Predicting Ad Appeal and Buying Interest: Large Scale Experiment of Facial Responses to Advertisements

By assistant prof. Heta Desai

Abstract—Billion online video ad views occur every month. We did a comprehensive study of measuring facial responses to video material gathered through the Internet and how this corresponds to advertisement efficacy. We recorded more than 12,000 facial reactions from 1,223 test subjects for 170 commercials on different industries and product categories. Reactions were computer-scored frame by frame, amounting to 3.7 million frames—a ratio unthinkable for traditional research techniques. Our results show that single-shot expressions are rare, but group responses have rich emotional patterns over time. By investigating the interaction between face response and ad performance, we show that ad liking can be reliably predicted (ROC AUC = 0.85) from webcam-recorded face data. In addition, changes in purchase intention can be predicted (ROC AUC = 0.78). Ad liking is motivated by eliciting expressions, particularly positive ones. But inducing buy intent is harder than getting crowds to laugh: high- performing ads generate strong positive responses with immediate exposure of a brand. These results identify a strong and scalable approach for predicting ad effectiveness from automatic facial reactions, in isolation of viewer self-reported feedback. They offer informative clues regarding the structure of effective ads as well.

Index Terms—Facial expressions, emotion, market research

1 Introduction

Non-verbal signals, particularly facial expressions, have the ability to communicate dense and subtle information regarding one's emotional state. The face in particular is particularly known for its ability to communicate valence—whether something is good or bad. For instance, heightened activity in the zygomatic major muscle (AU12, which is responsible for smiling) is usually observed when one is exposed to positive content, whereas heightened activity in the corrugator muscle (AU4, which is responsible for furrowing the brow) is observed when one is exposed to negative content [1].Facial Action Coding System (FACS) [2] is an extensive system that recognizes 44 different action units (AUs) on the basis of the face's 27 muscles' movements. This system makes it possible to measure facial expressions systematically, objectively, and quantitatively. These AUs can blend together in numerous ways to create an unimaginably large number of meaningful expressions. However, manual FACS coding is technically demanding and requires hours—coding one minute of video takes five or six hours. Fortunately, computer vision has made significant advances recently to enable automatic detection of most of these facial expressions with high accuracy [3]. More recent studies also indicate that such systems can perform well in real, everyday environments—not merely under the controlled conditions of a laboratory [4], [5].

Watching videos on the internet is growing extremely fast. In the United States alone, more than 189 million individuals viewed online video in November 2013, and each viewed approximately 19 hours of them on average. The video advertising market is also seeing enormous growth, with billions of dollars spent every year. During the same month, close to 27 billion video ads were watched, and over half of the population of the United States was targeted—close to three times as many as ad views in November 2012. Video ads accounted for 36.2% of total online videos watched. Hulu and Netflix are among the streaming providers that routinely invite viewers to assess the usefulness or enjoyment of ads. In the meantime, video-sharing is becoming more common; 72% of adult Internet users employed video- sharing websites in 2013. Brands increasingly employ websites such as YouTube to disseminate their promotions, prompting users to socially share content. The Internet allows advertisers not only to reach more people but to target them with more accuracy. Studies have demonstrated that advertisements such as these are good for both consumers and advertisers since they maximize relevance and increase profits [6].

But it has been challenging to see the connection between affective reaction to content and ad impact via methods of the past, such as surveys. These tend to be time-consuming, laborious, and fail to recognize the dynamic, real-time affective response during an ad. It also may be impractical to seek viewer reaction where viewers are otherwise engaged—e.g., surfing the web or watching television—so automated prediction solutions are a useful alternative.

Figure 1 This study presents a large-scale investigation of facial reactions to web video commercials. Top Sample frames from the web video ads used in testing. Bottom Frames that were captured from webcam recordings of participants' facial responses. All photos

The recent advances in Internet connectivity and the ubiquity of webcams have facilitated the mass, opt-in recording of facial responses on a wide variety of online media [4]. It is very effective in a way that it can record facial responses from a large, geographically dispersed group of individuals. It avoids many of the weaknesses of conventional market research by enabling volunteers to see material in their own natural settings, not a lab, without attaching physical electrodes, and gathering huge

quantities of information at low cost—less than $10 per volunteer for seeing 10 ads and filling out a 30-minute questionnaire. Figure 1 shows sample ad frames together with corresponding facial responses. Though face reactions in uncontrolled natural environments are challenging to record due to the problematic light, camera placement, and social effects, these can be overcome through careful experimental design.

Several cycles of web-based facial video collection [4], [7] have optimized the procedure adopted here. One of the central measures of advert success is likability [8], [9], and early analysis of over 3,000 facial response videos by our initial analysis suggests that facial responses can be applied to predict advert liking [10].

However, findings in McDuff et al. [10] were limited to three adverts and therefore did not address question of applicability on a broad scale. More research has related facial expressions with advert recall [11] and likelihood of users skipping adverts (\"zapping\") [12]. One of the most important metrics for ads is purchase intent (PI), or whether a consumer intends to purchase from the advertised firm. Teixeira et al. [13] investigated facial signals and PI.

In this paper, facial responses to ads presented over the last 12 years are examined, and what we introduce is what we believe is the first model to automatically determine the ad's potential to express purchase intent from facial responses. Similar traditional laboratory-based approaches— involving thousands of people all over the world and hand-annotating expressions in 3.7 million frames— would have been unaffordable on an economic as well as logistic level. Figure 2 indicates our process of facial response collecting, machine-driven expression measurement analysis, and comparing them to ad performance. An ad's likability prediction and resulting purchase intention has strong utility value for ad content pre-testing and video target optimization on Internet TV and web-sharing sites. This research's contributions are: 1) constructing the largest-ever facial response corpus to ads, 2) investigating facial response correlation with ad likability and purchase intent, and 3) identifying aggregate emotional response features that increase an ad's impact.

2 Related Work

2.1 Facial Expression Recognition

Over the last 15 years, the automatic analysis of facial expressions has garnered significant interest from the fields of computer vision, psychology, and affective computing. Early efforts predominantly concentrated on staged and rehearsed facial behaviors. More recently, however, the focus has shifted toward capturing naturalistic and spontaneous expressions [4], [14], [15], as well as subtle facial cues [16].

Figure 2 provides an overview of the framework employed in this study. 1) Spontaneous and natural facial reactions to video advertisements are gathered through software integrated into an online survey. 2) Leveraging the Internet's connectivity and the widespread.

