Recognition of Advertisement Emotions with Application to Computational Advertising

Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Mohan Kankanhalli, *Fellow, IEEE*, Stefan Winkler, *Fellow, IEEE*, and Ramanathan Subramanian, *Senior Member, IEEE*

Abstract—Advertisements (ads) often contain strong emotions to capture audience attention and convey an effective message. Still, little work has focused on affect recognition (AR) from ads employing audiovisual or user cues. This work (1) compiles an affective video ad dataset which evokes coherent emotions across users; (2) explores the efficacy of content-centric convolutional neural network (CNN) features for ad AR vis-ã-vis handcrafted audio-visual descriptors; (3) examines user-centric ad AR from Electroencephalogram (EEG) signals, and (4) demonstrates how better affect predictions facilitate effective computational advertising via a study involving 18 users. Experiments reveal that (a) CNN features outperform handcrafted audiovisual descriptors for content-centric AR; (b) EEG features encode ad-induced emotions better than content-based features; (c) Multi-task learning achieves optimal ad AR among a slew of classifiers and (d) Pursuant to (b), EEG features enable optimized ad insertion onto streamed video compared to content-based or manual insertion, maximizing *ad recall* and *viewing experience*.

Index Terms—Affect Recognition; Advertisements; Perception; Content-centric Features; Convolutional Neural Networks; EEG; Multimodal; Multi-task Learning; Ad Insertion

1 Introduction

Advertising is a pivotal industry in today's digital world, where advertisers showcase their products and services as highly worthy and rewarding. Emotions play a crucial role in conveying an effective message to viewers, and are known to mediate consumer attitudes towards brands [1]–[3]. Emotions are also critical for spreading public health and safety awareness, where certain personal choices are portrayed as *beneficial*, while others are portrayed as *deleterious* and possibly *fatal*. Therefore, the ability to objectively characterize video advertisements (ads) in terms of their emotional content has multiple applications–*e.g.*, inserting appropriate ads at optimal temporal points within a video stream can benefit both advertisers and consumers of video streaming websites such as *YouTube* [4], [5]. Subjective experience of pleasantness (***valence***) and emotional intensity (***arousal***) are important affective dimensions [6], and both influence responses to ads in distinct ways [7]. Stimulus valence and arousal are also known to influence recall of images [8], movie scenes [9] and videos [4].

Only a few works have attempted prediction of ad emotions despite the popularity of affective computing, and interest in inferring emotions elicited by image [10], [11], speech [12], audio [13], music [14] and movie [15], [16] content. Affect recognition (AR) from video ads is non-trivial as with music and movie clips [14], [15], [17], [18] since human emotional perception is subjective. In lieu of detecting discrete emotion categories such as *joy* and *sorrow*, many AR works model emotions based on the valence (val) and arousal (asl) dimensions [6], [19]. Overall, AR methods are broadly classified as *content-centric* or *user-centric*. *Content-centric* AR models emotions by examining textual, audio and visual cues [17], [18]. In contrast, *user-centric* AR identifies elicited emotions from facial [20] or physiological [9], [14]–[16] measurements acquired from users or multimedia consumers. While enabling the study of various emotional states, user-centered methods nevertheless suffer from subjectivity issues.

This work expressly studies emotions conveyed by ads, and employs (i) explicit user ratings and (ii) associated content and user-centric measurements for ad AR. Firstly, we examine the efficacy of 100 diverse, carefully curated video ads to coherently evoke emotions across viewers. To this end, we analyse affective first impressions of five experts and 23 novice annotators to note that the two groups are concordant. Secondly, we explore the utility of Convolutional Neural Networks (CNNs) and domain adaptation for encoding emotional audiovisual features. As our ad dataset is relatively small and insufficient for CNN training, we apply knowledge gained from the large-scale, annotated LIRIS-ACCEDE movie dataset [21] for decoding ad emotions. Extensive experimentation confirms that CNN descriptors outperform handcrafted audio-visual descriptors proposed in [17], with a substantial improvement observed for val recognition.

Thirdly, we perform user-centric ad AR with EEG recordings from annotators, and show that a three-layer CNN trained on EEG features performs best for both asl and val recognition. To our knowledge, this is the first work to elaborately compare content and user-centric methods for ad AR. In addition, we explore multi-task learning and feature/decision fusion for asl and val classification. Lastly, we examine if accurate encoding of ad emotions facilitates optimized insertion of ads onto a video stream, as ads

Abhinav Shukla is with the Imperial College, London.
Shruti Shriya Gullapuram is with the Computer Science Department at the University of Massachusetts, Amherst.
Harish Katti is with the Centre for Neuroscience at the Indian Institute of Science, Bangalore.
Mohan Kankanhalli and Stefan Winkler are with the School of Computing at the National University of Singapore.
Ramanathan Subramanian is with Indian Institute of Technology (IIT), Ropar.

chiefly contribute to the revenue of video hosting websites such as *YouTube*. A study with 18 users confirms that insertion of ads identified via EEG-based cues maximizes both ad recall and viewing experience. In summary, we make the following contributions:

1. We present one of the few works to examine ad emotions extending prior findings [22], [23]. We also characterize ad emotions in terms of explicit human opinions, and underlying (content-centric) audiovisual plus (user-centric) EEG features.
2. We present a carefully curated affective dataset of 100 ads and associated affective ratings. Based on statistical analyses, we note that the ad dataset is capable of evoking coherent emotions across disparate users.
3. We show that CNN-based transfer learning, achieved by fine-tuning the *Places205* Alexnet [24], effectively captures audiovisual emotions. CNN features outdo handcrafted descriptors proposed in [17].
4. We compare and contrast AR achieved with content and user-based CNN features. An EEG-based CNN model best encodes emotional attributes. Also, multi-task learning to exploit similarities among emotionally similar ads considerably benefits ad AR.
5. We show how improved AR enables better ad-embedding onto a video stream. Optimized ad insertion results in greater *ad recall* and an enhanced *viewing experience* for users.

2 Related Work

We review related work on (a) affect recognition (b) influence of ads on consumer behavior, and (c) computational advertising to highlight research gaps and position our work with respect to the literature.

2.1 Affect Recognition

Both *content-centric* and *user-centric* approaches have been proposed to infer emotions evoked by multimedia stimuli. Content-centric methods [17], [18] predict the likely elicited emotions by examining image, audio and video-based emotion correlates [17], [23], [25]. In contrast, user-centric AR methods [14]–[16] estimate the stimulus-evoked emotion based on user behavioral (head pose and facial emotion) and physiological cues. Physiological signals indicative of emotions include pupillary dilation [26], eye-gaze [9], [27] and neural activity [14], [15], [28]. Some works also demonstrate the benefit of combining content and user cues– audio-visual features extracted from video clips plus facial responses of users watching those clips are combined in [29] to achieve five-class affective categorization.

Both content and user centric methods require labels denoting stimulus emotion; such labels are compiled from annotators whose opinions are deemed *acceptable* [30], [31], as emotion perception is highly subjective. We curate a set of 100 ads which elicit similar emotions from both *experts* and *novice annotators*.

2.2 Emotional impact of ads

Ad-induced emotions significantly influence consumer behavior both explicitly and implicitly [1]–[3], especially for hedonistic products. While many works examine the correlation between ad emotions and user behavior, few works utilize these findings for targeted advertising. The only work to incorporate emotional information for advertising is CAVVA [4], where ad-in-video insertion is modeled as a discrete optimization problem based on emotional relevance between video scenes and ads. Based on consumer psychology rules, asl and val scores of video scenes and ads are matched to determine (a) suitable ads for presentation and (b) optimal ad insertion points that maximize user engagement. Two recent related works [22], [23] show how improved ad AR positively impacts viewing experience will watching an ad-embedded video.

