Image and Vision Computing 65 (2017) 49–57

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Behavioral cues help predict impact of advertising on future sales ☆

Gábor Szirtes[a,*], Javier Orozco[a], István Petrás[a], Dániel Szolgay[a], Ákos Utasi[a], Jeffrey F. Cohn[b]

[a] Realeyes OÜ, Tölgyfa utca 24, Budapest 1027, Hungary
[b] University of Pittsburgh, 4322 Sennott Square, Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 1 May 2016
Received in revised form 7 March 2017
Accepted 13 March 2017
Available online 22 March 2017

Keywords:
Market research
Behavioral cue
Predictive modeling
Facial expression analysis

ABSTRACT

Advertising aims to influence consumer preferences, appraisals, action tendencies, and behavior in order to increase sales. These are all components of emotion. In the past, they have been measured through self-report or panel discussions. While informative, these approaches are difficult to scale to large numbers of consumers, fail to capture moment-to-moment changes in appraisals that may be predictive of sales, and depend on verbal mediation. We used web-cam technology to sample non-verbal responses to television commercials from four product categories in six different countries. For each participant, head pose, head motion, and more frequent facial expressions like smiling, surprise and disgust were automatically measured at each video frame and aggregated across subjects. Dynamic features from the aggregated series were input to simple linear ensemble classifier with 10-fold cross-validation to predict product sales. Sales were predicted with ROC AUC = 0.75, 95% CI [0.727,0.773] and predictions for *unseen* categories were consistent for all, but one product groups (ROC AUC varies between 0.74 and 0.83, except for Confections with 0.61). Predictions for *unseen* countries showed similar pattern: ROC AUC varied between 0.71 and 0.89, with the exception of Russia with ROC AUC 0.53. In comparison with previous attempts, our approach yielded higher overall performance and greater generalization over not modeled factors like country or category. These findings support the feasibility, efficiency, and predictive validity of sales predictions from large-scale sampling of viewers' moment-to-moment responses to commercial media.

1. Introduction

Advertising is about influencing consumer preferences, appraisals, action tendencies, and purchases. Television and increasingly online video commercials are a key component. Over 80 billion dollars is spent annually on television commercials in the US alone [1]. For the companies that produce commercials and for their clients, there is great interest in evaluating the effectiveness of commercials they produce and distribute. One approach is to correlate television advertisements with product sales (online shopping in a short time window around the time of tv ad) [2]. This approach enables a gross estimate of direct influence of advertising on sales but is blind to consumer reactions to individual commercials. For that, it is necessary to assess consumer responses to specific commercials in relation to product sales.

One solution is to ask viewers to report on their responses to commercials. Focus groups, personal interviews, random-digit phone surveys, and online surveys have been used for this purpose. While providing useful information, these methods have notable limitations. They pull for rational thinking rather than emotional responses that may be more predictive of purchase behavior; respondents must verbally represent what often are non-verbal, often unconscious cognitive-emotional reactions; and the dynamics of their responses may be compromised by recency effects. Demand characteristics and social desirability effects may bias reports as well. Focus groups, surveys, and related methods further assume that verbal reports are necessarily the best indices of purchasing influences. Evidence suggests otherwise. People's preferences often are outside of their awareness and strongly influenced by emotion [3,4].

Emotions consist of multiple components that include subjective feelings, action tendencies and physiological arousal. All are prime candidates for influencing likelihood of purchase decisions. During emotion episodes, these components become correlated [5].

Automated facial expression analysis using web-cam video acquisition is a promising alternative. Using computer vision and machine learning, facial expressions of emotion to television advertisements can be measured on a moment-to-moment basis. This approach avoids the necessity for viewers to verbally report their experience, captures fine-grained information about the timing of behavior, and

☆ This paper has been recommended for acceptance by Mohammad Soleymani.

* Corresponding author.
E-mail address: gabor.szirtes@realeyesit.com (G. Szirtes).

http://dx.doi.org/10.1016/j.imavis.2017.03.002
0262-8856/

can be scaled to large numbers of viewers from multiple geographical regions and countries. In seminal work, Ref. [6] found that facial expression measured in this way was predictive of (self-reported) ad liking purchase intent. Given the wide availability of web-cam technology and the efficiency of this approach, it becomes possible to plan population-based research for more accurate investigation of viewer's reactions to commercial presentations and their relation to product sales.

