User-centric Affective Video Tagging from MEG and Peripheral Physiological Responses

Mojtaba Khomami Abadi[1,5], Seyed Mostafa Kia[1,3], Ramanathan Subramanian[2], Paolo Avesani[3,4], Nicu Sebe[1,4]
[1]University of Trento, Italy
[2]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
[3]NeuroInformatics Laboratory (NILab), Fondazione Bruno Kessler, Trento, Italy
[4]Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy
[5]Semantic, Knowledge and Innovation Lab (SKIL), Telecom Italia

***Abstract*—This paper presents a new multimodal database and the associated results for characterization of affect (*valence*, *arousal* and *dominance*) using the Magnetoencephalogram (MEG) brain signals and peripheral physiological signals (horizontal EOG, ECG, trapezius EMG). We attempt single-trial classification of affect in movie and music video clips employing emotional responses extracted from eighteen participants. The main findings of this study are that: (i) the MEG signal effectively encodes affective viewer responses, (ii) clip arousal is better predicted by MEG, while peripheral physiological signals are more effective for predicting valence and (iii) prediction performance is better for movie clips as compared to music video clips.**

I. Introduction

Representing, measuring and predicting emotion in multimedia content adds significant value to multimedia systems [1]. Nevertheless, only a handful of works have attempted to categorize multimedia content based on the emotion(s) they evoke, and they typically involve (i) analysis of the content [8] with simple models scarcely reflective of human perception, or (ii) recognition of the viewer's facial expressions [11], which denote a subjective/circumstantial manifestation of the true affect, or (iii) based on physiological responses [15], which capture only a limited aspect of human emotion. Recently, cognitive-based approaches employing means such as fMRI or EEG to map brain signals with the induced affect [9], [10], [13], [22] have gained in popularity, but the experimental set-up used by these methods only allows for the use of simple, well-focused signals or time-limited, complex stimuli such as movie clips.

In this paper, we present the first work to perform single-trial affect decoding using the Magnetoencephalogram (MEG) signal, which is an emerging technology for capturing functional brain activity. Unlike fMRI, MEG responses can be acquired with the subject seated comfortably in a quiet, shielded room and different from EEG, MEG does not require the use of gel for positioning dozens of electrodes. The MEG device allows for naturalistic user reaction as the user is seated on an armchair, having little physical contact with the sensing coil. The advantages of the MEG setup are that (i) the acquired physiological signals are more meaningful, since they are not affected by psychological stressors, and (ii) MEG responses can be recorded with higher spatial/temporal resolution, as compared to EEG and fMRI.

The non-invasiveness of MEG enables reliable acquisition of users' brain responses over longer time durations, and so it is possible to use full-length movies for analysis– as a first step, we present preliminary classification results on a large dataset of 40 music videos and 36 movie clips in this paper. In addition to MEG, we also use complementary physiological signals in the form of electrooculogram (EOG), electrocardiogram (EEG) and trapezium electromyogram (EMG) for characterizing affect, making our approach user-centric and multi-modal. This MEG-based affect sensing approach can be used to generate affective curves for movie content as in [8], and therefore, a user-centered framework, more informative and reliable than a content-based one, can be used to drive important consumer applications such as personalization of user-delivered content and automated highlights compilation in the near future.

In summary, this paper makes the following contributions: (1) It represents the first work to employ single-trial MEG decoding for affect characterization, and classification results are presented for a dataset of 40 music video clips and 36 movie clips, which is one of the largest reported in affective computing literature. Also, given the large variability in subjects' affective ratings and brain activations, some studies (*e.g.*, [13]) attempt classification with participant-specific stimulus labels. Instead, given that our final objective is affective video tagging which is stimulus-specific rather than subject-specific, our classification assumes a single label per stimulus, as determined from the mean affective rating given by participants. (2) We also compare the suitability of two stimuli types (music and movie clips) for studying affect. While the various studies reported in literature have used a variety of stimuli (images, music videos and movies) to study affect, no comparisons to determine which stimulus is better suited for affective analysis have been reported to our knowledge. This work represents a first step in that direction.

