Multimodal Emotion Recognition from Advertisements with Application to Computational Advertising

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Abhinav Shukla
201302135
abhinav.shukla@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2018

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled " **Multimodal Emotion Recognition from Advertisements with Application to Computational Advertising** " by **Abhinav Shukla**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Ramanathan Subramanian

To my family for always being there for me

Acknowledgments

This research was partially supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative (SeSaMe Centre at the National University of Singapore), Google India's Research Travel Grant Program (to present the paper at ACM Multimedia 2017 in Mountain View, CA, USA) and an ACM ICMI 2017 conference grant (to present at ICMI 2017 in Glasgow, UK).

Abstract

Advertisements (ads) are often filled with strong affective content covering a gamut of emotions intended to capture viewer attention and attempt to convey an effective message. However, most approaches to computationally analyze the emotion present in ads are based on the text modality and only a limited amount of work has been done on affective understanding of advertisements videos from the content and user-centric perspectives.

This work attempts to bring together recent advances in deep learning (especially in the domain of visual recognition) and affective computing, and use them to perform affect estimation of advertisements. We first create a dataset of 100 ads which are annotated by 5 experts and are evenly distributed over the valence-arousal plane. We then perform content-based affect recognition via a transfer learning based approach to estimate the affective content in this dataset using prior affective knowledge gained from a large annotated movie dataset. We employ both visual features from video frames and audio features from spectrograms to train our deep neural networks. This approach vastly outperforms the existing benchmark.

It is also very interesting to see how human physiological signals, such as that captured by Electroencephalography (EEG) data are able to provide useful affective insights into the content from a user-centric perspective. Using this time series data of the electrical activity of the brain, we train models which are able to classify the emotional dimensions of this data. This also enables us to benchmark this user-centric performance and compare it to the content-centric deep learning based models, and we find that the user-centric models outperform the content-centric models, and set the state-of-the-art in ad affect recognition.

We also combine the two kinds of modalities (audiovisual and EEG) using decision fusion and find that the fusion performance is greater than either single modality, which shows that human physiological signals and the audiovisual content contain complementary affective information. We also use multi task learning (MTL) on top of the features of each kind to exploit the intrinsic relatedness of the data and boost the performance.

Lastly, we validate the hypothesis of better affect estimation being able to enhance a real world application by supplying the affective values computed by our methods to a computational advertising framework to get a video program sequence with ads inserted at emotionally relevant points, determined to be appropriate based on the affective relevance between the the program content and the ads. Multiple user studies find that our methods significantly outperform the existing algorithms and are very close

(and sometimes better than) human level performance. We are able to achieve much more emotionally relevant and non disruptive advertisement insertion into a program video stream.

In summary, this work (1) compiles an affective ad dataset capable of evoking coherent emotions across users; (2) explores the efficacy of content-centric convolutional neural network (CNN) features for affect recognition (AR), and find that CNN features outperform low level audio-visual descriptors; (3) study user-centric ad affect recognition from Electroencephalogram (EEG) responses (with conventional classifiers as well as a novel CNN architecture for EEG) acquired while viewing the content that outperform content descriptors; (4) Examine a multi-task learning framework based on CNN and EEG features which provides state of the art AR from ads; (5) Demonstrates how better affect predictions facilitates more effective computational advertising in a real world application.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Advertising is a huge and profitable industry and advertisers intend to portray their products or services as not only useful, but also highly desirable and rewarding. 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]. Emotions are critical for conveying an effective message to viewers, and have been found to mediate consumer attitudes towards brands [15, 14, 33]. 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. Similar objectives are at play in messages for public health and safety, where certain life choices are portrayed as beneficial and improving one's quality of life, while others are portrayed as harmful and potentially fatal. The ability to objectively quantify advertisements (ads) in terms of emotional content therefore has a wide variety of applications– *e.g.*, inserting the right type of ads at optimal temporal points within a video stream can beneficially impact both advertisers and consumers in video streaming websites such as YouTube [53, 52].

Subjective experience of pleasantness (valence) and emotional intensity (arousal) are important affective dimensions [37], and both modulate emotional responses to ads in distinct ways [8]. Affective content has also been shown to modulate recall of key concepts and episodes in movies [47] and video ads [53]. Strongly emotional ads can result in improved brand recall, which can directly and positively impact product sales.

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 [24, 6], speech [30], audio [2], music [26] and movie [1, 48] content. However, affect characterization in ads is a non-trivial problem as with other stimuli such as music and movie clips examined by

[1] http://www.cnbc.com/2016/12/05/global-ad-spend-to-slow-in-2017-while-2016-sales-were-nearly-500bn.html
[2] http://www.pwc.com/gx/en/industries/entertainment-media/outlook/segment-insights/internet-advertising.html

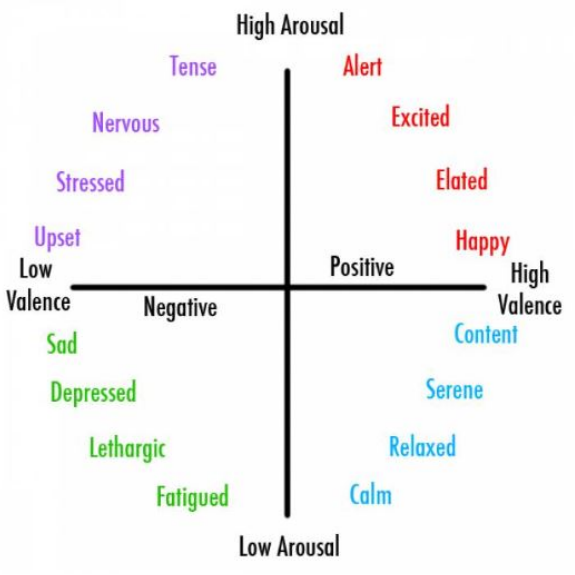


Figure 1.1 The affective dimensions of valence and arousal

prior works [13, 51, 26, 1]. Given that human emotional perception is subjective and the detection of specific emotions such as *joy*, *sorrow* and *disgust* is relatively hard, popular affect recognition (AR) works represent emotions along the valence and arousal dimensions [37, 12]. 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 [13, 51]. In contrast, *user-centric* AR methods aim to recognize the elicited emotions by monitoring the user or multimedia consumer via facial [19] or physiological [26, 1, 48, 47] measurements. While enabling a fine-grained examination of emotional perception, which is a transient phenomenon, user-centered methods nevertheless suffer from subjectivity limitations.

This work expressly investigates the modeling of emotions conveyed by ads, and employs subjective human opinions and objective multimedia features to this end. Firstly, upon carefully compiling a diverse set of 100 ads, we examine the efficacy of this ad dataset to coherently evoke emotions across viewers. To this end, we compare the affective opinions of five experts and 23 novice annotators[3] , and find that the two groups are highly concordant. Secondly, we explore the utility of Convolutional Neural Networks (CNNs) for encoding audio-visual emotional features. As the compiled ad dataset is relatively small and insufficient for CNN training, we employ *domain adaptation* to transfer affective knowledge gained from the LIRIS-ACCEDE movie dataset [4] for modeling ad emotions. Extensive experimentation confirms that the synthesized CNN descriptors outperform popular audio-visual features proposed in [13] especially for valence recognition. Thirdly, we examine user-centric ad AR using EEG responses and find that they outperform audiovisual CNN features. This work expressly examines and compares the utility of *content-centric* and *user-centric* approaches for ad AR. As emotion is a subjective human

[3]unfamiliar with emotional attributes, and representing the general population.

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. Lastly, we show how accurate encoding of the ad emotions can facilitate optimized insertion of ads into streaming video, used for income generation by online websites such as YouTube. A user study with 18 viewers confirms that the insertion of emotionally relevant ads within the streamed video can maximize viewer experience.

In summary, we make the following research contributions:

1. Ours is one of the few works to examine AR in ads, and the only work to characterize ad emotions in terms of subjective human opinions and objective audio-visual features.

2. We 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*.

3. We explore the utility of CNNs for encoding ad emotions. We show the effectiveness of a new CNN, ***AdAffectNet*** (AAN) generated by fine-tuning the *Places205* CNN architecture [55] for AR. For fine-tuning and domain adaptation, we have employed the extensively annotated LIRIS-ACCEDE movie dataset [4]. Extensive experiments reveal that the AAN features outperform emotional audio-visual descriptors proposed in [13], and the best content-centric AR performance is achieved with multi-task learning which exploits audio-visual similarities among emotionally homogeneous ads.

4. We explicitly compare and contrast content-centered ad AR thrugh audio-visual CNN features and user-centered ad AR through EEG features and find that human centered EEG features are more effective for the prediction of both valence and arousal.

5. We 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.

The thesis is organized as follows. Chapter II reviews related literature and overviews the compiled ad dataset. Chapter III presents the techniques adopted for content-centered ad AR, while Chapter IV discusses user-centered ad AR. Chapter V discusses the fusion of the content and user-centered modalities and shows how they carry complementary information. Each chapter about methods also describes a user study to establish how improved AR facilitates computational advertising. Chapter VI summarizes the main findings and concludes the thesis.

Chapter 2

Related Work and the Dataset

2.1 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.

2.1.1 Affect Recognition

Many approaches have been devised to infer the emotions evoked by multimedia stimuli in a *content-centric* or *user-centric* manner. Content-centric approaches [13, 51] predict the elicited emotion by examining audio-visual cues in the analyzed stimuli. In contrast, user-centric AR methods [26, 1, 48] predict the stimulus-evoked emotion by measuring physiological changes in users (or content consumers). Nevertheless, both content and user centric methods require labels identifying stimulus emotion, and these labels are compiled from reliable annotators whose affective opinions are generally *acceptable*, given human subjectivity in emotion perception.

Building on the circumplex emotion model that represents emotions in terms of valence and arousal [37], 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 [13, 49, 38] to recognize affect in a supervised manner; or *user-centric*, which measure stimulus-driven variations in specific physiological signals such as pupillary dilation [21], gazing patterns [47, 34] and neural activity [26, 1, 54]. 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.

2.1.2 Emotional impact of ads

Ad-induced emotions have been shown to shape consumer behavior in a significant manner [15, 14]. Although this key observation was made nearly three decades ago [14], computational advertising methods till recently have matched low-level visual and semantic properties between video segments and candidate ads [31]. Recent work [33] indicates a shift form the traditional thinking by emphasizing that ad-evoked emotions can change brand perception among consumers. While a body of works have examined the correlation between ad emotions and user behavior, very few works have exploited these findings for developing targeted advertising mechanisms. The only work that incorporates emotional information for modeling context in advertising is CAVVA [53], where arousal and valence evoked by video scenes are estimated via [13] to identify optimal ads and corresponding insertion points which maximize user engagement. Twi very recent and closely related works to ours [38, 39] 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 [53]. A user study shows that better ad AR translates to better ad memorability and enhanced user experience while watching an ad-embedded video stream.