Since the ultimate goal of marketing is to increase sales (which is, unfortunately, not directly linked to ad liking or even expressed purchase intent), the strongest evidence of the usefulness of automated behavior analysis in market research would come from studies that could identify correlation between observable behavior (like elicited emotion responses) and changes in sales due to a particular ad campaign. Using sales data (sales lift or increase in sales in a given observation window) from MARS, Incorporated and web-cam recordings of viewer responses to commercials, McDuff [7,8] obtained mixed results. In Ref. [8] facial expression based analysis outperformed survey based methods in predicting sales performance only, *when average commercials (half of the data) were discarded*. Combination of self-reports and expression analysis did not bring about significant improvement, either. In Ref. [7] sales performance was about random, when four product categories were represented in both the training and test sets. When removing one category, performance increased for both survey and facial expression based methods (both achieving moderate accuracy), but their combination performed significantly better implying that different methods reveal complementary information. In spite of the mixed results, these works suggested the efficacy of combining "crowd-sourcing" methodology, automated facial expression analysis, and supervised machine learning algorithms to differentiate between high and low performing ads where labels are based on sales lift data (i.e., increase in sales). These initial findings suggested that automated behavioral cue based analysis has great practical value and with further improvements this approach can become a viable alternative to traditional, survey based methods. In comparison with alternative methods, it scales well to large samples of respondents and can be executed more quickly. A critical challenge is to achieve higher overall accuracy and greater consistency in performance across product categories. In addition, a viable method must be able to cope with real "in the wild" conditions, like varying video quality or non-cooperating respondents.

McDuff's approach assumes prototypical response behavior, thus individual responses are aligned and aggregated into a well defined temporal response (frequency of eye brow raise or smile at a given moment of time). Descriptive statistics of these unified panel responses (e.g. max, mean or slope of a fitted line) are then used to form a concise representation of the facial expressions observed in a panel of respondents corresponding to a particular ad. Correlation between the panel responses and the sales performance of the ads was then learned by a non-linear classifier method (Support Vector Machine with Radial Basis Function kernel, RBF-SVM), but the complexity of the obtained models was not reported.

Recent work in automatic facial expression analysis suggests that dynamic features (e.g., velocity or acceleration of facial expression and head pose) strongly encode emotion. Velocity and acceleration of head and facial movement, in particular, can express emotion and related affective states [9,10]. In a series of studies, the "packaging" of non-verbal behavior (e.g., co-occurring head pose and motion) and dynamics have proven critical to the meaning of facial expression. Orientation or timing of head movement, for instance, encodes meaning. Smiles of enjoyment and embarrassment, for example, have similar static features (e.g., contraction of the zygomatic major and orbicularis oculi), but differ in head pose and movement. For enjoyment, head pose is frontal or slightly raised, while for embarrassment it pitches down and to the side [11,12].

We wondered whether dynamic features and alternative representations of facial expressions that do not rely on temporal alignment and are robust under abovementioned real conditions would result in higher and more consistent accuracy across product categories. In addition, the complexity of the chosen classifier may question the validity of the entire approach, due to the limited data (163 ads) and high dimensional representation (16 dimensions).

Our goal is thus to test the hypothesis that correlation between dynamics of facial expressions and head motion and product sales can be learnt from real observations of varying quality via parsimonious models that generalize well. We evaluate this hypothesis for the product categories and countries examined by McDuff and include additional countries (from now on reference model will be referred to as McDuff-model to facilitate comparison). To facilitate comparison of our proposed approach with McDuff's reference model [7,8], we list the main similarities and differences in Table 1. The differences will be detailed out in the subsequent sections.

The paper is organized as follows. In Section 2 we define the task and sales data. We then describe the data acquisition method, the data representation, and the classification model. For each part, comparisons with the McDuff-model [7,8] are also given. Performance of the proposed model is then presented in Section 3 together with numerical comparisons. In Section 4, we discuss findings and future work informed by those findings.

2. Proposed method

2.1. Objectives

Our first objective was to collect in a fast and economic way a large number of spontaneous behavioral responses via web-cams to a given set of commercials for which sales lift data is provided by MARS, Incorporated. Our second objective was then to design, implement and validate a parsimonious and transparent model that can accurately predict sales performance from the available

Table 1
The table describes major differences in the reference study of McDuff and the proposed approach.