The paper is organized as follows: section II overviews related affective studies. The experimental protocol employed for recording viewers' affective responses, and the feature extraction methodology are respectively described in sections III and IV respectively. Experimental results are discussed in section V, while conclusions are stated in section VI.

II. Related work

While affective content creators *intend* to convey a certain emotion (or a set of emotions) through the created stimulus, the *actual* emotion induced upon perceiving the stimulus is influenced by a number of psychological and contextual factors, and can therefore be highly subjective. Consequently, correlating the *observed* emotional response with the *expected* response is a non-trivial problem which is typically simplified

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.102

in practice employing the following ideas: (1) Most affective studies assume that the entire gamut of human emotions can be represented as a set of points on the valence-arousal[1] plane as demonstrated by Greenwald *et al.* [6], and (2) To largely ensure that the elicited and expected emotions are consistent, the presentation stimuli are carefully selected based on previous studies, or based on 'ground truth' valance-arousal ratings compiled from a large population that evaluates the stimuli prior to the actual experiment.

Emotional states have been found to produce specific types of physiological responses- *e.g.*, excitement is associated with increased heart-beat and respiration rates, and this correlation is exploited in a number of physiology-based affect studies. Heart-rate, skin temperature and conductance level, blood pressure and facial EMG are recorded as subjects view affective imagery in [20]. Their experiments indicate that the responses for anger and fear are uniquely different from responses to neutral images.

Among physiology-based affective studies with video stimuli, Lisetti and Nasoz [15] employ a two-pronged approach to elicit frustration along with other emotions from 29 participants. They use movie clips proposed by [7] to evoke sadness, anger, amusement, fear and surprise, and induce frustration by asking subjects to solve difficult mathematical questions without pencil and paper. GSR, heart rate, temperature, EMG and heat flow responses are recorded using an armband, and over 80% accuracy is obtained in classifying the aforementioned emotions using extracted features.

Employing physiological signals for emotion recognition from audio music clips is described by Kim and André in [12]. In a study conducted with three subjects, the authors employ four bio-channel sensors to measure electromyogram, electrocardiogram, skin conductivity and respiration changes and correlate them with the subjects' emotional state. Using an emotion-specific multilevel dichotomous classification scheme, the authors achieve 95% and 70% recognition accuracy for subject-dependent and independent classification for the four (high/low valence/arousal) musical emotions.

In the DEAP dataset, Koelstra *et al.* [13] record EEG, GSR, blood volume pressure, respiration rate, skin temperature and Electrooculogram (EOG) patterns as 32 viewers are presented with 40 one-minute music video segments. These responses are correlated with arousal, valence, liking and dominance ratings provided by participants during the experiment. A mean accuracy of over 60% is obtained for single-trial binary classification with EEG and peripheral physiological signals. The MAHNOB-HCI multimodal database compiled by Soleymani *et al.* [22] contains face videos, audio and physiological signals as well as eye-gaze data of 27 participants who watched 20 emotional movie/online clips in one experiment, and 28 images and 14 short videos in another. Their database facilitates affect computation using single or multiple modalities and determination of the most suitable modalities.

III. Experimental Protocol and Data Analysis

In this section, we present a brief description of (a) MEG and peripheral physiological signals employed in the study and (b) stimuli selection procedure before detailing the (c) experimental set-up and protocol, and (d) the analysis of self-assessment ratings.

A. MEG and peripheral physiology signals

To collect users' affective responses we employed (i) magnetoencephalogram (MEG), (ii) horizontal electrooculogram (hEOG), (iii) electrocardiogram (ECG), and (iv) trapezius electromyogram (Trapezius EMG) signals that are described below:

Magnetoencephalogram: MEG is a technology that 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 pico-Tesla and since sensors are really sensitive to noise, the MEG equipment is located in a 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.