2.3 Computational advertising

Exploiting AR models for commercial applications has gained interest lately. ***Computational advertising*** focuses on presenting contextually relevant ads to multimedia users for commercial benefits or social good. Despite the fact that ads are emotional, computational advertising methods have mainly matched low-level features between video segments and candidate ads [32], ignoring emotional relevance. A paradigm shift in this regard was introduced by CAVVA [4], which proposes optimized ad insertion based on emotional relevance. CAVVA achieves *content-based* asl and val matching of video scenes and ads; such matching can also be done via a *user-centric* framework [26]. We explore both *content* and *user-centric* cues for performing ad insertions.

2.4 Analysis of related work

Related literature reveals that (1) AR studies are hampered by emotional subjectivity, and a control dataset coherently evoking user emotions is critical to this end; (2) Even though ads are emotional and ad emotions significantly impact user behavior, little effort has been devoted towards incorporating emotional video-ad relevance for computational advertising. To address these issues, we present a control set of affective ads which elicit concordant opinions from both *experts* and *naive users*. We then leverage CNNs for learning audiovisual and EEG-based emotion predictors. Optimal AR is achieved with a CNN classifier employing EEG features, while CNN-based audiovisual descriptors outperform handcrafted features proposed in [17]. We also show how better affect encoding facilitates ad-to-video insertion through the CAVVA mechanism [4] via a user study.

3 Advertisement Dataset

This section describes our ad dataset and protocol employed for collecting user ratings plus EEG responses.

3.1 Dataset Description

The ***circumplex*** emotion model [6] defines ***valence*** as the feeling of *pleasantness/unpleasantness* and ***arousal*** as the *intensity of emotional feeling*. Following this definition, five experts (authors of this work) carefully evaluated and compiled a dataset of 100, roughly 1-minute long commercial ads such that they were uniformly distributed over the arousal–valence plane (Figure 1). The 100 ads are publicly

TABLE 1
Summary statistics for quadrant-wise ads.

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

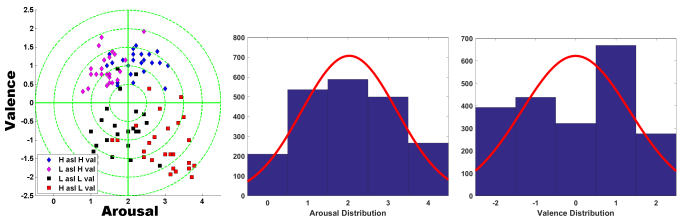


Fig. 1. (left) Scatter plot of mean (discrete) asl, val ratings provided by annotators, color-coded with expert labels. (middle) Asl and (right) Val rating distribution with Gaussian pdf overlay (view under zoom).

available on video hosting websites[1], and an ad was chosen only if there was consensus among all experts on its valence and arousal labels (categorized as either *high* (H)/*low* (L)). High valence ads typically involved product promotions, while low valence ads were awareness messages depicting ill effects of smoking, alcohol, drug abuse, *etc*. Expert labels were considered as ***ground-truth*** in all experiments.

We then examined if the compiled ads could coherently evoke emotions across viewers. To this end, the 100 ads were independently rated by 23 annotators for val and asl. Annotators were familiarized with these emotional attributes prior to the rating task. All ads were rated on a 5-point scale, which ranged from -2 (*very unpleasant*) to 2 (*very pleasant*) for val and 0 (*calm*) to 4 (*highly aroused*) for asl. Table 1 presents summary statistics for the ads. Our low val ads are longer, and are found to elicit stronger emotional reactions from viewers based on the compiled asl scores.

To assess whether the compiled ads evoked coherent emotions, we computed inter-rater agreement via the (a) Krippendorff's α, (b) Fleiss κ and (c) Cohen's κ scores. The α coefficient is applicable when multiple raters rate items ordinally. We obtained $\alpha = 0.62$ (*substantial* agreement) and 0.36 (*fair* agreement) respectively for val and asl, implying that val impressions were more consistent across raters. We also computed the Fleiss κ agreement among annotators. The Fleiss κ statistic (generalization of Cohen's κ) applies when multiple raters assign categorical values (*high*/*low* in our case) to items. Upon thresholding each rater's asl, val scores by their mean rating to assign *high*/*low* labels for each ad, we observed a Fleiss κ of 0.56 (*moderate*) for val and 0.27 (*fair*) for asl among raters. Finally, computing Cohen's κ agreement between each annotator and groundtruth labels (denoting expert opinion), we obtained a mean Cohen's κ of 0.86 (*excellent* agreement) and 0.68 (*substantial* agreement) across annotators for val and asl respectively. Overall, these observations convey (a) considerably higher agreement for val than for asl and (b) consistent affective impressions the compiled ads evoke in the *annotator* and *expert* groups.

Another desirable property of an affective dataset is the relative independence of the asl and val dimensions [6], [33]. To examine asl-val relations for our dataset, we (i) examined scatter plots of annotator ratings, and (ii) computed correlations amongst ratings. Scatter plot of the mean asl, val annotator ratings, and the distribution of these ratings are shown in Fig. 1. The scatter plot color-coded based on expert labels, is different from the 'C' shape observed in prior works [14], [15], [34], and attributed to the hypothesis that only high asl evokes high val ratings. Examination of the scatter plot reveals that a number of ads are rated as moderate asl, but high/low val. Also, roughly uniform asl and val distributions are observed resulting in Gaussian fits with large variance, especially for val. This is plausible as ads are designed to convey a strong positive or negative message, while images and movie scenes may convey a relatively neutral emotion. Wilcoxon rank sum tests on ratings expectedly revealed different asl values for high vs. low asl ads ($p < 0.0001$), and distinctive val scores for high vs. low valence ads ($p < 0.0001$).

Pearson correlation between asl and val ratings with correction for multiple comparisons [35] revealed a negative and insignificant correlation of 0.17, implying that ad asl and val impressions were largely unrelated. Consequently, our 100 ads constitute a *control* affective dataset as (i) they induce a large range of asl and val impressions, which are also found to be largely independent; Different from the 'C'-shape characterizing the asl-val relationship for other stimulus types, asl and val ratings are more uniformly distributed for the ad stimuli, and (ii) there is fair-to-substantial concordance among annotators in addition to the high level of agreement between novice raters and experts on affective labels, implying that our ads evoked coherent emotions.

3.2 EEG acquisition protocol

We recorded Electroencephalogram (EEG) brain signals of the 23 raters via the *Emotiv* wireless, wearable headset. The Emotiv device comprises 14 electrodes and has a sampling rate of 128 Hz. To maximize engagement and minimize fatigue during the rating task, raters took a break after every 20 ads, and viewed the 100 ads over five sessions spanning two hours. Each ad was preceded by a 1s fixation cross to orient user attention, and to measure resting state EEG for baseline power subtraction. Raters had to record their asl and val impressions via mouse clicks within 10s upon ad viewing. EEG recordings were segmented into *epochs*, denoting the ad viewing trials. Upon removing corrupted and aborted recordings, we totally obtained 1738 epochs.

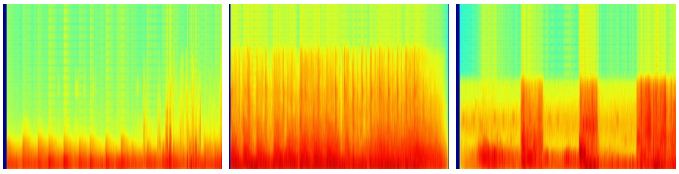
From 1738 recorded epochs, we manually rejected epochs with movement-related artifacts. EEG was band-limited between 0.1–45 Hz, and independent component analysis (ICA) was performed to remove eye movement, eye blink and muscle movement artifacts. This process removed 212 epochs to leave us with 1526 *clean* epochs. Hereon, **clean** EEG data will refer to the 1526 pre-processed epochs, whereas **raw** EEG data will denote the original 1738 epoch data– we attempted CNN-based AR on both data.