Most facial expression recognition systems are similarly organized. Facial registration is first performed to normalize the face and reduce the impact of its position and pose. Some recent face registration techniques are active appearance models (AAM) [17] and constrained local models (CLM) [18].
Subsequently, shape and/or appearance features are derived from a pre-defined region of interest (ROI) and fed into a computational model that maps the features to expression or action unit labels. Common features are histograms of oriented gradients (HOG) and local binary patterns (LBP), with support vector machines (SVM) being the most commonly used model type. For an overview of facial expression recognition methods, refer to [3]

Smile detection is among the most widespread and stable applications of facial expression recognition. Whitehill et al. [15] developed a state-of-the-art smile detector that was learned from images collected from the internet, offering an effective way of collecting training data for the classifier. Our own work has also shown effective smile detection in uncontrolled online environments [10]. Facial behavior can be distinguished according to the Facial Action Coding System (FACS), discrete labels (i.e., the six essential emotional states), or continuous measurements of emotion from facial behavior, including valence and arousal/activation. Some alternative newer means of dimensionally quantifying emotion from facial behavior have been proposed, including valence [19]. We utilize here in this work custom classifiers that we use to detect single action units (e.g., AU02) and discrete emotional labels (e.g., disgust).

2.2 Media and Emotion

Kassam's [20] research on facial expression summarizes that facial expressions and self-reported reactions both share significant variability, with expression analysis giving new information regarding emotional experience that is different from that given by measures of self-report. It has been shown to predict the emotional content of media viewing [21], and make inferences about media preference based on automatically derived facial reactions [10], [22]. Joho et al. [23] showed that personal moments of viewing pleasure can be detected through automated facial behavior analysis. Zhao et al. [24] proposed a video recommendation and indexing scheme based on automatically extracted expressions of six universal emotions: amusement, sadness, disgust, anger, surprise, and fear.

The temporal nature of facial responses is essential in relating such measurements to preference [25]. Facial assessments of emotion allow one to record accurate temporal information on an individual's affective responses. This article supports research by [25], generalizing their validity to unconstrained, naturalistic settings and large-scale validating them.

2.3 Market Research

Micu and Plummer [26] further employed facial electromyography (EMG) to evaluate zygomatic major (AU12) activity while participants viewed TV commercials. The findings revealed that physiological data reflect results different from self-reported emotions, which is consistent with the results of Kassam's [20]. Our evidence from daily life also corroborates these results.

Hazlett and Hazlett [11] used facial EMG in capturing viewers of commercials and found that face muscle activity provided a more stringent measure of recall than self-report, with highest EMG activity at emotionally engaging points in commercials. Teixeira et al. [12] showed that affect inclusion minimizes "zapping" (skipping) of internet commercials. Berger and Milkman [27] found that positive affect-evoking content was shared more than negative affect-evoking content, and that virality was also linked to highly arousing content. Recall, a widely accepted measure of ad success, is affected by emotion [28]. Ambler and Burne [29] discovered that ads that were more affectively intense were more likely to be recalled, and beta-blockers dampening affect diminished recall capability.

However, these tests were carried out in a laboratory environment instead of natural environments and generally analyzed just 10 to 20 commercials. Big-scale analysis is vital to confirm whether the results hold true for the vast population. Our previous work [10] was the initial to prove that automatically collected facial reactions to internet commercials can be used to predict advertising effectiveness measures, such as ad likability and re-watch intention. Teixeira et al. [13] demonstrated that the entertaining activities connected to a brand (evaluated through smiling behavior) was more universal in achieving to increase purchase intention than nonconnected entertaining activities, in turn relating that more universal smiles don't necessarily mean best for the execution of an ad. But both [10] and [13] used only smiling and not a wider range of facial responses.

3 Data and Data Collection

3.1 Video Ads

We analyzed 170 video ads in four nations: 30 French, 50 German, 60 UK, and 30 US. The ads originally aired between 2001 and 2012 and averaged 32 seconds (standard deviation = 14 seconds) in length.
The sample contained a varied set of ads that the brands promoted as successful or unsuccessful within different product categories.

3.1.1 Product Categories

The majority of the tested video advertisements promoted three wide product categories: food (instant rice and pasta, for example), confectionary (gum, chocolate, and candy), and pet care. Among the 170 advertisements, 23 were of other product categories. Notably, these are most often purchased products by consumers from these categories and do not involve long-term purchasing decisions (such

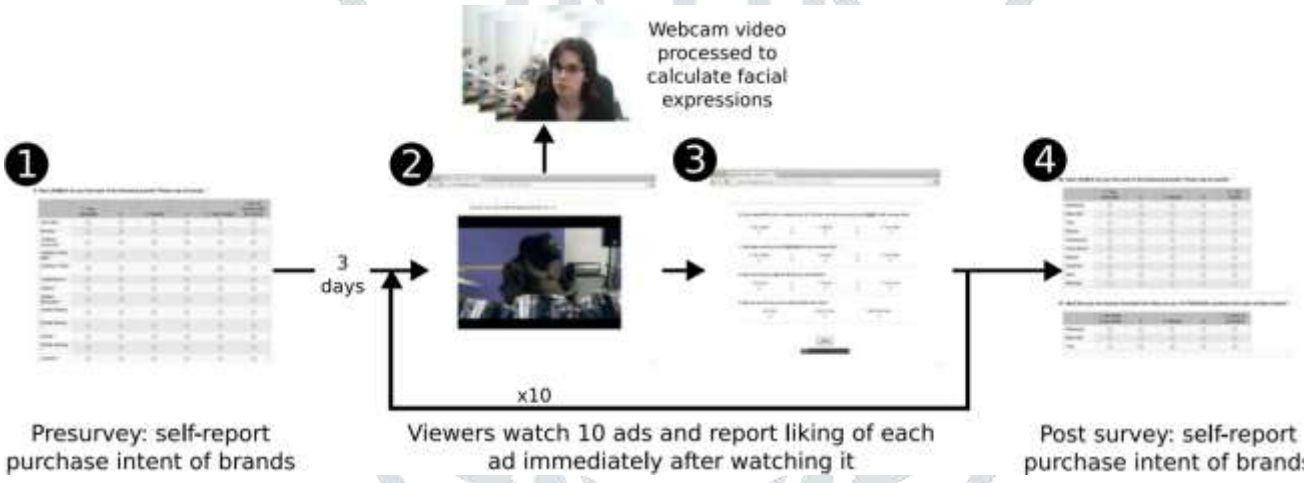
as, for example, purchasing an automobile). Table 1 outlines the number of commercials by product category.

Table 1 Number of Videos Tested from Each Product Category and Emotion Category (Categorized Using MTurk Labelers)

		Product Category				
		Petcare	Confec.	Food	Other	Total
Emotion Category	Amusement	14	46	7	8	75
	Heart-warming	7	2	0	4	13
	Cute	11	1	2	0	14
	Exciting	3	5	2	3	13
	Inspiring	2	3	2	2	9
	Sentimental	5	1	3	0	9
	No Majority	11	17	3	6	37
	Total	53	75	19	23	170

The largest proportion of ads were intentionally amusing. Total number of ads: 170. Total number of amusing ads: 75.