Horizontal electrooculogram: Electrooculography is the measurement of eye activities (*i.e.*, movements, fixations and blinks). In this study, we used an horizontal EOG which reflects mostly the horizontal eye movement of users. We placed two electrodes on the left and right side of the users' face close to their eyes. Zygomatic muscle activities produce high frequency components in the bipolar EOG signal, and hence the EOG signal captures general information about facial activation. In comparison to other facial characteristics, EOG is shown to be more effective for measuring affect across different cultures [24].

Electrocardiogram: Electrocardiogram signal is well known for its relevance in emotion recognition studies [22], [12]. ECG signals were recorded using three sensors attached on the participants' body. Two of the electrodes were placed on the wrist pulses and a reference was placed on a boney part of the arm (ulna bone). This setup allows for precise detection of heart beats, and consequently, accurate computation of heart rate (HR) and heart rate variability (HRV).

Trapezius electromyogram: Different people exhibit varying movement patterns while experiencing emotions. However, some movements are involuntary and not controlled by the person– *e.g.*, nervous twitches produced when someone is becoming anxious, nervous, or excitable. Trapezius EMG is shown to correlate effectively with users' stress level in [23].

B. Stimuli selection procedure

Many affective studies are conducted with image stimuli, and there exist standard affective image datasets (*e.g.* [14]) for researchers to evaluate their findings. However, in spite of studies confirming that reliable emotion elicitation and characterization is feasible with complex video stimuli such as movies [9], there exist few affective video datasets. An affective music video dataset comprising 40 one-minute affective

[1] *Valence* indicates the type of emotion induced by the stimulus in the viewer (*e.g.*, pleasant or unpleasant), while arousal denotes the intensity of emotion (*e.g.*, exciting or boring) [8].

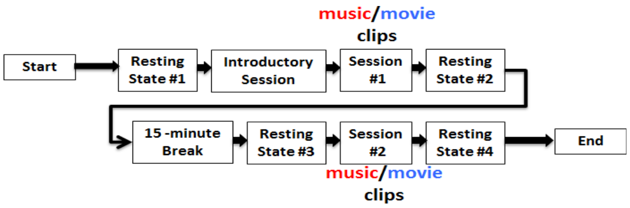

Fig. 1. Timeline for experimental protocol.

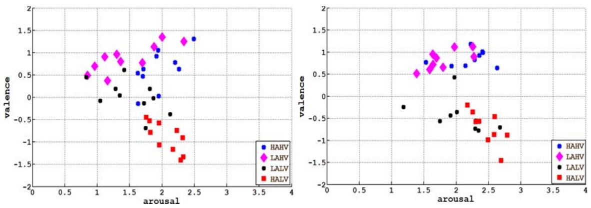

Fig. 2. Mean self-assessment AV ratings for music clips (left) and movie clips (right).

highlights has recently been presented in [13]. Our endeavor was to create a large-sized affective movie dataset along those lines since: (1) Temporal context, whose importance in emotion perception has been acknowledged [3], is conveyed effectively by both audio and visual content in movies, whereas context in music videos is predominantly conveyed by the audio, and supplemented by visuals; (2) Movies can effectively elicit a larger range of emotions as compared to music videos.

To this end, we initially compiled a set of 52 Hollywood movie clips, most of which are suggested as suitable for affective studies in [7], [4]. These clips were shown to 42 subjects, who self-assessed their emotional state to provide arousal and valence (AV) ratings as well as the most appropriate emotion tag for each movie clip. These AV ratings were considered as '***ground truth***' ratings in our study. We finally chose 36 movie clips which obtained the most consistent and representative scores and were uniformly distributed over the arousal–valence plane. These clips were 51s–128s long ($\mu = 80, \sigma = 20$), and were associated with diverse emotional tags such as *funny*, *amusing*, *exciting*, *sad*, *disgusting* and *angering*. To investigate whether MEG-based affect recognition is effective across stimuli types, we also used the 40 one-minute music video highlights suggested in [13] in our experiments.

