To Skip or not to Skip? A Dataset of Spontaneous Affective Response of Online Advertising (SARA) for Audience Behavior Analysis

Songfan Yang*[1] and Le An*[2] and Mehran Kafai[3] and Bir Bhanu[4]
[1] College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China
[2] BRIC, University of North Carolina at Chapel Hill, NC 27599, USA
[3] HP Labs, Palo Alto, CA 94304, USA
[4] Center for Research in Intelligent Systems, University of California, Riverside, CA 92521, USA
syang@scu.edu.cn, lan004@unc.edu, mehran.kafai@hp.com, bhanu@cris.ucr.edu

***Abstract*— In marketing and advertising research, "zapping" is defined as the action when a viewer skips a commercial advertisement. Researchers analyze audience's behavior in order to prevent zapping, which helps advertisers to design effective commercial advertisements. Since emotions can be used to engage consumers, in this paper, we leverage automated facial expression analysis to understand consumers' zapping behavior. To this end, we collect 612 sequences of spontaneous facial expression videos by asking 51 participants to watch 12 advertisements from three different categories, namely *Car*, *Fast Food*, and *Running Shoe*. In addition, the participants also provide self-reported reasons of zapping. We adopt a data-driven approach to formulate a zapping/non-zapping binary classification problem. With an in-depth analysis of expression response, specifically smile, we show a strong correlation between zapping behavior and smile response. We also show that the classification performance of different ad categories correlates with the ad's intention for amusement. The video dataset and self-reports are available upon request for the research community to study and analyze the viewers' behavior from their facial expressions.**

I. INTRODUCTION

Advertisements (a.k.a. ads) on diverse categories of commercial goods are ubiquitous and have a strong impact on people's shopping behavior. Commercial ads on multimedia devices are more effective as they deliver contents through both verbal and visual communication. On one hand, commercial ads on TV are the most common ways to broadcast certain products or service since they can reach a large number of people. On the other hand, the marketing expense for TV commercial ads keeps increasing, especially during major events or prime time. For example, the cost of a 30-second commercial ad during the Super Bowl in US has hit 4 million US dollars in 2013[1].

As seen in the past decade, a surge of online multimedia data has a great influence on the public. For example, 72 hours of video data are uploaded to YouTube every minute [1]. As a result, more and more people tend to spend time watching videos or interacting with other people on the Internet instead of sitting in front of TV. In addition, the vastly growing popularity of mobile devices such as smart phones and tablets enables easy access of multimedia at any time or location.

With the increased advertising cost on TV and decreased audiences, online advertising has become more popular to marketers as more audience can be reached with a lower cost. For example, Google utilizes the TrueView in-stream advertising tool [2] on YouTube to present the ad to the viewer prior to the display of the video content. The viewers have the option to skip the ad and move directly to the desired video after five seconds of viewing the ad or continue to watch the ad after five seconds. In this case, zapping occurs when the viewer chooses to skip the ad. Zapping is an important topic in marketing and advertising research to evaluate the attention an ad receives from the viewer [3]. The action of zapping indicates that the viewer is no longer interested in the commercial ad. This behavior means the loss of a potential consumer for the advertiser. In the case of online advertising such as on YouTube, zapping directly influences the advertising cost and impact since the advertisers are only billed if a viewer watches the ad for at least 30 seconds without skipping. Thus, viewer attention, or zapping analysis, are of great importance both to the online media provider (e.g., Google) to obtain the maximum profit from the advertiser, and to the advertiser to achieve the advertising goal without excessive cost.

There are different ways to evaluate the effectiveness of advertising. For example, self-report has been used to record the subjective feeling of the viewer. This approach has an important limitation referred to as "cognitive bias", and may not always be able to capture lower-order emotions in an accurate way [4]. Recently, facial expression has been used to analyze the effect of advertising [5], [6], [7]. With the help from recent advances in computer vision and pattern recognition research, facial expressions can be automatically detected, which reveal the implicit emotion from the viewer when watching the ads [6]. Accurate facial expression analysis facilitates the marketing and advertising researchers in understanding a viewer's emotional state and behavior. This has the potential to improve the effectiveness of advertising or even design interactive commercial ads to enhance the advertising experience. Moreover, acquiring facial expression is non-intrusive and does not interrupt viewer's watching experience. Besides, widely available cameras on personal

*S. Yang and L. An contributed equally to this work and are both first authors. This work was supported in part by NSF grants 0905671 and 0727129.

[1] http://www.forbes.com/sites/alexkonrad/2013/02/02/even-with-record-prices-10-million-spot/

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

devices such as smart phones and laptop makes the data acquisition easy and inexpensive.