4 Content & User-centered AR

4.1 Content-centered Analysis

For content-centered analysis, we extracted and examined audio-visual descriptors from the ads to predict the emotion

1. details @ http://abhinav95.github.io/projects/mm17/dataset



LAHV HAHV HALV

Fig. 2. SGs computed for an exemplar (left) low asl, high val, (middle) high asl, high val and (c) high asl, low val ad. x denotes time (0-10s with 40ms resolution), while y denotes frequency (range of 0 to 14 kHz). Higher spectral intensities are encoded in yellow and red, while lower intensities are shown in blue and green.

(in terms of *high*/*low* asl and val) they are likely to evoke. To this end, we employed a deep convolutional neural network (CNN), and the popular handcoded audio-visual descriptors (such as motion activity, audio pitch, *etc*.) proposed by Hanjalic and Xu [17]. CNNs are employed for many recognition problems, particularly visual [36] and audio [37], but require vast labeled datasets. As our dataset comprised only 100 ads, we fine-tuned the pre-trained *Places205* [36] model via the large-scale and labeled LIRIS-ACCEDE movie dataset [21], and utilized the resulting model to extract emotional ad descriptors– this process is termed *domain adaptation* in machine learning literature.

The ***Places205*** model [36] is originally designed for scene understanding. It is trained on the Places-205 dataset comprising 2.5 million images from 205 scene categories. Places-205 contains a wide variety of scenes captured under varying illumination, viewpoint and field of view, and we hypothesized a coherent relation between scene perspective, lighting and the scene emotion. To find-tune the *Places205* CNN, we employed the LIRIS-ACCEDE dataset [21] which contains asl, val ratings for 9800 $\approx$ 10s long movie snippets. Our ads, contrastingly, range from 30–120s.

4.1.1 FC7 feature extraction via CNNs

To extract deep audio-visual features for ads, we input to the *Places205* CNN *key-frame* images for video, and *spectrograms* for audio. We fine-tuned *Places205* via LIRIS-ACCEDE [21], and extracted features output by the penultimate fully connected (FC7) CNN layer.

Keyframes as Visual Descriptors: From each training video, we uniformly sampled one *keyframe* every 3s– this generated a continuous video profile for AR, and 1791 keyframes were sampled from our 100 ads.

Spectrograms as Audio Descriptors: Spectrograms (SGs, see Fig 2) are visual representations of the audio frequency spectrum and have been employed for music and speech AR [38]. Transforming audio content to a spectrogram image translates audio classification to a visual recognition problem. We extracted SGs over the 10s long LIRIS-ACCEDE clips, and consistently from 10s ad segments. This process generated 610 SGs for our ad dataset. Following [38], we combined multiple tracks to obtain a single SG (instead of two for stereo). Each SG is generated using a 40ms window short time Fourier transform (STFT), with 20ms overlap. Larger densities (denoted by red and yellow shades) of high frequencies can be noted in the SGs for high asl ads, as these intense scenes are often characterized by high frequency (*e.g.*, sudden loud sounds). Conversely, low asl ads tend to sustain the audio, and therefore contain high densities of low frequency sounds.

CNN Training for audio-visual features: We used the Caffe [39] 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 trained four binary classifiers to recognize high and low asl/val from audio/visual features. To fine-tune *Places205*, we used only the top and bottom 1/3rd-ranked LIRIS-ACCEDE videos based on asl and val ratings; we expected these extreme-rated clips will better train a model for inferrring ad emotions. 4096-dimensional FC7 layer descriptors extracted for our ads from the four networks were used for ad AR.

4.1.2 AR with low level audio-visual features

We benchmark CNN features against handcrafted features proposed by Hanjalic and Xu [17] for ad AR. [17] still remains a very popular AR baseline as seen from recent works [14], [15]. In [17], asl and val are modeled via low-level descriptors describing motion activity, colorfulness, shot change frequency, voice pitch and sound energy. These predictors are intuitive and interpretable, and are employed to estimate time-continuous asl and val levels. Table 2 summarizes the content-centric AR features. We perform AR at the keyframe/spectrogram level, while per-frame class probabilities are aggregated to obtain ad-level scores for the computational advertising application (Section 6).

4.2 User-centered Analysis

1738 *raw* or alternatively, 1526 *clean* EEG epochs were used for user-centered experiments, to examine how noise impacted EEG-based AR. These epochs were of different lengths as ad durations were variable. For dimensional consistency and to examine temporal effects, we utilized the (a) *first* 3667 samples ($\approx 30s$ of EEG data), (b) *last* 3667 samples and (c) *last* 1280 samples (10s EEG data) from each epoch. Epochs comprising data from 14 EEG channels were input to classifiers upon vectorization. On top of conventional classifiers, we also used a deep neural network to classify EEG epochs as described below.

4.2.1 EEG Feature Extraction for CNN Training

As we used relatively few (1738) epochs with high dimensionality (14 channels $\times$ 3667 time points = 51338 dimensions), a CNN trained on this data is susceptible to overfitting. Therefore, we applied Principal Component Analysis (PCA) to reduce epoch dimensionality. PCA has been successfully employed to obtain a *good* input for CNN-based EEG classification [40]–[46].

CNN Training for EEG features: Dimensionality-reduced EEG features (preserving 90% data variance) were passed to a CNN for val, asl recognition. We used a CNN employed for time-series data classification [47] and implemented with the Keras [48] library. This three-layer network has two 1D convolutional layers and a fully-connected layer. Training was performed with 64 1×3 filters in the 1D convolutional layers and 128 nodes in the fully connected layer. Other

TABLE 2
Extracted features for content-centric AR. +ve class proportions (as %) for val/asl in the audio and visual modalities are specified.

Attribute Descriptors	Valence/Arousal Audio	Video	aud+vid (A+V)
CNN Features	4096D FC7 features from 10s SGs.	4096D FC7 features extracted from keyframes sampled every 3 seconds.	8192D FC7 features from SGs + keyframes over 10s intervals.
Hanjalic [17] Features	Per-second sound energy and pitch statistics [17].	Per-second shot change frequency and motion statistics [17].	Concatenation of audio-visual features.
+ve class prop (%)	43.8/51.9	43.4/51.6	43.8/51.9

parameters include a momentum of 0.9, weight decay of 0.0005 and a base learning rate of 0.0001. A dropout level of 0.5 was used to prevent overfitting. The model was trained for 100 epochs, and *early stopping* was enforced if validation loss increased over five successive iterations. For content and user-centric AR, 80% training and 20% test data were used with 10-fold repetition (10 $\times$ 5-fold cross validation). Training and test sets were mutually exclusive in both cases, *i.e.*, data corresponding to each ad video were either part of the training or test set.

5 Experiments and Results

We first describe the settings and classifiers employed for *binary* content and user-centric AR, where the objective is to assign a H/L label for the asl and val evoked by each ad, with fc7/low-level audiovisual/EEG features. Ground truth labels are as provided by the *experts*.

Metrics and Experimental Settings: We used the F1-score (F1) defined as the harmonic mean of precision and recall as our performance metric. F1 is apt for our setting due to the slightly imbalanced class distribution. We compare our audiovisual fc7 and EEG features against Han features [17], which are interpretable, and employed to dynamically estimate scene emotions. As [17] inherently uses audiovisual emotional features, we only consider feature and decision fusion for Han. User-centered AR employs PCA-applied EEG features (Sec. 4.2.1).

As we perform AR on a small dataset, results obtained over ten repetitions of 5-fold cross validation (CV), *i.e.*, totally 50 runs are presented. CV is used to overcome *overfitting* on small datasets, and SVM parameters are tuned from the range $[10^{-3}, 10^{3}]$ via an inner five-fold CV on the training set. In order to examine temporal variation in AR, we present F1 obtained over (a) all ad frames ('All'), (b) last 30s (L30) and (c) last 10s (L10) for *content-centered* AR. Similarly, results are presented for (a) first 30s (F30), (b) last 30s (L30) and (c) last 10s (L10) for *user-centered* AR. These settings were chosen as the sampling rate is much higher for EEG as compared to audio/video.

