http://eprints.gla.ac.uk/151479/

Deposited on: 21 November 2017

Evaluating Content-centric vs User-centric Ad Affect Recognition

Abhinav Shukla
International Institute of Information Technology
Hyderabad, India
abhinav.shukla@research.iiit.ac.in

Shruti Shriya Gullapuram
International Institute of Information Technology
Hyderabad, India
shruti.gullapuram@students.iiit.ac.in

Harish Katti
Centre for Neuroscience, Indian Institute of Science
Bangalore, India
harish2006@gmail.com

Karthik Yadati
Delft University of Technology
Delft, Netherlands
n.k.yadati@tudelft.nl

Mohan Kankanhalli
School of Computing, National University of Singapore
Singapore, Singapore
mohan@comp.nus.edu.sg

Ramanathan Subramanian
School of Computing Science, University of Glasgow
Singapore, Singapore
ramanathan.subramanian@glasgow.ac.uk

ABSTRACT

Despite the fact that **advertisements** (ads) often include strongly emotional content, very little work has been devoted to affect recognition (AR) from ads. This work explicitly compares **content-centric** and **user-centric** ad AR methodologies, and evaluates the impact of enhanced AR on computational advertising via a user study. Specifically, we (1) compile an affective ad dataset capable of evoking coherent emotions across users; (2) explore the efficacy of *content-centric* convolutional neural network (CNN) features for encoding emotions, and show that CNN features outperform low-level emotion descriptors; (3) examine *user-centered* ad AR by analyzing Electroencephalogram (EEG) responses acquired from eleven viewers, and find that EEG signals encode emotional information better than content descriptors; (4) investigate the relationship between objective AR and subjective viewer experience while watching an ad-embedded online video stream based on a study involving 12 users. To our knowledge, this is the first work to (a) expressly compare user vs content-centered AR for ads, and (b) study the relationship between modeling of ad emotions and its impact on a real-life advertising application.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; *User centered design*;

KEYWORDS

Affect recognition, Ads, Content-centric vs User-centric, CNNs, EEG, Multimodal analytics, Computational Advertising

1 INTRODUCTION

 Advertising is a rapidly evolving global industry that aims to induce consumers into preferentially buying specific products or services. In this digital age, audio-visual content is increasingly becoming the preferred means of delivering advertising campaigns. The global advertising industry is estimated to be worth over US $500 billion[1], and web advertising is expected to be a key profit-making sector with video advertising playing a significant role[2]. Advertisements (ads) often contain strongly emotional content to convey an effective message to viewers. Ad *valence* (pleasantness) and *arousal* (emotional intensity) are key properties that modulate emotional values and consumer attitudes associated with the advertised product [10, 11, 27]. In the context of Internet video advertising (as with *YouTube*), modeling the emotional relevance between ad and program content can improve program comprehension and advertisement brand recall, as well as optimize user experience [36].

Even though automated mining of ad emotions is beneficial, surprisingly very few works have attempted to computationally recognize ad emotions. This is despite the field of ***affective computing*** receiving considerable interest in the recent past, and a multitude of works modeling emotions elicited by image [6, 18], speech [24], audio [2], music [20] and movie [1, 33] content. Overall, affect recognition (AR) methods can be broadly classified as *content-centric* or *user-centric*. *Content-centric* AR approaches characterize emotions elicited by multimedia content via textual, audio and visual cues [9, 35]. In contrast, *user-centric* AR methods aim to recognize the elicited emotions by monitoring the user or multimedia consumer via facial [14] or physiological [1, 20, 32, 33] measurements.

This paper expressly examines and compares the utility of *content-centric* and *user-centric* approaches for ad AR. As emotion is a subjective human feeling, most recent AR methods have focused on a variety of human behavioral cues. Nevertheless, ads are different from conventional media such as movies, and are compact representations of themes and concepts which aim to impact the viewer within a short span of time. Thus, it would be reasonable to expect that ads contain powerful audio-visual content to convey the intended emotional message. While some works have compared content and user-centric features for AR, an explicit comparison has not been performed for ads to our knowledge. Another question

---

that we try to answer in this work, perhaps for the first time in affective computing, is *whether improved AR as given by objective measures, directly impacts subjective human experience* while using a multimedia application.

We first present a carefully curated affective ad dataset, capable of evoking coherent emotions across viewers as seen from emotional impressions reported by *experts* and *novice annotators*. On ensuring that the ads are able to reliably evoke target emotions (in terms of arousal and valence levels), we examine the efficacy of content and user-based methods for modeling ad emotions–specifically, high-level convolutional neural network (CNN) features and low-level audio visual descriptors [9] are explored for content-centered analysis, while EEG measurements are employed for user-centered AR. CNN features outperform low-level audio-visual descriptors, but are inferior to EEG signals implying that user-centric cues enable superior ad AR. We then show how improved AR achieved by the CNN and EEG features reflects in terms of better ad memorability and user experience for a computational advertising application [36].

To summarize, this work makes the following contributions: (1) To our knowledge, this is the first work to explicitly compare and contrast content-centered and user-centered ad AR; (2) This is also the first work to demonstrate how an improvement in objective AR performance improves subjective ad memorability and user experience while watching an ad-embedded online video stream. Our findings show that enhanced AR can facilitate better ad insertion onto broadcast multimedia content; (3) The compiled dataset of 100 affective ads along with accompanying subjective ratings and EEG responses is unique for ad-based AR.

The paper is organized as follows. Section 2 reviews related literature, while Section 3 overviews the compiled ad dataset and the EEG acquisition protocol. Section 4 presents the techniques adopted for content and user-centered ad AR, while Section 5 discusses AR results. Section 6 describes a user study to establish how improved AR facilitates computational advertising. Section 7 summarizes the main findings and concludes the paper.

2 RELATED WORK

To position our work with respect to the literature and highlight its novelty, we review the related work examining (a) affect recognition (b) the impact of affective ads on consumer behavior (c) computational advertising.