Source of signal	McDuff-model	Proposed approach
Advertisements	163, 4 countries	147, 6 countries
Participants	About 1200	18,793, only category users
Recording conditions	In the wild, fixed frame rate recordings, long sessions of watching 10 ads	In the wild, varying and inconsistent frame rate, shorter sessions with 4 ads
Representation	Summary statistics on a mix of discrete emotion classifier outputs and Facial Action Unit detectors using average panel responses	Aggregation of individual response statistics on discrete emotion classifier outputs and head pose
Modeling	Nonlinear SVM	Parsimonious ensemble model of independent linear regressors
Analysis	ROC AUC with Leave One Out validation	ROC AUC with more reliable K-fold cross-validation, model complexity analysis

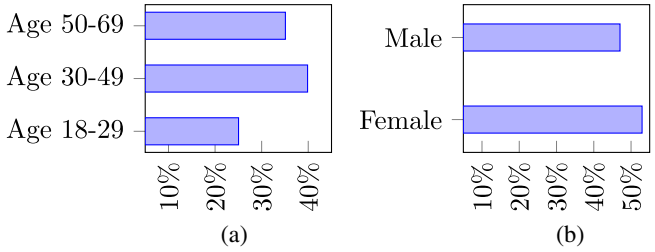

Fig. 1. Demographics of the participants (a) age distribution, (b) gender distribution.

observations. Finally we wanted to make a fair and thorough comparison with the McDuff-model thus demonstrating that our method indeed provides a reliable and practical tool for analyzing behavioral responses at scale for market research purposes.

2.2. Participants

National census based panels of paid participants (average panel size was 277 subjects) were recruited in six countries (Australia, France, Germany, Russia, United Kingdom and USA) by third-party field agencies. All subjects gave informed consent to participate in a panel, provide demographic information, and have their faces recorded while viewing commercials on their home computers. Informed consent followed the procedures delimited in our privacy policy (https://www.realeyesit.com/privacy). Additional IRB approval is not required for commercial research. The total number of participants was 18,793, but for quality reasons described in Subsection 2.4 only 12,262 sessions were finally used in the analysis. The age and gender distribution of participants is shown in Fig. 1 with small differences between countries.

Responses of the participants were recorded remotely via their own home computer and web-cam. Since we cannot control recording conditions, our data collection is indeed "in the wild" acquirement. One of the strongest technical limitations we face is that in addition to relatively low average frame rates of the recordings, the frame rate is not consistent for individual recordings. The participants are used to take part in various studies. Asking people to view videos in the way we did is a well-validated procedure for eliciting emotion (e.g., [13]). The data collection was conducted the same way as it is done in our business service and we haven't experienced major deviation in the response statistics compared to other collections. An implicit evidence for spontaneous behavior is that participants often forget that they are being recorded and leave the room or are getting engaged in unrelated activities like talk, eating, etc. In addition to demographics constraints, there were 2 more selection criteria. The technical requirement was that each participant has internet access and web-cam attached to her home computer. Importantly, we screened participants by their category use thus showing them only relevant ads (relevance criterion). This is in contrast to the McDuff studies where only 70% of the participants were actual category users.

2.3. Stimuli and class membership

The commercials represented four product categories: confections, food, pet care, and chewing gum. They were originally aired between 2013 and 2015 in six different countries. The commercials varied in duration between 10 and 30 s. As all commercials are relatively recent, in 44% of the sessions participants claimed that they had seen the ad before. Ideally, respondents should be filtered by past exposure, such filtering is infeasible as it would double the cost of data acquisition.

Sales lift data were provided by MARS, Incorporated. Target score was derived from the actual contribution of the ad campaign to "sales lift". To measure sales lift for each commercial, exposed and control (unexposed) comparison groups were identified by MARS, Incorporated and their actual purchases were traced. The ratio of purchase propensity in the exposed group to the comparison group was then averaged over the set of exposed/comparison groups. Sales lift rating was quantified on a four-point ordinal scale for training classifiers. Similar to previous attempts [7] we simplified the regression task into a binary problem: commercials with ratings 1 and 2 are converted into low performance class, while high performance class is made of ads with ratings 3 and 4. Let us note, however, that the additional information encoded in the ordinal scale was used in training part of our predictive model. The distribution of commercials across categories and regions and the score distributions are plotted in Fig. 2.

Complicating analysis, about a third of the commercials were variations of each other. We considered two commercials as variations if differences between them were due to small edits in length or content. As an example, some commercials had the same storyline, but displayed a different brand label or were produced in a different language. We report results separately for all commercials and for the case in which related ads are combined into a single label.

The study design was comparable to Ref. [7], except for the following differences. We included two additional countries. The commercials they used aired in 2002–2012; ours aired more recently. Their set contained 163 unique commercials; ours contained 116 unique ones out of the available 147 commercials. Sales lift in their study was quantified on a 3-point ordinal scale and ours on a 4-point ordinal scale.