C. Experimental set-up

1) Materials and set-up: All MEG recordings were performed in a shielded room with controlled illumination. Due to the sensitivity of the MEG equipment, all other devices used for data acquisition were placed in an adjacent room, and were controlled by the experimenter. Two PCs were used, one (Intel i7, 8 GB RAM) for stimulus presentation and the other for MEG data recording. The stimulus presentation protocol was developed using MATLAB's Psychtoolbox (http://psychtoolbox.org/) along with some functions adapted from the ASF stimulus presentation framework [19]. Also, synchronization markers were sent from the stimulus presenter PC to the MEG recorder at the beginning and end of each stimulus display. All stimuli were shown at a resolution of 1024 × 768 pixels and at a screen refresh rate of 60 Hz, and this display was projected onto a screen placed about a meter in front of the subject inside the MEG acquisition room. All music/movie clips were played at 20 frames/second, upon normalizing the audio volume to have a maximum power amplitude of 1. Also, participants were provided with a microphone to communicate with the experimenters. The ***Neuromag*** device, which outputs 306 channels (corresponding to 102 magnetometers and 204 gradiometers) with a sampling frequency of 1 KHz, is used for recording MEG responses.

2) Protocol: 18 university graduate students (8 male, 10 female, age range 27.3 ± 4.3) participated in the experiments. Data acquisition for each participant was spread over two sessions– movie clips were presented in one session, while music videos were presented in the other. The order of presentation of the music/movie clips was counterbalanced across subjects. During each session, the music/movie clips were shown in random order, such that two clips of similar valence and arousal did not follow one another. To avoid fatigue, each acquisition session was split into two halves (with 20 music or 18 movie clips shown in each half) and lasted for one hour. We recorded 1 minute of resting state brain activity before and after each session (see Fig.1 for protocol timeline).

Prior to each recording session, the participant was briefed about the experiment and was asked to remove any metallic objects he/she was wearing before entering the MEG room- this was mandatory as metals would interfere with the magnetic field. Then, a practice trial was conducted so that the subject could acquaint him/herself with the protocol. Each acquisition session involved a series of trials. During each trial, a fixation cross was first shown for 4 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 by the stimulus 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), (iii) Dominance ('How much are you in control of your emotions?') on a scale of 0 (in full control of emotions) to 4 (overwhelmed with emotions). A maximum of 15 seconds was available to the participant to convey each rating.

D. Self- assessment ratings– Music vs movie clips

In this section, we compare the valence-arousal ratings provided by participants for music and movie clips. Participant ratings are (i) a conscious reflection of their emotional state upon viewing the stimuli, and therefore, should be significantly correlated with their physiological responses (ii) ultimately used for valence and arousal classification, and the variance in ratings can provide vital cues regarding the best case classification results and (iii) also indicative of whether the presented stimuli can effectively evoke the expected emotional responses from viewers.

Fig.2 presents plots of the mean AV ratings obtained from 18 participants for the music and movie clips respectively. The blue, magenta, black and red colors are used to 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. A C-shape is observed for both movie and music clips, consistent with previous studies [14], [13]– it is attributed to the difficulty in evoking low arousal but strong valence responses. This phenomenon is particularly obvious in the case of music clips, where there is considerable overlap between the four clusters. For movie clips however, overlapping of clusters is observed only along the arousal dimension.

To further investigate this observation, we performed a Wilcoxon signed-rank test as in [13] to check if high and low arousal stimuli induced different valence ratings. The test showed that the valence ratings for high and low arousal movie stimuli were significantly different (p <0.005 in both cases). While this observation also holds for music clips, the valence ratings for low arousal stimuli vary less significantly as compared to movie clips (p <0.005 for high arousal music clips and p <0.01 for low arousal clips). Therefore, valence-arousal distinction is clearer for movie clips as compared to music clips.