2.1 Affect recognition

Building on the circumplex emotion model that represents emotions in terms of valence and arousal [29], many computational methods have been designed for affect recognition. Typically, such approaches are either *content-centric* which employ image, audio and video-based emotion correlates [9, 30, 34] to recognize affect in a supervised manner; or *user-centric*, which measure stimulus-driven variations in specific physiological signals such as pupillary dilation [15], gazing patterns [28, 32] and neural activity [1, 20, 37]. Performance of these models is typically subject to the variability in subjective, human-annotated labels, and careful affective labeling is crucial for successful AR. We carefully curate a set of 100 ads such that they are assigned very similar emotional labels by

two independent groups comprising experts and novice annotators. These ads are then mined for emotional content via content and user-based methods. User-centered AR is achieved via EEG signals acquired via the wireless and wearable *Emotiv* headset, while facilitates naturalistic user behavior and can be employed for large-scale AR.

2.2 Emotional impact of ads

Ad-induced emotions have been shown to shape consumer behavior in a significant manner [10, 11]. Although this key observation was made nearly three decades ago [10], computational advertising methods till recently have matched low-level visual and semantic properties between video segments and candidate ads [25]. Recent work [27] indicates a shift form the traditional thinking by emphasizing that ad-evoked emotions can change brand perception among consumers. A very recent and closely related work to ours [30] discusses how efficient affect recognition from ads via deep learning and multi-task learning can lead to improved online viewing experience. In this work, we show how effectively recognizing emotions from ads via content and user-based methods can achieve optimized insertion of ads onto streamed/broadcast videos via the CAVVA framework [36]. A user study shows that better ad AR translates to better ad memorability and enhanced user experience while watching an ad-embedded video stream.

2.3 Computational advertising

Exploiting affect recognition models for commercial applications has been a growing trend in recent years. The field of **computational advertising** focuses on presenting contextually relevant ads to multimedia users for commercial benefits, social good or to induce behavioral change. Traditional computational advertising approaches hae worked by exclusively modeling low-level visual and semantic relevance between video scenes and ads [25]. A paradigm shift in this regard was introduced by the CAVVA framework, which proposed an optimization-based approach to insert ads onto a video stream based on the emotional relevance between the video scenes and candidate ads. CAVVA employed a *content-centric* approach to match video scenes and ads in terms of emotional valence and arousal. However, this could be replaced by an interactive and **user-centric** framework as described in [15]. We explore the use of both *content-centric* (via CNN features) and **user-centric** (via EEG features) methods for formulating an ad-insertion strategy. A user study shows that CNN-based ad insertion results in better ad memorability, while an EEG-based strategy achieves the best user experience. The following section describes the compiled ad dataset, and the EEG acquisition protocol.

3 ADVERTISEMENT DATASET

This section presents details regarding the ad dataset used in this study along with the protocol employed for collecting EEG responses for user-centric AR.

3.1 Dataset Description

Defining **valence** as the feeling of *pleasantness*/*unpleasantness* and **arousal** as the *intensity of emotional feeling* while viewing an audio-visual stimulus, five experts carefully compiled a dataset of 100,

Table 1: Summary statistics for quadrant-wise ads.

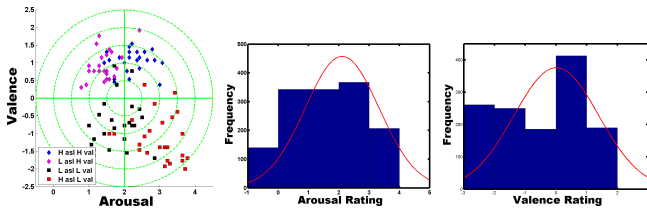| Quadrant	Mean length (s)	Mean asl	Mean val
H asl, H val	48.16	2.17	1.02
L asl, H val	44.18	1.37	0.91
L asl, L val	60.24	1.76	-0.76
H asl, L val	64.16	3.01	-1.16

Figure 1: (left) Scatter plot of mean asl, val ratings color-coded with expert labels. (middle) Asl and (right) Val rating distribution with Gaussian pdf overlay (view under zoom).

roughly 1-minute long commercial advertisements (ads) which are used in this work. These ads are publicly available[3] and found to be uniformly distributed over the arousal–valence plane defined by Greenwald *et al.* [8] (Figure 1). An ad was chosen if there was consensus among all five experts on its valence and arousal labels (defined as either *high* (H)/*low* (L)). The high valence ads typically involved product promotions, while low valence ads were social messages depicting the ill effects of smoking, alcohol and drug abuse, *etc.*. Labels provided by experts were considered as **ground-truth**, and used for all recognition experiments in this work.

To evaluate the effectiveness of these ads as affective control stimuli, we examined how consistently they could evoke target emotions across viewers. To this end, the ads were independently rated by 14 annotators for valence (val) and arousal (asl)[4]. All ads were rated on a 5-point scale, which ranged from -2 (*very unpleasant*) to 2 (*very pleasant*) for val and 0 (*calm*) to 4 (*highly aroused*) for asl. Table 1 presents summary statistics for ads over the four quadrants. Evidently, low val ads are longer and are perceived as more arousing than high val ads suggesting that they evoked stronger emotional feelings among viewers.

Furthermore, we computed agreement among raters in terms of the (i) Krippendorff's α and (ii) Cohen's κ scores. The α coefficient is applicable when multiple raters code data with ordinal scores–we obtained $\alpha = 0.60$ and 0.37 for val and asl implying valence impressions were most consistent across raters. We then computed the κ agreement between annotator and ground-truth labels to determine concordance between the annotator and expert groups. To this end, we thresholded each rater's asl, val scores by their mean rating to assign H/L labels for each ad, and compared them against ground-truth labels. This procedure revealed a mean agreement of 0.84 for val and 0.67 for asl across raters. Computing κ between the annotator and expert *populations* by thresholding the mean asl, val score per ad across raters against the grand mean gave a $\kappa = 0.94$ for val and 0.67 for asl[5]. Clearly, there is good-to-excellent

---

[3]On video hosting websites such as YouTube.
[4]Annotators were familiarized with emotional attributes prior to the rating task.
[5]Chance agreement corresponds to a κ value of 0.

agreement between annotators and experts on affective impressions with considerably higher concordance for val. The observed concordance between the independent expert and annotator groups affirms that the compiled 100 ads are effective control stimuli for affective studies.