2.4. Collection of behavioral responses

All commercials were viewed by participants on their own computer while their face was recorded by web-cam and streamed to a server. Image resolution was 640 × 480. This "in the wild" setting ensures more ecologically valid spontaneous behavior than would be

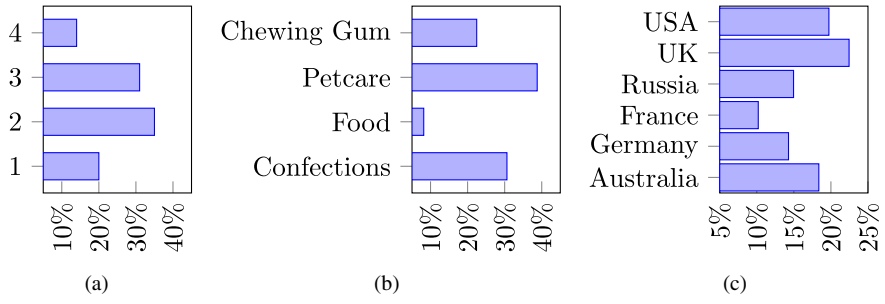

Fig. 2. Statistics of the commercials: (a) distribution of sales-lift ratings from 1 (lowest) to 4 (highest); (b) distribution of product categories; and (c) distribution of countries.

Fig. 3. Example data.

possible in a laboratory at the cost of image quality and frame rate. Average frame rate was about 13 fps. Videos were omitted if face was occluded or subjects were engaged in unrelated activities like talking or eating. Fig. 3 displays some examples of varying illumination, pose, distance from the web-cam, and facial expression.

Subjects viewed up to four commercials presented in a random order. Session length was approximately 10 min. By contrast, in Refs. [7,8], subjects watched 10 commercials presented in a random sequence and completed self-report ratings between them; session length averaged 36 min. We chose a shorter format because Refs. [14,15] found a negative correlation between session length and data quality. In addition, we used larger samples (on average 277 subjects viewed each ad versus 100) to counter the impact of video quality as well as large variations in the observability of the viewers' responses. Even after applying our conservative quality filtering (see below), the effective mean sample size was 164, still significantly larger than 100 (54) as reported in Ref. [7] (effective sample size was not explicitly reported, but it was stated that only about 54% of the recordings contain small, but observable responses.).

2.5. Data representation

2.5.1. Preprocessing

First, we discarded recordings that did not match in duration with the ad (maximum difference was set to 1.5 s). We also dropped recordings where delay between any of the subsequent frames was longer than 2 s. Second, color frames were converted into grayscale intensities. Third, facial features were extracted and input to classifiers for emotion detection. Fourth, the raw features as well as the output of the emotion algorithms were used to form time series signals for predictive modeling.

In order to compensate for noise and to help temporal alignment of time series corresponding to the same ad we planned to apply zero phase smoothing and resampling on all observations. However, we realized that some of the descriptive statistics like variance distribution in a given time window are quite sensitive to the realization of the preprocessing steps. Although the proposed preprocessing was not optimized for the subsequent task (task independent preprocessing), frame rate variability between and across panels interfered in a complex way with the subsequent signal processing steps, thus making our approach strongly dependent on some irrelevant parameters. In turn we decided to keep data intact and use more robust descriptors instead.

2.5.2. From facial expressions to signals

On each frame we detect the location and estimate pose (yaw, pitch and roll in degrees) of the head/face using an improved version of the method of Ref. [16] and locate the precise position of a set of facial landmarks (alignment of key points) based on a modified implementation of the algorithm proposed in Ref. [17]. Local geometry of the landmarks as well as texture patches around them are then used as descriptors by our in-house emotion classification system trained to classify facial expressions into discrete expression categories such as smile, surprise or disgust.

The most frequent facial expression is the smile [18]. Smiles may convey enjoyment, favorable appraisal, anticipation, and action tendencies to approach [19]. From perspective of automated detection, smiles often involve relatively large geometric and textural deformations that are advantageous (see e.g. [20]). Since most of the advertisements in our data set were designed to be amusing or joyful, it is expected that signals derived from smile carry information about the elicited emotional states. While "confused" expression is also relatively frequent on our own training data set, we found that surprise and disgust related signals are more informative for the sales prediction task.

The output is then a multi-dimensional time series of estimated head pose and three facial expression classifier output together with their corresponding probability output (posterior probability that a class label is chosen for a given set of descriptors).

In searching optimal representations for the sales prediction task we wanted to identify features that display temporal changes that correlate with the evolution of the response eliciting stimulus (ad). In addition, we wanted to avoid temporal alignment and averaging of individual responses and we also wanted to follow a common procedure for all signals thus avoiding the need of additional parameter optimization. This way we can ensure that the resulting model will be robust and less sensitive to the particular properties of the training data set. The common approach was the following.

1. For each time series obtained from the head pose estimator and the facial expression classifiers, we calculate temporal differences between subsequent frames (detrending) in a given recording. ($dx_i^j = \frac{dx^j}{dt}_i$ denotes temporal difference at time i for subject j.)
2. Temporal normalization: $dx_i^{j*} = dx_i^j{}_* < dt^j >$, where $< dt^j >$ denotes the average time step in a given recording.