Classifiers: We considered both shallow and deep classifiers for AR. Among shallow classifiers, we employed linear discriminant analysis (LDA), linear SVM (LSVM) and radial basis function SVM (RSVM). LDA and LSVM partition training data via a separating hyperplane, while RSVM transforms input data to a high-dimensional feature space where the samples are linearly separable. Audiovisual fc7 descriptors were input to shallow classifiers for content-centered analysis, while EEG features were fed to the shallow and CNN classifiers for user-centered AR.

Apart from the above *single-task learning* methods, we also explored *multi-task learning* (MTL) for AR. While learning multiple *related* tasks, MTL jointly learns a set of task-specific classifiers by modeling task relations, which is highly beneficial when learning with few data. Among the MTL methods available as part of MALSAR [49], we employed the sparse graph-regularized MTL (SR-MTL) where *a-priori* knowledge regarding task-relatedness is modeled via a graph. MTL is inherently suited for dimensional AR, as one can expect audio-visual similarities among emotionally similar ads. We model each asl-val quadrant as a *task* (*e.g.*, all H asl, H val ads will have identical labels). Also, quadrants with same asl/val labels are deemed *related*, while those with dissimilar labels are considered *unrelated*. Task relatedness is modeled via graph edge weights with weights of 1 and 0 respectively for related and unrelated tasks.

The graph then guides learning of task weights as shown in Fig. 3, where SR-MTL is fed with the specified features computed over the final 30s of all ads. Darker shades denote large MTL weights. Shot change frequency is a key predictor of asl [17], and one can notice high weights for H asl, H val ads in particular. The attributable reason is that H asl H val ads involve frequent shot changes to convey emotional intensity, while the mood of H asl, L val ads is mainly influenced by scene semantics (depicting drug and alcohol abuse, and overspeeding). Likewise, pitch amplitude is a key val predictor, and salient weights can be consistently seen over the 30s temporal window for HV ads. Finally, higher weights for H val ads with the motion activity feature are indicative of accentuated motion.

For content-centric AR, apart from unimodal (audio (A) or visual (V)) fc7 features, we also employed *feature fusion* (A+V entries in Table 3). Probabilistic *decision fusion* (DF) of the unimodal classifier outputs are denoted by 'A+V DF' entries in Table 3, and by 'A+ V + EEG DF entries' in Table 5. Audiovisual *feature fusion* (A+V) involved concatenation of fc7 A and V features over 10s windows (see Table 2), while the W_{est} technique [50] was employed for *decision fusion*. In DF, the test label is assigned the index j, $j \in \{\text{H(1) ,L(0)}\}$, corresponding to maximum class probability $P_j = \sum_{i=1}^{2} \alpha_i^* t_i p_i$, where i denotes the constituent modalities, p_i's denote posterior class probabilities and $\{\alpha_i^*\}$ are the weights maximizing test F1, 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 here that (i) the use of a validation set for parameter tuning is precluded by the small dataset size as with [1,18] and (ii) DF F1 scores denote 'maximum possible' performance.

5.1 Results Overview

Tables 3 and 4 respectively present content and user-centric AR results, while Table 5 presents the audiovisual+EEG fusion results. Highest F1 achieved for a given modality over all classifiers and temporal settings is denoted in bold. We elaborate the results as follows.

Content-centric Analysis: Focusing on unimodal descriptors in Table 3, video fc7 features predict val (peak F1 = 0.79) considerably better than asl, while audio fc7 features encode asl (peak F1 = 0.68) slightly better than val (peak F1 = 0.66). Also, MTL (peak F1 = 0.96 for val, 0.94 for asl) starkly

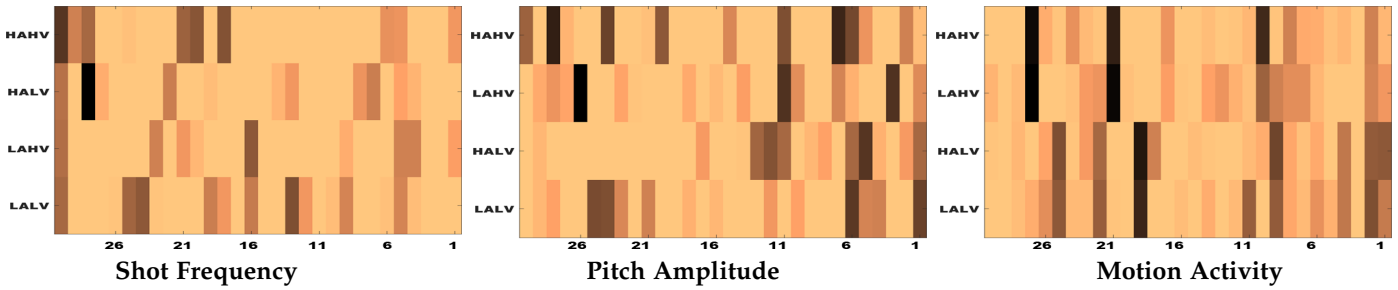


Fig. 3. Learned MTL weights for the four quadrants (tasks) when fed with the specified low-level features computed over the final 30s of 100 ads.

outperforms single task classifiers. With both single and multi-task classifiers, higher F1-scores are noted with video fc7 features, implying that the raw video frames achieve better AR than spectrograms.

We then note that multimodal methods perform comparable or better than unimodal ones. For val, the best fusion F1 (0.75 with feature fusion + RSVM) is superior to audio (F1 = 0.66), but inferior to video (F1 = 0.79). Contrastingly for asl, decision fusion (F1 = 0.75) considerably outperforms unimodal methods (0.68 with A, and 0.67 with V). Examining feature fusion approaches, A+V fc7 features outdo Han features on both single and multi-task methods. Performance difference is prominent for val (F1 = 0.75 with fc7 vs 0.65 with Han), while comparable AR is achieved for asl (F1 of 0.63 with fc7 vs 0.59 with Han).

Examining AR performance with DF, DF (F1 = 0.75) substantially outperforms feature fusion (F1 = 0.59) for asl, while underpeforming for val (F1 = 0.72 with DF and 0.75 with feature fusion). Among classifiers, RSVM produces the best F1-scores among single-task classifiers with all features. This indicates that the A+V fc7 features are not easily linearly separable as such. Nevertheless, the linear MTL model beats all single-task methods with both fc7 and Han features. A+V fc7 + MTL F1-scores are second best after video fc7 + MTL, which achieves optimal AR. Therefore, learning feature similarities among *emotionally similar* ads enables better separability of H and L asl/val data.

EEG-based AR: Here, we mainly examine whether (a) better AR is achievable with EEG vis-ã-vis audiovisual features; (b) the three-layer CNN (Section 4.2.1) better encodes emotions as compared to shallow classifiers, and (c) whether the considered CNN architecture can perform similarly with *clean* vs *noisy* EEG signals (*i.e.*, epochs involving muscle movement artifacts), as CNNs can effectively learn target encodings from disparate data.

Observing Table 4, we firstly note that EEG-based results are generally superior to content-centric ones. The best EEG-based val and asl F1's are considerably higher than the best content-based unimodal results. As with audiovisual features, EEG achieves better val recognition different from prior findings [14]–[16]. In this regard, we observe that positive val correlates with increased activity in the frontal lobes [51], and the *Emotiv* device efficiently captures frontal lobe activity despite its limited spatial resolution.

Among the three shallow classifiers, RSVM again performs best with EEG. However, the three-layer CNN achieves optimal AR, considerably outperforming other classifiers. Also, while very comparable results are achieved with the *raw* and *clean* EEG data with shallow classifiers, larger differences are noted with the CNN and MTL methods. Therefore, while all the methods are able to work with noisy EEG data, the CNN and MTL methods can better utilize the cleaned EEG data. As with audiovisual descriptors, highest F1-scores with EEG (close-to-ceiling performance for both val and asl) are also obtained with the MTL classifier, reinforcing its utility for emotion recognition.