Another desirable property of an affective dataset is the independence of the asl and val dimensions. We (i) examined scatter plots of the annotator ratings, and (ii) computed correlations amongst those ratings. The scatter plot of the mean asl, val annotator ratings, and the distribution of asl and val ratings are presented in Fig. 1. The scatter plot is color-coded based on expert labels, and is interestingly different from the classical 'C' shape observed with images [23], music videos [20] and movie clips [1] owing to the difficulty in evoking medium asl/val but strong val/asl responses. The distributions of asl and val ratings are also roughly uniform resulting in Gaussian fits with large variance, with modes observed at the median scale values of 2 and 0 respectively. A close examination of the scatter plot reveals that a number of ads are rated as moderate asl, but high/low val. This is owing to the fact that ads are designed to convey a strong positive or negative message to viewers, which is not typically true of images or movie scenes. Finally, Wilcoxon rank sum tests on annotator ratings revealed significantly different asl ratings for high and low asl ads ($p < 0.00005$), and distinctive val scores for high and low valence ads ($p < 0.000001$), consistent with expectation.

Pearson correlation was computed between the asl and val dimensions with correction for multiple comparisons by limiting the false discovery rate to within 5% [5]. This procedure revealed a weak and insignificant negative correlation of 0.19, implying that ad asl and val scores were largely uncorrelated. Overall, (i) Our ads constitute a control affective dataset as asl and val ratings are largely independent; (ii) Different from the 'C'-shape characterizing the asl-val relationship for other stimulus types, asl and val ratings are uniformly distributed for the ad stimuli, and (iii) There is considerable concordance between the experts and annotators on affective labels, implying that the selected ads effectively evoke coherent emotions across viewers.

3.2 EEG acquisition protocol

As 11 of the 14 annotators rated the ads for asl and val upon watching them, we acquired their Electroencephalogram (EEG) brain activations via the *Emotiv* wireless headset. To maximize engagement and minimize fatigue during the rating task, these raters took a break after every 20 ads, and viewed the entire set of 100 ads over five sessions. Upon viewing each ad, the raters had a maximum of 10 seconds to input their asl and val scores via mouse clicks. The Emotiv device comprises 14 electrodes, and has a sampling rate of 128 Hz. Upon experiment completion, the EEG recordings were segmented into *epochs*, with each epoch denoting the viewing of a particular ad. Upon removal of noisy epochs, we were left with a total of 804 *clean* epochs. Each ad was preceded by a 1s fixation cross to orient user attention, and to measure resting state EEG power used for baseline power subtraction. The EEG signal was band-limited between 0.1–45 Hz, and independent component analysis (ICA) was performed to remove artifacts relating to eye movements,

eye blinks and muscle movements. The following section describes the techniques employed for content and user-centered AR.

4 CONTENT & USER-CENTERED ANALYSIS

This section presents the modeling techniques employed for content-centered and user-centered ad affect recognition.

4.1 Content-centered Analysis

For content centered analysis, we employed a convolutional neural network (CNN)-based model, and the popular affective model of Hanjalic and Xu based on low-level audio visual descriptors [9]. CNNs have recently become very popular for visual [22] and audio [12] recognition, but they require vast amounts of training data. As our ad dataset comprised only 100 ads, we fine-tuned the pretrained *places205* [22] model via the affective LIRIS-ACCEDE movie dataset [4], and employed the fine-tuned model to extract emotional descriptors for our ads. This process is termed as **domain adaptation** in machine learning literature.

In order to learn deep features for ad AR, we employed the *Places205* CNN [19] originally trained for image classification. *Places205* is trained using the Places-205 dataset comprising 2.5 million images involving 205 scene categories. The Places-205 dataset contains a wide variety of scenes captured under varying illumination, viewpoint and field of view, and we hypothesized a strong relationship between scene perspective, lighting and the scene mood. The **LIRIS-ACCEDE** dataset contains asl, val ratings for ≈ 10 s long movie snippets, whereas our ads are about a minute-long with individual ads ranging from 30–120 s.

4.1.1 FC7 Feature Extraction via CNNs.
For deep CNN-based ad AR, we represent the *visual* modality using *key-frame* images, and the *audio* modality using *spectrograms*. We fine-tune *Places205* via the LIRIS-ACCEDE [4] dataset, and employ this model to compute the fully connected layer (fc7) visual and audio ad descriptors.

Keyframes as Visual Descriptors. From each video in the ad and LIRIS-ACCEDE datasets, we extract one *key frame* every three seconds– this enables extraction of a continuous video profile for affect prediction. This process generates a total of 1791 key-frames for our 100 ads.

Spectrograms as Audio Descriptors. Spectrograms (SGs) are visual representations of the audio frequency spectrum, and have been successfully employed for AR from speech and music [3]. Specifically, transforming the audio content to a spectrogram image allows for audio classification to be treated as a visual recognition problem. We extract spectrograms over the 10s long LIRIS-ACCEDE clips, and consistently from 10s ad segments. This process generates 610 spectrograms for our ad dataset. Following [3], we combine multiple tracks to obtain a single spectrogram (as opposed to two for stereo). Each spectrogram is generated using a 40 ms window short time Fourier transform (STFT), with 20 ms overlap. Larger densities of high frequencies can be noted in the spectrograms for high asl ads, and these intense scenes are often characterized by sharp frequency changes.

CNN Training. We use the Caffe [13] deep learning framework for fine-tuning *places205*, with a momentum of 0.9, weight decay

Table 2: Extracted features for content-centric AR.

| Attribute | Valence/Arousal | | |
	Audio	Video	aud+vid (A+V)
CNN Features	4096D Alexnet FC7 features obtained with 10s SGs.	4096D Alexnet FC7 features by extracted from keyframes sampled every 3 seconds.	8192D FC7 features with SGs + keyframes over 10s intervals.
Hanjalic [9] Features	Per-second sound energy and pitch statistics [9].	Per-second shot change frequency and motion statistics [9].	Concatenation of audio-visual features.

of 0.0005, and a base learning rate of 0.0001 reduced by $\frac{1}{10}^{th}$ every 20000 iterations. We totally train four binary classification networks to recognize high and low asl/val from audio/visual features. To fine-tune *places205*, we use only the top and bottom 1/3rd LIRIS-ACCEDE videos in terms of asl and val rankings under the assumption that descriptors learned for the extreme-rated clips will effectively model affective concepts. 4096-dimensional *fc7* layer outputs extracted from the four networks for our 100 ads are used in the experiments.