This transformation maps observations on the same scale by compensating for varying sampling rates.

3. 2 second long bins are used as temporal segments for each recording, regardless of the frame rate or the duration of the entire recording.
4. For each bin we calculate the 90*th percentile* of the normalized differences. ($mx_k^j = 90th \quad percentile_{i \in k}(dx_i^{j*})$, where the notation $i \in k$ means that the *ith* value (frame) falls in bin k. Number of values may vary in the different bins).
5. The bin values are then weighted and summed up to yield one number that describes the differences between the last 3–4 s and the rest. The applied weight vector is a simple zero sum step function. ($dx^j = \sum_{k=1}^{n} w_k * mx_k^j$, $\sum_i w_i = 0$, where n is the number of bins in the given recording and there is no need to use frame or segment indices).
6. Aggregation of individual values. From the set of the previously obtained normalized, individual signals, a particular descriptive statistics (75^{th} or 90^{th} percentile) is calculated. These signal values then describe the sample response to a given stimulus and do not depend on time or subject indices.
7. To diminish aliasing effects due to the arbitrary segment boundaries, bins were shifted in both directions up to 0.25 s and all steps above were repeated. The finally obtained sample signal is then the average of these calculations. This step, while not necessary, seems to make our approach more robust.

Additional optimization of this procedure (like varying time bins, various forms of normalization, use of different weight functions, etc.) would likely yield better performance, but such fine tuning would raise concerns about overall robustness and feasibility of our approach. Bin size, for example was defined based on the average frame rate and the duration distribution and onset dispersion of the annotated events in our proprietary training data set. If small perturbations of the select parameters show graceful degradation in the correlation, then we consider the parameter robust. While the McDuff-model relied on simple summary statistics of aggregate sample responses, such as maximum or gradient of a linear fit, we hypothesized that dynamics of the elicited emotional responses analyzed at the subject level *before* aggregation would be more robust and distinctive. In addition, our approach does not assume uniform video frame rate, which is often difficult to achieve with remote recordings. While we do not seek continuous rise in the measured signal, we also assume measurable change in the temporal responses to the ads that motivates the use of differentiation, normalization and weighting.

Of several candidate features we selected three signals derived from various facial expressions and one signal derived from head pose. The source of the signals, the descriptive statistics used in the signal and their Pearson correlation with the binary sales lift scores are shown on Table 2.

Fig. 4 displays the major steps of the proposed signal generation process from observations on individual subjects to sample distribution (aggregate panel response) and to the final signal value assigned to the corresponding advertisement.

Interestingly we found positive correlation between the scores and the disgust based signal.

Table 2
The table shows the selected signals (simple summary statistics), the corresponding source and the Pearson correlation with the sales lift score.

Source of signal	Descriptive statistics	Correlation
Smile	75*th* percentile	0.31
Disgust	75*th* percentile	0.32
Surprise	75*th* percentile	0.29
Head pose (roll)	90*th* percentile	0.26

It is also somewhat surprising that head pose related signal indicates more frequent or larger head pose changes near the end of the sessions. We compared signals derived from yaw, pitch and roll and roll based features showed the highest correlation with the rating. Previous work has found that gaze direction strongly correlates with head pose [21,22] so larger head pose variations may reflect a lasting effect of the stimulus content and do not correspond to the very last segment of the stimulus, since subjects with extreme head pose do not look at the direction of the screen.

We also found that for all signals the optimal weight function assumes a step shape, assigning positive value for the last 3–4 s (that is all signals measure differences between the very end and the rest of the recordings). For Head Roll we found that even higher correlation can be achieved by assigning positive weight for the last 6–8 s. This deviation may indicate that head pose changes are less synchronized (temporal onsets are dispersed) and duration may also vary.

We believe that due to the small data size (number of commercials to be tested), it is difficult to give a more thorough and plausible interpretation of the findings other than emphasizing the fact that both facial expressions and head pose related signals carry complementary information about sales performance.

In comparison, the signals of the McDuff model were extracted from a mix of facial action unit activations which are strongly related to particular discrete expressions (eye brow raise is often associated with surprise), discrete expressions (smile) as well as "valence" which was derived from the estimated intensity of all discrete facial expressions. We instead used a simpler mix of 2 signal types, one related to discrete emotion categories (smile, disgust and surprise), while the other one related to head pose changes which is less difficult to measure than facial action units.