General Observations: Relatively small σ values, denoting minimal variance in cross-validated AR performance, are observed in the 'All' condition with both audiovisual and EEG CNN features in Tables 3 and 4. These trends reveal that the classification models are minimally impacted by overfitting. Examining temporal windows for audiovisual AR, significantly higher σ's are noted with Han features as well as with the L30 and L10 temporal segments, conveying that those models do not generalize well. Higher σ's are observed for the L30 and L10 conditions reveal the greater variance in AR performance on terminal ad frames. Contrastingly, very similar σ's are noted for the different temporal windows considered with EEG data in Table 4.

Interestingly in Table 3, one can note a considerable dip in asl recognition for the L30 and L10 conditions with A and V features, while val F1-scores are more consistent with the 'All' condition. Also, a sharp fall in MTL performance is noted for L30 and L10. The above trends reveal that (1) Greater heterogeneity in ad content towards endings is highlighted by the large variance in the L30 and L10 conditions with uni/multimodal features; conversely, consistent AR performance is noted with EEG features for the different temporal segments. Therefore, while audiovisual content conveying ad emotion may significantly vary over time, human viewers typically grasp the conveyed emotion rather instantaneously; (2) Fusion models synthesized with Han features are most prone to overfitting, given the larger σ values seen with respect to other models. (3) Lower asl recognition 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 in Table 3, implying that asl is a more transient phenomenon than val (this also explains the lower agreement for asl in Section 3.1), while coherency between val features and labels sustains over time.

Fusion of Content and User-Centric Modalities: Given the difference in AR performance for and user-based features (especially the variance across temporal segments), one could argue that the audiovisual and EEG modalities encode complementary information. Therefore, we examined if fusing the content (A+V fc7) and EEG outputs resulted in better asl/val recognition. Corresponding results are tabulated in Table 5.

Comparing Table 5 against Tables 3 and 4 clearly reveals

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

Method	Valence			Arousal		
	F1 (all)	F1 (L30)	F1 (L10)	F1 (all)	F1 (L30)	F1 (L10)
Audio FC7 + LDA	0.61±0.04	0.62±0.10	0.55±0.18	0.65±0.04	0.59±0.10	0.53±0.19
Audio FC7 + LSVM	0.60±0.04	0.60±0.09	0.55±0.19	0.63±0.04	0.57±0.09	0.50±0.18
Audio FC7 + RSVM	0.64±0.04	**0.66±0.08**	0.62±0.17	**0.68±0.04**	0.60±0.10	0.53±0.19
Video FC7 + LDA	0.69±0.02	0.79±0.08	0.77±0.13	0.63±0.03	0.58±0.10	0.57±0.18
Video FC7 + LSVM	0.69±0.02	0.74±0.08	0.70±0.15	0.62±0.02	0.57±0.09	0.52±0.17
Video FC7 + RSVM	0.72±0.02	**0.79±0.07**	0.74±0.15	**0.67±0.02**	0.62±0.10	0.58±0.19
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

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

Method	Valence			Arousal		
	F1 (F30)	F1 (L30)	F1 (L10)	F1 (F30)	F1 (L30)	F1 (L10)
Raw 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
Raw 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
Raw 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
Clean EEG + LDA	0.79 ± 0.03	0.79 ± 0.03	0.77 ± 0.03	0.76 ± 0.03	0.75 ± 0.03	0.71 ± 0.04
Clean 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
Clean EEG + RSVM	**0.82 ± 0.03**	**0.82 ± 0.03**	0.81 ± 0.03	**0.78 ± 0.02**	0.77 ± 0.03	0.75 ± 0.04
Raw 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
Clean 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
Raw 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
Clean 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

TABLE 5
Probabilistic fusion of audiovisual & EEG classifier outputs. 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)
(Raw EEG + RSVM) + (A+V fc7 RSVM) 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
(Raw EEG + CNN) + (A+V fc7 RSVM) 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 + RSVM) + (A+V fc7 RSVM) 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) + (A+V fc7 RSVM) 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

that fusing information from the two sources is beneficial. Fusion-based asl and val F1-scores are consistently better than individual counterparts, and considerably superior when shallow classifiers are employed for prediction (rows 1 and 3). These findings reveal the potential for fusion of *content-centric* and *user-centric* cues, as in [52]–[54].

6 Computational Advertising - User Study

Obtained AR results reveal that the audiovisual fc7 and EEG descriptors outdo Han features. This section demonstrates that improved AR positively impacts computational advertising– specifically, *better* ad AR facilitates *optimized* insertion of ads onto streamed (*e.g.*, YouTube) video. We hypothesize that optimized ad insertion will result in: (1) maximal ad recall, and (2) the best viewing experience.

The question that we seek to answer here is *Whether better affect estimation, achieved by the CNNs harnessing audiovisual and EEG descriptors, leads to optimal insertion of ads at appropriate scene transition points in a video sequence?* A principled methodology to insert ads in video is proposed by the CAVVA algorithm [4]. CAVVA is a genetic algorithm-based optimization for inserting ads onto streamed video. On top of low-level context matching by advertising frameworks such as VideoSense [32], CAVVA models affective relevance between video scenes and inventory ads to determine the (a) *suitable* ads to insert, and (b) the *best* temporal positions where the chosen ads should be inserted.

Based on insights from consumer psychology, CAVVA proposes ad insertion rules that seek to strike a balance between (a) maximizing brand memorability (ad recall), and (b) minimally disrupting (or enhancing) viewer engagement and experience. To examine the above research question, we performed a study with CAVVA and 18 users to compare ad recall and subjective quality of advertising schedules generated with affective scores estimated via (a) the audiovisual CNN model, (b) the EEG CNN model and (c) ratings

TABLE 6
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

provided by experts. Details on the (i) ad and video data employed, (ii) ad insertion strategies and (iii) user study and associated results are as follows.

6.1 Ad and Video Datasets

For the user study, we used 28 ads (out of the original 100), and three program videos. The 28 ads were equally distributed among the four val-asl quadrants based on the *expert labels*. Program videos were scenes from a television sitcom (*Friends*) and two movies (*In pursuit of Happiness* (ipoh) and *Children of Heaven* (coh)), which predominantly depicted social themes and situations invoking high-to-low valence and moderate arousal (see Table 6 for statistics). Each program video comprised eight scenes implying that there were seven possible ad-insertion or scene transition points. The average scene length in the considered program videos was 118 seconds.

6.2 Advertisement insertion strategy

We used three affect estimation models (audiovisual CNN, EEG CNN and manual) to provide asl, val scores for the ads. Asl, val scores for the 24 program video scenes (8 scenes $\times$ 3 videos in Table 6) were computed as mean of the ratings (between [-2,2] for val and [0,4] for asl) acquired from three experts, and rescaled to [0,1] via min-max normalization. Ad affect scores were computed as follows. For the **content-centric** method, we used normalized softmax class probabilities output by the video CNN model [55] for val estimation, and similarly probabilities from the audio CNN for asl estimation. The mean score computed over all video/audio frames was used to denote an ad's affective score. Similarly, mean of the normalized softmax probabilities over all EEG epochs for an ad was used to denote the user-centric **EEG** asl, val scores. Average of continuous val and asl ratings annotated by five experts via *FeelTrace* [56] was used to denote ad affect scores in the **Manual** method.

We then adopted the CAVVA optimization framework [4] to obtain nine unique **video program sequences** (VPSs with an average length of 19.6 minutes) comprising the inserted ads. These VPSs constitute the different combinations of the three program videos and the employed affect estimation approach (audiovisual/EEG/manual). Exactly five (out of seven possible) ads were inserted 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 rates of 5 and 2.14 respectively. Among the 21 inserted ads, 13 had been labeled as *high val* by experts, while 10 were labeled as *high asl*.

6.3 Experiment and Questionnaire Design

To evaluate the subjective quality of the generated VPSs and thereby the efficacy of the affect estimation techniques for computational advertising, we recruited 18 users (7 female, mean age 20.1 years) who were university students. Each user viewed three VPSs in random order such that *each VPS was generated via a unique affect estimation approach*. We used a randomized 3×3 Latin square design to cover all the nine VPSs with every three users. Thus, each VPS was seen by six of 18 users, and we have a total of 54 unique user responses (18 users $\times$ three video modes per user).