2.1.3 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 [31]. 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 is a framework for **C**omputational **A**ffective **V**ideo in **V**ideo **A**dvertising [53], which is used as the computational advertising framework to evaluate the effectiveness of all affect recognition methods discussed in the subsequent chapters. 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 [21].

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 an EEG-based strategy achieves the best user experience and is best for ad memorability. The following section describes the compiled ad dataset, and the EEG acquisition protocol.

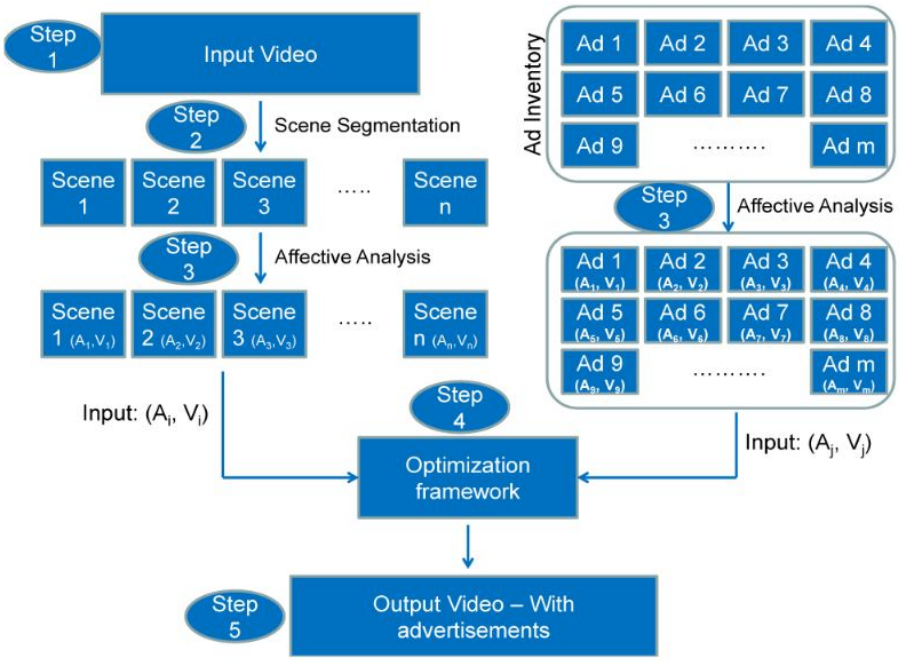


Figure 2.1 A schematic diagram of the CAVVA ad insertion framework

2.1.4 Analysis of related work

Examination of the literature reveals that (1) AR studies are hampered by the subjectivity in emotion perception, and a control dataset that can coherently evoke emotions across users is necessary for effectively learning content or physiology-based emotion predictors; (2) Despite the well-known impact of ad emotions on user behavior, there has hardly been any attempt to incorporate emotion-related findings in a computational advertising framework.

In this regard, we present the first work to compile a control set of affective ads, which elicit concordant affective opinions from experts and naive users. Also, we synthesize CNN-based emotion descriptors which are found to outperform audio visual features proposed in [13]. We also perform a user study and show how better affect encoding can facilitate the ad-insertion framework [53] to improve viewing experience. Details pertaining to our ad dataset are presented below.

2.2 Advertisement Dataset

This section presents details regarding the ad dataset used in this study.

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, roughly 1-minute long commercial advertisements (ads) which are used in this work. These ads are

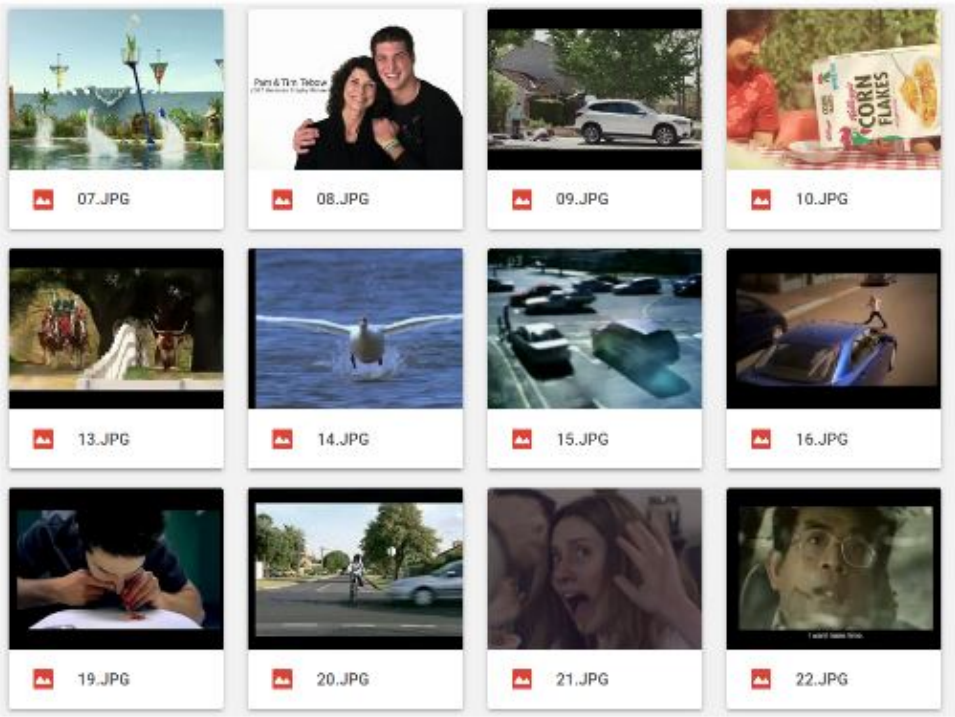


Figure 2.2 Representative thumbnails images from the ads in our dataset

publicly available[1] and found to be uniformly distributed over the arousal–valence plane defined by Greenwald *et al.* [12] (Figure 2.3). 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)[2]. 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 2.1 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

[1] On video hosting websites such as YouTube.

[2] Annotators were familiarized with emotional attributes prior to the rating task.

Table 2.1 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

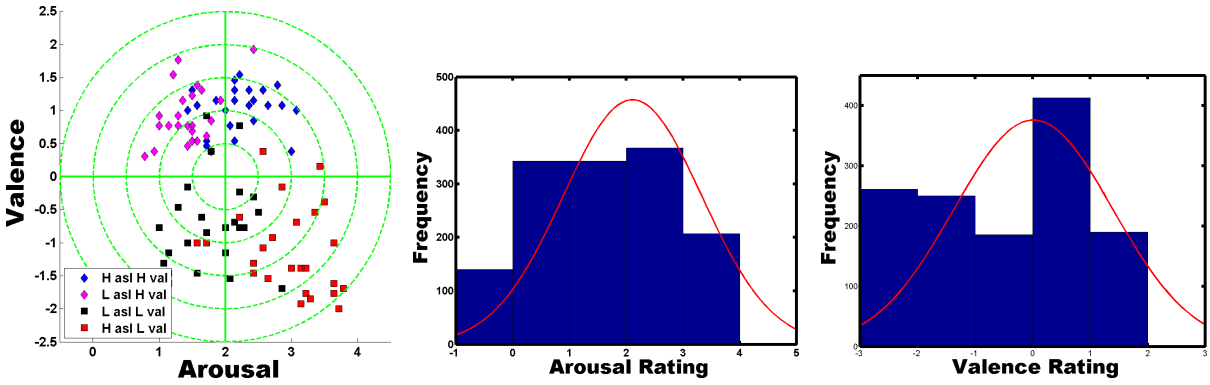


Figure 2.3 (left) Scatter plot of mean asl, val ratings color-coded with expert labels. (middle) Asl and (right) Val rating distribution with Gaussian pdf overlay (view under zoom).

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[3]. Clearly, there is good-to-excellent 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. 2.3. The scatter plot is color-coded based on expert labels, and is interestingly different from the classical 'C' shape observed with images [29], music videos [26] 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

[3]Chance agreement corresponds to a κ value of 0.

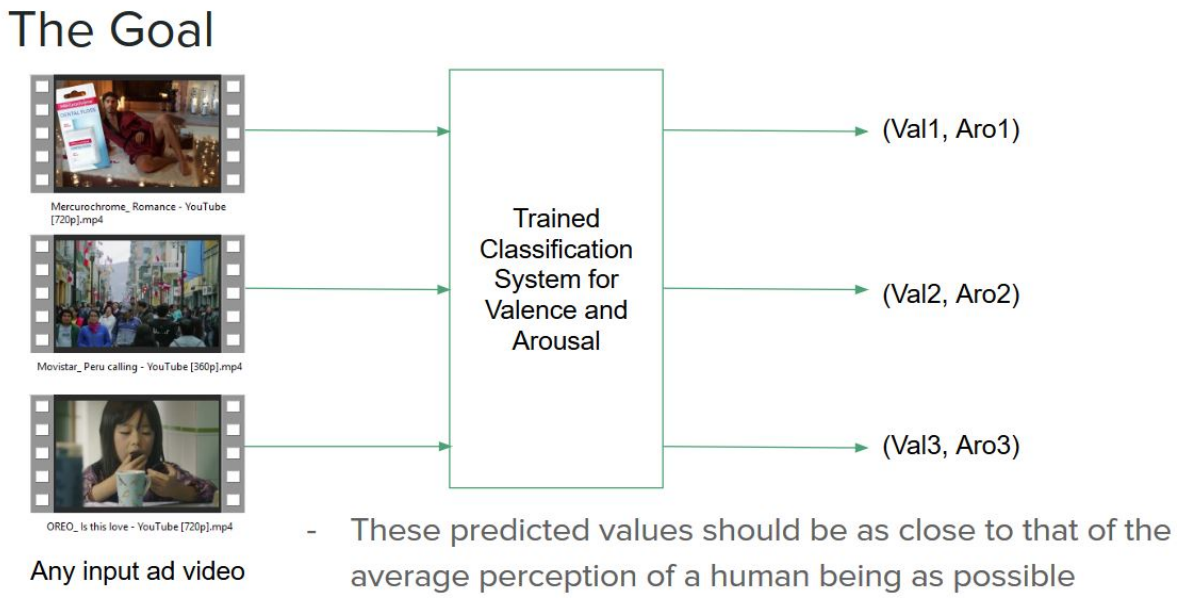


Figure 2.4 An example of the target of affect recognition from ads in our dataset

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. Figure 2.4 shows what the target of affect recognition methods discussed in subsequent chapters will be.