2.6. Modeling

Limited sample size and potential label noise make modeling difficult or even impossible if complexity of the used approach is high. So we opted for simple ensemble modeling with averaging [23,24] with the following assumptions. We treat signals as independent and do not consider higher order interactions between them. This assumption allows for training simple (weak) experts whose votes can be summarized in an ensemble model. The second assumption is that we seek linear relationships between signals and target score and non-linearity is induced by thresholding (binarization of the individual experts' output). Such thresholding supports signal denoising. The workflow of our model is shown in Fig. 5. The ensemble model is composed of standard linear regressors, nonlinear terms (binarization), summation and final thresholding. For ROC AUC (receiver operating characteristics area under curve) calculation, the output of the summation is used instead. The processing is the same for all signals and incorporates the following steps. The input x to the linear regressor at the first stage is one of the selected features described above. The target variable is the original 4 point rating as described in Section 2.4. The weight and bias parameters (w, β) are trained on the training set in a stage-wise manner (instead of applying joint optimization of all parameters simultaneously in the two stages). As next step the output y of the regressor is binarized. This step enables noise suppression by learning a threshold α. After this stage the outputs $\hat{y}$ of the individual signal modeling paths are combined by simple summation and thresholding

In the McDuff-model the classifier of choice was RBF-SVM [25,26]. After training the decision boundary is represented by "support vectors" which are the most difficult cases from both classes to be distinguished. An advantage of that method is that it can learn complex interactions between features and is not sensitive to class imbalance or skew. A disadvantage is that the required sample size depends on