Noting that significant inter-subject differences could have influenced the observed distribution of the mean ratings, we also performed a second experiment. Assuming that the ground-truth AV ratings were provided by an 'ideal' annotator, we compared the mean agreement between the participant and ground truth ratings using the Cohen's Kappa measure. To this end, we thresholded each user's ratings based on their mean rating, to assign a stimulus to either high/low valence/arousal. Then, we computed κ between the ground-truth and these labels. The mean κ over all subjects for the music-valence, music-arousal, movie-valence and movie-arousal were found to be 0.5111, 0.1728, 0.6574 and 0.2917 respectively. This again demonstrates that inter-rater agreement is higher for movie stimuli, especially for valence, between the two subject populations that provided the ground truth and performed the experiment. Therefore, movie stimuli are found to evoke similar emotions across viewers more effectively as compared to music videos. Likewise, we also obtain better affect classification with physiological features for movies as detailed in section V. The next section details the steps involved in the extraction of MEG and peripheral physiology features.

IV. Data Analysis

This section describes the procedure for (i) data preprocessing and feature extraction, (ii) classification. Compiled user ratings and affective response signals for both music and movie stimuli are processed in an identical manner.

A. Preprocessing the MEG data and feature extraction

The MEG data preprocessing consists of three main steps that are handled using the MATLAB Fieldtrip toolbox [17]:

Trial Segmentation: Participant responses corresponding to each trial are extracted by segmenting the MEG signal from 4 second prior to stimulus presentation (pre-stimulus) to the end of stimulus presentation. In this way, for each subject, we extract 36 and 40 trials for the movie clips and music videos respectively.

Frequency domain filtering: After downsampling the signal to 300 Hz, low-pass and high-pass filtering with cut-off frequencies of 95 Hz and 1 Hz respectively are performed. Applying the high-pass filter, low frequency noise in the MEG signal generated by moving vehicles is removed. Conversely, the low-pass filter removes some high frequency artifacts generated by muscle activities (between 110 Hz- 150 Hz).

Channel correction: Dead and bad channels are removed from the MEG data and replaced with interpolated values. Dead channels have zero value over time, while bad channels are outliers with respect to metrics such as signal variance and z-value of signal power over time. To preserve the consistency of MEG data over each trial and subject, removed channels are replaced with signals obtained from the interpolation of neighboring channels.

Upon segmenting the MEG response for each trial, the most informative content for affect classification needs to be extracted. In MEG studies, the spectral power of certain frequencies is the popularly used feature. There are several methods for computing spectral power of signals like Hanning tapers, multitapers and wavelet. Multitapers and wavelet are typically used in order to achieve a better control over the frequency smoothing. In these methods, high frequency smoothing has been found to be principally beneficial when dealing with brain signals above 30 Hz [16], [18]. Therefore, we use the variable-width wavelet method to transform our signal to the time-frequency domain for spectral power analysis.

We use a time-step of 1 second for temporal processing of the signal corresponding to each trial and a frequency step of 1 Hz to scan through a frequency range of 1-45 Hz. We linearly vary 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 perform the following steps: (a) We use a standard Fieldtrip function for combining the two planar gradiometers' spectral power for each sensor. (b) In order to better elucidate the MEG response dynamics following stimulus presentation, each trial power is divided by the baseline power between 2 and 1 second pre-stimulus interval. (c) To increase dynamic range of the spectral power, the time-frequency output is logarithm transformed.

Per subject and per movie clip, the output of the above time-frequency spectral power analysis is a 3-dimensional matrix with the following dimensions: synthetic information of 102 gradiometers × clip length time points × 45 frequencies. Similarly, for each of the 40 music clips, the output dimensions are 102×60×45.