Chapter 3

Content-Centric Ad Affect Recognition with Convolutional Neural Networks and Multi Task Learning

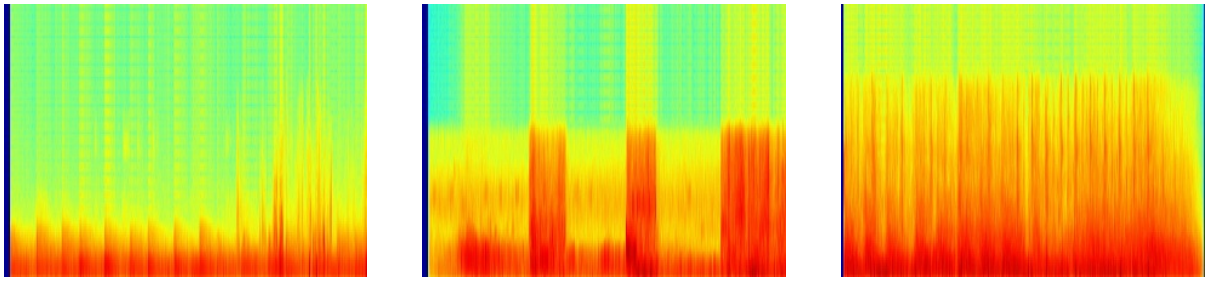
3.1 Method

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 [13]. Due to subjective variance in emotion perception, careful affective labeling is imperative for effectively learning content-centric [13, 51] or user-centric [26, 1] affective correlates, which is why we analyze ads that evoked perfect consensus among experts. CNNs have recently become very popular for visual [28] and audio [16] recognition, but they require vast amounts of training data. As our ad dataset comprised only 100 ads, we fine-tuned the pre-trained *places205* [28] 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.

To learn deep features for modeling ad affect, we employed the *Places205* CNN [55] intended for image classification. *Places205* is trained using the Places-205 dataset comprising 2.5 million images and 205 scene categories. The Places-205 dataset contains a wide variety of scenes with varying illumination, viewpoint and field of view, and we hypothesized a strong relationship between scene perspective, lighting and the scene mood. **LIRIS-ACCEDE** contains asl, val ratings for $\approx$ 10 s long movie snippets, whereas our ads are about a minute-long (ranging from 30–120 s).

3.1.1 FC7 Feature Extraction via CNNs

For 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 to synthesize ***AdAffectNet*** (AAN), and use the fully connected layer (fc7) AAN descriptors for our analysis.



L asl, H val **H asl, L val** **H asl, H val**

Figure 3.1 Exemplar spectrograms for varied emotional ads. x denotes time (0-10 s), while y denotes spectral magnitude observed at each time instant. High and low frequency densities are respectively shown in red and green shades.

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 10 s long LIRIS-ACCEDE clips, and consistently from 10 s 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. Fig. 3.1 shows three exemplar spectrograms indicative of emotional ad content. Note greater densities of high frequencies in high asl ads, and such intense scenes are often characterized by sharp frequency changes.

CNN Training We use the Caffe [17] deep learning framework for fine-tuning *places205*, with a momentum of 0.9, weight decay 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.

3.1.2 AR with audio-visual features

We will mainly compare our CNN-based AR framework against the algorithm of Hanjalic and Xu [13] in this work. Even after a decade, this algorithm remains one of the most popular AR baselines as noted from recent works such as [26, 1]. In [13], asl and val are modeled via low-level descrip-

Table 3.1 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 [13] Features	Per-second sound energy and pitch statistics [13].	Per-second shot change frequency and motion statistics [13].	Concatenation of audio-visual features.

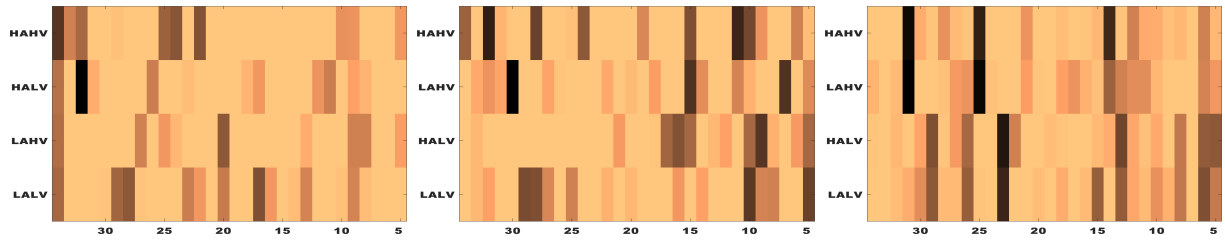
tors describing motion activity, colorfulness, shot change frequency, voice pitch and sound energy in the scene. These hand-crafted features are intuitive and interpretable, and employed to estimate time-continuous asl and val levels conveyed by the scene. Table 3.1 summarizes the audio-visual features used for content-centric AR.

3.2 Experiments and Results

We first provide a brief description of the classifiers used and settings employed for our binary AR experiments, 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 features. Experimental results will be discussed thereafter.

Classifiers: We employed the Linear Discriminant Analysis (LDA), linear SVM (LSVM) and Radial Basis SVM (RSVM) classifiers for AR. LDA and LSVM attempt to 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.

In addition to the above *single-task learning* methods which do not exploit the underlying structure of the input data, we also explored the use of *multi-task learning* (MTL) for AR. When posed with the learning of multiple *related* tasks, MTL seeks to jointly learn a set of task-specific classifiers on modeling task relationships, which is highly beneficial when learning with few examples. Among the MTL methods available as part of the MALSAR package [56], we employed the sparse graph-regularized MTL (SR-MTL) where *a-priori* knowledge regarding task-relatedness is modeled in the form of a graph R. Given tasks $t = 1...T$, with X_t denoting training data for task t and Y_t their labels, SR-MTL jointly learns a weight matrix $W = [W_1..W_T]$ such that the objective function $\sum_{t=1}^{T} \|W_t^T X_t - Y_t\|_F^2 + \alpha\|WR\|_F^2 + \beta\|W\|_1 + \gamma\|W\|_F^2$ is minimized. Here, α, β, γ are regularization parameters, while $\|.\|_F$ and $\|.\|_1$ denote the matrix Frobenius (ℓ_2) and ℓ_1-norms respectively.



Shot Frequency **Pitch Amplitude** **Motion Activity**

Figure 3.2 Learned MTL weights for the four quadrants (tasks) when fed with the specified low-level features computed over the final 30 s of the 100 ads.

MTL is particularly suited for dimensional AR, and one can expect similarities in terms of audio-visual content among high val or high asl ads. We exploit underlying similarities by modeling each asl-val quadrant as a task (*i.e.*, all H asl, H val ads will have identical task labels). Also, quadrants with same asl/val labels are deemed as related tasks, while those with dissimilar labels are considered unrelated. The graph R guides the learning of W_t's, as shown in the three examples in Fig.3.2, where SR-MTL is fed with the specified features computed over the final 30 s of all ads. Darker shades denote salient MTL weights. Shot change frequency is found to be a key predictor of asl in [13], and one can notice salient weights for H asl H val ads in particular. The attributable reason is that our H asl H val ads involve frequent shot changes to maintain emotional intensity, while the mood of our H asl L val ads[1] is strongly influenced by semantics. Likewise, pitch amplitude is deemed as a key val predictor, and salient weights can be consistently seen over the 30 s temporal window for HV ads. Finally, more salient weights for H val ads with the motion activity feature implies that our positive val ads involve accentuated motion.

Metrics and Experimental Settings: We used the F1-score (F1), defined as the harmonic mean of precision and recall as our performance metric, given our unbalanced dataset (Table 3.1). 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 3.1), while the W_{est} technique [27] is employed for decision fusion (DF). In DF, the test label is computed as $\sum_{i=1}^{2} \alpha_i^* t_i p_i$, where i indexes 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 i^{th} modality, then $t_i = \alpha_i F_i / \sum_{i=1}^{2} \alpha_i F_i$ for given α_i.

As the Hanjalic (Han) algorithm [13] uses audio plus visual features to model asl and val, we only consider (feature and decision) fusion performance in this case. As we evaluate AR performance on a small dataset, we present AR results over 10 repetitions of 5-fold cross validation (CV). CV is typically used to overcome the *overfitting* problem on small datasets, and the optimal classifier parameters (including regularization parameters for MTL) are determined from the range $[10^{-3}, 10^{3}]$ via an inner

[1] which depict topics like drug and alcohol abuse, and overspeeding.

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 three frames (L3) and (c) last frame (L)[2].

3.2.1 Results Overview

Table 3.2 presents the asl, val F1-scores under the various settings. The highest F1 over all the considered temporal windows, achieved with the single-task and multi-task classifiers, and via feature/decision fusion are denoted in bold. Based on the observed results, we make the following claims.

Focusing on ***unimodal fc7 features and single-task classifiers***, 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. Much higher asl, val recognition scores are achievable with the ***MTL classifier*** (F1 of 0.96 for val and 0.94 for asl) due to its ability to exploit the underlying similarities among audio and visual features among similarly labeled ads. MTL F1-scores are consistently higher with V features for both asl and val.

Concerning recognition with ***single-task classifiers*** and ***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[3]. 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). Focusing on the ***MTL classifer***, MTL F1-scores in the **A+V FC7 + MTL** condition are considerably higher than single-task F1-scores[4] analogous to the unimodal case, even though the best F1-scores are still less than those achieved with video fc7 features + MTL.

Comparing A+V ***fc7 vs Han*** features, fc7 descriptors clearly outperform Han features with both single and multi-task approaches. The difference in performance is prominent for val, while comparable recognition is achieved with both features for asl. RSVM produces the best F1-scores for both asl and val among ***single-task classifiers*** with unimodal and multimodal approaches. However, the linear ***MTL*** model considerably outperforms all single-task methods with both fc7 and Han features. These observations suggest that while the H and L asl/val features for *all ads* are difficult to linearly classify *per se*, exploiting underlying similarities among *quadrant-specific ads* enables better linear separability.

Relatively small σ values are observed for the 'All' condition with the adopted five-fold CV procedure in Table 3.2, especially with fc7 features suggesting that the AR results are ***minimally impacted*** by ***overfitting***. Examining ***temporal windows*** considered for AR, higher σ's are observed for the L3 and L cases, which denote model performance on the terminal ad frames. Surprisingly, one can note a general

[2]This equates to estimating the ad asl/val typically over the terminal 30/10 s, when one would expect the conveyed emotion to be strongest.

[3]The relative efficacy of feature or decision fusion depends on the specific problem and features on hand.

[4]We are not aware of MTL-based decision fusion methods.

degradation in asl recognition for the L3 and L conditions with A/V features, while val F1-scores are more consistent. Also, a sharp degradation in performance is noted with MTL for the L3 and L conditions. 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 and MTL-based approaches; (2) Fusion models synthesized with Han features appear to be more prone to overfitting, given the generally larger σ values seen with the corresponding models; (3) That asl recognition is lower in the L3 and L 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 val features and labels is sustainable over time.

Table 3.2 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
Audio FC7 + MTL	0.85±0.02	0.83±0.10	0.78±0.20	0.78±0.03	0.62±0.14	0.45±0.16
Video FC7 + MTL	**0.96±0.01**	0.94±0.07	0.82±0.25	**0.94±0.01**	0.87±0.12	0.63±0.29
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
A+V FC7 + MTL	**0.89±0.03**	0.88±0.11	0.77±0.26	**0.87±0.03**	0.68±0.17	0.46±0.20
A+V Han + MTL	0.77±0.04	0.79±0.07	0.74±0.15	0.78±0.04	0.73±0.11	0.58±0.22

3.2.2 Discussion

As ads are inherently emotional and have great influencing/monetizing capacity [14, 33], the ability to infer ad emotions and make optimal ad insertions within video streams would be highly advantageous for multimedia systems. Therefore, it is surprising that very few works [31, 53] have attempted to mine visual and emotional content in ads. In this regard, our work expressly sets out to model the emotion conveyed by 100 ads based on subjective human opinions and objective audio-visual features.