3.1.2 Emotion Categories



Advertisements did not all share a similar emotional intent or content. Various ads would elicit differing responses in audience members—any given one would seek to amuse, while another might seek to move one to the point of affect using pathos. Knowledge about intended emotional effects from an advertisement supplies useful information regarding how one is interpreting face movements. Labeling the emotion was achieved with the aid of Amazon's Mechanical Turk (MTurk) interface, crowd-source labeling with hiring a minimum of three coders for every clip. All of the coders saw a video and answered the question: "CHOOSE the words that best describe the type of FEELINGS you think this video was designed to induce." Potential answers were Sentimental, Inspiring, Exciting, Romantic, Heart-warming, Amusing, and Cute, and the coders could choose more than one. Each video was assigned the most frequent label. Table 1 shows the number of ads per emotion category. These seven categories were originally developed by the first author, who screened all ads and selected words deemed best to convey their affective tone.

3.2 Participants

Users were recruited from four nations (the US, UK, Germany, and France) to view the ads and respond to a survey. Recruitment attempted to balance gender, age group, and economic status (in terms of yearly salary) as equally as possible, while limiting potential self-selection bias. In addition, in each ad, at least 70 percent of the audience consisted of frequent consumers of the advertised product category. Figure 4 shows the distribution of participants by gender, age, and economic status.

Not everyone who was contacted had a working webcam or would agree to have their responses recorded;

those with neither were not allowed to continue with the survey. Of the respondents who started the survey, 48 percent had a functional webcam, and of those, 49 percent consented to have their facial reactions tracked—leaving 23.5 percent of initial respondents eligible to participate. These statistics reflect the necessity of reaching a high number of individuals in order to create a dataset through this method. Luckily, it is inexpensive to contact potential participants, so it is possible to reach large numbers. A greater concern is the self-selection bias introduced by restricting participation to individuals with webcams who agree to be recorded. To mitigate this, we ensured a balanced demographic mix, as described above (see Fig. 4). Future studies could investigate how to quantify the effect of this self- selection bias.

Overall, 1,223 people successfully filled out the survey. All participants saw 10 ads and collected 12,230 facial responses. Each ad was viewed by an average of 72 viewers. Participants could only take the survey once and could not retake it, even with a new set of ads.

3.3 Survey

The web-based survey utilized both video material and face expression capture software. All participants in the study viewed 10 videos that were particular to their country. The survey design is illustrated in Figure 3. Before participating, participants were requested to allow webcam videos to be streamed onto the server, as illustrated in Figure 5, which is a screenshot of the permission question asking.

3.3.1 Pre-survey

Respondents were initially approached to fill in a pre-survey aimed at creating baseline purchase intention. They filled in a purchase intention question across several brands.

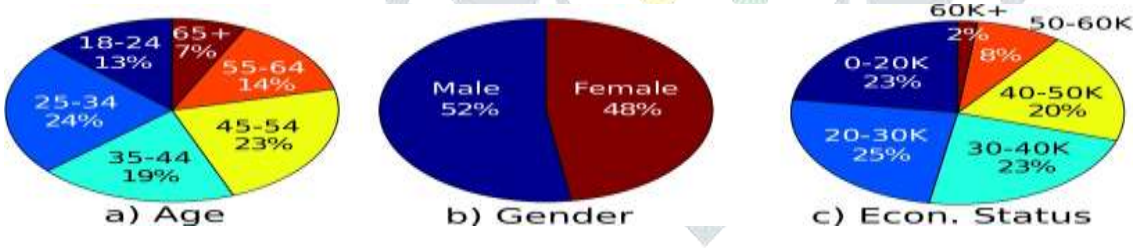


Figure 3 Demographic composition of the 1,223 respondents in this survey: (a) age, (b) gender, (c) economic status (approximate annual income in thousands of US$).

Q. The following time you shop for [product category] how probable is that you WILL BUY from each of the above companies?

Not likely		Neutral		Very likely
1.	2.	3.	4.	5.

They were then contacted a further three days later to fill in most of the survey.

3.3.2 Main Survey

After invitation to take part in the main survey, participants watched 10 videos in a randomised order (in order to offset ordering bias). Participants were briefly shown their webcam view prior to beginning to position the camera and set lighting to best quality, markedly enhancing the quality of the facial expression data recorded. Participants were also requested to remove headgear and not chew gum or eat during the session.

After every video, they were prompted to rate the subsequent item on enjoyment:

Q. How much you LIKED the AD you just watched?

Not at all		Neutral		Very much
1.	2.	3.	4.	5.

Following exposure to all of the ads, they were again prompted with a purchase intention item:

Q. The next time you're buying [product category], how likely are you TO BUY products from each of these brands?

Not likely		Neutral		Very likely
1.	2.	3.	4.	5.

And in a last item on the final page, we queried the respondent:

Q. How COMFORTABLE did you feel during the study?

88 percent rated as "very comfortable" to "neutral," 3 percent "very uncomfortable.".

Q. Did you act any differently than you would have if you were watching these ads NOT for a research project?

Figure 5 Participants were given permission forms prior to watching the commercials and initiating the webcam feed. Moreover, they will be required to provide standard Flash webcam access permission to activate their camera.

In this, 71 percent answered "no differently," 25 percent answered "a little differently," and 4 percent answered "very differently.".

Such responses, along with feedback from the videotaped recording, indicate participants' responses leaned towards natural responses. But demands to agree to be recorded may have affected their

behavior to some extent. Nevertheless, we find these responses to be more accurate than those received while employing a lab environment, leaning towards inducing false behavior.

Members were paid around $8.10 as a gesture of thanks for their time (in their local currency), and the survey took 36 minutes on average.

4 AUTOMATED FACIAL ANALYSIS

The facial action classifiers of Affectiva were employed to analyze the facial videos. The facial expression analysis pipeline is shown in Figure 6.

4.1 Face Detection

The Nevenvision facial feature tracker was used to automatically locate the face and locate 22 facial feature points per video frame. Figure 6 shows the location of these facial landmarks. Throughout the 12,230 facial videos, a total of 4,767,802 frames were processed, with a face found in 3,714,156 frames (77.9 percent). In non-face detected frames, the classifiers did not produce a value.

4.2 Expression Detectors

Affectiva custom algorithms were used in an effort to compute expression probabilities. Eyebrow raise, smile, disgust expression, and positive and negative valence expression classifiers (a few examples appear in Figure 7, aligning original frames with cropped facial areas) were used. These expressions were selected due to their appropriateness for advertising and audience reaction,

according to previous research [11]. In all cases, classification was performed using support vector machines (SVMs) with radial basis function (RBF) kernels. Signed distance from the classifier's hyperplane was computed and normalized using a monotonic function, having been trained to normalize values into [0, 1] range.