We designed the user evaluation so as to reveal whether the generated VPSs (a) included seamless ad insertions, (b) facilitated user engagement (or alternatively, resulted in minimal disruption) towards the VPS content and (c) ensured a pleasant overall viewing experience and maximized brand memorability. To this end, we evaluated whether a particular ad insertion strategy resulted in (i) increased brand recall (both *immediate* and *long-term* recall) and (ii) minimally disruptive (or improved) viewing experience.

Recall evaluation is intended to verify if the inserted ads were attended to and remembered by viewers, and the immediate and day-after recall were ***objective*** measures quantifying the impact of ad insertion on short-term (immediate) and long-term (day-after) memorability of the VPS-embedded ads. Specifically, we measured the proportion of (i) inserted ads that were correctly recalled (*Correct* recall or *hit rate*), (ii) inserted ads that were not recalled (*Forgotten* or *miss rate* $= 1 -$ *hit rate*) and (iii) non-inserted ads incorrectly recalled as seen (*Incorrect* recall or *false alarm*). For inserted ads which were correctly recalled, we also assessed whether viewers perceived them to be contextually (or emotionally) relevant to the program content (*i.e.*, whether the ad insertions were perceived to be *appropriate*).

Upon viewing a VPS, each user was provided with a representative visual frame from each of the 28 ads and a sequence-specific response sheet to test ad recall and impression concerning ad insertion quality. All recall and insertion quality-related responses were acquired 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 0–4 Likert scale for the questions below with 4 implying *best* and 0 denoting *worst*:

1. Were the advertisements uniformly distributed across the video program?
2. Did the inserted advertisements blend well with the program flow?
3. Whether the inserted ads matched with the surrounding scenes with respect to *content* and *mood*?
4. What was the overall VPS viewing experience?

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

6.4 User study results

As mentioned previously, program video scenes were assigned asl, val scores manually by three experts, while the content-centric CNN (denoted as 'Content' hereon), EEG and Manual methods were employed to estimate affective scores for ads. The overall quality of the CAVVA-generated VPS is influenced by the affective ratings assigned to both

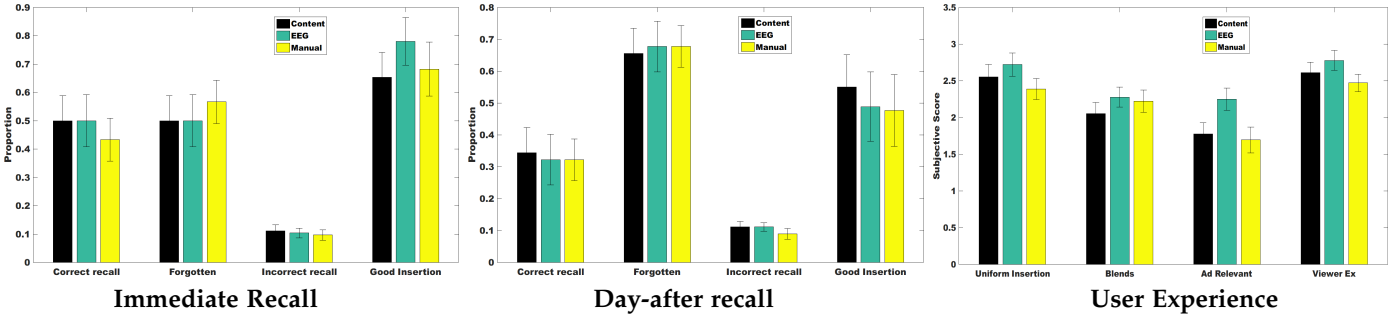


Fig. 4. Summary of user study results in terms of recall and user experience-related measures. Error bars denote standard error of mean.

the *video scenes* and *ads*. In this regard, we hypothesized that better ad affect estimation would result in *optimized* ad insertions maximizing brand recall and viewing experience.

Firstly, we examined if there were any similarities in the ad asl and val scores estimated by the Content, EEG and Manual approaches in terms of Pearson correlations. We found that (1) there was significant and positive correlation between asl scores generated by the Manual and EEG approaches ($\rho = 0.55, p < 0.005$), while asl scores computed via the Manual and Content methods ($\rho = 0.13$, n.s.) as well as via Content and EEG ($\rho = -0.22$, n.s.) were largely uncorrelated. A similar pattern was noted for val scores with a highly positive and significant correlation observed between Manual and EEG ($\rho = 0.80, p < 0.000001$), while the Content–Manual ($\rho = 0.33, p = 0.08$) and the Content–EEG ($\rho = 0.19$, n.s.) scores showed an insignifcant positive correlation. These results are indicative of the fact that neural responses, which represent an *implicit* manifestation of emotional perception/expression, best reflect *explicit* affective impressions reported by humans. It is therefore unsurprising that a large number of recent affect prediction approaches [14]–[16], [57] have employed neural sensing as one of the modalities incorporating emotional information.

Based on the user responses, we computed the mean proportions for correct recall, ad forgottenness, incorrect recall and good insertions *immediately* and a *day after* the experiment. Similarly, subjective experience scores were also collated for the three VPS generation schemes. Fig. 4 summarizes the response results.

A key measure indicative of a successful advertising strategy is *high brand recall* [1], [4], [26], and the immediate and day-after recall rates observed with the three ad affect estimation methods are presented in Fig. 4 (left),(middle). A surprising result observed from Fig. 4 (left) and (middle) is that ads from the content and EEG-based VPSs are better recalled (or less forgotten) than manual-based. Content-based ad insertions were best recalled both immediately and the day-after, even though recall rates for the three ad-insertion approaches were not statistically different. Given the extensive literature connecting affective attributes and memorability, we examined if any such relationships could be inferred from the user study. Overall, we found a significant and positive correlation between ad *val rating* and *recall* ($\rho = 0.44, p < 0.05$) consistent with prior findings [9], in addition to about $\frac{2}{3}$rds of the VPS ads having positive val.

Ad recall rate was much worse for the *day-after* condition with a high proportion of ads being forgotten. Also, the proportion of *incorrectly recalled* ads was minimal in both the immediate and day-after conditions. Some discernible differences were observed in the proportion of *good insertion* impressions for the three methods– we remark here that ad recall and viewing experience are not necessarily positively correlated (some ads may be memorable simply because they disrupt viewing experience); however, embedding ads at optimal temporal locations can enhance both ad recall and viewing experience. Post-hoc independent and right-tailed t-tests revealed that the proportion of immediate 'good insertion' impressions was maginally higher for EEG as compared to manual ($t_{34} = 1.337, p = 0.095$).

A number of significant differences were nevertheless observed with respect to subjective user impressions of the VPSs generated via the three ad insertion methods (4 (right)). The EEG-based mechanism scored highest for all the considered criteria. Specifically, *uniform insertion* scores were marginally higher for EEG with respect to manual ($t_{34} = 1.5646, p = 0.063$). A one-way balanced ANOVA on *ad relevance* scores revealed the significant effect of the ad-insertion strategy ($p < 0.05$). Post-hoc t-tests further revealed that EEG-based ad relevance was significantly higher than manual ($t_{34} = 2.3785, p < 0.05$) or content-based ($t_{34} = 2.1893, p < 0.05$). EEG-based VPSs were also found to have the highest *viewing experience* scores, which were significantly higher than manual-based VPSs ($t_{34} = 1.7033, p < 0.05$). No differences were noted regarding how the inserted ads *blended* with the program flow.

7 Discussion

Even though ads routinely employ emotions to attract viewer attention and convey an effective message, very few works predict ad emotions. Our work is the first elaborate attempt to (a) perform AR on video ads, and (b) demonstrate via a user study that improved ad AR enables optimized ad insertion onto streamed video, as measured in terms of *brand recall* and *viewing experience*.