Motivated by interests in analyzing audience behavior on online commercials, we have collected and released a dataset called Spontaneous Affective Response of online Advertising (SARA), containing spontaneous facial expression data in video sequences. These videos are recorded non-intrusively when the participants are watching online commercial ads. In total 12 commercial ads from different categorizes are presented to each viewer and a zapping (i.e., skip) option is available, which mimics the real-world online advertising setting such as that on YouTube. The term zapping implies that the viewer of a commercial ad is no longer interested in its content/presentation, thus opts not to continue watching the ad. In addition, we provide a baseline study by predicting the audience's zapping behavior from their facial expressions. Specifically, smile response is used, which has proven to be an useful indicator of a viewer's preference of commercial ads [8], [6]. Although other datasets such as AM-FED [7] that includes facial expression response from viewers have been published, SARA is the first publicly available facial expression dataset that includes audience's zapping behavior and self-report feedback, which are of great importance to analyze the advertising effectiveness. Note that in [6] facial expression data are collected to study the emotion-induced engagement in online ads. However, this is a private dataset and cannot be accessed by other researchers.

The main contributions of this paper are summarized as follows:

1) A dataset of spontaneous affective (*i.e.*, facial expression) response of viewers watching online commercial ads is collected and published. As far as we know, this is the first publicly available dataset containing both facial expression response and zapping behavior.
2) In contrast to the AM-FED dataset [7] in which each viewer only watched one commercial by his/her choice, each participant in our experiments are presented with 12 commercials in different categories. In this way, subjective bias is avoided. In addition, participants have the option to either watch or skip the ads, which is an important behavior for further analysis.
3) Besides the spontaneous video data recorded, each participant fills in a self-report form, indicating his/her reason of zapping a particular ad. This self-report is also released together with the dataset to facilitate in-depth analysis.
4) We provide a baseline for predicting zapping based on automatically detected smile from the video sequences and point out the effectiveness of smile-based zapping prediction for different kinds of ads.

The rest of the paper is organized as follows. Section II reviews related work. In Section III we provide details of the design and protocol of data collection. A data-driven zapping analysis using smile response is presented in Section IV. Baseline methods and experimental results of automatic zapping prediction are provided in Section V together with in-depth analysis. Finally, Section VI concludes this paper and indicates future directions.

II. Related Work

A. Automatic Facial Expression Recognition

Typically six discrete facial emotion states are often considered in the literature, namely amusement, fear, anger, disgust, surprise, and sadness. These emotions are triggered by the action units (AU) corresponding to independent motion of the facial muscle [9]. In [10] a challenge of facial expression recognition is presented and baseline methods for AU detection and emotion detection are given together with video data and evaluation protocol for standardized evaluation. Instead of performing a score level fusion, Yang *et al.* [11] generate a holistic face representation called Emotion Avatar Image (EAI) from entire video sequence and achieve an improved recognition result compared to the frame based approach. Zheng [12] synthesizes multi-view facial feature vectors and proposes a group sparse reduced-rank regression model to select the optimal sub-regions on a face for expression recognition. A Boosted Deep Belief Network (BDBN) is proposed in [13] to perform feature learning, feature selection, and classifier construction iteratively, achieving dramatic improvements on two public datasets.

Instead of trying to detect all possible emotion types, some approaches focus on detecting specific emotion. For example, Shan [14] uses pixel intensity difference as features for smile detection and adopts AdaBoost to form a strong classifier from weak classifiers. Inspired by the face alignment approach in [11], an efficient smile detection system with automatic face alignment is developed in [15]. A survey of expression recognition methods can be found in [16].

B. Online Advertising Analysis using Facial Expression

To study the spontaneous facial expression of people watching commercials, a comprehensively labeled dataset called AM-FED is introduced in [7]. Both facial videos with labeled frames and self-report responses are provided, together with a smile detection baseline. In the setting of data collection, participants may choose one of the three commercials to watch. An interesting study is conducted by McDuff *et al.* [8] who have collected 611 naturalistic and spontaneous facial response from people watching the presidential debate in 2012. It is found out that voter preference can be predicted with an average accuracy of over 73% based on facial expression and self-reported preference.

As the consumers have a choice to switch away from either a TV commercial ad or an online ad, it is challenging for the advertisers to retain consumers' attention during the course of a commercial ad [3]. Recently, smile has been demonstrated as an useful indicator of a viewer's preference of commercial ads [5]. Teixeira *et al.* [6] incorporate joy and surprise expression recognition from a Bayesian neural network classification system to analyze the viewer's zapping decision. They conclude that the velocity of the joy response

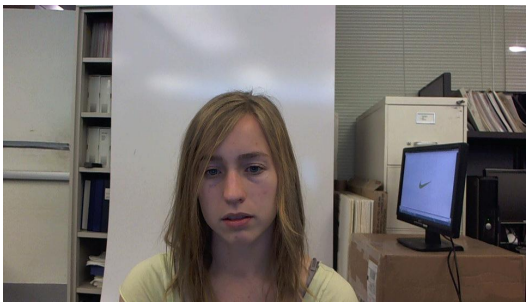


Fig. 1. The data collection setting. A duplicate monitor in the back is used for data synchronization.

highly impacts the viewer's zapping behavior. The study suggests that attention paid to a commercial ad determines the interest of the viewer and retaining a viewer's attention can produce desirable communication effects.