4.1.2 AR with audio-visual features.
We will mainly compare our CNN-based AR framework against the algorithm of Hanjalic and Xu [9] in this work. Even after a decade, this algorithm remains one of the most popular AR baselines as noted from recent works such as [1, 20]. In [9], asl and val are modeled via low-level descriptors describing motion activity, colorfulness, shot change frequency, voice pitch and sound energy in the scene. These hand-crafted features are intuitive and interpretable, and employed to estimate time-continuous asl and val levels conveyed by the scene. Table 2 summarizes the audio-visual features used for content-centric AR.

4.2 User-centered analysis

The 804 clean epochs obtained from the EEG acquisition process were used for user-centered analysis. However, these 804 epochs were of different lengths as the duration of each ad was variable. To maintain dimensional consistency, we performed user-centric AR experiments with (a) the *first* 3667 samples (≈ 30s of EEG data), (b) the *last* 3667 samples and (c) the *last* 1280 samples (10s of EEG data) from each epoch. Each epoch sample comprises data from 14 EEG channels, and the epoch samples were input to the classifier upon vectorization.

5 EXPERIMENTS AND RESULTS

We first provide a brief description of the classifiers used and settings employed for binary content-centric and user-centric AR, where the objective is to assign a binary (H/L) label for asl and val evoked by each ad, using the extracted fc7/low-level audio visual/EEG features. The ground truth here is provided by the experts, and has a substantial agreement with the user ratings in Sec. 3.1. Experimental results will be discussed thereafter.

Classifiers: We employed the Linear Discriminant Analysis (LDA), linear SVM (LSVM) and Radial Basis SVM (RSVM) classifiers in our AR experiments. LDA and LSVM separate H/L labeled training data with a hyperplane, while RSVM is a non-linear classifier which separates H and L classes, linearly inseparable in the input space, via transformation onto a high-dimensional feature space.

Metrics and Experimental Settings: We used the F1-score (F1), defined as the harmonic mean of precision and recall as our performance metric, due to the unbalanced distribution of positive

Table 3: Ad AR from content analysis. F1 scores are presented in the form $\mu \pm \sigma$.

Method	Valence			Arousal		
	F1 (all)	F1 (L30)	F1 (L10)	F1 (all)	F1 (L30)	F1 (L10)
Audio FC7 + LDA	0.61±0.04	0.62±0.10	0.55±0.18	0.65±0.04	0.59±0.10	0.53±0.19
Audio FC7 + LSVM	0.60±0.04	0.60±0.09	0.55±0.19	0.63±0.04	0.57±0.09	0.50±0.18
Audio FC7 + RSVM	0.64±0.04	**0.66±0.08**	0.62±0.17	**0.68±0.04**	0.60±0.10	0.53±0.19
Video FC7 + LDA	0.69±0.02	0.79±0.08	0.77±0.13	0.63±0.03	0.58±0.10	0.57±0.18
Video FC7 + LSVM	0.69±0.02	0.74±0.08	0.70±0.15	0.62±0.02	0.57±0.09	0.52±0.17
Video FC7 + RSVM	0.72±0.02	**0.79±0.07**	0.74±0.15	**0.67±0.02**	0.62±0.10	0.58±0.19
A+V FC7 + LDA	0.70±0.04	0.66±0.08	0.49±0.18	0.60±0.04	0.52±0.10	0.51±0.18
A+V FC7 + LSVM	0.71±0.04	0.66±0.07	0.49±0.19	0.56±0.04	0.49±0.10	0.47±0.19
A+V FC7 + RSVM	**0.75±0.04**	0.70±0.07	0.55±0.17	**0.63±0.04**	0.56±0.11	0.49±0.19
A+V Han + LDA	0.59±0.09	0.63±0.08	0.64±0.12	0.54±0.09	0.50±0.10	0.58±0.08
A+V Han + LSVM	0.62±0.09	0.62±0.10	0.65±0.11	0.55±0.10	0.51±0.11	0.57±0.09
A+V Han + RSVM	**0.65±0.09**	0.62±0.11	0.62±0.12	**0.59±0.12**	0.58±0.11	0.56±0.10
A+V FC7 LDA DF	0.60±0.04	0.66±0.04	0.70±0.19	0.59±0.02	0.60±0.07	0.57±0.15
A+V FC7 LSVM DF	0.65±0.02	0.66±0.04	0.65±0.08	0.60±0.04	0.63±0.10	0.53±0.13
A+V FC7 RSVM DF	**0.72±0.04**	0.70±0.04	0.70±0.12	0.69±0.06	**0.75±0.07**	0.70±0.07
A+V Han LDA DF	0.58±0.09	0.58±0.09	**0.61±0.09**	0.59±0.06	0.59±0.07	0.61±0.08
A+V Han LSVM DF	0.59±0.10	0.59±0.09	0.60±0.10	**0.61±0.05**	0.61±0.08	0.60±0.09
A+V Han RSVM DF	0.60±0.08	0.56±0.10	0.58±0.09	0.58±0.09	0.56±0.06	0.58±0.09

Table 4: Ad AR from EEG analysis. F1 scores are presented in the form $\mu \pm \sigma$.

Method	Valence			Arousal		
	F1 (F30)	F1 (L30)	F1 (L10)	F1 (F30)	F1 (L30)	F1 (L10)
LDA	0.79 ± 0.03	0.79 ± 0.03	0.75 ± 0.03	0.75 ± 0.03	0.74 ± 0.03	0.71 ± 0.04
LSVM	0.77 ± 0.03	0.76 ± 0.04	0.77 ± 0.05	0.74 ± 0.03	0.73 ± 0.02	0.69 ± 0.04
RSVM	**0.83 ± 0.03**	**0.83 ± 0.03**	**0.81 ± 0.03**	**0.80 ± 0.02**	**0.80 ± 0.03**	**0.76 ± 0.04**

and negative samples. For content-centric AR, apart from unimodal (audio (A) or visual (V)) fc7 features, we also employed feature fusion and probabilistic decision fusion of the unimodal outputs. Feature fusion (A+V) involved concatenation of fc7 A and V features over 10 s windows (see Table 2), while the W_{est} technique [21] was employed for decision fusion (DF). In DF, the test label is assigned the index i corresponding to maximum $P_i = \sum_{i=1}^{2} \alpha_i^* t_i p_i$, where i denotes the A,V modalities, p_i's denote posterior A,V classifier probabilities and $\{\alpha_i^*\}$ are the optimal weights maximizing test F1-score, and determined via a 2D grid search. If F_i denotes the training F1-score for the ith modality, then $t_i = \alpha_i F_i / \sum_{i=1}^{2} \alpha_i F_i$ for given α_i. Note that the use of a validation set for parameter tuning is precluded by the small dataset size as with [1,18] and that the DF results denote 'maximum possible' performance.