Classifier outputs gave continuous, probabilistic, moment-by-moment ratings per frame, giving one-dimensional measurements per video. Figure 8 presents sample response tracks with screenshots of two subjects.

Eyebrow Raise (E): The detector takes HOG features [30] extracted from the whole ROI of the face as input to SVM. It offers a continuous probability of an eyebrow raise (from 0 to 1), with training samples marked as 1 for occurrence of AU01 or AU02, and 0 otherwise.

Smile (S): Taking the HOG features of the entire facial ROI as input to the SVM, this detector provides a continuous probability of a smile in the range 0 to 1. Unlike an AU12 detector, this will detect smiles in general, and training samples will be marked as 1 for a smile occurrence and 0 otherwise.

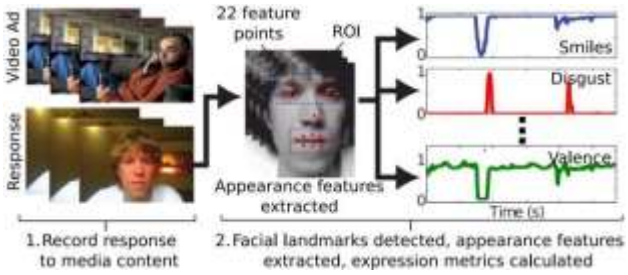
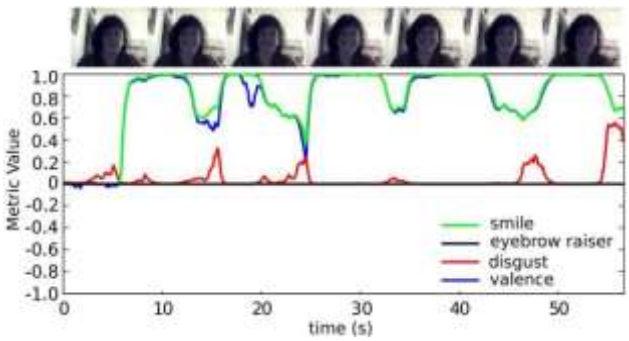


Figure 4 Diagram of facial expression analysis pipeline: 1) Facial videos were recorded as media content was being streamed. 2) The Nevenvision facial feature tracker detected the facial landmarks in the video frames. 3) Histogram of oriented gradient (HOG) feat ures, extracted from the region of interest within each frame, were employed to calculate the expression metrics.

Disgust (D): This detector utilizes HOG features of the entire facial ROI as input to SVM, providing a continuous probability of a disgust expression (0 and 1). Training samples were labeled as 1 for disgust expressions and 0 otherwise. In another experiment, videos predicted to express disgust were recorded by exposing subjects to disgust-evoking material.

Figure 5 Facial expression drawings of eyebrow raise, smile, disgust, and positive and negative valence, shown with both original video frames and cropped frames emphasizing the facial region.



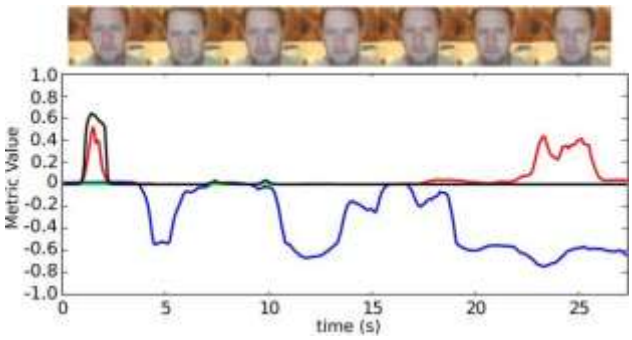


Figure 6 Expression tracks: (Top) Sample response with high smiling and positive valence. (Bottom) Sample response with high disgust and negative valence. Corresponding video frames are shown above each track, cropped to expand the face for easier viewing.

Valence (V): HOG features obtained from the complete facial region of interest (ROI) are employed by the detector. The output is a scalar value in between -1 to 1 such that -1 is a neg-valenced face and 1 is a pos-valenced face. Three-class labeling was done for valence classification with regard to the below rules:

If a smile is detected, valence = +1.

If AU04, AU09, or AU15 occurs, valence = -1. Otherwise, valence = 0.

4.3 Training and Testing of Expression Detectors

Over 5,000 unstructured images labelled by human annotators trained under the Facial Action Coding System (FACS) were utilized for constructing and experimenting with the classifiers. Even if the same expressions in a sequence, efforts have been made towards utilizing a varied set of instances per action. Images were from webcam videos collected over 20 diverse studies being conducted in Asia, Europe, and America. Though similar to the webcam videos in this study, they were not exactly the same. Human coders coded the pictures for the presence or absence of a particular expression. Three FACS-trained coders coded each video for the presence or absence of AU01, AU02, AU04, AU09, AU15, disgust, and smile, and the most frequent label was taken as the final classification. Table 2 shows the area under the receiver operating characteristic curves for the smile, disgust, and valence classifiers (defined above). For three-category valence classification, performance is given for each pair: positive vs. negative, positive vs. neutral, and neutral vs. negative examples.

Table 2 Area under the Receiver Operating Characteristic Curves for the

	Classifier					
	Eye. R.	Smile	Disgust	Valence		
				+ve/-ve	+ve/neut.	neut./-ve
AUC	77.0	96.9	86.7	97.3	92.2	71.5

5 FACIAL ACTIVITY CHARACTERISTICS

5.1 Expressiveness of Viewers

To evaluate viewer expressiveness, we examined the metrics across all videos. Frames where the expression classifier output was below 0.1 were categorized as having no expression present. Of the 3,714,156 frames where a face was detected, 82.8 percent showed no detectable eyebrow raise, smile, disgust, or non-neutral valence expression.

Figure 9 presents histograms depicting the number of frames with probabilities for each expression type (smiles, disgust, and positive and negative valence), accompanied by example frames from selected probability ranges. The vast majority of frames lacked a detectable eyebrow raise, smile, disgust expression, or positive/negative valence. Table 3 details the percentage of frames exhibiting each expression metric above the 0.1.

threshold: only 6 percent featured an eyebrow raise > 0.1, 7.9 percent showed a smile > 0.1, and 5.5 percent displayed a disgust expression > 0.1.