III. Data Collection

To collect the facial expression data when the participants are watching ads, we seat the participants in front of a PC monitor, which is used to display the ad contents. Each participant is presented with 12 ads selected from the three categories listed in Table I in random order. The facial expression of the participants is captured using a Logitech c910 webcam mounted on the top of the monitor. The resolution of the webcam is set to 960×544 with a frame rate of 24 fps. The average resolution on face is approximately 220×220 pixels. The length of each ad ranges from 30 to 90 seconds.

During the display of an ad, participants have the choice to either watch an ad until the end or zap at any moment by clicking on the skip button. In either case, they are given a 30 seconds break after each ad to reduce the aroused emotion and prepare for the next ad with neutral emotion. The entire data collection procedure for one participant takes about eight minutes on average and during this process the participants are not interrupted or distracted.

As shown in Fig. 1, a secondary monitor behind the participant is utilized as an easy solution for data synchronization. This monitor displays the same content watched by the participants. The webcam is able to capture a participant's facial expression as well as the corresponding content he/she is watching. In this setting, we are able to separate the expression data according to the ad contents shown on the secondary monitor. This enables the analysis of facial expression responses with respect to the individual ad. The facial expression response from the 30 seconds break between the ads are not used in this work.

The 12 ads shown in Table I are selected based on the following criteria:

1) Popularity: The *Car*, *Fast Food*, and *Running Shoe* are the categories that almost everyone is familiar with and is well connected to.
2) Minimum gender bias: We wish to eliminate the gender bias of the ad in this research. The selected ad categories have less gender bias compared with categories such as *Beer* or *Makeup*.
3) Recognizable brand: Since online video viewers are the targeted ad receivers, we select the ad from brands that either have their official YouTube Channel or participated in the YouTube ad campaign. In this way, we have access to the ad for this research.
4) Varying entertainment levels: We have carefully evaluated the entertainment information in each ad. Our final ad selection includes both kinds of ads that are very amusing and that are less entertaining.

TABLE I

Selected Advertisements for Data Collection

Category	Brand	Ad Name	Length (in *s*)
Car	Toyota	I Wish	60
	Honda	We Know You	90
	Chevy	Wind Test	30
	Nissan	Enough	30
Fast Food	Jack In The Box	Hot Mess	30
	Subway	New Footlong	30
	Carl's Jr.	Oreo Ice Cream	32
	Pizza Hut	Make It Great	30
Running Shoe	Nike	Flyknit Lunar 1+	47
	Adidas	Boost	30
	Puma	Mobium and Adaptive	30
	Under Armour	I Will Innovation	60

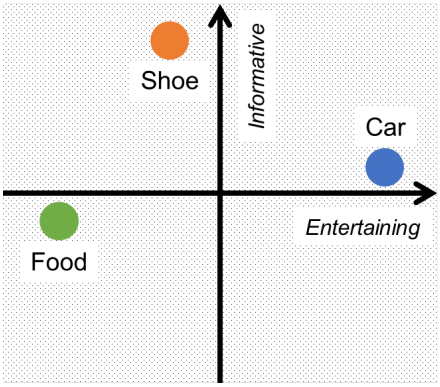


Fig. 2. The property of each ad category in use.

Elper *et al.* [3] show that the lack of *entertainment* and *information* factors are the two major reasons for zapping. Therefore, it is essential to carefully select the ad to include both aspects, which broaden the feasible analysis in the future. As seen in Fig. 2, the "Car" category is very entertaining, and the "Fast Food" is less entertaining. The ads in "Running Shoe" aims at demonstrating the technology of sport gear. Therefore, they attempt to draw the audience's attention by the *information* factor.

The ads in Table I can be accessed on Youtube by directly searching the brands and the ad names. In total 51 people have participated in our data collection. In terms of gender, there are 31% female and 69% male. In terms of ethnicity, the dataset includes 40% Asian, 25% Euro-American, 16% African-American, and 19% other ethnicity groups.

A. The Zapping Distribution

Since the participants are given the option of zapping at anytime, the fraction of an ad being watched varies for different participants. Fig. 3 shows the distribution of ad fraction that is being watched. To distinguish zapped and non-zapped cases, a Gaussian mixture model with two components is fitted to the data. We find that 90% of the ad fraction is the best value to separate the two components of the mixture. As seen in Fig. 3, the probability of zapping is dramatically higher in the range of 90% to 100%. This is plausible since most ads end up with the product or brand logo, which does not attract audience's further attention. Practically, zapping at 90% to 100% indicates that an ad has been watched until the end. In the 0% to 90% range, the first half (0% to 45%) has a slightly higher probability than the second half on the average. This tells us that participants in our experiments tend to zap early if they are not attracted early during an ad.

One interesting fact worth noting is that the popular TrueView advertisement publisher only bills the advertiser if an ad has been watched for more than 30 seconds [2]. However, a billing mechanism based on the percentage of the ad being watched may be more feasible. If we have a better understanding of the zapping behavior, a win-win-win situation can be proposed: the audience receives more desirable video content; the advertiser obtains more attention from the audience; and the publisher (such as YouTube) gains more revenue.