1) MEG 3D DCT features: According to Ahmed *et al.* [2], signal information can be approximated effectively with few low-frequency DCT components. DCT is often used in signal, image and speech compression applications due to its strong energy compaction ability. For example, Davis *et al.* [5] showed that perceptually related aspects of the short-term speech spectrum contributed to superior performance of the mel-frequency cepstrum coefficients in speech applications. Inspired by these works, we employ DCT for compressing information encoded in time-frequency domain over channels. The 3D DCT coefficient matrix, B, is calculated as:

$$\begin{gathered} B_{pqr} = \alpha_p \alpha_q \alpha_r \sum_{l=0}^{L-1} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{lmn} \cos \frac{\pi \left(2l+1\right) p}{2L} \\ \cos \frac{\pi \left(2m+1\right) q}{2M} \cos \frac{\pi \left(2n+1\right) r}{2N}, \\ 0 \leq p \leq L-1, 0 \leq q \leq M-1, 0 \leq r \leq N-1 \end{gathered}$$

where $\alpha_p = \begin{cases} \frac{1}{\sqrt{L}}, & p=0 \\ \sqrt{\frac{2}{L}}, & 0<p \leq L-1 \end{cases}$ $\alpha_q = \begin{cases} \frac{1}{\sqrt{M}}, & q=0 \\ \sqrt{\frac{2}{M}}, & 0<q \leq M-1 \end{cases}$

$$\alpha_r = \begin{cases} \frac{1}{\sqrt{N}}, & q=r \\ \sqrt{\frac{2}{N}}, & 0<r \leq N-1 \end{cases}$$

Here, L, M, N represent the number of spatial, time and frequency steps, respectively. A is the 3-dimensional matrix of power spectral features. Therefore, after computing 3D DCT coefficients, we only use a sub-cube constructed by the first n coefficients from each of 3 feature dimensions. The feature vector for each trial will now contain n^3 entries. In our experiments, we assign $n = 3$ and the leading 27 coefficients are used as MEG 3D-DCT features. These DCT features incorporate information from the spatial, time and frequency dimensions.

2) MEG temporal events (MEG–TE): Spectral power analysis for the output of the 102 magnetometer channels is performed in an identical fashion as the gradiometer signals to obtain the 3-D matrix as above. Then, we average the spectral power over space and four major frequency bands that are: theta (3-7 Hz), alpha (8-13 Hz), beta (14-29 Hz) and gamma (30-45 Hz). 'Interesting' MEG temporal events are then calculated as the number of times the averaged spectral power is above/below its mean ± standard deviation (std) for each of the four frequency bands, and hence we get 8 features in total.

3) 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. In this study, we employ simple features like statistical measures of the signal and signal energy as adopted from [13] and [22] (see table I) .

4) ECG features: Common features that are extracted from the ECG signal are inter-beat interval (IBI), heart rate (HR), and heart rate variability (HRV). HR generally reflects emotional involvement, and it is also a commonly used feature for detecting valence. HRV refers to the temporal changes in consecutive IBI and it is useful for estimating the level of stress in adults. We adopt the procedures employed in Kim and André's work [12] to localize the heart beats and using this information, we extract IBI, HR, and HRV. We use the statistical measure (see table I) over the three heart activity information as features for classification.

5) Trapezius EMG: EMG is often used to analyze the correlation between cognitive emotion and physiological reactions [21]. In our experiment, we use bipolar electrodes that are placed in contact with the skin above the the trapezius muscle to measure the mental stress of users as in [12], [13]. We also use some of the features employed by Kim and André [12] and Koelstra *et al.* [13] (see table I).

B. Classification procedure

For each of the 18 subjects, we have 36 trials for movie clips and 40 trials for music videos. Having extracted two sets of MEG features, and three sets of peripheral physiology features for each trial, we need to determine the affective tag of a stimulus from these features. To this end, we solve three binary classification problems- employing MEG features to differentiate between (i) *low* versus *high* arousal, (ii) *low* versus *high* valence and (iii) *low* vs *high* dominance.