As the Hanjalic (Han) algorithm [9] uses audio plus visual features to model asl and val, we only consider (feature and decision) fusion performance in this case. User-centered AR uses only EEG information. As we evaluate AR performance on a small dataset, AR results obtained over 10 repetitions of 5-fold cross validation (CV) (total of 50 runs) are presented. CV is typically used to overcome the *overfitting* problem on small datasets, and the optimal SVM parameters are determined from the range $[10^{-3}, 10^3]$ via an inner five-fold CV on the training set. Finally, in order to examine the temporal variance in AR performance, we present F1-scores obtained over (a) all ad frames ('All'), (b) last 30s (L30) and (c) last 10s (L10) for *content-centered* AR, and (a) first 30s (F30), (b) last 30s (L30) and (c) last 10s (L10) for *user-centered* AR. These settings were

chosen bearing in mind that EEG sampling rate is much higher than the audio or video sampling rate.

5.1 Results Overview

Tables 3 and 4 respectively present content-centric and user-centric AR results for the various settings described above. The highest F1 score achieved for a given temporal setting across all classifiers and either unimodal or multimodal features is denoted in bold. Based on the observed results, we make the following claims.

Superior val recognition is achieved with both *content-centric* and *user-centric* methods. Focusing on *content-centric* results, **unimodal fc7 features**, val (peak F1 = 0.79) is generally recognized better than asl (peak F1 = 0.68) and especially with video features. A and V fc7 features perform comparably for asl. Concerning recognition with **fused fc7 features**, comparable or better F1 scores are achieved with multimodal approaches. In general, better recognition is achieved via decision fusion as compared to feature fusion[6]. For val, the best fusion performance (0.75 with feature fusion and RSVM classifier) is superior compared to A-based (F1 = 0.66), but inferior compared to V-based (F1 = 0.79) recognition. Contrastingly for asl, fusion F1-score (0.75 with DF) considerably outperforms unimodal methods (0.68 with A, and 0.67 with V). Comparing ***A+V fc7 vs Han*** features, fc7 descriptors clearly outperform Han features and the difference in performance is prominent for val, while comparable recognition is achieved with both features for asl. The

---

[6]To our knowledge, either of feature or decision fusion may work better depending on the specific problem and available features.

RSVM classifier produces the best F1-scores for both asl and val with unimodal and multimodal approaches.

User-centric or ***EEG***-based AR results are generally better than *content-centric* results achieved under similar conditions. The best *user-centric* val and asl F1-scores are considerably higher than the best *content-centric* results. Again, val is recognized better than asl with EEG data (as with the *content-centric* case), which is interesting as EEG is known to correlate better with asl rather than val. Nevertheless, positive val is found to correlate with higher activity in the frontal lobes as compared to negative val as noted in [26], and the Emotiv device is known to efficiently capture frontal lobe activity despite its limited spatial resolution. Among the three classifiers considered with EEG data, RSVM again performs best while LSVM performs worst.

Focusing on the different temporal conditions considered in our experiments, relatively small σ values are observed for the 'All' *content-centric* condition with the five-fold CV procedure (Table 3), especially with fc7 features. Still lower σ's can be noted with EEG-based classification results, suggesting that our overall AR results are **minimally impacted** by **overfitting**. Examining **temporal windows** considered for *content-centered* AR, higher σ's are observed for the L30 and L10 cases, which denote model performance on the terminal ad frames. Surprisingly, one can note a general degradation in asl recognition for the L30 and L10 conditions with A/V features, while val F1-scores are more consistent.

Three inferences can be made from the above observations, namely, (1) Greater heterogeneity in the ad content towards endings is highlighted by the large variance with fusion approaches; (2) Fusion models synthesized with Han features appear to be more prone to overfitting, given the generally larger σ values seen with the models; (3) That asl recognition is lower in the L30 and L10 conditions highlights the limitation of using a *single* asl/val label (as opposed to dynamic labeling) over time. Generally lower F1-scores achieved for asl with all methods suggests that asl is a more transient phenomenon as compared to val, and that coherency between *content-based* val features and labels is sustainable over time.

User-centered AR results obtained over the first 30, last 30 and final 10 s for the ads are relatively more stable than *content-centered* results, especially for val. However, there is a slight dip in AR performance for asl over the final 10s. As the ads were roughly one minute long, we can infer that (a) the consistent F1 scores achieved for the firs and last 30s suggests that humans tend to perceive the ad mood rather quickly. This is in line with the objective of ad makers, who endeavor to convey an effective message within a short time duration. However, the dip in asl performance over the final 10s as with *content centered* methods again highlights the limitation of using a single affective label over the entire ad duration.

5.2 Discussion

We now summarize and compare the *content-centric* and *user-centric* AR results. Between the *content-centric* features, the deep CNN-based ***fc7*** descriptors considerably outperform the audio-visual ***Han*** features. Also, the classifiers trained with ***Han*** features are more prone to over-fitting than ***fc7***-based classifiers, suggesting that the CNN descriptors are more robust as compared to low-level ***Han*** descriptors. Fusion-based approaches do not perform much

Name	Scene length (s)	Manual Rating	
		Valence	Arousal
coh	127±46	0.08±1.18	1.53±0.58
ipoh	110±44	0.03±1.04	1.97±0.49
friends	119±69	1.08±0.37	2.15±0.65

better than unimodal methods. However, ***EEG***-based AR achieves the best performance, considerably outperforming content-based features and thereby endorsing the view that emotions are best characterized by human behavioral cues.