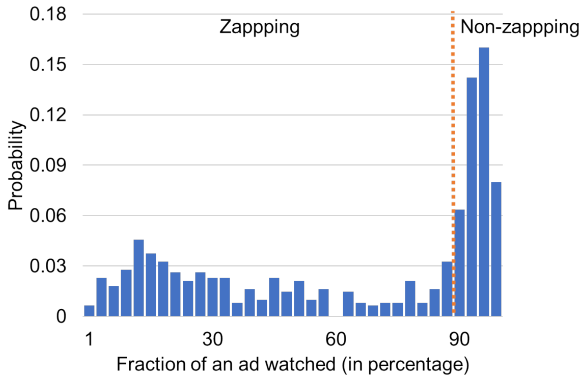


Fig. 3. The zapping distribution. The data-driven threshold at 90% is used to separate the data into zapping and non-zapping classes.

B. Audience Feedback Analysis

During the break of watching each ad, subjects were also asked to fill out a questionnaire which contains the following questions:

1) Did you skip the ad?
2) Why did you skip? Mark all that apply:
   - The ad is not funny.
   - The ad is not informative.
   - I have seen this ad before.

Table II shows a sample of the questionnaire that the participants are asked to fill in.

TABLE II
SAMPLE FEEDBACK QUESTIONNAIRE TO BE FILLED BY THE PARTICIPANTS AFTER WATCHING EACH AD

Ad	I didn't skip	Not entertaining	Not informative	Watched before
1				
2				
...				
12				

Fig. 4 plots the histogram of the questionnaires answered by the participants during data collection for all three ad categories. As observed from Fig. 4, *informativeness* and *previously-watched* factors play a less important role compared to the *entertainment* factor. The "Car" category is indeed designed to be more entertaining than the other two categories. The "Fast Food" category ads are intentionally selected to be the least amusing among all categories.

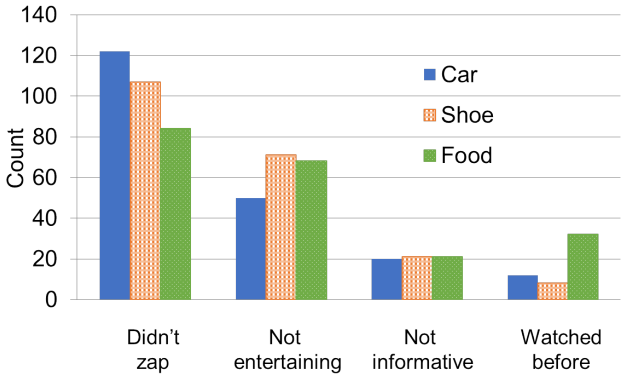
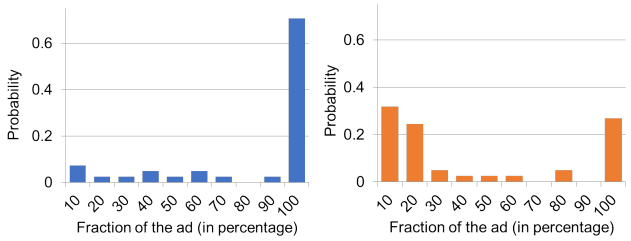


Fig. 4. The distribution of the questionnaire answers of three ad categories.

One interesting fact worth pointing out is that there is a discrepancy of participants' feedback in Fig. 4 and ad designers' intention in Fig. 2. The ad designers of "Running Shoe" intend to engage the audience by showing the technology related information. However, the audience do not view this as more informative, but rather less entertaining. As evidenced in Fig. 4, *not informative* is not a major factor for zapping while *not entertaining* is the dominant reason for zapping. In addition, even "Running Shoe" ads are supposed to be more informative, the counts of zapping due to *not informative* are very close to the counts of the other two categories. This interesting observation may help in the analysis of the audience behavior and mitigate the discrepancy of information perceived by the sender and receiver. Although not studied in this work, the analysis of *informativeness* of the ads will be pursued in the future work.

In Fig. 5 we show the most and the least funny ads, namely "Toyota" and "Subway", from participants' zapping behavior. The self-reported feedback of these participants is summarized in Fig. 6. Similar observation holds in both Fig. 5 and Fig. 6, revealing that the participants think "Toyota" is more entertaining than "Subway".



(a) Toyota (b) Subway

Fig. 5. The zapping distribution of two exemplar ads: "Toyota" and "Subway". Most participants watched the entire "Toyota" ad, while large amount of participants zapped in the beginning of "Subway".

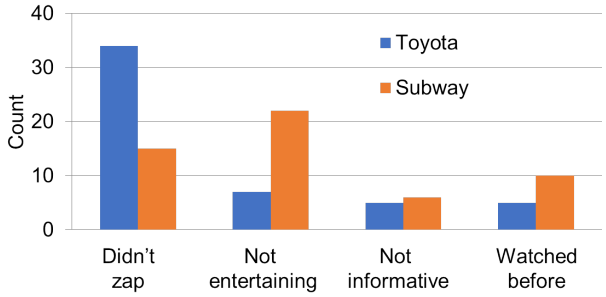


Fig. 6. The distribution of the questionnaire answers of the two ads "Toyota" and "Subway". Most participants think "Toyota" is more entertaining than "Subway".