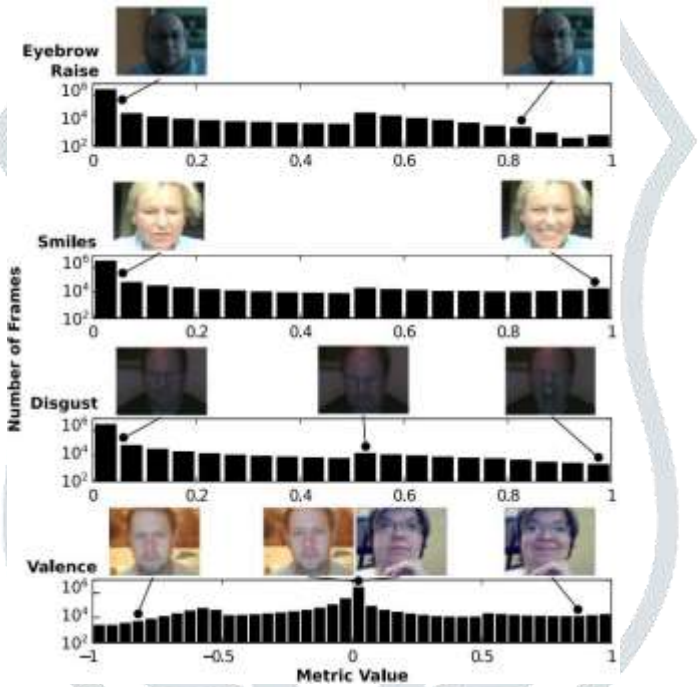


Figure 7 Frame distribution histograms over expression probabilities: (a) Smile, (b) Disgust, (c) Valence. No eyebrow raise, smile, disgust, or positive/negative valence expression of magnitude larger than 0.1 was present in 82.8 percent of frames. While ad responses in natural situations were not frequent, expressive responses did occur among the 70+ audience viewers for each ad. Note: y-axis is scaled logarily.

Table 3 Percentage of the 3,714,156 frames with expression measured in bins of 10 evenly spaced classifier output bins, centered at the listed values.

54.5 percent of the face videos during a period never exhibited signs higher than a range of 0.1, but 36.9 percent of the videos reported signs higher than 0.5. But with an average audience size of more than 70 per ad, we found significant reactions—levels of 0.5 or higher—to be exhibited at least by one viewer per ad. In addition, positive valence words were found to be far more common than negative valence words, a trend that is in line with the general intent of advertisements to create positive sentiment.

5.2 Aggregate Characteristics

Figure 10 shows the mean valence ratings of the ads that were tested, ranked from lowest to highest positive valence. Interestingly, the majority of ads had mean expression ratings with negative valence. The above findings, and the findings prior, show that watching is not typical, but viewers respond differently to adverts, and the adverts elicit

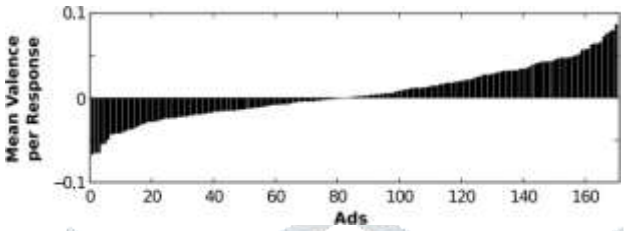
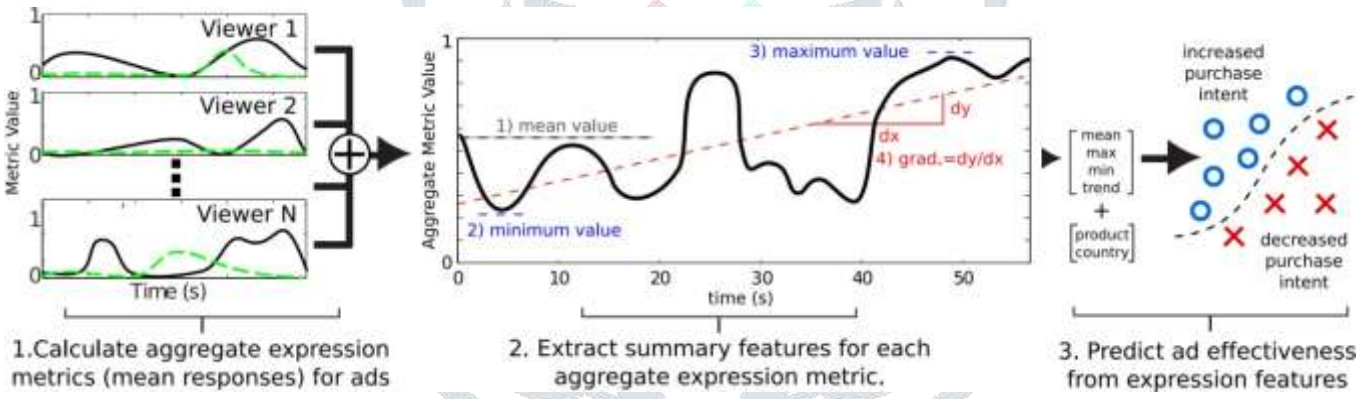


Figure 8 Mean expression valence metrics for the 170 ads sorted by

a range of expressions (from high positive to negative valence). There were not many ads without overall positive or negative valence. For the rest of this paper, we change the scope to predict aggregate level outcomes (an ad's effectiveness to everyone) instead of individual-level outcomes (its effectiveness to a given viewer). Individual-level predictions appear in [10].



Summary features calculated from the facial expression metrics, used to forecast ad success.

6 CLASSIFICATION

To examine the predictive validity of facial reactions, we created and validated classifiers for predicting ad performance (self-reported liking and purchase intent reactions) based on the facial response measures we captured along with contextual factors (product category and country). Here, we outline the process of feature calculation, labeling, and validation, training, and testing model. Figure 11 illustrates a flow diagram of aggregate metric calculation, feature extraction, and classification procedures.

6.1 Aggregate Metrics Calculation

Aggregate expression strengths for every advertisement were inferred from isolated facial reactions, illustrated in Figure 11 (step 1). These are the average intensity of expressions over all observers of an advertisement, excluding frames with no face detected. Mean tracks were calculated for the eyebrow raise, smile, disgust, and valence classifiers.

6.2 Summary Feature Extraction

6.2.1 Facial Metric Features

Summary features were derived from the aggregate facial expression measures using, for example, the mean, maximum, minimum, and gradient for each measure. Figure 11 (step 2) shows the extraction process from an aggregate measure trace. This provided four features per classifier, resulting in a 16-element feature vector per ad for the four facial measures.

6.2.2 Contextual Features

Product category and country of origin were included as contextual features, in the form of a binary matrix. The matrix contained columns for each of the five product categories and each of the four countries.

6.3 Computing Labels

Labels were derived from participants' self-reported responses to survey questions (see Section 3.3). We framed the task as a two-class classification problem, due to the challenge of predicting ad performance from natural facial responses. Distinguishing between ads liked more or less than average, or that impacted purchase intent more or less than average, was an appealing challenge.

6.3.1 Liking Score

Each ad was assigned a liking score from the mean reaction to the question, "How much did you LIKE the AD you just saw?" Ads were divided into two classes: ads with a mean liking score above the median and ads at or below it. With the use of the median, class sizes equal in number were possible. Five ads did not have complete labels, leaving 165 liking examples.