Superior val recognition is achieved with both *content-centric* and *user-centric* AR methods. Also, temporal analysis of classification results reveals that content-based val features as well as user-based val impressions are more stable over time, but asl impressions are transient. Cumulatively, the obtained results highlight the need for fine-grained and dynamic AR methods as against most contemporary studies which assume a single, *static* affective label per stimulus.

6 COMPUTATIONAL ADVERTISING- USER STUDY

Given that superior ad AR is achieved with user EEG responses (see Table 4), we examined if enhanced AR resulted in the insertion of appropriate ads at vantage temporal positions within a streamed video, as discussed in the CAVVA video-in-video ad insertion framework [36]. CAVVA is an optimization-based framework for ad insertion onto streamed videos (as with *YouTube*). It formulates an advertising schedule by modeling the emotional relevance between video scenes and candidate ads to determine (a) the subset of ads for insertion, and (b) the temporal positions (typically after a scene ending) at which the chosen ads are to be inserted. In effect, CAVVA aims to strike a balance between (a) maximizing ad impact in terms of brand memorability, and (b) minimally disrupting (or enhancing) viewer experience while watching the program video onto which ads are inserted. We hypothesized that better ad affect recognition should lead to optimal ad insertions, and consequently better viewing experience. To this end, we performed a user study to compare the subjective quality of advertising schedules generated via ad asl and val scores generated with the content-centric **Han** [9] and **Deep** CNN models, and the user-centric **EEG** model.

6.1 Dataset

For performing the user study, we used 28 ads (out of the 100 in the original dataset), and three program videos. The ads were equally divided into four quadrants of the valence-arousal plane based on asl and val labels provided by experts. The program videos were scenes from a television sitcom (*friends*) and two movies (*ipoh* and *coh*), which predominantly comprised social themes and situations capable of invoking high-to-low valence and moderate arousal (see Table 5 for summary statistics). Each of the program videos comprised eight scenes implying that there were seven candidate ad-insertion points in the middle of each sequence. The average scene length was found to be 118 seconds.

6.2 Advertisement insertion strategy

We used the three aforementioned models to perform ad affect estimation. For the 24 program video scenes (3 videos × 8 scenes), the average of asl and val ratings acquired from three experts was used to denote affective scores. For the ads, affective scores were computed as follows. For the **Deep** method, we used normalized softmax class probabilities [7] output by the video-based CNN model for val estimation, and probabilities from the audio CNN for asl estimation. The mean score over all video/audio ad frames was used to denote the affective score in this method. The average of the per-second asl and val level estimates over the ad duration was used to denote affective scores for the **Han** approach. Mean of SVM class posteriors over all EEG epochs was used for the **EEG** method. We then adopted the CAVVA optimization framework [36] to obtain nine unique **video program sequences** (with average length of 19.6 minutes) comprising the inserted ads. These video program sequences comprised ads inserted via the three affect estimation approaches onto each of the three program videos. Exactly 5 ads were inserted (out of 7 possible) onto each program video. 21 of the 28 chosen ads were inserted at least once into the nine video programs, with maximum and mean insertion frequencies of 5 and 2.14 respectively.

6.3 Experiment and Questionnaire Design

To evaluate the subjective quality of the generated video program sequences and thereby the utility of the three affect estimation techniques for computational advertising, we recruited 12 users (5 female, mean age 19.3 years) who were university undergraduates/graduates. Each of these users viewed a total of three video program sequences, corresponding to the three program videos with ad insertions performed using *one* of the three affect estimation approaches. We used a randomized 3×3 Latin square design in order to cover all the nine generated sequences with every set of three users. Thus, each video program sequence was seen by four of the 12 viewers, and we have a total of 36 unique responses.

We designed a questionnaire for the user evaluation so as to reveal whether the generated video program sequences (a) included seamless ad insertions, (b) facilitated user engagement (or alternatively, resulted in minimum disruption) towards the streamed video and inserted ads and (c) ensured good overall viewer experience. To this end, we evaluated whether a particular ad insertion strategy resulted in (i) increased brand recall (both *immediate* and *day-after* recall) and (ii) minimal viewer disturbance or enhanced user viewing experience.

Recall evaluation to intended to verify if the inserted ads were attended to by viewers, and the immediate and day-after recall were **objective** measures that quantified the impact of ad insertion on the short-term (immediate) and long-term (day-after) memorability of advertised content, upon viewing the program sequences. Specifically, we measured the proportion of (i) inserted ads that were recalled correctly (*Correct* recall), (ii) inserted ads that were not recalled (*Forgotten*) and (iii) non-inserted ads incorrectly recalled as viewed (*Incorrect* recall). For those inserted ads which were correctly recalled, we also assessed whether viewers perceived them to be contextually (emotionally) relevant to the program content.

Upon viewing a video program sequence, the viewer was provided with a representative visual frame from each of the 28 ads to test ad recall along with a sequence-specific response sheet. In addition to the recall related questions, we asked viewers to indicate if they felt that the recalled ads were inserted at an appropriate position in the video (*Good insertion*) to verify if ad positioning positively influenced recall. All recall and insertion quality-related responses were acquired from viewers as binary values. In addition to these objective measures, we defined a second set of **subjective** user experience measures, and asked users to provide ratings on a Likert scale of 0–4 for the following questions with 4 implying *best* and 0 denoting *worst*:

1. Were the advertisements uniformly distributed across the video program?
2. Did the inserted advertisements blend with the program flow?
3. Whether the inserted ads were relevant to the surrounding scenes with respect to their content and mood?
4. What was the overall viewer experience while watching each video program?

Each participant filled the recall and experience-related questionnaires immediately after watching each video program. Viewers also filled in the day-after recall questionnaire, a day after completing the experiment.

6.4 Results and Discussion

As mentioned previously, scenes from the program videos were assigned asl, val scores based on manual ratings from three experts, while the Deep, Han and EEG-based methods were employed to compute affective scores for ads. The overall quality of the CAVVA-generated video program sequence hinges on the quality of affective ratings assigned to both the video scenes and ads. In this regard, we hypothesized that better ad affect estimation would result in optimized ad insertions.