We carefully curated a small but diverse set of ads based on consensus among experts, and examined if those ads could coherently evoke emotions across viewers by acquiring asl and val ratings from 14 annotators. A good-to-excellent agreement on asl and val impressions is noted between the expert and annotator groups. Also, annotator ad ratings are found to be uniformly distributed over the asl-val plane, with only a weak negative correlation noted between asl and val ratings.

As the compiled ads are found to constitute a control affective dataset, we modeled the emotion conveyed by these ads in terms of audio-visual features. Specifically, we extracted fc7 layer outputs from the *AdAffectNet* CNNs fed with key frames and spectrograms for video and audio-based affect modeling. While CNNs have been previously used for video and audio-based AR [4, 11], the modeled scenes are only a few second long snippets. In contrast, we have explored the validity of CNN-based AR for full-length ads, some of which are over a minute long, in this work.

Obtained AR results confirm that the synthesized fc7 features are effective predictors of asl and val. They outperform the audio-visual features proposed by Hanjalic and Xu [13] with both single and multi-task classifiers. In particular, while fc7 features are considerably better for val, Han features provide competitive performance for asl. Optimal AR is achieved with the MTL classifier, which is able to effectively exploit the underlying similarities among emotionally homogeneous ads in terms of audio visual content. Nevertheless, a significant drop in recognition performance is generally noted for the terminal ad portion with most methods, and especially for asl, implying that (a) asl is a more transient phenomenon as compared to val, and there is less coherence between the employed AV features and asl labels towards the ad endings, and (b) the use of a single asl/val over the entire ad duration may be inappropriate, and one may need to acquire time-varying labels for affective studies. The next section will evaluate whether the superior AR achieved with AAN fc7 features translates to optimized ad insertion in a computational advertising task via a user study.

3.3 Computational Advertising- User Study

Given the superior AR achieved by our *AdAffectNet* CNN features, we hypothesized that this should in turn enable optimized selection and insertion of affective ads within streamed video content, as discussed in the CAVVA ad insertion framework [53]. Video-in-video advertising is complex, as it aims to strike a balance between (a) maximizing ad impact, and (b) minimally disrupting (or ideally, enhancing) viewing experience while watching a program video into which the ads are embedded. Also, while ad insertion strategies have modeled ad-video relevance in terms of low-level visual context [31]

Table 3.3 Summary of program video statistics.

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

and high-level emotional context [53], their performance has not been compared against human context assessment.

To this end, we performed a user study to evaluate whether the ad insertion framework formulated in [53], which employs the affect estimation methodology of Hanjalic and Xu [13], would benefit from better affect prediction via our deep AAN features. Better estimation of the affect induced by the video content and candidate ads can enable optimized selection of ads and corresponding insertion points. Specifically, we compared video program sequences generated via the CAVVA framework by estimating arousal (asl) and valence (val) scores for the ads and video scenes via (a) the baseline method of Hanjalic and Xu [13], (b) our deep AAN model and (c) human annotators.

3.3.1 Dataset

For the user study, we chose 28 ads (from the 100 used in this work) and three program videos. The program videos were scenes from a television sitcom *Friends* (***friends***) and two movies *The Pursuit of Happyness* (***ipoh***) and *Children of Heaven* (***coh***), with predominantly social themes and situations invoking high-to-low val and asl. Summary statistics of the three program videos are presented in Table 4.2. Each program video was segmented into 8 scenes, and the average scene length was 118 seconds. We obtained val and asl scores for the video scenes and 28 ads using (a) normalized softmax class probabilities [7] output by our AAN model, with video and audio fc7 features respectively used for val and asl estimation (b) the baseline model (Han) [13] and (c) ratings from three experts (Manual). We then inserted ads into each program video based on method-specific affect scores with the optimization strategy described in [53], and obtained 9 unique ***video program*** sequences (mean length 19.4 min) comprising the inserted ads. Exactly 5 ads were inserted in each program video, and 21 of the 28 chosen ads were cumulatively inserted at least once onto the 9 video programs (upon being selected via any of the three methods), with an average insertion frequency of 2.14.

3.3.2 Experiment and Questionnaire Design

We recruited a total of 17 users (5 female, mean age 20.5 years) to evaluate the video program sequences. Each user saw one exemplar sequence corresponding to the three affect prediction strategies. We followed a randomized 3×3 Latin square design so that all nine video programs were covered with three users.

Our user evaluation was in two parts; In line with the twin goals underlying seamless ad insertion within streaming video, we evaluated whether the ad insertion strategy resulted in (a) increased brand recall, and (b) minimal disturbance and improved viewing experience. We performed recall evaluation by measuring the impact of the ad insertion strategy on *immediate* and *day-after* recall. These *objective* measures quantified the impact of ad insertion on short-term (immediate) and long-term (day-after) memory of viewers, on viewing the video programs. 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, perhaps owing to their inherent salience (*Incorrect* recall). For those ads that were inserted into program sequences and were correctly recalled, we also assessed whether viewers perceived them to be contextually appropriate with respect to program content.

The viewer was provided with a key-frame visual from each of the 28 ads, as well as a response sheet for every video program sequence. In addition to the recall related questions, we asked viewers to indicate whether they perceived the correctly recalled ads as being inserted at an appropriate position in the video stream (*Good insertion*)[5]. All recall and insertion quality-related responses were acquired from users as binary values. We pooled responses from viewers after they had watched video sequences generated via deep AAN, Han and manual affective scores for analyses.

While increased ad recall reflects a key desired effect of a successful ad insertion strategy, ads that are out of sync with the video program flow may disrupt viewer experience. In some cases, this disruptiveness may indirectly contribute to the recall, but would adversely impact viewing experience. So, mere recall alone does not indicate optimal ad insertion, relevance of the ad to the program or an enhanced viewer experience. To address these issues, we defined a second set of *subjective* experience evaluation 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. Whether advertisements were uniformly distributed across the video program?

2. Whether the inserted ads blended well with the flow of the video program?

3. Whether the inserted ads had a content and mood similar to surrounding program?

4. What was the overall viewer experience while each video program?

Each participant filled the recall and experience-related questionnaires (provided in supplementary material) after watching each video program. They also filled in the day-after recall questionnaire, a day after completing the experiment.

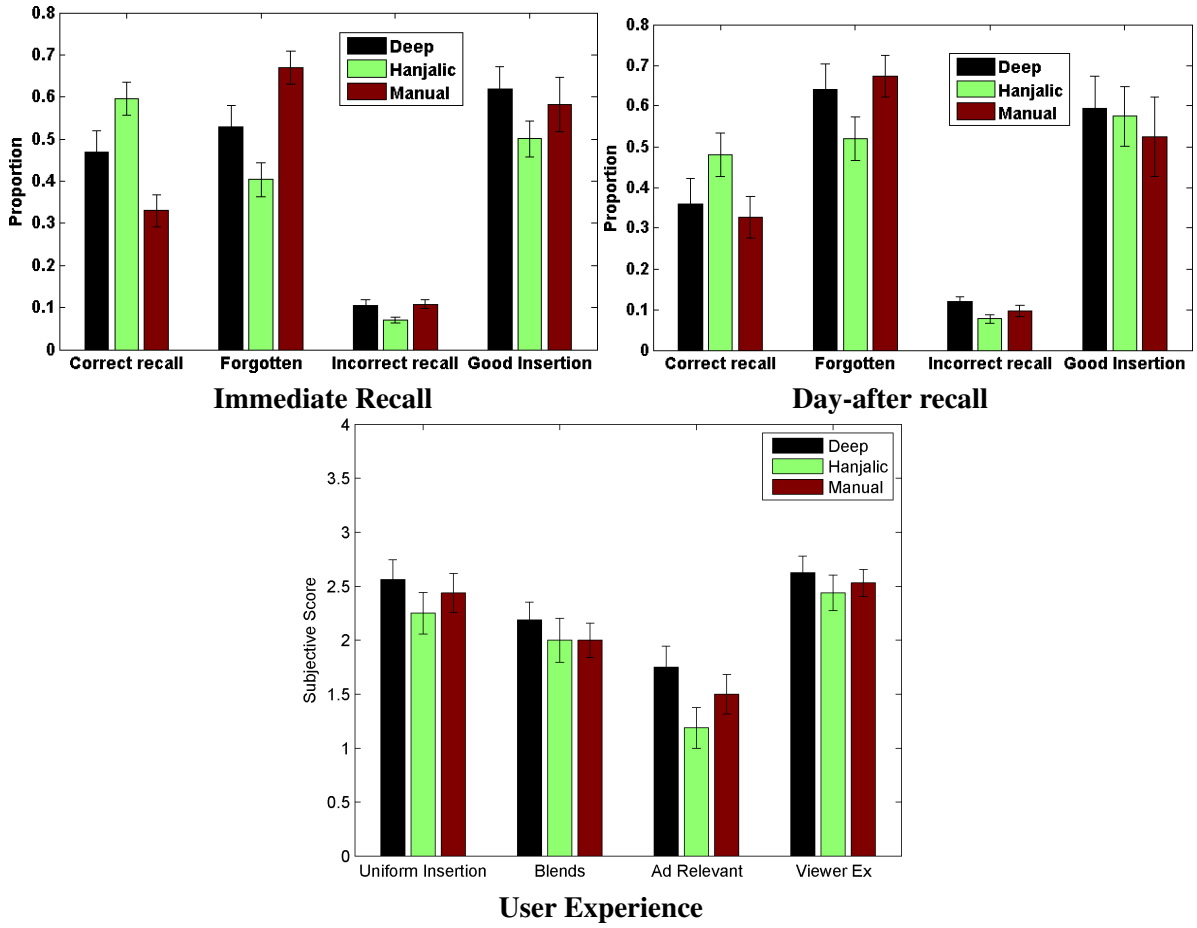


Figure 3.3 Summary of user study results in terms of recall and user experience-related measures. Error bars denote unit standard deviation.

3.3.3 Results and Discussion

We evaluate the effectiveness of affective scores obtained from (a) our AAN-fc7 features, (b) Han features [13] employed in CAVVA [53], and (c) human assessments, for modeling contextual relevance based on user recall and experience responses.

Fig. 3.3 depicts user recall and experience-related measures as obtained with the three affect measurement approaches. Focusing on recall, although ad insertion via Han method [13] results in higher immediate and day-after ad recall, lower incorrect recall and forgottenness ($p < 0.05$ in all cases), video programs generated with AAN-based affective scores are found to maximize user experience ($p < 0.05$ for insertion uniformity and ad relevance, and $p < 0.1$ for non-disruptive ad insertion). The significance of these effects was assessed by comparing the proportion mean distributions for each question and method, via independent t-tests. Ads placed via the AAN model and correctly recalled by users were

[5]This was one way of inferring if the ad placements facilitated their recall.