We firstly show that a curative set of 100 diverse ads can coherently evoke emotions across disparate users. The *expert* and *novice annotator* group concurred substantially on the asl and val ratings for the chosen ads, as noted from Cohen's κ scores quantifying inter-rater agreement. A scatter plot of the mean annotator asl, val ratings resembles a close-to-uniform distribution as envisioned by the experts. The asl-val ratings were also uncorrelated, conveying the relative independence of these affective dimensions.

We then evaluate the efficacy of *content* and *user-centric* techniques for ad AR. At the outset, it needs to be stressed that content and user-centered cues encode complementary

emotional information. While content-centric methods examine audiovisual content for extracting emotion descriptors, they are typically limited in their ability to encode *context* which is crucial for conveying emotions. Context may induce in the viewer an emotion very different from expectation based on content, and therefore we hypothesized that examining user cues could be more effective as evidenced by many of the recent AR approaches.

While CNNs have been previously used for video and audio-based AR [21], [58] on short snippets, we attempt CNN-based AR for full-length ads, some of which are over a minute long. Our content-centered AR experiments confirm that: (1) The proposed fc7 audio and visual CNN descriptors better predict val, and F1-scores reveal that video features are better at encoding emotions than spectrograms; (2) Multimodal methods generally achieve comparable or better AR than unimodal ones, and the fused (A+V) fc7 features produce substantially better results than audiovisual Han features for val; Probabilistic decision fusion achieves superior results with respect to feature fusion for asl, but inferior results for val. Conversely, EEG-based AR experiments reveal that (1) EEG features achieve substantially better AR than audiovisual descriptors; (2) The three-layer CNN classifier outperforms shallow classifiers trained on EEG data, and (3) Very comparable F1-scores are achieved with CNNs on both raw (or noisy) and clean EEG data, even though EEG data cleaning benefits shallow classifiers.

Some empirical trends are however unclear, notably the performance of multimodal vs. unimodal methods. Observing Tables 3 and 5, multimodal cues appear to elicit better AR performance than unimodal ones in some cases– *e.g.*, (A+V) DF in Table 3 performs better than audio fc7 or video fc7 for asl classification. Likewise in Table 5, some combinations of content+user cues outperform individual counterparts. However, feature/classifier fusion routinely produces results inferior to the unimodal cases. These results motivate the need for better fusion strategies.

AR experiments nevertheless unambigously confirm that ad emotions are better conveyed by user cues, which are inherently modulated by context [9] than content cues. Content-centric AR results over multiple temporal windows reveal that ad contents coherently reflect val labels over time, but not asl labels. There are two possible explanations: (a) Many studies have noted that user impressions of stimulus val are more stable over time than asl; also audiovisual ads are routinely designed to convey surprise/shock, and are hence likely to exhibit significant changes over time. (b) Given these content changes, the use of a *single* asl label over the *entire* ad duration may be inappropriate, and seeking to dynamically estimate asl levels could be more apt. Conversely, EEG-based AR results (Table 4) show only a minor deviation between the F30 and L30 conditions even for asl (lower F1s for the L10 condition can be attributed to fewer training data) suggesting that users grasp the general mood of ads fairly quickly and consistently.

General remarks regarding experiments include the following. Cumulatively, our AR results convey that the minimal impact of model overfitting– small variations in F1 are noted across the 50 runs in the 'All' condition for content-centric and over all conditions for user-centered AR. Among classifiers, RBF SVM consistently produces the best results among single-task classifiers, implying that both audiovisual and EEG features may not be trivially linearly separable in their respective feature spaces. However, the fact that the linear multi-task learning classifier achieves close-to-ceiling performance suggests that exploiting commonalities among emotionally similar ads greatly benefits AR.

We then proceeded to check if improved emotion estimation enabled optimized ad insertion for a computational advertising application. Based on data compiled from 18 users, we observe that video program sequences generated via audiovisual and EEG-based affective scores are more effective in terms of *ad recall* and eliciting a better *viewing experience* than manually generated VPSs. Specifically, ads from content-based VPSs are recalled marginally better, both in the immediate and day-after conditions. EEG-based VPSs nevertheless received the highest scores for different attributes relating to viewing experience. Ads in EEG-based VPSs are perceived to be (a) more uniformly distributed, and (b) more emotionally matched (relevant) with the preceding and succeeding video scenes. Finally, EEG-based VPSs are also found to produce the best viewing experience.

The surprising finding that the audiovisual and EEG-based VPSs are superior to manually generated VPSs can be explained as follows. Despite correlation analyses revealing that the general trends of the EEG and manual-based affective score estimates are similar, user study results reveal that the EEG and content-based CNNs generated *more accurate* scores, which enabled better video scene–ad matching. Audiovisual and EEG-based asl and val scores are estimated via CNN models, and deep CNNs have recently outperformed humans in tasks such as object recognition [59] and facial expression recognition [60] due to their ability to capture fine details from data. The CAVVA framework [4] comprises two components– one for selecting *ad insertion points*, and another for selecting the *ads* themselves. Asl scores only play a role in the choice of insertion points, whereas val scores influence both selections. As the EEG-based CNN performs best for both asl and val recognition, it also enables the most optimal ad insertions, and consequently the best viewing experience. Also, one should note that humans are better at rating attributes in relative than absolute terms [61], [62]; this possibly explains why the ad-level asl and val scores obtained from per-frame manual ratings may not be accurate.

This study has multiple limitations in terms of the *algorithms* and *hardware* employed for ad AR. In terms of algorithms, given the limited available data we only explore CNNs which do not encode temporal data dependencies. Given that emotion perception and expression evolves over time, modeling time dependencies for content and user-centric AR should be beneficial. In this respect, recurrent neural networks (RNNs) have shown much promise for emotion [63] and mental state [64] recognition. Likewise, due to the paucity of large-scale, labeled training datasets in the AR domain, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) provide an avenue to generate synthetic labeled data. Also, power spectral density analysis could be employed for EEG-based AR.

In terms of hardware, *Emotiv* headsets, which are wearable and enable naturalistic user responses are used for recording EEG data. While the EEG signal quality of gaming

headsets is inferior with respect to lab-grade EEG equipment, recent AR studies [16], [57] have resorted to wearable sensors since they enable user data capture at scale. Facial emotions [29] have also been successfully employed for large-scale AR. With ever-falling sensor costs, one envisages a combination of facial emotion capture, neural sensing and eye-tracking to be employed in *crowdsourced* studies.

8 Conclusion

The presented study (a) examines AR on a curated set of 100 ads, and (b) demonstrates that improved ad AR enables better ad insertion onto streamed video via a user study. As part of future work, multi-task feature selection to determine the best emotion predictors [65] and deep multi-task learning [66] will be explored for optimized ad AR. In addition, we will also work on efficient fusion-based approaches and addressing aforementioned limitations.