To begin with, each stimulus needs to be assigned a classification label. Given the large variability in the affective ratings provided by different subjects, a number of studies such as [13] adopt a subject-specific classification approach, with participant-specific stimulus labels. While participant-specific classification is also adopted in this work, we however use a *single label per stimulus* based on the mean affective rating provided by participants. To this end, the mean valence/arousal/dominance rating provided by all subjects is used as a threshold, and those ratings greater than or equal to the threshold are assumed to correspond to the 'high' label.

TABLE I. EXTRACTED FEATURES FROM EACH MODALITY

Modality	Extracted Features
MEG 3D-DCT (27)	the first 27 coefficients of the 3D-DCT transform over the time-frequency spectral power.
MEG temporal events (8)	the percentage of times the spectral power outcome is above/below its mean ± std for each of the four frequency bands.
ECG (18)	mean, standard deviation, skewness, kurtosis of inter-beat intervals (IBI), heart rate (HR), and heart rate variability (HRV) over time. Moreover, the percentage of times each of the IBI, HR, and HRV measurements had a value above/below their mean ± std. is employed
Horizontal EOG (7)	energy of the signal, mean, standard deviation, skewness, kurtosis of the signal samples over time as well as the percentage of times the signal is above/below its mean ± std.
Trapezius EMG (8)	energy of the signal, mean and variance of the signal over time as well as the percentage of times the signal is above/below its mean ± std.

TABLE II. DISTRIBUTION (NUMBERS AND %) OF SAMPLES IN EACH CLASS WITH MEAN RATING-BASED STIMULUS LABELING.

	Music video clips		**Movie video clips**	
Dimension	**High**	**Low**	**High**	**Low**
Arousal	22 (55%)	18 (45%)	22 (61.1%)	14 (38.9%)
Valence	22 (55%)	18 (45%)	19 (52.8%)	17 (47.2%)
Dominance	16 (40%)	24 (60%)	14 (38.9%)	22 (61.1%)

The distribution of 'high' and 'low' valence/arousal labels for the music and movie stimuli is presented in Table II. For both music and movie clips, the distribution of the 'high' and 'low' classes is unbalanced for both valence and arousal-valence distribution for movies is the most balanced. Given this unbalanced distribution of stimuli, we use F1-scores alongside classification accuracies to report our classification results. For classification, we use a linear SVM classifier with cost parameter C =1. We also adopt the leave-one-out cross validation scheme– for each participant, we train the model with all-but-one stimulus ratings and the corresponding physiological responses, and use this model to predict the label of the remaining stimulus.

V. EXPERIMENTAL RESULTS

Table III shows measured classification accuracy and F1-scores for music videos and movie clips, respectively. To test for significance, the F1-score distribution over participants is compared to the 0.5 baseline using an independent one-sample t-test. For movie clips, employing MEG 3D-DCT features, above-chance F1 scores are achieved for arousal, while statistically significant F1-scores are obtained for valence and dominance using the MEG-TE features. These results demonstrate that MEG signals effectively encode affective user responses, and suggest that the information extracted from the gradiometers and magnetometers are complementary in nature.

Employing a single label per stimulus, none of the classification results obtained with MEG features for music clips are significant. In order to compare the results using our framework with competing methods such as [13], we repeated the classification experiments with subject-specific stimulus labeling (denoted using SS). Classification results achieved in [13] are also listed. In this situation, comparable results are achieved for arousal classification– it is to be noted here that 32 subjects are part of the study reported in [13], while this work involves 18 subjects. Also, significant F1 scores are obtained with peripheral physiological features in this case.

Superior classification performance is achieved with both peripheral and MEG features for movie clips as compared to music clips. We also noted in section III that the affective ratings for movie stimuli are more consistent– both these trends suggest that movie clips are more effective stimuli than music clips for affective studies. Finally, better classification performance is achieved with MEG features for arousal, while peripheral features are found to be more effective for predicting valence– these results are consistent with the trends observed in [13].