'well inserted' (difference with respect to manual scores significant at $p < 0.05$) based on responses compiled immediately upon viewing.

Notably, ad insertions via AAN-based affective scores were opined to be (i) 'uniformly distributed' across the streamed video, and (ii) 'most relevant' in terms of emotional context with the video (Fig.4.2). These observations imply that our AAN model is more accurate than Han in capturing the mood of the ads and video scenes. In contrast, while the Han method achieves the best recall from viewers, it also scores the least with respect to insertion-point distribution and relevance, implying an adverse impact on viewing experience.

To examine ***how affective attributes influenced ad recall***, we correlated the ad recall measures with their (manually assigned) mean val, asl ratings. A meaningful relationship was noted between *estimated ad valence* and the *forgottenness rate* (Pearson $\rho = 0.45, p < 0.05$), indicating that positive val ads tend to be forgotten more easily. This observation agrees with the prior findings of Rimmele *et al.* [36], who discovered that recall performance was maximum for negative valence images in a memory study.

Surprisingly, ad insertions based on manual affective scores resulted in lowest recall and highest forgottenness among the three methods, while performing second best with respect to experience measures. This can be partly attributed to the higher val ratings observed for the selected ads based on manual scores ($\mu_{val} = 0.6$ for 12 unique inserted ads) as compared to ads selected based on AAN-based val estimates ($\mu_{val} = 0.3$ over 11 unique inserted ads). Ads selected based on Han val estimates had the lowest mean val ($\mu_{val} = 0.23$ over 12 unique inserted ads), suggesting that more low val ads were selected via the Han approach, resulting in least forgottenness. Nevertheless, viewers forgot nearly half the ads immediately and most ads a day later with all the considered methods. This reveals the scope for improving the ad-placement strategy by placing specific emphasis on ad retention.

To examine the ***relationship between affective scores*** estimated by the three methods ***and the inserted ads***, we first examined if there was any relationship between manual ratings and computationally estimated scores. We found a significant correlation between Han-predicted and manual asl scores (Pearson $\rho = 0.4, p < 0.05$), but only a weakly significant correlation between manual ratings and AAN-based asl scores (Pearson $\rho = 0.24, p = 0.09$) on considering the 24 program video scenes and 28 ads (52 scores in total). Conversely, manual val ratings correlated significantly with our AAN model (Pearson $\rho = 0.45, p < 0.001$), but only weakly with Han estimates (Pearson $\rho = 0.25, p = 0.08$).

The CAVVA optimization framework [53] 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 valence scores influence both components. Our results suggest that accurate val prediction, as accomplished by our AAN model, plays a critical role in enhancing the subjective user experience. Although this improved experience seems to come at the expense of ad recall, we note that the Han method results in a disruptive experience despite high recall, and hence solely emphasizing on recall may not necessarily lead to the optimal ad placement strategy. Figure 3.4 illustrates the entire system used so far for content-centric AR and application to computational advertising.

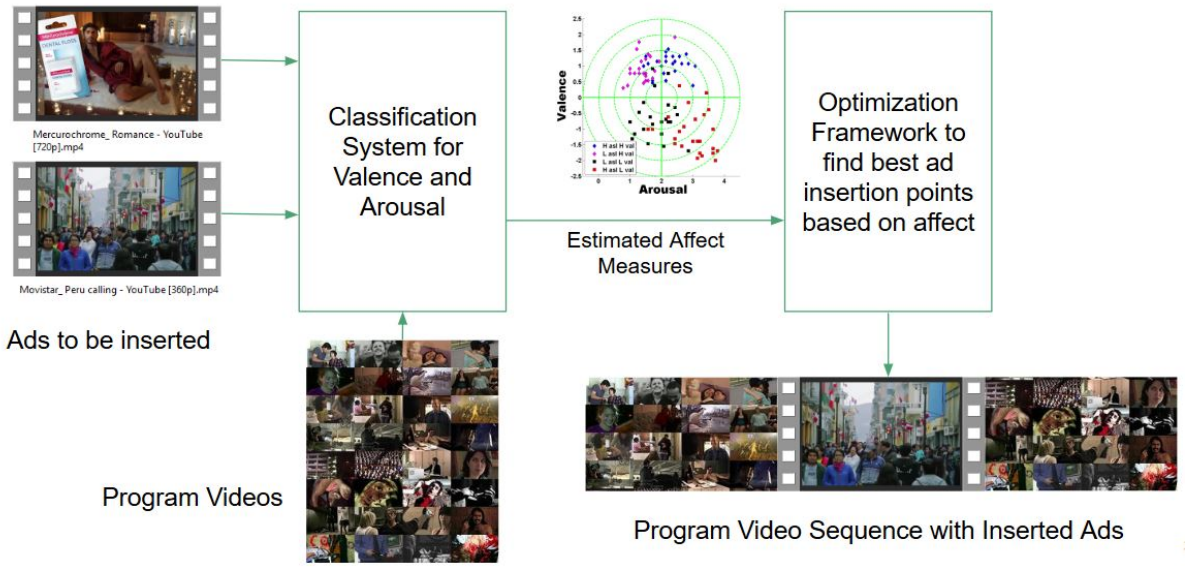


Figure 3.4 An illustration of the content-centric affect recognition system with application to computational advertising

Chapter 4

User-Centric Ad Affect Recognition using Electroencephalography (EEG)

This chapter studies user-centric ad affect recognition using Electroencephalography responses of viewers.

4.1 EEG acquisition protocol

As the 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. We recorded a total of 1738 epochs.

4.1.1 Clean vs Unclean EEG Data

Out of the 1738 epochs that we have, we perform manual visual rejection for those epochs which have a clear artifact in them. Each ad was preceded by a 1s fixation cross to orient user attention,

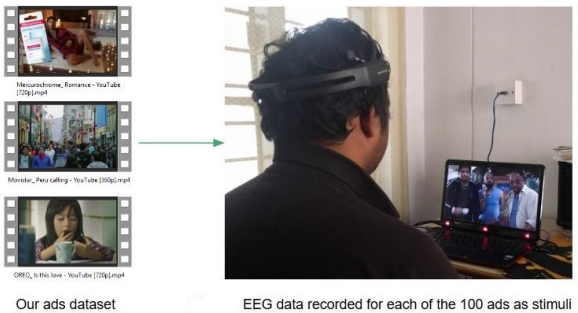


Figure 4.1 An illustration of the recording protocol for the EEG data

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. This process results in the removal of 212 epochs to leave us with 1526 *clean* epochs. For the rest of the paper, **clean** EEG data refers to the 1526 preprocessed epochs after visual rejection and ICA whereas the **unclean** EEG data refers to the raw data in the 1738 epochs. We perform our experiments separately on both these sets.

4.2 Method

The 1738 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 ($\approx 30s$ 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.

4.3 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 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 3.1), while the W_{est} technique [27] 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 i^{th} 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 [13] 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.

4.3.1 Results Overview

Tables 3.2 and 4.1 respectively present content-centric and user-centric AR results for the various settings described above. For the user centric EEG methods, the *Clean* and *Unclean* EEG data have the same meaning as discussed earlier. 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[1]. 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 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

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

negative val as noted in [32], 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 3.2), 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.

4.3.2 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 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.

Table 4.1 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)**
Clean + 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
Unclean + LDA	0.79 ± 0.02	0.78 ± 0.02	0.76 ± 0.03	0.76 ± 0.02	0.76 ± 0.02	0.72 ± 0.04
Clean + 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
Unclean + LSVM	0.78 ± 0.03	0.77 ± 0.04	0.77 ± 0.05	0.75 ± 0.03	0.74 ± 0.02	0.70 ± 0.04
Clean + RSVM	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.81 \pm 0.03}$	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.80 \pm 0.03}$	$\mathbf{0.76 \pm 0.04}$
Unclean + RSVM	0.80 ± 0.03	0.79 ± 0.03	0.79 ± 0.03	0.77 ± 0.03	0.77 ± 0.04	0.74 ± 0.04

4.4 Computational Advertising - User Study

Given that superior ad AR is achieved with user EEG responses (see Table 4.1), 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 [53]. 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** [13] and **Deep** CNN models, and the user-centric **EEG** model.

4.4.1 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 4.2 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.

Table 4.2 Summary of program video statistics.

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

4.4.2 Advertisement insertion strategy

We used the three aforementioned models to perform ad affect estimation. For the 24 program video scenes (3 videos $\times$ 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 [53] 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.

4.4.3 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.

4.4.4 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

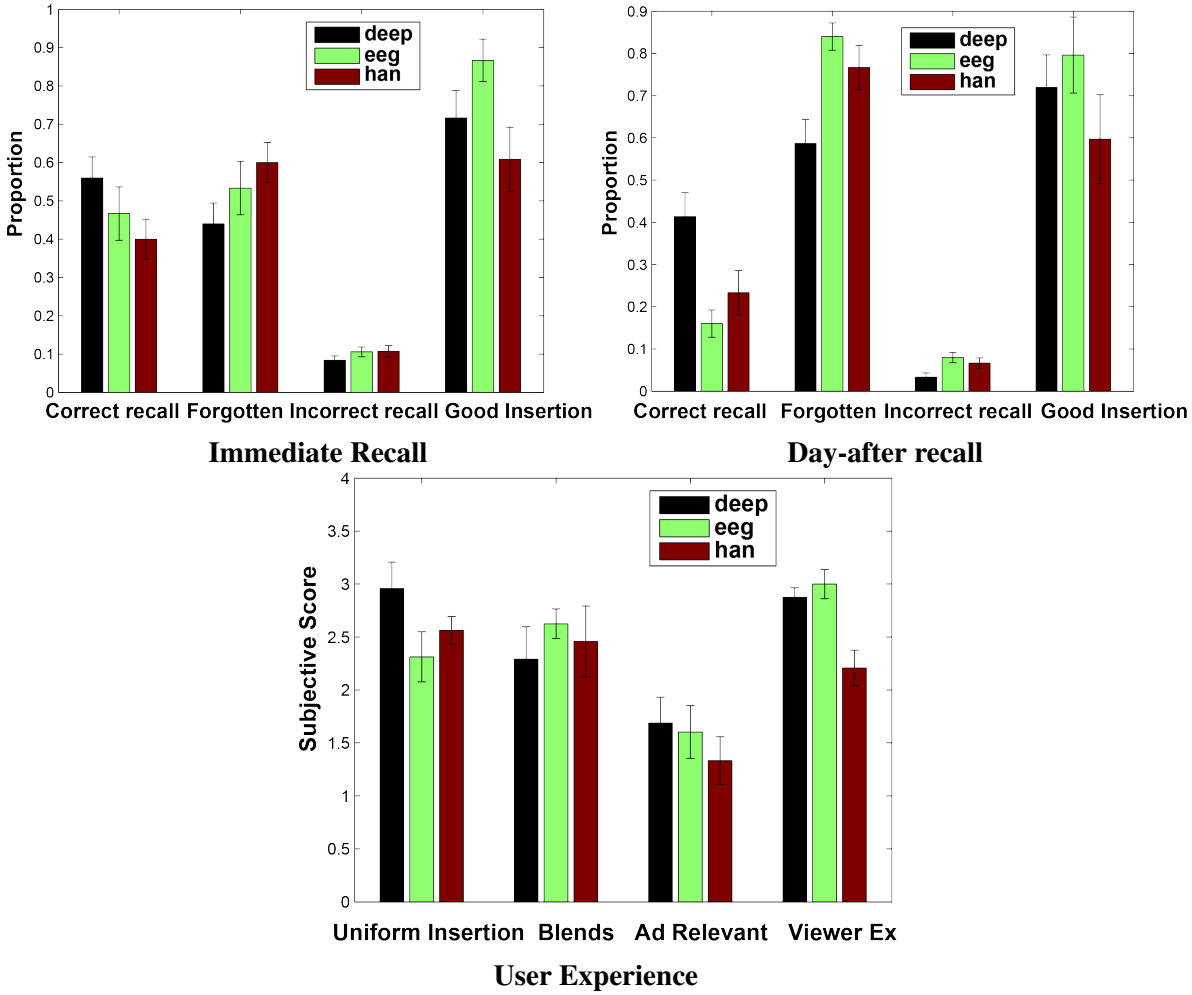


Figure 4.2 Summary of user study results in terms of recall and user experience-related measures. Error bars denote unit standard deviation.