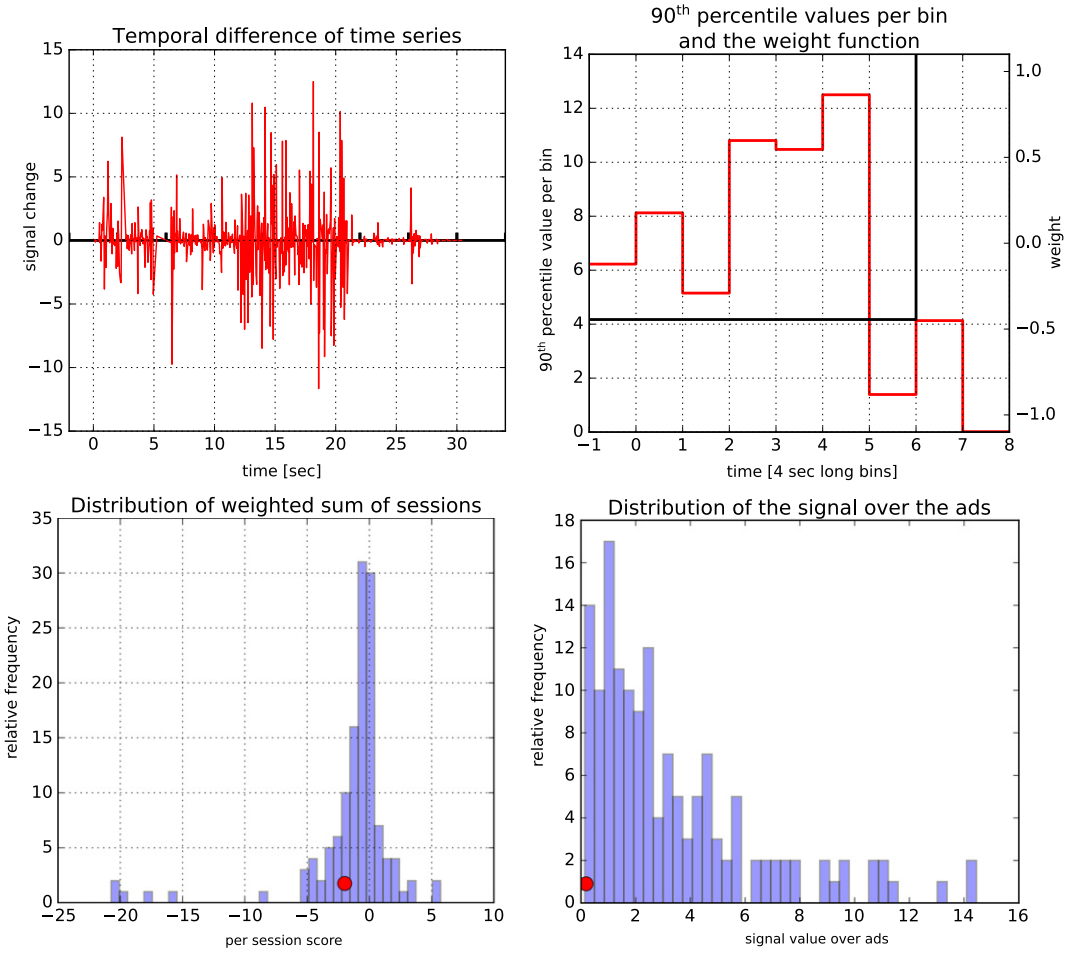

Fig. 4. Main signal processing steps from individual data to "Surprise" signal for a given advertisement. Sub-figures are the visual representation of signal processing described in Section 2.5.2. Top row left: temporal difference calculated from the output of the surprise classifier for a given subject. Top row right: red curve shows the maximum values of the normalized temporal differences for each time segment (binning). Black curve shows the weight function that assigns a positive or negative weight to each bin. The weighted sum of the bin values characterizes the surprise response of one subject. Bottom row left: panel level description. Distribution of the individual surprise response as calculate in the previous step. The particular example is denoted by a red dot. For the final "Surprise" signal of the given advertisement we selected the maximum value over the subjects in a given panel. Bottom row right: the distribution of the "Surprise" signals over the advertisements. The calculated signal of the given ad is denoted by a red dot. Since we found positive correlation between this signal and sales lift data, this ad most likely belongs to the low performing class.

the representation. High ratio of support vectors over sample size indicates that the requirement is not met and the resulting model will have large generalization error on unseen data. In Ref. [8] time series were segmented into 10 parts and summary statistics (max, mean, min) were calculated for each segment. The resulting high dimensional representation was then input to the SVM classifier. In the more recent report of Ref. [7] segmentation was dropped and the same summary statistics were calculated over the entire time series of the facial expression estimates (presence of AUs, intensity of given discrete expression, etc.). The resulting representation still had 16 dimensions. We speculate that one of the reasons for the relatively low and variable accuracy was that sample size was too small relative to dimensionality. We opted for a simpler linear ensemble model of lower dimensionality.

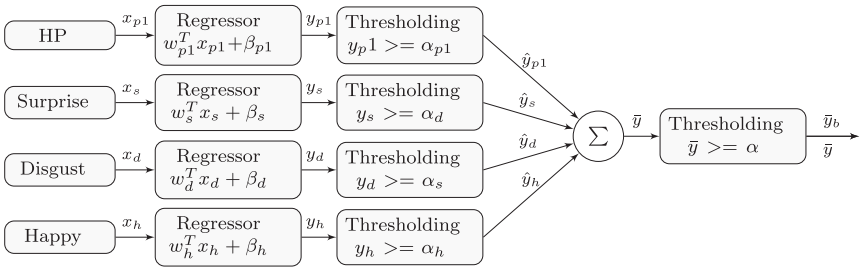

Fig. 5. Ensemble predictor. Inputs of the model are one head pose signal (x_{p1} and 3 facial expression related signals (x_s, x_d, x_h) as described above. An independent linear regressor is trained on each one dimensional signal using the *original* 4 point rating. The regressor outputs are binarized via thresholding for which optimal threshold value is learnt from the data. This binarization step acts as strong non-linear denoising. At the next stage the thresholded values are simply summed up and binarized again. To keep modeling simple, each input is assigned the same weight, but further optimization would yield signal specific weights. All of the model parameters are learned on the training set.

3. Results and discussion

We first report test results across all commercials, countries and product categories. We then report results for more fine-grained comparisons. These are models that 1) include only a single variant for related commercials, which eliminates any bias due to correlation among the sample commercials but may be influenced by the reduced number of commercials; and 2) models that differentiate between product categories and countries. We then compare the current findings with that of the McDuff-model. This comparison addresses our hypothesis that dynamic features enable increased accuracy and greater consistency across product categories.

For all comparisons, we report both accuracy and area under the receiver operating characteristics curve (ROC AUC). Accuracy is the sum of true positives and true negatives divided by all cases. It is intuitively appealing but difficult to interpret when distributions are imbalanced. In such cases, accuracy becomes a biased estimator of agreement between a classifier and ground truth [27]. ROC AUC quantifies the continuous relation between true and false positives. If higher rank is assigned to the "positive class" (in our case commercials that scored higher) then the area under the curve gives the probability that a randomly selected positive instance will be ranked higher than a randomly selected negative one. By definition ROC AUC is 0.5 for a random classifier. ROC AUC is unaffected by imbalance between positive and negative cases, although it may mask differences between classifiers in precision and recall [28,29]). In our data, class imbalance is mild when comparing across product categories and countries (56%), but often is larger when comparing between categories or countries. Thus, accuracy should be interpreted with caution.