6.3.2 Purchase Intent Score

Purchase intent score was derived by quantifying average delta of answers to the question, "Next time you are buying [product category], how likely are you TO BUY products from each of these brands?" posed in pre-survey and post-survey. Advertisements were split into two categories: one with above-

median average purchase intent delta and one with at or below median. Seven incomplete labels were removed, resulting in 163 purchase intent examples.

7 RESULTS AND DISCUSSION

7.1 Ad Liking Prediction

Figure 13 shows the receiver operating characteristic (ROC) and precision-recall (PR) curves of the ad liking prediction model with the decision threshold of the SVM different in both situations. Table 4 gives an overview of the area under the curve (AUC) values of the ROC and PR curves for the ad liking prediction model for performance comparison when facial features are used independently versus when a combination of facial and contextual features is used. Table 5 shows confusion matrix of the SVM classifier (on best decision threshold, i.e., the point nearest to (1,0) on ROC curve) using both contextual features and facial features for ad liking prediction. Table 4 also shows Cohen's kappa values for best classifier with each feature combination. Median parameters chosen at validation time were C = 10 and γ = 0.1.

It could be that other types of ads (e.g., comedy ads to amuse vs. charity ads to elicit sympathy) would have varying correlations between viewer facial responses and ad effectiveness. To investigate, we applied the same analysis to only ads MTurk coders had labeled as intended to be funny.

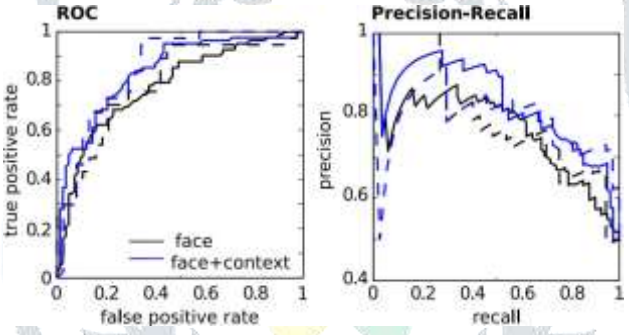


Figure 9 Receiver operating characteristic (ROC) and precision-recall (PR) curves of ad liking models with various SVM decision threshold are shown. The curves are colored: black for performance using facial features alone, and blue for combined performance using facial and contextual features. Solid lines show results for all adverts, while dashed lines show results for those annotated as amusing.

The ROC AUC and PR AUC for both ROC and PR curves are presented in Table 4, and the confusion matrix is presented in Table 5. The models of funny ads marginally outperform others with greater ROC AUC and PR AUC in three out of four cases. For the funny ad model, only 18 out of 75 ads were incorrectly classified with a 76 percent accuracy rate.

Some instances of true positives, true negatives, false positives, and false negatives for the top performer are illustrated in Figure 14. Steep slopes with high peak values.

on positive expressions (smiles and valence) characterize high-liking ads. Ads with lower liking either have minimal activity across all measures (most viewers expressed nothing that was detectable) or higher negative expressions (disgust) than positive expressions (smiles). These findings accord with prior work [25], [32], and suggest that peak and final emotions disproportionately influence the manner in which people recall their experiences.

Exceptions do occur, though. For example, one advertisement (Fig. 14h) produced high smiling and low disgust but a mean liking response of 3.36—below the class threshold of 3.44—placing it at risk of misclassification due to its closeness to the class threshold. There were also other instances in the positive class with very low facial responses (e.g., Figs. 14i and 14l) and, as a result, were misclassified.

Because the leave-one-ad-out training and test procedure is not participant-independent—10 ads were viewed by each participant—we also performed an extra analysis with a leave-ten-ads-out procedure to be participant-independent. For this participant-independent case, the ROC AUC was 0.821 and the PR AUC was 0.752.

Table 4 Area under the ROC and PR curves for ad liking classifier: (Top) All ads (N = 165), (Bottom) Only humorous ads (N = 75).

Ads	Features	ROC AUC	PR AUC	Cohen's κ
All	Naive	0.5	0.5	0.5
	Face	0.779	0.762	0.72
	Face Context	0.840	0.828	0.76
Amusing	Naive	0.5	0.5	0.5
	Face	0.790	0.798	0.73
	Face Context	0.850	0.797	0.76

Cohen's kappa values of top classifiers and self-reported labels are presented.

Table 5 Best liking classifier confusion matrices: (Top) All the ads (N = 165), (Bottom) Only the humorous ads (N = 75), with the threshold from the point on the ROC curve nearest to (0,1).

Ads		Actual +ve (High Liking)	Actual −ve (Low Liking)
All	Predict +ve	66	24
	Predict -ve	16	59
Amusing	Predict +ve	26	7
	Predict -ve	11	31

7.2 Purchase Intent Prediction

Figure 15 shows receiver operating characteristic (ROC) and precision-recall (PR) curves for the purchase intent (PI) score prediction model. Area under the curve (AUC) values for the ROC and PR curves, performance comparison between facial features alone and combined facial and contextual features are shown in Table 6. Results of all advertisements and results of funny advertisements are shown separately.

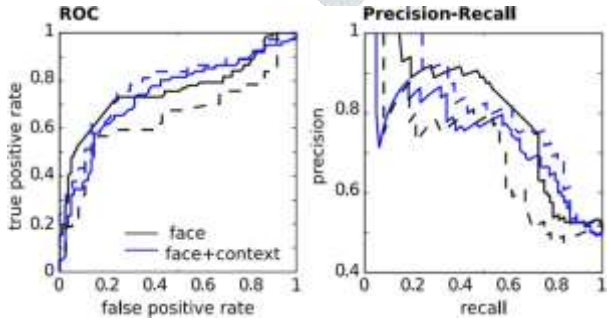


Figure 10 Receiver operating characteristic (ROC) and precision-recall (PR) curves for purchase intent models, with various SVM decision threshold. Black lines indicate performance on facial features alone, and blue lines indicate performance on facial and context features. Solid lines indicate results for all ads, and dashed lines indicate results for humorous ads only.

Cohen's kappa for the best classifier for each feature set is in Table 4. Confusion matrix for the best performing SVM classifier for PI score prediction is in Table 7. For the humorous adverts, there were only 18 misclassified out of 74, with a 76 percent accuracy and 0.757 F1-score. For validation purposes, the median SVM parameters were C = 1.0 and γ = 1.0.

Figure 16 shows some of the best-performing PI model's true positives, true negatives, false positives, and false negatives, with brand appearance times indicated (dashed grey line). Prediction performance for the PI model was weaker than for the liking model, suggesting a more complex relationship between facial responses and purchase intent change, as expected. Increasing the number of positive statements is not as effective to stimulate purchase intent as increasing liking for ads. But in Figure 16, all true positives and false negatives all exhibit an increase in overall smiling after an appearance by a brand, with no such trend in true negatives. The results are consistent with Teixeira et al. [13], which demonstrates that emotions generated by exposure to a brand are more intense. Our results also show that brand sightings shortly before peak positive emotions triggered purchase .