TABLE III. ACC AND F1-SCORES OVER PARTICIPANTS FOR MUSIC AND MOVIE CLIPS. STARS INDICATE WHETHER THE F1-SCORE DISTRIBUTION OVER SUBJECTS IS SIGNIFICANTLY HIGHER THAN 0.5.($^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$).

Movie clips	**Arousal**		**Valence**		**Dominance**	
Feature Type	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
MEG 3D-DCT	0.61	0.60^{**}	0.49	0.49	0.52	0.51
MEG–TE	0.50	0.49	0.57	0.57^{***}	0.59	0.59^{***}
Peripheral	0.54	0.52	0.63	0.63^{***}	0.54	0.53
Music clips	**Arousal**		**Valence**		**Dominance**	
Feature Type	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
MEG 3D-DCT	0.51	0.50	0.53	0.53	0.52	0.51
MEG–TE	0.51	0.51	0.49	0.49	0.51	0.50
Peripheral	0.51	0.51	0.58	0.58^{***}	0.53	0.52
MEG 3D-DCT (SS)	0.57	0.56^{***}	0.52	0.52	0.51	0.49
MEG–TE (SS)	0.52	0.51	0.49	0.48	0.52	0.51
EEG [13]	0.62	0.58^{**}	0.58	0.56^{**}		
Peripheral (SS)	0.51	0.50	0.56	0.55*	0.49	0.46
Peripheral [13]	0.57	0.53^{*}	0.63	0.61^{**}		

VI. CONCLUSION AND THE FUTURE WORK

In this work, we have attempted affective tagging of music and movie videos through a user-centric approach employing MEG and peripheral physiology signals. Classification results are reported on data compiled from 18 users for a database comprising 40 music and 36 movie clips– one of the largest reported in literature. Obtained classification results suggest that MEG signals effectively encode affective viewer responses and are more useful for predicting stimulus arousal. As part of future work, we intend to (i) extend the current study by involving more subjects and using full-length movies, and (ii) develop an effective methodology for fusing the information from MEG and peripheral signals for better affect prediction.

Acknowledgments. This work was partially supported by A*STAR Singapore under the Human Sixth Sense Program (HSSP) grant and by the FIRB 2008 project S-PATTERNS.

REFERENCES

[1] Emotion and multimedia content. In B. Furht, editor, *Encyclopedia of Multimedia*. Springer, 2006.

[2] N. Ahmed, T. Natarajan, and K. Rao. Discrete cosine transfom. *IEEE Transactions on Computers*, 23:90–93, 1974.

[3] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

[4] E. E. Bartolini. Eliciting emotion with film: Development of a stimulus set. Master's thesis, Wesleyan University, 2001.

[5] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:357–366, 1980.

[6] 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.

[7] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.

[8] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[9] U. Hasson, R. Malach, and D. J. Heeger. Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1):40 – 48, 2010.

[10] J. V. Haxby, S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(8):404–416, 2011.

[11] 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 Appl.*, 51(2):505–523, 2011.

[12] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.

[13] 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.

[14] P. Lang, M. Bradley, and B. Cuthbert. IAPS: Affective ratings of pictures and instruction manual. Technical report, University of Florida, 2008.

[15] 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.

[16] P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophysical Journal*, 76(2):691 – 708, 1999.

[17] 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.

[18] D. Percival and A. Walden. *Spectral Analysis for Physical Applications*. 1993.

[19] 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.

[20] R. Sinha and O. A. Parsons. Multivariate response patterning of fear and anger. *Cognition and Emotion*, 10(2):173–198, 1996.

[21] D. M. Sloan. Emotion regulation in action: emotional reactivity in experiential avoidance. *Behaviour Research and Therapy*, 42(11):1257 – 1270, 2004.

[22] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *T. Affective Computing*, 3(1):42–55, 2012.

[23] 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.

[24] M. Yuki, W. W. Maddux, and T. Masuda. *Journal of Experimental Social Psychology*, 43(2):303–311, 2007.