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 4.2 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* [15, 53, 21], and the immediate and day-after recall rates observed for three considered approaches are presented in Fig. 4.2 (left),(middle). Video program sequences obtained with Deep affective scores result in 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. 4.2(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 [53] 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.

Chapter 5

Deep Neural Network for EEG & Fusion of Content and User-centric Approaches

This chapter studies user-centric ad affect recognition using Electroencephalography responses of viewers using a convolutional neural network and a multi-task learning framework. It also discusses the fusion of content-centric and user-centric modalities.

5.1 EEG acquisition protocol

As the 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. We recorded a total of 1738 epochs.

5.1.1 Clean vs Unclean EEG Data

Out of the 1738 epochs that we have, we perform manual visual rejection for those epochs which have a clear artifact in them. 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. This process results in the removal of 212 epochs to leave us with 1526 *clean* epochs. For the rest of the paper, **clean** EEG data refers to the 1526 preprocessed epochs after visual rejection and ICA whereas the **unclean** EEG data refers to the raw data in the 1738 epochs. We perform our experiments separately on both these sets.

5.2 Method

The 1738 clean epochs obtained from the EEG acquisition process were used for user-centered analysis. However, these 1738 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 ($\approx 30s$ 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.

5.2.1 EEG Feature Extraction for CNN Training

We have a relatively small number of epochs in our dataset (1738), and a very high dimensionality for each epoch in the time domain (14 channels * 3667 time points = **51338** dimensions in the vectorized epoch). Training a CNN in such a scenario is highly susceptible to problems associated with overfitting. To alleviate these problems, we apply Principal Component Analysis (PCA) on the vectorized epochs in order to both reduce the dimensionality and retain the most important components of the initial feature set. PCA has found success in classfication of EEG signals in recent times [18, 40]. Other works also stress the necessity of an EEG signal representation as a preprocessing step prior to neural network training [25, 42, 44, 45, 43]. [18] in particular highlights the effectiveness of PCA for the EEG signal that is needed to get a good representation that is suitable for a deep learning network.

5.2.2 CNN Training for EEG features

The extracted EEG features were then passed to a CNN for valence and arousal recognition. The architecture that we used was based on a recent work on time series sensor data classification [35]. The network is three layers deep with two 1-D convolutional layers followed by a fully connected layer. We used the Keras [10] library for our implementation. For training, we used 90% of the collected dataset as the training set and the remaining 10% as the testing set. The training was performed with a momentum of 0.9, weight decay of 0.0005, and a base learning rate of 0.0001, and the best performance was achieved with 64 filters in the 1-D convolutional layers and 128 nodes in the fully connected layer. A dropout level of 0.5 was used to add regularization and combat overfitting. The model was trained for a maximum of 100 epochs and the early stopping criterion was used to prevent model degradation in case the validation loss increased for 5 successive training iterations. We used a 10-fold cross-validation scheme to evaluate the performance of our model on the dataset.

5.3 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 Chapter II

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 slightly unbalanced distribution of positive and negative samples. We attempted probabilistic decision fusion of the unimodal outputs for audiovisual features with the EEG features. The W_{est} technique [27] 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 i^{th} 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.

Experimental results will be discussed thereafter.

Table 5.1 Ad AR from EEG analysis with CNN and MTL. 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)**
CNN Clean	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.88 \pm 0.04}$	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.87 \pm 0.03}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.80 \pm 0.06}$
CNN Unclean	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.79 \pm 0.04}$
MTL Clean	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.90 \pm 0.04}$
MTL Unclean	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.87 \pm 0.04}$	$\mathbf{0.85 \pm 0.05}$

Table 5.2 Ad AR from Fusion of Content Analysis & EEG. 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)**
(Unclean EEG + SVM) + (AV fc7) DF	0.85 ± 0.03	0.84 ± 0.03	0.84 ± 0.03	0.84 ± 0.03	0.83 ± 0.03	0.80 ± 0.04
(Unclean EEG + CNN) + (AV fc7) DF	0.87 ± 0.03	0.87 ± 0.03	0.86 ± 0.02	0.86 ± 0.01	0.85 ± 0.03	0.83 ± 0.04
(Clean EEG + SVM) + (AV fc7) DF	0.86 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.85 ± 0.02	0.83 ± 0.04	0.82 ± 0.04
(Clean EEG + CNN) + (AV fc7) DF	0.91 ± 0.03	0.89 ± 0.03	0.88 ± 0.02	0.88 ± 0.02	0.87 ± 0.02	0.84 ± 0.04

5.3.1 Results Overview

Tables 3.2 and 4.1 respectively present content-centric and user-centric AR results for the various settings described above, whereas Table5.2 presents the results for fusion of the content and user-centric modalities. 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.

In addition to the above *single-task learning* methods which do not exploit the underlying structure of the input data, we also explored the use of *multi-task learning* (MTL) for AR from EEG features. MTL seeks to jointly learn a set of task-specific classifiers on modeling task relationships, which is highly beneficial when learning with few examples. As described earlier in Chapter III, we employed

the sparse graph-regularized MTL (SR-MTL) where *a-priori* knowledge regarding task-relatedness is modeled in the form of a graph R. Given tasks $t = 1...T$, with X_t denoting training data for task t and Y_t their labels, SR-MTL jointly learns a weight matrix $W = [W_1..W_T]$ such that the objective function $\sum_{t=1}^{T} \|W_t^T X_t - Y_t\|_F^2 + \alpha\|WR\|_F^2 + \beta\|W\|_1 + \gamma\|W\|_F^2$ is minimized. Here, α, β, γ are regularization parameters, while $\|.\|_F$ and $\|.\|_1$ denote the matrix Frobenius (ℓ_2) and ℓ_1-norms respectively. MTL is particularly suited for dimensional AR, and one can expect similarities in terms of audio-visual content among high val or high asl ads. We exploit underlying similarities by modeling each asl-val quadrant as a task (*i.e.*, all H asl, H val ads will have identical task labels). Also, quadrants with same asl/val labels are deemed as related tasks, while those with dissimilar labels are considered unrelated. The graph R guides the learning of W_t's. We find that the EEG features combined with MTL give the best results obtained in this thesis, with a peak F1 score of 0.97 for valence and 0.96 for arousal.

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. 4.4.2). 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 [46, 22, 23] problems. We attempt the fusion of content-centric audiovisual models with the user-centric EEG based models via the DF technique. We find that the fusion results are superior to either unimodal result (Best F1 score of 0.91 for val and 0.88 for asl). This leads us to believe that both the different kinds of modalities carry complementary affective information.

Overall, *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 [32], and the Emotiv device is known to efficiently capture frontal lobe activity despite its limited spatial resolution.

5.4 Computational Advertising - User Study

Given our observation of human centric EEG features leading to superior AR from ads, we studied whether improved AR results in a better computational advertising application by virtue of insertion of ads at appropriate break points in a video (as discussed in the CAVVA ad insertion framework [53]). CAVVA is a genetic algorithm optimization based framework for inserting ads onto streamed videos. It models the affective relevance between video scenes and candidate ads to determine (a) which ads to insert, and (b) the position after a scene ending at which the chosen ads are to be inserted. CAVVA tries to strike a balance between (a) maximizing ad impact (recall) in terms of brand memorability, and (b)

Table 5.3 Summary of program video statistics.

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

minimally disrupting (or even enhancing) the overall viewer experience. To validate our hypothesis, we performed a user study to compare the recall as well as the subjective quality of advertising schedules generated via affective scores from the content-centric Audiovisual CNN model, the user-centric EEG CNN model and manual ratings given by the experts.

5.4.1 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 4.2 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.

5.4.2 Advertisement insertion strategy

We used the three aforementioned models to perform ad affect estimation. For the 24 program video scenes (3 videos $\times$ 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 **content-centric** 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. Similarly, the mean of normalized softmax class probabilities over all EEG epochs for a particular ad was used for the user-centric **EEG** method. The average of continuous valence and arousal ratings of 5 experts was used for the **Manual** method. We then adopted the CAVVA optimization framework [53] 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.

5.4.3 Experiment and Questionnaire Design

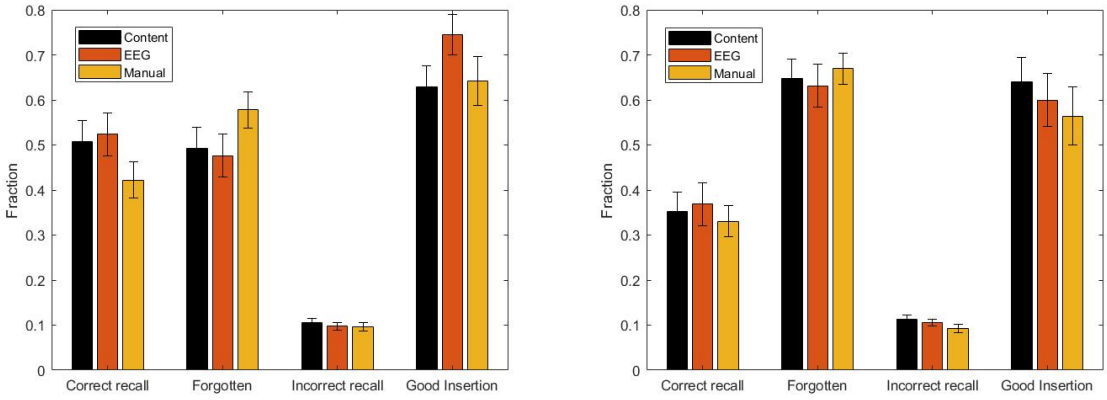
To evaluate the subjective quality of the generated video program sequences and thereby the utility of the three affect estimation techniques for computational advertising, we recruited 18 users (7 female, mean age 20.1 years) who were university undergraduates/graduates. Each of these users viewed a 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 six of the 18 viewers, and we have a total of 54 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.