Firstly, we computed the similarity in terms of the ad asl and val scores generated by the three approaches in terms of Pearson correlations, and found that (1) there was significant and positive correlation between asl scores generated by the Han–EEG ($\rho = 0.5, p < 0.01$) as well as the Han–Deep methods ($\rho = 0.42, p < 0.05$). However, the Deep and EEG-based asl scores did not agree significantly ($\rho = 0.22$, n.s.). For val, the only significant correlation was noted between the Han and Deep approaches ($\rho = 0.41, p < 0.05$), while the Han and EEG ($\rho = 0.07$, n.s.) as well as the Deep and EEG val scores ($\rho = 0.20$, n.s.) were largely uncorrelated. This implies that while methods content-centric and user-centric methods agree well on asl scores, there is significant divergence between the val scores generated by the two approaches.

Based on the questionnaire responses received from viewers, we computed the mean proportions for correct recall, ad forgottenness, incorrect recall and good insertions immediately and a day after the experiment. Figure 2 presents the results of our user study and there are several interesting observations. A key measure indicative of a successful advertising strategy is *high brand recall* [11, 15, 36], and the immediate and day-after recall rates observed for three considered approaches are presented in Fig. 2 (left),(middle). Video program sequences obtained with Deep affective scores result in
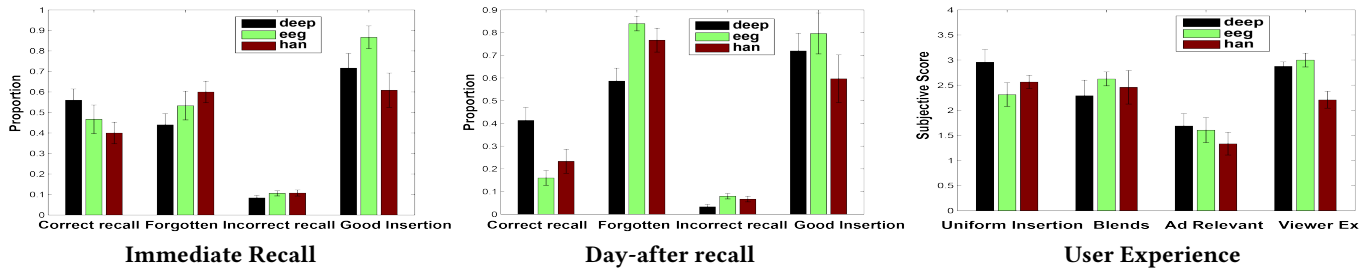
Figure 2: Summary of user study results in terms of recall and user experience-related measures. Error bars denote unit standard deviation.

high immediate and day-after recall, least ad forgottenness and least incorrect recall. Ads inserted via the EEG method are found to be the best inserted, even if they have relatively lower recall rates as compared to the Deep approach ($p < 0.05$ for independent t-test). Ads inserted via Han-generated affective scores have the least immediate recall and are also forgotten the most, and are also perceived as the the worst inserted. The trends observed for immediate and day-after recall are slightly different, but the various recall measures are clearly worse for the day-after condition with a very high proportion of ads being forgotten. Nevertheless, the observed results clearly suggest that the Deep and EEG approaches which achieve superior AR compared to the Han method also lead to better ad memorability.

However, it needs to be noted that higher ad recall does not directly translate to a better viewing experience. On the contrary, some ads may well be remembered because they disrupted the program flow and distracted viewers. In order to examine the impact of the affect-based ad insertion strategy on viewing experience, we computed the mean subjective scores acquired from users (Fig. 2(right)). Here again, the Deep method scores best in terms of uniform insertion and ad relevance, while the EEG method performs best with respect to blending and viewer experience ($p < 0.05$ with two-sample t-tests in all cases). Interestingly, the Han method again performs worst in terms of ad relevance and viewer experience. The CAVVA optimization framework [36] has two components– one for selection of ad insertion points into the program video, and another for selecting the set of ads to be inserted. Asl scores only play a role in the choice of insertion points, whereas val scores influence both components. In this context, the two best methods for val recognition, which also outperform the Han approach for asl recognition, maximize both ad recall and viewing experience.

7 DISCUSSION AND CONCLUSION

This work evaluates the efficacy of content-centric and user-centric techniques for ad affect recognition. At the outset, it needs to be stressed that content and user-centered AR methods encode complementary emotional information. Content-centric approaches typically look for emotional cues from low-level audio-visual (or textual) features, and do not include the human user as part of the computational loop; recent developments in the field of CNNs [22] have now made it possible to extract high-level emotion descriptors. Nevertheless, emotion is essentially a human feeling, and best manifests via user behavioral cues (*e.g.*, facial emotions, speech and physiological signals), which explains why a majority of contemporary

AR methods are user-centered [1, 20, 37]. With the development of affordable, wireless and wearable sensing technologies such as *Emotiv*, AR from large scale user data (termed *crowd modeling*) is increasingly becoming a reality.

We specifically evaluate the performance of two content-centered methods, the popular **Han** baseline for affect prediction from low-level audio-visual descriptors, and a **Deep** CNN-based framework which learns high-dimensional emotion descriptors from video frames or audio spectrograms, against the user-centered approach which employs **EEG** brain responses acquired from eleven users for AR. Experimental results show that while the deep CNN framework outperforms the Han method, it nevertheless performs inferior to an SVM-based classifier trained on EEG epochs for asl and val recognition. A study involving 12 users to examine if improved AR facilitates computational advertising reveals that (1) Ad memorability is maximized with better modeling of the ad affect via the Deep and EEG methods, and (2) Viewing experience is also enhanced by better matching of affective scores among the ads and video scenes. To our knowledge, this paper represents the first affective computing work to establish a direct relationship between objective AR performance and subjective viewer opinion.

Future work will focus on the development on effective alternative strategies to CAVVA for video-in-video advertising, as CAVVA is modeled on *ad-hoc* rules derived from consumer psychology literature. Also, we observe that EEG-encoded affective information is complementary to representations learned by the Han and Deep CNN approaches, as EEG signals are derived from human users and there is little correlation between the val scores computed via the content and user-centered methods (Sec. 6.2). This reveals the potential for fusion strategies where *content-centric* and *user-centric* cues can be fused in a cross-modal decision making framework, as successfully attempted in prior [16, 17, 31] problems.

REFERENCES

[1] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. 2015. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Trans. Affective Computing* 6, 3 (2015), 209–222.