Figure 11 Aggregate ad response statistics correctly and incorrectly estimated by the ad likability model are reported, i.e., true positives, true negatives, false positives, and false negatives. The statistics are color-coded: eyebrow raise (black), smiles (green), and disgust (red). Ad likability is forecast by high peaks in positive expressions, strong overall expressiveness, and high upward trend in positive expressions, and low expressiveness with low ad likability. Individual plot boundaries identify product category for the advertised item.

intention. Moreover, Figure 16 also shows a false positive with a strong response feature (brand sighting before peak positive emotion), but the occurrence of a peak in disgust close to another brand sighting shows that negative emotions could dominate positive ones and associate with the brand. This subtlety would have been lost if smile responses in isolation had been explored, e.g., in [13].

To measure participant-independent performance, we again ran the PI analysis on a leave-ten- ads-out training, validation, and test protocol. The ROC AUC here fell to 0.680, illustrating the difficulty of this process owing to the massive amount of data left out during training and validation.

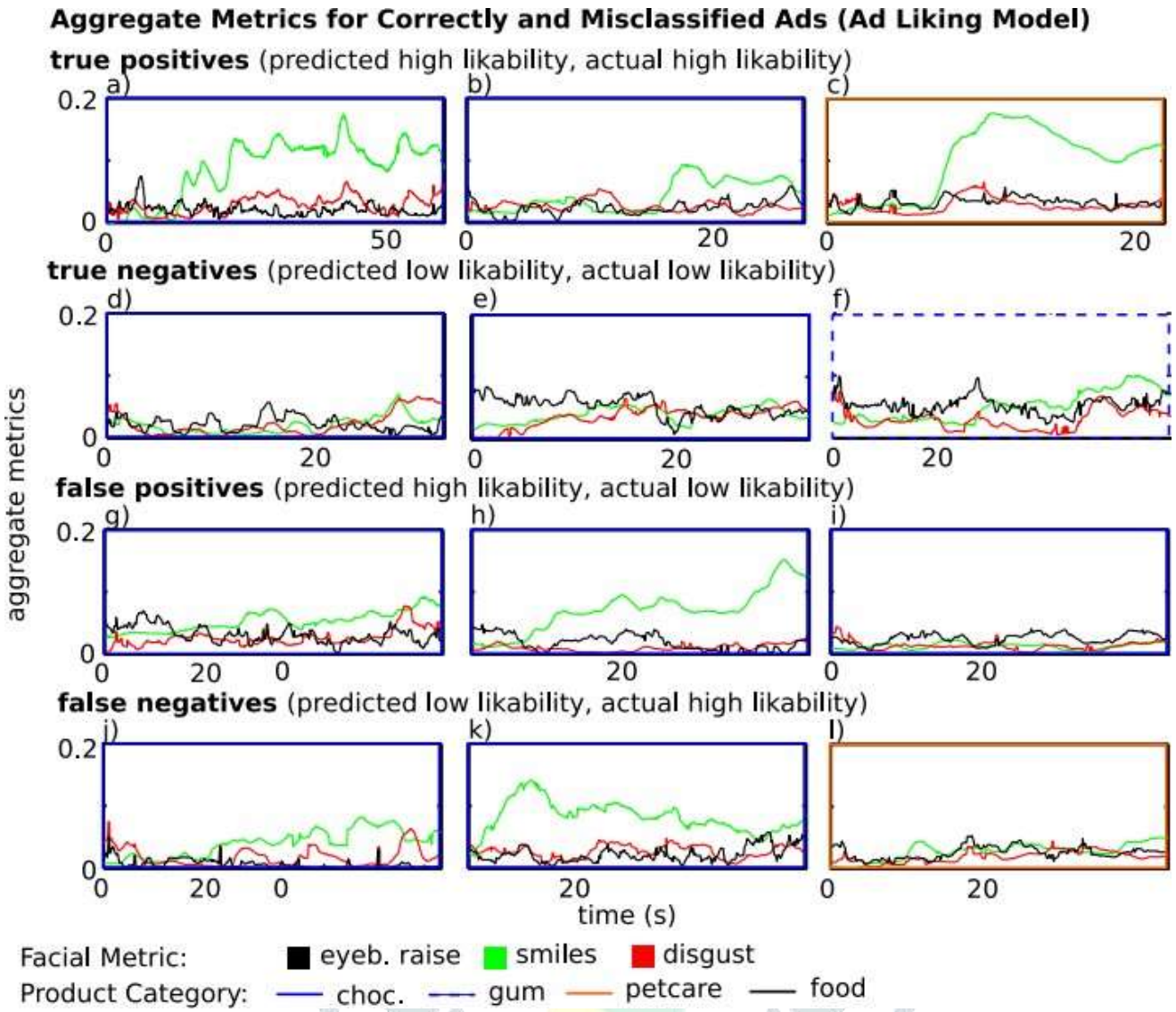


Table 6 Area under ROC and PR curves for Purchase Intent Classifier: (Top) All ads (N = 163), (Bottom) Only funny ads (N = 74).

Ads	Features	ROC AUC	PR AUC	Cohen's κ
All	Naive	0.5	0.5	0.5
	Face	0.755	0.804	0.74
	Face Context	0.739	0.741	0.71
Amusing	Naive	0.5	0.5	0.5
	Face	0.647	0.696	0.69
	Face Context	0.781	0.811	0.76

8 CONCLUSION AND FUTURE WORK

We carried out the largest study to date of facial reactions to web ads. Using an internet site and cutting-edge analysis of facial emotion, we monitored and compared 12,230 facial reactions to 170 ads in four nations (France, Germany, UK, US) and reviewed more than three million frames of film. Such scale of

analysis would have been impossible using standard lab-based data gathering and frame-by-frame manual coding.

We measured viewers' eyebrow raises, smiles, disgust, and positive and negative valence expressions frame by frame and related them to two best measures of ad effectiveness: liking for the ad and changes in purchase intent toward the brand. Facial responses to ads watched under conditions of real-world viewing were small, with only 17.2 percent of frames evidencing discernible eyebrow raises, smiles, disgust, or valence expressions. Almost half of the facial response videos contained no discernible expressions. However, aggregate measures suggested that each ad induced measurable responses in segments of the audience, delivering useful temporal emotion data.

