computer Science, University of Trento, Italy. His research interests include: user centric affective computing in human computer interaction and affective multimedia analysis.



**Ramanathan Subramanian** received his PhD degree in Electrical and computer engineering from the National University of Singapore in 2008. Ramanathan Subramanian is a Research Scientist at ADSC, Singapore, and his research focuses on Human-centered computing. In particular, he is interested in developing methods to understand human visual and emotional perception as well as behavior, and modeling the same in HCI applications.



**Seyed Mostafa Kia** received his M.Sc. in cognitive neuroscience from the University of Trento, Italy, 2013. He is currently a PhD student in Computer science at the University of Trento. His research interests include machine learning and brain decoding on MEG data.



**Paolo Avesani** is a researcher at Fondazione Bruno Kessler (FBK), in Trento, Italy, where he is leading NILab, a Neuroinformatics Laboratory raised as joint initiative of FBK and the Center for Mind/Brain Sciences (CIMeC) of the University of Trento. He received his Dr. Degree in Information Science from the University of Milan, Italy. His current research interests include machine learning, functional neuroimaging analysis, computational methods for brain decoding and brain

mapping.



**Ioannis (Yiannis) Patras** received the the Ph.D. degree from the Department of Electrical Engineering, Delft University of Technology, The Netherlands, in 2001. He is a Senior Lecturer in Computer Vision in the School of Electronic Engineering and Computer Science in the Queen Mary University of London. He is/has been in the organizing committee of IEEE SMC2004, FGR2008, ICMR2011, ACM2013 and was the general chair of WIAMIS2009. He is associate editor in the Image and Vision Computing Journal. His research interests lie in Computer Vision, Pattern Recognition and Multimodal Human Computer Interaction.



**Nicu Sebe** is a professor at the University of Trento, Italy, where he is leading research in multimedia information retrieval and human-computer interaction in computer vision applications. He has served as a general co-chair of the IEEE FG08, CIVR2007,2010 and WIAMIS09 and was one of the initiators and a program co-chair for the Human-Centered Multimedia track in ACM2007. He was general chair of ACM2013 and program chair of ACM2011. He is a senior member of

the IEEE.