[2] Tuka AlHanai and Mohammad Ghassemi. 2017. Predicting Latent Narrative Mood Using Audio and Physiologic Data. In *AAAI Conference on Artificial Intelligence*.

[3] Yoann Baveye. 2015. *Automatic prediction of emotions induced by movies.* Theses. Ecole Centrale de Lyon.

[4] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Trans. Affective Computing* 6, 1 (2015), 43–55.

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Series B (Methodological)* 57, 1 (1995), 289–300.

[6] Maneesh Bilalpur, Seyed Mostafa Kia, Tat-Seng Chua, and Ramanathan Subramanian. 2017. Discovering Gender Differences in Facial Emotion Recognition via Implicit Behavioral Cues. In *Affective Computing & Intelligent Interaction*.

[7] Christopher M. Bishop. 2013. *Pattern Recognition and Machine Learning*. Vol. 53. Springer.

[8] M. K. Greenwald, E. W. Cook, and P. J. Lang. 1989. Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology* 3 (1989), 51–64.

[9] Alan Hanjalic and Li-Quan Xu. 2005. Affective Video Content Representation. *IEEE Trans. Multimedia* 7, 1 (2005), 143–154.

[10] Morris B Holbrook, Rajeev Batra, and Rajeev Batra. 1987. Assessing the Role of Emotions as Mediators of Consumer Responses to Advertising. *Journal of Consumer Research* 14, 3 (1987), 404–420. DOI : https://doi.org/10.1002/job.305

[11] Morris B Holbrook and John O Shaughnessy. 1984. The role of emotlon in advertising. *Psychology & Marketing* 1, 2 (1984), 45–64.

[12] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech Emotion Recognition Using CNN. In *ACM Multimedia*. 801–804.

[13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. CAFFE: Convolutional Architecture for Fast Feature Embedding. In *ACM Int'l Conference on Multimedia*. 675–678.

[14] Hideo Joho, Jacopo Staiano, Nicu Sebe, and Joemon M Jose. 2011. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications* 51, 2 (2011), 505–523.

[15] Mohan Kankanhalli Karthik Yadati and, Harish Katti and. 2013. Interactive Video Advertising: A Multimodal Affective Approach. *Multimedia Modeling (MMM 13)* (2013).

[16] Harish Katti, Marius V. Peelen, and S. P. Arun. 2016. Object detection can be improved using human-derived contextual expectations. *CoRR* abs/1611.07218 (2016).

[17] Harish Katti, Anoop Kolar Rajagopal, Kalapathi Ramakrishnan, Mohan Kankanhalli, and Tat-Seng Chua. 2013. Online estimation of evolving human visual interest. *ACM Transactions on Multimedia* 11, 1 (2013).

[18] Harish Katti, Ramanathan Subramanian, Mohan Kankanhalli, Nicu Sebe, Tat-Seng Chua, and Kalpathi R Ramakrishnan. 2010. Making computers look the way we look: exploiting visual attention for image understanding. In *ACM Int'l conference on Multimedia*. 667–670.

[19] Aditya Khosla, Wilma A. Baingridge, Antonio Torralba, and Aude Oliva. 2013. Modifying the memorability of face photographs. *International confernece on computer vision (ICCV)* (2013).

[20] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Trans. Affective Computing* 3, 1 (2012), 18–31.

[21] Sander Koelstra and Ioannis Patras. 2013. Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing* 31, 2 (2013), 164–174.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems*. 1097–1105.

[23] Peter J. Lang, Margaret M. Bradley, and B. N. Cuthbert. 2008. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-8. The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

[24] Chul Min Lee and Shrikanth S Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* 13, 2 (2005), 293–303.

[25] Tao Mei, Xian-Sheng Hua, Linjun Yang, and Shipeng Li. 2007. VideoSense: Towards Effective Online Video Advertising. In *ACM Int'l Conference on Multimedia*. 1075–1084.

[26] Danny Oude Bos. 2006. EEG-based emotion recognition - The Influence of Visual and Auditory Stimuli. In *Capita Selecta (MSc course)*. University of Twente.

[27] Michel Tuan Pham, Maggie Geuens, and Patrick De Pelsmacker. 2013. The influence of ad-evoked feelings on brand evaluations: Empirical generalizations from consumer responses to more than 1000 {TV} commercials. *International Journal of Research in Marketing* 30, 4 (2013), 383 – 394.

[28] Hamed R.-Tavakoli, Adham Atyabi, Antti Rantanen, Seppo J. Laukka, Samia Nefti-Meziani, and Janne Heikkila. 2015. Predicting the Valence of a Scene from Observers' Eye Movements. *PLoS ONE* 10, 9 (2015), 1–19.

[29] James Russell. 1980. A circumplex model of affect. (1980).

[30] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Ramanathan Subramanian. 2017. Affect Recognition in Ads with Application to Computational Advertising. In *ACM Int'l conference on Multimedia*.

[31] Ramanathan Subramanian, Harish Katti, Kalapathi Ramakrishnan, Mohan Kankanhalli, Tat-Seng Chua, and Nicu Sebe. 2010. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision*.

[32] Ramanathan Subramanian, Divya Shankar, Nicu Sebe, and David Melcher. 2014. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of vision* 14, 3 (2014), 1–18.

[33] Ramanathan Subramanian, Julia Wache, Mojtaba Abadi, Radu Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* (2016).

[34] Vassilios Vonikakis, Ramanathan Subramanian, Jonas Arnfred, and Stefan Winkler. 2017. A Probabilistic Approach to People-Centric Photo Selection and Sequencing. *IEEE Transactions on Multimedia* (2017).

[35] Hee Lin Wang and Loong-Fah Cheong. 2006. Affective understanding in film. *IEEE Trans. Circ. Syst. V. Tech.* 16, 6 (2006), 689–704.

[36] Karthik Yadati, Harish Katti, and Mohan Kankanhalli. 2014. CAVVA: Computational affective video-in-video advertising. *IEEE Trans. Multimedia* 16, 1 (2014), 15–23.

[37] Wei-Long Zheng, Jia-Yi Zhu, Yong Peng, and Bao-Liang Lu. 2014. EEG-based emotion classification using deep belief networks. *IEEE International Conference on Multimedia & Expo* (2014).