We designed and tested an ad liking predictor based on viewers' emotional response and obtained excellent performance (ROC AUC = 0.850). Steep upward patterns of valence and high highs on positive words were indicators of high liking ratings, as established by earlier individual-level research. We also developed and validated a model for forecasting purchase intent changes from automatically coded facial reactions (ROC AUC = 0.781). The two models performed well at forecasting effectiveness, providing insights into effective ad setups—such as best brand placement. The results indicate that brand sightings immediately before peak positive feelings increase purchase intent.

Our analysis is limited to observing, but perhaps does not include all facial activity. We concentrated on a limited number of action units (AU02, AU04, AU09, AU10, AU12, AU15) because it is hard to identify naturalistic expressions in low-resolution video and because they are media- measurement relevant. Other action units or groups can be included in future studies. Previous research has shown the potential of facial responses for short-term sales effect prediction [33], which we intend to further elaborate.

There are some directions for future research that emerge in this research. We worked with short video material (30-60 second commercials), so it is an open question to what extent this approach applies to longer forms. Fleureau et al. [34] measured physiological responses to 2-hour films and saw high levels of arousal variation, suggesting that facial responses to longer material may have different frequency and duration. Our research included ads from four nations; applying this cross-nationally to markets such as China or India might uncover cultural effects on emotional reactions to advertising. Our ads also captured single, one-off purchase decisions (chocolate bars, for example), but more products with longer-term decisions (cars, for example) need to be examined further to establish these differences. A combination of our method with audio-visual content analysis—looking at scenes cuts, music, or brand presence—would add to such findings.

ACKNOWLEDGMENTS

Affectiva provided its cloud-based expression analysis platform. This effort was funded by the MIT Media Lab Members Consortium and MARS. Daniel McDuff's funding was facilitated through an NEC fellowship. Appreciative acknowledgement goes to the diligence of the reviewers

References

[1] P. D. Bolls, A. Lang, and R. F. Potter, “The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements,” J. Commun. res., vol. 28, no. 5, pp. 627–651, 2001.

[2] P. Ekman and W. V. Friesen, Facial Action Coding System. Palo Alto, CA, USA: Consulting Psychologists Press, 1977.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.

[4] D. McDuff, R. el Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," IEEE Trans. Affective Comput., vol. 3, no. 4, p. 456–468, Fourth Quarter 2012.

[5] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard, "Moodmeter: Counting smiles in the wild," in Proc. ACM Conf. Ubiquitous Comput., 2012, pp. 301–310.

[6] J. P. Johnson, "Targeted advertising and advertising avoidance," RAND J. Economics, vol. 44, no. 1, pp. 128–144, 2013.

[7] D. McDuff, R. el Kaliouby, E. Kodra, and R. W. Picard, "Measuring voter's candidate preference based on affective responses to election debates," in Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction, 2013, pp. 369–374.

[8] R. I. Haley, "The ARF copy research validity project: Final report," presented at the 7th Annu. ARF Copy Res. Workshop, New York, NY, USA, 1990.

[9] E. G. Smit, L. Van Meurs, and P. C. Neijens, "Effects of advertising likeability: A 10-year perspective," J. Advertising Res., vol. 46, no. 1, pp. 73–83, 2006.

[10] D. McDuff, R. el Kaliouby, T. Senechal, D. Demirdjian, and R. W. Picard, "Automatic measurement of ad preferences from facial responses gathered over the internet," J. Image Vis. Comput., vol. 32, no. 1, pp. 630–640, 2014.

[11] R. L. Hazlett and S. Y. Hazlett, "Emotional response to television commercials: Facial EMG vs. self-report," J. Advertising Res., vol. 39, pp. 7–24, 1999.

[12] T. Teixeira, M. Wedel, and R. Pieters, "Emotion-induced engagement in Internet Video Ads," J. Marketing Res., vol. 49, no. 2, pp. 144–159, 2012.

[13] T. Teixeira, R. W. Picard, and R. el Kaliouby, "Why, when and how much to entertain consumers in advertisements? A webbased facial tracking field study," J. Marketing Sci., vol. 33, no. 6, pp. 809–827, 2014.

[14] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," Philosophical Trans. Royal Soc. B: Biol. Sci., vol. 364, no. 1535, pp. 3505–3513, 2009.

[15] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[16] T. Senechal, J. Turcot, and R. el Kaliouby, "Smile or smirk automatic detection of spontaneous asymmetric smiles to understand viewer experience," in Proc. 10th IEEE Int. Conf. Autom. Face Gesture Recognit., 2013, pp. 1–8.

[17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.

[18] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in Proc. Brit. Mach. Vision Conf., 2006, pp. 929–938.

[19] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," Int. J. Synthetic Emotions, vol. 1, no. 1, pp. 68–99, 2010.

[20] K. S. Kassam, "Assessment of emotional experience through facial expression," Ph.D. dissertation, Harvard Univ., Cambridge, MA, USA, 2010.

[21] D. McDuff, R. el Kaliouby, K. Kassam, and R. W. Picard, "Affect valence inference from facial action unit spectrograms," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, 2010, pp. 17–24.

[22] E. Kodra, T. Senechal, D. McDuff, and R. el Kaliouby, "From dials to facial coding: Automated detection of spontaneous facial expressions for media research," in Proc. IEEE Int. Conf. Autom, Face Gesture Recognit. Workshops, 2013, pp. 1–6.

[23] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents," J. Multimedia Tools Appl., vol. 51, pp. 505–523, 2011.

[24] S. Zhao, H. Yao, and X. Sun, "Video classification and recommendation based on affective analysis of viewers," J. Neurocomputing, vol. 119, pp. 101–110, 2013.

[25] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes," J. Personality Soc. Psychol., vol. 65, no. 1, pp. 45–55, 1993.

[26] A. Micu and J. T. Plummer, "Measurable emotions: How television ads really work," J. Advertising Res., vol. 50, no. 2, pp. 137–153, 2010.

[27] J. Berger and K. Milkman, "What makes online content viral?" Univ. Pennsylvania, Philadelphia, PA, USA, unpublished manuscript, 2011.

[28] A. Mehta and S. C. Purvis, "Reconsidering recall and emotion in advertising," J. Advertising Res., vol. 46, no. 1, pp. 49–56, 2006.

[29] T. Ambler and T. Burne, "The impact of affect on memory of advertising," J. Advertising Res., vol. 39, pp. 25–34, 1999.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005, pp. 886–893.

[31] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27, 2011.

[32] C. Varey and D. Kahneman, "Experiences extended across time: Evaluation of moments and episodes," J. Behav. Decision Making, vol. 5, no. 3, pp. 169–185, 1992.

[33] D. McDuff, R. el Kaliouby, E. Kodra, and L. Larguinet, "Do emotions in advertising drive sales? Use of facial coding to understand the relationship between ads and sales effectiveness," in Proc. Eur. Soc. Opin. Marketing Res. Congr., 2013, pp. 1–13.

[34] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction, 2013, pp. 73–78.