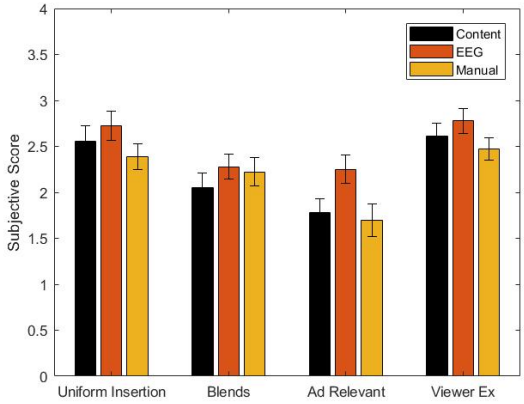
Recall evaluation to intended to verify if the inserted ads were attended to by viewers, and the immediate and day-after recall were ***objective*** measures that quantified the impact of ad insertion on the short-term (immediate) and long-term (day-after) memorability of advertised content, upon viewing 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?



Immediate Recall **Day-after recall**



User Experience

Figure 5.1 Summary of user study results in terms of recall and user experience-related measures. Error bars denote unit standard deviation.

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.

5.4.4 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 content-centric CNN, EEG and Manual 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.

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 ex-

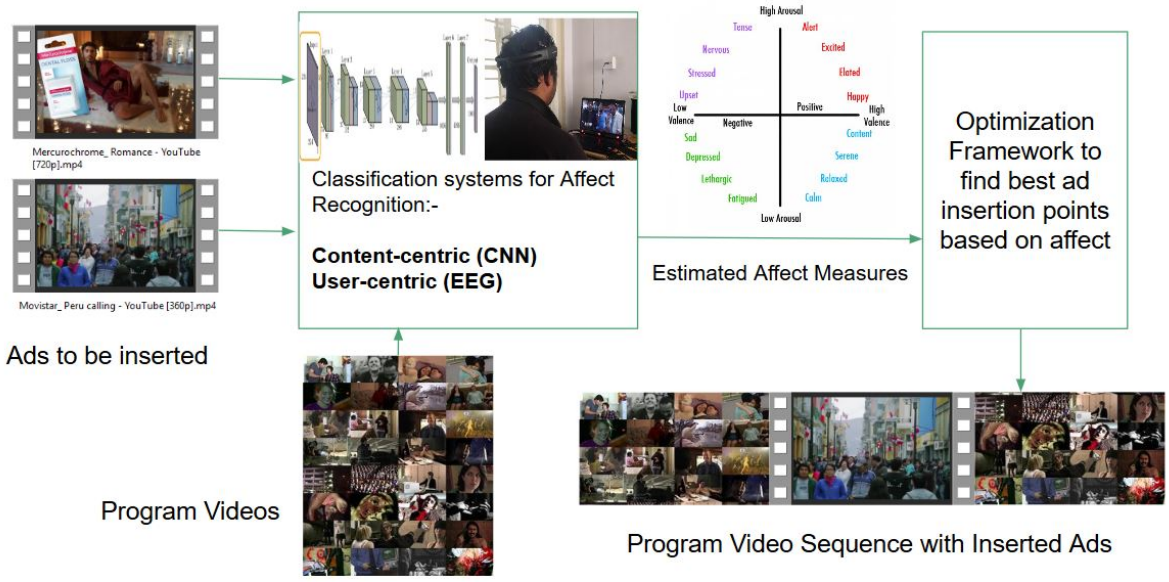


Figure 5.2 An illustration of the entire system

periment. Figure 5.1 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* [15, 53, 21], and the immediate and day-after recall rates observed for three considered approaches are presented in Fig. 5.1 (left),(middle). Video program sequences generated using EEG affective scores result in the highest immediate and day-after recall and the least ad forgottenness, with the content-centric CNN scores coming second. Ads inserted by the EEG method are also found to be the best inserted. The trends observed for immediate and day-after recall are slightly different, but the rate of recall is much worse for the day-after condition with a high proportion of ads being forgotten. Nevertheless, the observed results clearly indicate that the user-centric EEG approach that achieves superior AR as compared to the content-centric CNN method also leads to better ad memorability. Figure 5.2 shows a complete schematic representation of the entire system.

Chapter 6

Conclusions

This work discusses affect prediction from ads, and the utility of better ad affect estimation is demonstrated via a computational advertising application. A curative set of 100 diverse ads is compiled based on expert consensus, and its effectiveness as an affective dataset is examined based on ratings acquired from 14 raters. Dataset suitability is confirmed by (1) excellent agreement between the expert and annotator groups, and (2) uniform distribution of the asl and val ratings with minimal correlation between them. The dataset in itself is a potentially very important contribution to the community.

We evaluate 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 [28] 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 [26, 54, 1]. 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.

As ads are inherently emotional and have great influencing/monetizing capacity [14, 33], the ability to infer ad emotions and make optimal ad insertions within video streams would be highly advantageous for multimedia systems. Therefore, it is surprising that very few works [31, 53] have attempted to mine visual and emotional content in ads. In this regard, our work expressly sets out to model the emotion conveyed by 100 ads based on subjective human opinions and objective audio-visual features.

6.1 A Suitable Control Dataset for Affective Studies

We carefully curated a small but diverse set of ads based on consensus among experts, and examined if those ads could coherently evoke emotions across viewers by acquiring asl and val ratings from 23 annotators. A good-to-excellent agreement on asl and val impressions is noted between the expert and

annotator groups. Also, annotator ad ratings are found to be uniformly distributed over the asl-val plane, with only a weak negative correlation noted between asl and val ratings.

As the compiled ads are found to constitute a control affective dataset, we modeled the emotion conveyed by these ads in terms of audio-visual features. Specifically, we extracted fc7 layer outputs from the *AdAffectNet* CNNs fed with key frames and spectrograms for video and audio-based affect modeling. While CNNs have been previously used for video and audio-based AR [4, 11], the modeled scenes are only a few second long snippets. In contrast, we have explored the validity of CNN-based AR for full-length ads, some of which are over a minute long, in this work. We also fully compare content-centric and user-centric AR, and design a CNN based approach for EEG responses recorded while viewing the ads.

6.2 Content-Centric Ad Affect Recognition

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. AAN-based audio-visual features are proposed for encoding ad affect, and are found to significantly outperform features proposed by Hanjalic and Xu [13] for val recognition. Best results with the AAN features are achieved with the MTL classifier, which effectively exploits the underlying audio-visual similarities among emotionally homogeneous ads. Finally, a study involving 17 users confirms that better modeling of ad emotions facilitates insertion of contextually relevant ads onto a streamed video. Specifically, the proposed AAN model is able to estimate ad valence better than the baseline [13], resulting in enhanced viewing experience.

AAN-based audio-visual features are then proposed for encoding ad affect, and are found to significantly outperform features proposed by Hanjalic and Xu [13] for val recognition. Best results with the AAN features are achieved with the MTL classifier, which effectively exploits the underlying audio-visual similarities among emotionally homogeneous ads.

Finally, a study involving 17 users confirms that better modeling of ad emotions facilitates insertion of contextually relevant ads onto a streamed video. Specifically, the proposed AAN model is able to estimate ad valence better than the baseline [13], resulting in enhanced viewing experience.

6.3 User-Centric Ad Affect Recognition

We evaluate the performance of two content-centered methods, the popular **Han** baseline for affect prediction from low-level audio-visual descriptors, and our 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 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.

We also develop a novel CNN architecture for AR from EEG data and find that it gives us the best unimodal AR performance. We further boost this performance using multi task learning which exploits the intrinsic relatedness of the affective quadrants which is useful in predicting a particular affective attribute. We then fuse the content-centric and user-centric modalities using decision fusion and find that the fusion performance is greater than either unimodal performance, which confirms that the audiovisual and the EEG modalities carry complementary affective information.

Two studies involving 12 and 18 users each to examine if improved AR facilitates computational advertising reveal that (1) Ad memorability is maximized with better modeling of the ad affect via the audiovisual CNN 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 work represents the first affective computing work to establish a direct relationship between objective AR performance and subjective viewer opinion, i.e. **better accuracy in estimating valence and arousal from advertisement videos leads to a more emotionally relevant and non disruptive computational advertising experience**.

6.4 Conclusions from all Results

We now summarize and compare the entirety of 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. Audiovisual fusion-based approaches do not perform much 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. Even better performance is seen by the fusion of the best content-centric and user-centric methods, which points to the two kinds of modalities carrying complementary information about valence and arousal. Table 6.1 shows a complete comparision of all results. For the user centric EEG methods, the prefix **U** stands for *unclean* EEG data and the prefix **C** stands for *clean* EEG data, as discussed in chapter IV. DF stands for the decision fusion technique, and the other classifiers follow the same conventions followed in the rest of this work.

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.

Table 6.1 Ad AR from all tested approaches. 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
Audio FC7 + MTL	0.85±0.02	0.83±0.10	0.78±0.20	0.78±0.03	0.62±0.14	0.45±0.16
Video FC7 + MTL	**0.96±0.01**	0.94±0.07	0.82±0.25	**0.94±0.01**	0.87±0.12	0.63±0.29
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
A+V FC7 + MTL	**0.89±0.03**	0.88±0.11	0.77±0.26	**0.87±0.03**	0.68±0.17	0.46±0.20
A+V Han + MTL	0.77±0.04	0.79±0.07	0.74±0.15	0.78±0.04	0.73±0.11	0.58±0.22
C-EEG + 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
U-EEG + LDA	0.79 ± 0.02	0.78 ± 0.02	0.76 ± 0.03	0.76 ± 0.02	0.76 ± 0.02	0.72 ± 0.04
C-EEG + 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
U-EEG + LSVM	0.78 ± 0.03	0.77 ± 0.04	0.77 ± 0.05	0.75 ± 0.03	0.74 ± 0.02	0.70 ± 0.04
C-EEG + 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
U-EEG + RSVM	0.80 ± 0.03	0.79 ± 0.03	0.79 ± 0.03	0.77 ± 0.03	0.77 ± 0.04	0.74 ± 0.04
C-EEG + CNN	**0.89 ± 0.05**	0.88 ± 0.04	0.88 ± 0.05	**0.87 ± 0.03**	0.85 ± 0.04	0.80 ± 0.06
U-EEG +CNN	0.85 ± 0.03	0.85 ± 0.03	0.83 ± 0.03	0.84 ± 0.02	0.82 ± 0.03	0.79 ± 0.04
C-EEG + MTL	**0.97 ± 0.01**	**0.97 ± 0.01**	0.93 ± 0.03	**0.96 ± 0.01**	0.94 ± 0.02	0.90 ± 0.04
U-EEG + MTL	0.92 ± 0.01	0.91 ± 0.01	0.90 ± 0.01	0.90 ± 0.02	0.87 ± 0.04	0.85 ± 0.05
U-EEG-SVM AV DF	0.85 ± 0.03	0.84 ± 0.03	0.84 ± 0.03	0.84 ± 0.03	0.83 ± 0.03	0.80 ± 0.04
U-EEG-CNN AV DF	0.87 ± 0.03	0.87 ± 0.03	0.86 ± 0.02	0.86 ± 0.01	0.85 ± 0.03	0.83 ± 0.04
C-EEG-SVM AV DF	0.86 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.85 ± 0.02	0.83 ± 0.04	0.82 ± 0.04
C-EEG-CNN AV DF	**0.91** ± 0.03	0.89 ± 0.03	**0.88** ± 0.02	0.88 ± 0.02	0.87 ± 0.02	0.84 ± 0.04