IV. Zapping Analysis from Automatic Facial Expression Recognition

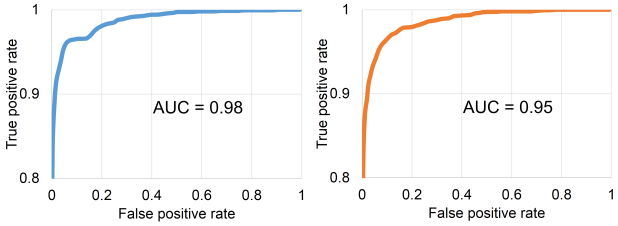
Based on the zapping distribution shown in Fig. 3, we quantify the audience's behavior into two different classes: zapping and non-zapping, and use 90% of the ad length as the threshold in separating these classes. We further propose to use the spontaneous facial expression response to analyze the zapping behavior. That is, given the facial expression response of a participant watching an ad, determine whether this sequence belongs to zapping or non-zapping class. We formulate this as a binary classification problem. Before presenting the detail of this baseline method for zapping prediction in Section V, we first study the spontaneous facial expression to better understand how expression relates to the zapping behavior.

Specifically, we analyze the characteristics of the dataset from the perspective of the audience's smile response. These characteristics are essential in motivating our zapping classification feature. We could potentially design systems that recognize different facial expressions. However, we observe during data collection that the dominant facial expressions of the participants are *neutral* and *smile*. Therefore, in light of the philosophy of Occam's razor, we choose to use smile expression for zapping analysis in this work.

A. High-fidelity Smile Response Measurement

The goal is to compute the probability of smile on a per-frame basis. The faces are first extracted using Viola-Jones face detector [17] and then aligned using a dense flow-based registration technique [18]. The aligned faces are resized to 200×200 pixels and then divided into 20×20 non-overlapping regions. The Local Phase Quantization [19] texture features are computed for each region and these features are then concatenated to form the feature representation of the entire face for smile detection.

Smile detection is formulated as a smile/neutral binary classification problem. We adopt the linear Support Vector Machine (SVM) [20] for classification. For accurate person-independent smile detection, the classifier is trained on multiple databases with a large number of subjects from: FEI [21], Multi-PIE [22], CAS-PEAL [23], CK+ [24], and data from Google image search [15]. In total, 3578 images (1543 smiling faces and 2035 neutral faces) are in the training set.



(a) Training data (b) Testing data

Fig. 7. ROC curve for our person-independent smile detection algorithm.

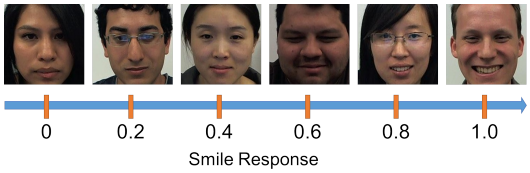


Fig. 8. Sample smile response results. The response value reflects the intensity of smile.

The following tests are carried out in a person-independent manner where no test subject is included during training. The Area Under Curve (AUC) is 0.98 for the 10-fold cross validation result across the entire training set (see Fig. 7(a)). To demonstrate the generalization of this classifier, we carry out a test on a selection of 10,000 frames from our SARA dataset. The AUC is 0.95 as shown in Fig. 7(b), which means that the smile classifier generalizes well on unseen data. The probability output of the SVM smile classifier is used as the smile response. Since we use *smile* vs. *neutral* instead of *smile* vs. *non-smile*, it is interesting to observe that under this setting, the classification rate of test data is not only superior, but also the probabilistic output of smile detector is able to capture the smile intensity as illustrated in Fig. 8. The reason we choose *smile* vs. *neutral* setup is that the participants are concentrated on the viewing experience. Most of the expressions other than smile are of neutral nature, and very few participants display excessive non-smile expression. In this case, the probabilistic outputs closely correlate with the smile intensity. Besides, there are neutral examples with open mouth in the training data, and therefore, the classifier is not

just naively predicting random mouth open motion but rather muscle motion caused by smiling.

For proof of concept, we verify the probabilistic outputs with the manually annotated smile intensity results. We gather three annotators expertized in face and facial expression recognition, and each annotator is given 500 frames sampled from the entire SARA dataset. The annotators score the smile intensity of each frame by comparing it with the reference figures similar to the ones in Fig. 8. The median value of all three annotators is selected as the ground-truth smile intensity to mitigate discrepancy among annotators. The resulting absolute mean error intensity is 0.216 between the prediction and the ground-truth.