Acknowledgment

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

[1] M. B. Holbrook and J. O. Shaughnessy, "The role of emotlon in advertising," *Psychology & Marketing*, vol. 1, no. 2, pp. 45–64, 1984.
[2] M. B. Holbrook and R. Batra, "Assessing the Role of Emotions as Mediators of Consumer Responses to Advertising," *Journal of Consumer Research*, vol. 14, no. 3, pp. 404–420, 1987.
[3] 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," *Int'l Journal of Research in Marketing*, vol. 30, no. 4, pp. 383 – 394, 2013.
[4] K. Yadati, H. Katti, and M. Kankanhalli, "CAVVA: Computational affective video-in-video advertising," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 15–23, 2014.
[5] K. Yadati, "Online Multimedia Advertising," Master's thesis, Natonal University of Singapore, Singapore, 2013.
[6] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, pp. 1161–1178, 1980.
[7] 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*, vol. 24, no. 4, pp. 45–54, 1995.
[8] A. Khosla, W. A. Baingridge, A. Torralba, and A. Oliva, "Modifying the memorability of face photographs," *ICCV*, 2013.
[9] 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*, vol. 14, no. 3, pp. 1–18, 2014.
[10] 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 Multimedia*, 2010, pp. 667–670.
[11] M. Bilalpur, S. M. Kia, T.-S. Chua, and R. Subramanian, "Discovering gender differences in facial emotion recognition via implicit behavioral cues," in *ACII*, 2017.
[12] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.
[13] T. AlHanai and M. Ghassemi, "Predicting latent narrative mood using audio and physiologic data," in *AAAI*, 2017.
[14] 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*, vol. 3, no. 1, pp. 18–31, 2012.
[15] 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*, vol. 6, no. 3, pp. 209–222, 2015.
[16] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affective Computing*, 2016.
[17] A. Hanjalic and L.-Q. Xu, "Affective Video Content Representation," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.
[18] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits, Systems & Video Technology*, vol. 16, no. 6, pp. 689–704, 2006.
[19] 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*, vol. 3, pp. 51–64, 1989.
[20] 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*, vol. 51, no. 2, pp. 505–523, 2011.
[21] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.
[22] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian, "Evaluating content-centric vs. user-centric ad affect recognition," in *ICMI*, 2017, pp. 402–410.
[23] ——, "Affect recognition in ads with application to computational advertising," in *ACM Multimedia*, 2017.
[24] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.
[25] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler, "A probabilistic approach to people-centric photo selection and sequencing," *IEEE Trans. Multimedia*, 2017.
[26] M. K. Karthik Yadati and, Harish Katti and, "Interactive video advertising: A multimodal affective approach," *Multimedia Modeling*, 2013.
[27] 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*, vol. 10, no. 9, pp. 1–19, 2015.
[28] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," *ICME*, 2014.
[29] D. McDuff and M. Soleymani, "Large-scale affective content analysis: Combining media content features and facial reactions," in *Automatic Face Gesture Recognition*, 2017, pp. 339–345.
[30] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.
[31] J. Ye, J. Li, M. G. Newman, R. B. A. Jr., and J. Z. Wang, "Probabilistic multigraph modeling for improving the quality of crowdsourced affective data," *IEEE Trans. Affective Computing*, vol. 1, no. 1, 2017.
[32] T. Mei, X.-S. Hua, L. Yang, and S. Li, "Videosense: Towards effective online video advertising," in *ACM Multimedia*, 2007, pp. 1075–1084.
[33] L. F. Barrett and J. A. Russell, "The structure of current affect: Controversies and emerging consensus," *Current Directions in Psychological Science*, vol. 8, no. 1, pp. 10–14, 1999.
[34] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, Tech. Rep. A-8, 2008.
[35] 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)*, vol. 57, no. 1, pp. 289–300, 1995.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[37] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *ACM Multimedia*, 2014, pp. 801–804.
[38] Y. Baveye, "Automatic prediction of emotions induced by movies," Theses, Ecole Centrale de Lyon, Nov. 2015.
[39] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "CAFFE: Convolutional architec-

ture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.
[40] 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*, vol. 2014, 2014.
[41] 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*. Springer, 2016, pp. 127–150.
[42] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," in *ACM Multimedia*. ACM, 2017, pp. 1809–1817.
[43] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *CVPR*. IEEE, 2017, pp. 6809–6817.
[44] S. Stober, D. J. Cameron, and J. A. Grahn, "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," in *NIPS*, 2014, pp. 1449–1457.
[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] S. Stober, "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *ICASSP*. IEEE, 2017, pp. 6175–6179.
[47] 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*, vol. 144, pp. 180–191, 2018.
[48] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.
[49] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-tAsk Learning via StructurAl Regularization*, Arizona State University, 2011. [Online]. Available: http://www.public.asu.edu/~jye02/Software/MALSAR
[50] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.
[51] D. Oude Bos, "Eeg-based emotion recognition - the influence of visual and auditory stimuli," in *Capita Selecta (MSc course)*. University of Twente, 2006.
[52] R. Subramanian, H. Katti, K. Ramakrishnan, M. Kankanhalli, T.-S. Chua, and N. Sebe, "An eye fixation database for saliency detection in images," in *ECCV*, 2010.
[53] H. Katti, M. V. Peelen, and S. P. Arun, "Object detection can be improved using human-derived contextual expectations," *CoRR*, vol. abs/1611.07218, 2016.
[54] H. Katti, A. K. Rajagopal, K. Ramakrishnan, M. Kankanhalli, and T.-S. Chua, "Online estimation of evolving human visual interest," *ACM Trans. Multimedia*, vol. 11, no. 1, 2013.
[55] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2013, vol. 53, no. 9.
[56] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Workshop on Speech and Emotion*, 01 2000.
[57] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affective Computing*, 2018.
[58] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *ICMI*, 2016, pp. 445–450.
[59] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *CVPR*, 2018, pp. 8697–8710.
[60] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv:1509.05371*, 2015.
[61] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *IEEE ISM*, 2008, pp. 228–235.
[62] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *EmoSPACE Workshop*, 2013, pp. 1–8.
[63] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. F. Gelbukh, and E. Cambria, "Dialoguernn: An attentive RNN for emotion detection in conversations," *CoRR*, vol. abs/1811.00405, 2018.
[64] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *ICLR*, 2016.
[65] D. Hernández-Lobato, J. M. Hernández-Lobato, and Z. Ghahramani, "A probabilistic model for dirty multi-task feature selection," in *ICML*, 2015, pp. 1073–1082.
[66] Y. Yang and T. M. Hospedales, "Trace norm regularised deep multi-task learning," *ArXiv*, 2016.

Abhinav Shukla is a PhD student in the iBUG group at Imperial College London. He completed his Masters in Computer Science at IIIT Hyderabad, India, from where he also received a Bachelors degree in Computer Science and Engineering in 2017. His research interests include self supervised and multimodal representation learning with applications in affective computing and audiovisual speech.

Shruti Shriya Gullapuram is an Applied Scientist at Microsoft, Redmond, WA. She received her Masters degree in Computer Science at the University of Massachusetts, Amherst, USA in 2019. She received her Bachelors degree in Electronics and Communication Engineering from IIIT Hyderabad, India in 2017. Her research interests broadly lie in the fields of Machine Learning, Computer Vision and Natural Language Processing.

Harish Katti received his PhD in computer science from the National University of Singapore, in 2012, a Masters degree in Bio-Medical Engineering from the Indian Institute of Technology, Bombay in 2006, and a B.Engg degree from Karnatak University, in 2000. His research interests lie at the intersection of cognition and media, and specifically in experimental and computational vision research. He is currently a post-doctoral fellow at the Center for Neuroscience, Indian Institute of Science, Bangalore

Mohan Kankanhalli is a Professor and Dean of the School of Computing at the National University of Singapore (NUS). Mohan obtained his BTech from IIT Kharagpur and MS and PhD from Rensselaer Polytechnic. His current research interests are in Multimedia Systems (content processing, retrieval) and Multimedia Security (surveillance and privacy). Mohan is actively involved in organizing major conferences, and serves on the editorial board of several journals.

Stefan Winkler is Deputy Director at AI Singapore and Associate Professor at the National University of Singapore. Prior to that, he was Distinguished Scientist and Program Director at the University of Illinois' Advanced Digital Sciences Center (ADSC). He has a Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, and a Dipl.-Ing. (M.Eng./B.Eng.) degree from the University of Technology Vienna, Austria. He is an IEEE Fellow and has published over 150 papers.

Ramanathan Subramanian received his Ph.D. in Electrical and Computer Engg. from NUS in 2008. He is an Associate Professor in Computer Science and Engg. at IIT Ropar. His past affiliations include IHPC (Singapore), U Glasgow (Singapore), IIIT Hyderabad (India) and UIUC-ADSC (Singapore). His research focuses on Human-centered computing, and especially on extracting and modeling non-verbal behavioral cues for interactive analytics. He is an IEEE Senior Member and a member of the ACM and AAAC.