Obtained AR results confirm that the synthesized fc7 features are effective predictors of asl and val. They outperform the audio-visual features proposed by Hanjalic and Xu [13] with both single and multi-task classifiers. In particular, while fc7 features are considerably better for val, Han features provide competitive performance for asl. Optimal AR is achieved with the MTL classifier (with both the audiovisual and EEG modalities), which is able to effectively exploit the underlying similarities among emotionally homogeneous ads in terms of audio visual content. Nevertheless, a significant drop in recognition performance is generally noted for the terminal ad portion with most methods, and especially for asl, implying that (a) asl is a more transient phenomenon as compared to val, and there is less coherence between the employed AV features and asl labels towards the ad endings, and (b) the use of a single asl/val over the entire ad duration may be inappropriate, and one may need to acquire time-varying labels for affective studies.

6.5 Major Takeaways

If there are three things that can be listed as the most important findings of this work, they would be the following:-

- **Deep audiovisual CNN features are more effective at affect recognition than traditional content centric baselines**
- **Human centered EEG responses are much more effective at affect recognition than any existing or proposed content centered method**
- **Better affect recognition leads to better computational advertising by virtue of inserting ads at more emotionally appropriate positions in videos**

6.6 Future Work

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. The rules [20] were designed more in the context of broadcast and print media and are slightly dated. They need to be more thoroughly validated in the modern online advertising context, and there is also the potential to incorporate new elements such as the length of the advertisement and how it impacts viewer experience into the optimization function.

We have established that human centered EEG responses are more effective at affect recognition than content centered methods. Building on EEG, it would also be interesting to study affect prediction via other user physiological measurements similar to [26, 1]. Other modalities including eye gaze, galvanic skin response, heart rate and a more portable EEG headset are all viable avenues for further exploration,

not just for computational advertising but for affect recognition problems in general even in other fields such as depression analysis or autism detection.

Another promising line of inquiry is to focus on the design of other informative (*e.g.*, recurrent neural network-based) multimedia features for modeling affect. The current content centric methods are based on convolutional neural network features alone. While they are effective at automatically being able to learn a good affective representation, they do not directly incorporate the temporal evolution of affect or the continuous sequence prediction of affect that a sequence based network such as an LSTM is able to do. Perhaps the most major limitation of this work is the prediction of a high or low binary label for valence and arousal. This limitation exists due to both the continuous annotation process being tedious and preliminary attempts at regression based models not performing well. However, there exist useful datasets such as the MAHNOB [41] HCI database which contain continuous affective annotations as well as multimodal sensor data that can be evaluated. There is also the potential to learn modality invariant representations of multimodal data to serve both the content centered and user centered modalities, such as recently illustrated by deep neural networks trained with adversarial techniques[50]. It could also be very useful to understand which modality carries useful information at a given point of time and subsequently prioritise its weight in the affect recognition process. A recent attempt to this end has been made by using reinforcement learning [9].

Finally, efficient approaches need to be designed for ad insertion within streamed video, so as to maximize both ad recall and viewing experience. Hopefully, this work can serve as a stimulus to spawn efforts into parallel directions and enhance both affect recognition and emotion driven computational advertising.

Related Publications

- **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 International Conference on Multimedia*
- **Abhinav Shukla**, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Ramanathan Subramanian. 2017. Evaluating Content-centric vs User-centric Ad Affect Recognition. In *ACM International Conference on Multimodal Interaction*
- **Abhinav Shukla**, Shruti Shriya Gullapuram, Harish Katti, Stefan Winkler, Mohan Kankanhalli, and Ramanathan Subramanian. Submitted. Emotion Recognition in Advertisement Videos with Application to Computational Advertising. In *IEEE Transactions on Affective Computing*

Bibliography

[1] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. DECAF: Meg-based multimodal database for decoding affective physiological responses. *IEEE Trans. Affective Computing*, 6(3):209–222, 2015.

[2] T. AlHanai and M. Ghassemi. Predicting latent narrative mood using audio and physiologic data. In *AAAI Conference on Artificial Intelligence*, 2017.

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

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

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

[6] M. Bilalpur, S. M. Kia, T.-S. Chua, and R. Subramanian. Discovering gender differences in facial emotion recognition via implicit behavioral cues. In *Affective Computing & Intelligent Interaction*, 2017.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 53. Springer, 2013.

[8] V. C. Broach, T. J. Page, and R. D. Wilson. Television Programming and Its Influence on Viewers' Perceptions of Commercials: The Role of Program Arousal and Pleasantness. *Journal of Advertising*, 24(4):45–54, 1995.

[9] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 163–171, New York, NY, USA, 2017.

[10] F. Chollet et al. Keras. https://github.com/keras-team/keras, 2015.

[11] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *International Conference on Multimodal Interaction*, pages 445–450, 2016.

[12] 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.

[13] A. Hanjalic and L.-Q. Xu. Affective Video Content Representation. *IEEE Trans. Multimedia*, 7(1):143–154, 2005.

[14] M. B. Holbrook, R. Batra, and R. Batra. Assessing the Role of Emotions as Mediators of Consumer Responses to Advertising. *Journal of Consumer Research*, 14(3):404–420, 1987.

[15] M. B. Holbrook and J. O. Shaughnessy. The role of emotlon in advertising. *Psychology & Marketing*, 1(2):45–64, 1984.

[16] Z. Huang, M. Dong, Q. Mao, and Y. Zhan. Speech emotion recognition using cnn. In *ACM Multimedia*, pages 801–804, 2014.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. CAFFE: Convolutional architecture for fast feature embedding. In *ACM Int'l Conference on Multimedia*, pages 675–678, 2014.

[18] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 2014.

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

[20] M. A. Kamins, L. J. Marks, and D. Skinner. Television commercial evaluation in the context of program induced mood: Congruency versus consistency effects. *Journal of Advertising*, 20(2):1–14, 1991.

[21] M. K. Karthik Yadati and, Harish Katti and. Interactive video advertising: A multimodal affective approach. *Multimedia Modeling (MMM 13)*, 2013.

[22] H. Katti, M. V. Peelen, and S. P. Arun. Object detection can be improved using human-derived contextual expectations. *CoRR*, abs/1611.07218, 2016.

[23] H. Katti, A. K. Rajagopal, K. Ramakrishnan, M. Kankanhalli, and T.-S. Chua. Online estimation of evolving human visual interest. *ACM Transactions on Multimedia*, 11(1), 2013.

[24] H. Katti, R. Subramanian, M. Kankanhalli, N. Sebe, T.-S. Chua, and K. R. Ramakrishnan. Making computers look the way we look: exploiting visual attention for image understanding. In *ACM Int'l conference on Multimedia*, pages 667–670, 2010.

[25] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1809–1817. ACM, 2017.

[26] 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 Trans. Affective Computing*, 3(1):18–31, 2012.

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

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1097–1105, 2012.

[29] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. 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, 2008.

[30] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.

[31] T. Mei, X.-S. Hua, L. Yang, and S. Li. Videosense: Towards effective online video advertising. In *ACM Int'l Conference on Multimedia*, pages 1075–1084, 2007.

[32] D. Oude Bos. Eeg-based emotion recognition - the influence of visual and auditory stimuli. In *Capita Selecta (MSc course)*. University of Twente, 2006.

[33] M. T. Pham, M. Geuens, and P. D. Pelsmacker. 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):383 – 394, 2013.

[34] H. R.-Tavakoli, A. Atyabi, A. Rantanen, S. J. Laukka, S. Nefti-Meziani, and J. Heikkila. Predicting the valence of a scene from observers' eye movements. *PLoS ONE*, 10(9):1–19, 2015.

[35] N. M. Rad, S. M. Kia, C. Zarbo, T. van Laarhoven, G. Jurman, P. Venuti, E. Marchiori, and C. Furlanello. Deep learning for automatic stereotypical motor movement detection using wearable sensors in autism spectrum disorders. *Signal Processing*, 144:180–191, 2018.

[36] U. Rimmele, L. Davachi, R. Petrov, S. Dougal, and E. Phelps. Emotion enhances the subjective feeling of remembering, despite lower accuracy for contextual details. *Emotion*, 11:553–562, 2011.

[37] J. Russell. A circumplex model of affect., 1980.

[38] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Affect recognition in ads with application to computational advertising. In *ACM Int'l conference on Multimedia*, 2017.

[39] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Evaluating content-centric vs. user-centric ad affect recognition. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 402–410, New York, NY, USA, 2017. ACM.

[40] S. Siuly, Y. Li, and Y. Zhang. Injecting principal component analysis with the oa scheme in the epileptic eeg signal classification. In *EEG Signal Analysis and Classification*, pages 127–150. Springer, 2016.

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

[42] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6809–6817, 2017.

[43] S. Stober. Learning discriminative features from electroencephalography recordings by encoding similarity constraints. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 6175–6179. IEEE, 2017.

[44] S. Stober, D. J. Cameron, and J. A. Grahn. Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings. In *Advances in neural information processing systems*, pages 1449–1457, 2014.

[45] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn. Deep feature learning for eeg recordings. *arXiv preprint arXiv:1511.04306*, 2015.

[46] R. Subramanian, H. Katti, K. Ramakrishnan, M. Kankanhalli, T.-S. Chua, and N. Sebe. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision*, 2010.

[47] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of vision*, 14(3):1–18, 2014.

[48] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 2016.

[49] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler. A probabilistic approach to people-centric photo selection and sequencing. *IEEE Transactions on Multimedia*, 2017.

[50] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 154–162, New York, NY, USA, 2017.

[51] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Trans. Circ. Syst. V. Tech.*, 16(6):689–704, 2006.

[52] K. Yadati. Online Multimedia Advertising. Master's thesis, Natonal University of Singapore, Singapore, 2013.

[53] K. Yadati, H. Katti, and M. Kankanhalli. CAVVA: Computational affective video-in-video advertising. *IEEE Trans. Multimedia*, 16(1):15–23, 2014.

[54] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu. Eeg-based emotion classification using deep belief networks. *IEEE International Conference on Multimedia & Expo*, 2014.

[55] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

[56] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.