B. The Mean Smile Response

We analyze the average smile response in the first 30 seconds (720 frames from 24 fps webcam) for both zapping and non-zapping sequences. Since the participant can zap at any time, the facial expression sequences are of various lengths. Therefore, the average smile response is computed as follows:

$$r_m(t) = \frac{\sum_{i=1}^{N} r_i(t)}{\sum_{i=1}^{N} I_i(t)}, \quad I_i(t) = \begin{cases} 1, & \text{if } r_i(t) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $r_i(t)$ is the smile response of sequence i at time t, N is the number of sequences, $I_i(t)$ is the indicator function that shows the existence of the smile response of sequence i at time t. In other words, for each frame, the average smile response is computed based on the available responses. Similarly, we compute the standard error of the mean (SEM) by:

$$SEM(t) = \frac{std(r(t))}{\sum_{i=1}^{N} I_i(t)} = \sqrt{\frac{\sum_{i=1}^{N}(r_i(t) - r_m(t))^2}{\sum_{i=1}^{N} I_i(t)(\sum_{i=1}^{N} I_i(t) - 1)}} \tag{2}$$

where $r_m(t)$ and $std(r(t))$ are the mean and the standard deviation of available smile responses at time t, respectively. In Fig. 9, the moment-to-moment mean smile response is bounded by the positive and negative SEM.

As observed in Fig. 9, the smile response level for the two classes is initially about the same. Thereafter, the response of the non-zapping class increases for the rest of the 30 seconds. On the contrary, for the zapping class, the response remains around 0.2 and decreases toward the end. Therefore, the moment-to-moment average smile response is potentially a good feature to separate zapping from non-zapping class. This observation is also in line with the conclusion in [6] that smile level largely correlates with the zapping behavior.

C. The Maximum Smile Response

The maximum smile response of the sequences is also different for zapping and non-zapping classes. Two examples are shown in Fig. 10. We plot the distribution of sequences from the two classes based on their maximum smile response in Fig. 11. The total probability of each group sums up

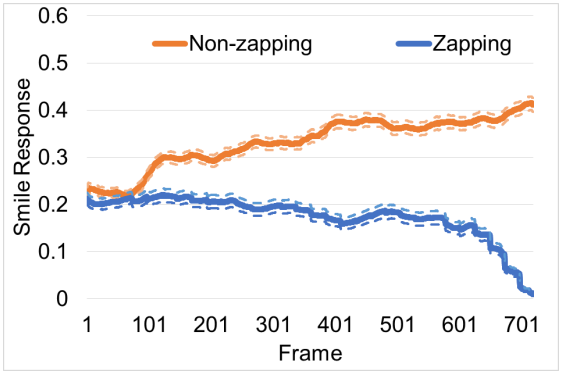
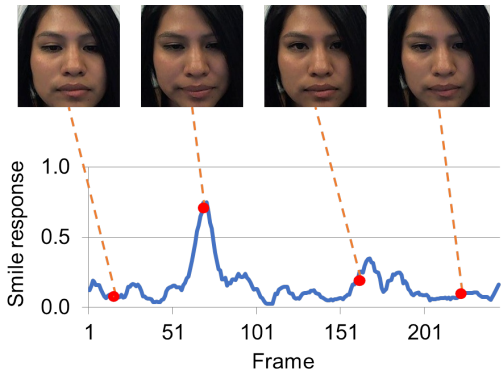
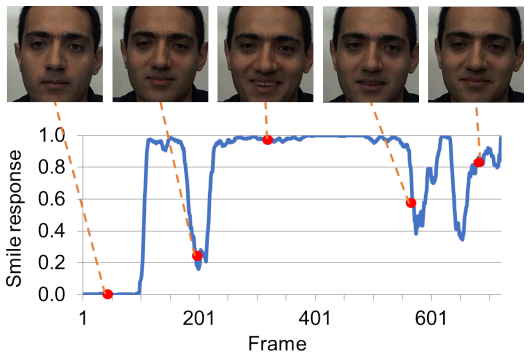


Fig. 9. The average smile response of zapping and non-zapping classes for the first 30 seconds (720 frames). Each sequence is bounded by its standard error of the mean (SEM).



(a) zapping class



(b) non-zapping class

Fig. 10. Sample frames of smile response from zapping and non-zapping classes.

to 1. As illustrated in Fig. 11, if a sequence's maximum smile response is above 0.5, then the chance is higher that it belongs to the non-zapping class, and vice versa for maximum smile response below 0.5. The probability reaches the highest for the non-zapping class if the maximum smile response is above 0.9. On the contrary for zapping class, majority of the sequences are with the maximum smile response less than 0.1.

For non-zapping class, the probability is the second highest (15.5%) when the smile response is less than 0.1. Observations on our data show that a few participants watch

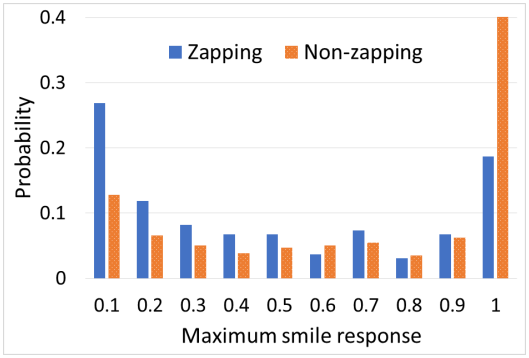


Fig. 11. The distribution of the zapping / non-zapping data based on their maximum smile response.

the entire ad but display minor smile expression. This means that entertaining content is not the only reason to keep the viewer engaged. Besides, we interviewed a few participants and found out that a small group of them enjoyed the ad but prefer not to show their feelings through facial expression.

For zapping class, the probability of zapping decreases as maximum smile response increases, and reaches the minimum when smile response is 0.8. However, the probability increases thereafter. After examining the data, we found that several participants were engaged by the ad and were smiling with high intensity in the beginning. However, they zapped immediately when the brand's logo or name showed up at the end of the ad. After interviewing with them, we found that most of the people behaved like this because they thought the ad is about to finish. From the advertiser's point of view, this scenario should be considered as a success. However, from the publisher's point of view, they will not get paid since they consider this scenario as zapping [2]. The aforementioned smile response features will be used for automatic zapping prediction in Section V.

V. Zapping Prediction Benchmark

As mentioned in Section IV, zapping behavior is correlated with a participant's smile response. In light of this observation, we aim at providing baseline methods and benchmark results for predicting zapping using automatically detected smile response. The class labels for different video sequences of participants' facial expression responses are annotated as "zapping" and "non-zapping". For features to be used in the classification, *mean smile response* (Mean) and *maximum smile response* (Max) as described in Section IV-B and IV-C are considered. In addition, as demonstrated in [5], dynamic information helps the classification of ad likability. Therefore, we also use the *dynamic smile response* (Dynamics) as feature for zapping classification. *dynamic smile response* is a time series consisting of smile responses from each frame. Since the sequences have different length, we normalize the original dynamic smile response to a predefined length (100 in the experiments) as our feature. For classification, we use a linear SVM classifier [20] and set $C = 1$ in the experiments. The setting of C is not sensitive in this classification task as we tested in the experiments.

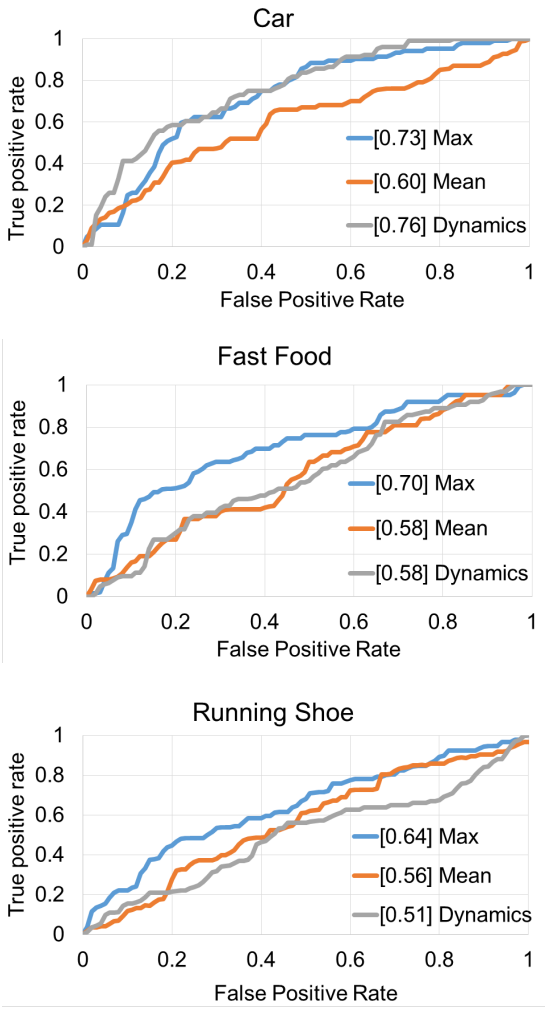


Fig. 12. ROC curves for zapping prediction of ads in three categories. Numbers are the AUC values.

Since the commercials are in three categories, the ROC curves for each category using different features are shown separately in Fig. 12. The number associated with each curve is the AUC value. Among the three ad categories, the classification performance of zapping is the highest for the "Car" category. In other words, zapping prediction using smile response is more accurate in this category. This is a plausible observation since the ad contents in the "car" category are composed of purposely designed humorous and joyful stories to solicit smile and attract attention. On the other hand, the contents in "Running Shoe" category aim at demonstrating the sport gear technology in order to attract attention, which do not necessarily trigger smile. This corresponds to the *information* factor [3] of an ad that engages the audience (see Fig. 2). As a result, only using smile response to predict zapping in this category is not sufficient. Although the ads in the "Fast Food" category include less informative content, zapping of this category depends more on the entertaining factor. Even more audiences report that they are the least entertaining ads (see Fig. 4), zapping prediction of "Fast Food" by smile response out-performs the "Running Shoe" as the classifier is build on the smile

response which reveals how entertaining an ad is. To sum up, the prediction performance correlates with the design and the entertaining content of an ad, which we demonstrated at the categorical level. The same correlation is also observed at individual ad level.

The performance of zapping prediction also depends on the chosen feature. The *mean smile response* does not lead to a good prediction since this "statistic" smooths out the discriminative but non-dominant high smile response by averaging it with the dominant low smile response scores. On the other hand, the *maximum smile response* better captures the discriminative high intensity smile triggered by the amusing contents of an ad, achieving better performance than the *mean smile response*. In the "Car" category where explicit entertaining contents are displayed, both *maximum smile response* and *dynamic smile response* outperform *mean smile response* by a large margin. In this category, using *dynamic smile response* achieves the best classification performance. We believe it is because when entertaining factor is more dominant compared with information factor, facial expression (specifically, smile) has a strong correlation with audience's zapping behavior. Under this circumstance, facial expression dynamics is a more discriminative feature compared with static features.

For the "Fast Food" category with limited amusing contents, viewer's facial response shows less engagement with the ads. The facial expression dynamics would not necessarily be discriminative enough for zapping prediction. Therefore, *maximum smile response* is able to better capture the zapping behavior than the other two features. In the "Running Shoe" category, due to its informative and less entertaining contents, zapping prediction based only on smile is less accurate. We observe that the prediction performance using expression dynamics degraded to almost random guess. This suggests that for ads that are intentionally designed to be informative but not entertaining, analyzing their zapping behavior from smile dynamics is less plausible and other emotion factors should be taken into account.

VI. Conclusions

To facilitate research and understanding in audience's response to commercial advertisements, we have collected and made publicly available a dataset called Spontaneous Affective Response of online Advertising (SARA) by recording the participants' facial expression and zapping behavior when they were shown 12 different commercials from three categories. This is the first public dataset that contains both facial expression, zapping behavior, as well as self-reported reasons for zapping. In addition, we have provided the benchmark method for predicting zapping based on automatically detected smile response. Experiments showed that for some commercial ads which intend to maintain audience's attention by joyful and interesting storyline, smile can be an important clue to predict zapping. Beyond smile, our current research also focuses on other emotions such as surprise to analyze audience's attention. In addition, more robust features derived from emotion response are to be investigated for a better prediction of zapping.

References

[1] Gabarron, E., Fernandez-Luque, L., Armayones, M., Lau, A.Y.: Identifying measures used for assessing quality of youtube videos with patient health information: A review of current literature. Interactive Journal of Medical Research (2013)

[2] Pashkevich, M., Dorai-Raj, S., Kellar, M., Zigmond, D.: Empowering online advertisements by empowering viewers with the right to choose. Journal of Advertising Research (2012)

[3] Elpers, J.L.W., Wedel, M., Pieters, R.G.: Why do consumers stop viewing television commercials? two experiments on the influence of moment-to-moment entertainment and information value. Journal of Marketing Research (2003)

[4] Poels, K., Dewitte, S.: How to capture the heart? reviewing 20 years of emotion measurement in advertising. Technical report, Katholieke Universiteit Leuven (2006)

[5] McDuff, D., Kaliouby, R., Picard, R.: Crowdsourcing facial responses to online videos. IEEE Trans. on Affective Computing (2012)

[6] Teixeira, T., Wedel, M., Pieters, R.: Emotion-induced engagement in internet video advertisements. Journal of Marketing Research (2012)

[7] McDuff, D., Kaliouby, R.E., Senechal, T., Amr, M., Cohn, J.F., Picard, R.W.: Affectiva-mit facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild. In: CVPR Workshops. (2013)

[8] McDuff, D., Kaliouby, R.E., Kodra, E., Picard, R.W.: Measuring voter's candidate preference based on affective responses to election debates. In: ACII. (2013)

[9] Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)

[10] Valstar, M., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: FG. (2011) 921–926

[11] Yang, S., Bhanu, B.: Understanding discrete facial expressions in video using an emotion avatar image. IEEE Trans. on Systems, Man, and Cybernetics, Part B (2012)

[12] Zheng, W.: Multi-view facial expression recognition based on group sparse reduced-rank regression. IEEE Trans. on Affective Computing (2014)

[13] Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: CVPR. (2014)

[14] Shan, C.: Smile detection by boosting pixel differences. IEEE Trans. on Image Processing (2012)

[15] An, L., Yang, S., Bhanu, B.: Efficient smile detection by extreme learning machine. Neurocomputing (2015)

[16] Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence (2009)

[17] Viola, P., Jones, M.: Robust real-time face detection. Int. Journal of Computer Vision (2004)

[18] Yang, S., An, L., Bhanu, B., Thakoor, N.: Improving action units recognition using dense flow-based face registration in video. In: FG. (2013)

[19] Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: ICISP. (2008)

[20] Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[21] Thomaz, C.E., Giraldi, G.A.: A new ranking method for principal components analysis and its application to face image analysis. Image and Vision Computing (2010)

[22] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing (2010)

[23] Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., Zhao, D.: The CAS-PEAL large-scale Chinese face database and baseline evaluations. IEEE Trans. on Systems, Man and Cybernetics, Part A (2008)

[24] Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: CVPR Workshops. (2010) 94–101