To ensure that the trained models do not overfit [30], in which case models learn to represent noise components in the training data and become unpredictable in new data, we applied different validation schemes to assess generalization capacity of the trained models. Appropriate for the sample size, we used K-fold cross-validation (Kx-CV) in which samples are iteratively split into K disjoint training and test sets and the final performance metrics are averaged over the test sets. In the tests we used $K = 10$ folds and the procedure was repeated $n = 10$ times. From the repeated measurements we calculate confidence intervals at 95% confidence using t-statistics, which is better suited for small sample size. To help interpret the results, we also report a baseline which is a random model with a prior of the class probability of the training data.

As ads can be grouped along model independent factors like regions and product category, particular cross validations can be run where splits are defined by these factors. We will refer to these validation scheme as Leave One Label Out (LOLO) validation. These experiments test robustness of model performance against variations in those factors.

To enable comparison with results in Ref. [7] we also conducted Leave One Out (LOO) where test folds contain only one sample. Let us note, however, that for some metrics (ROC AUC in particular) LOO displays strange behavior when sample sizes become small [31].

We also report results for the case when only one ad variation is selected. While this data filtering may reduce potential ambiguity in the class membership, it reduces sample size, making training more difficult. To avoid any bias induced by arbitrary selections we ran nested cross-validation for ad selection in each group of the ad variations. The reported metrics are then averages over random ad selections.

3.1. Test results on all samples

The proposed model was trained and cross-validated on all commercials (N = 147) without respect to product category or country. ROC AUC was 0.75 ± 0.023. See Table 3 for comparison with random baseline.

Table 3
Cross-validation test (emotion and head pose signals + ensemble model) using all sample points. Performance is expressed in Accuracy and ROC AUC. Where appropriate we report confidence interval at 95% confidence as well.

Repeated 10-fold CV	Accuracy	ROC AUC
Our model	$73.9 \pm 2.2\%$	0.747 ± 0.025
Random baseline	$53.4 \pm 2.5\%$	0.50

3.2. Robustness against ad variants

When the dynamic model was trained and cross-validated without inclusion of variants (N = 116), ROC AUC remained about the same and confidence interval decreased from ± 0.025 to ± 0.01 In this setting we kept only one variation out of several options in each ad group. To counter bias due to random selections we repeat the random ad selection 10 times and run 10-fold CV for each random selection. See Table 4.

Results obtained are quite similar to those obtained on all data points. It indicates that in contrast to our original hypothesis about ambiguity in the labels, the ad variations indeed elicit different behavioral responses. In turn, variations can be considered as independent sample.

3.3. Robustness against category and country differences

To test how well our model generalize we modified the training testing procedure as follows. Training was done on all but one product category, testing on the one omitted, and then iteratively repeating training and testing for each category. This is referred to as leave-one-label-out cross-validation (LOLO validation). Similarly, the same iterative LOLO can be performed for country.

ROC AUC was fairly consistent over all but one categories (when model was tested on Confections, it obtained low ROC AUC, indicating that this one category behaves differently).

ROC AUC was also fairly similar in all but one countries (the only exception with low ROC AUC value was Russia which does not have a single top performing ad with rating 4).

See Tables 5 and 6.

3.4. Comparison of approaches

The approach proposed by Ref. [7] and our model presented here involved web-cam assessments of subjects' responses to the same product categories in four of the same countries. In both cases, sales lift data were provided by MARS, Incorporated. In both cases results were quantified at ROC AUC, but in Ref. [7] only LOO validation was reported, while we reported repeated 10-fold cross-validation. The two major differences between the approaches are the features that represent data and the applied classification model. The two approaches differed in other respects, as well, unrelated to types of

Table 4
Cross-validation test of the proposed approach (mix of emotions and dynamic head pose signals + ensemble model) using random selections of unique variations of the ads. (Sample size N = 116). Performance is expressed in Accuracy and ROC AUC. Where appropriate we report confidence interval at 95% confidence as well.

10-Fold CV	Accuracy	ROC AUC
Our model	$72.0 \pm 0.8\%$	0.732 ± 0.01
Random baseline	$53.8 \pm 1.0\%$	0.50

Table 5
Generalization performance of the proposed sales prediction model on different product categories. The validation scheme is LOLO so train fold does not contain samples from the category the test ads belong to. #low and #high denote the number of samples in the low and high performing classes, respectively.

Category	Acc.	ROC AUC	#low	#high
Confections	55.6%	0.608	23	22
Food	83.3%	0.800	10	2
Petcare	82.5%	0.834	36	21
Chewing gum	69.7%	0.744	13	20
Average	72.8%	0.747		

features, products, or countries. These differences, such as the number of commercials (fewer for our model) and the viewing period (more recent and over fewer years for our model), and other procedural aspects are unrelated to type of features. In comparing the results for each model, we remain mindful of these other sources of variation.

3.4.1. Statistical analysis

With this caveat in mind, we report the influence of the features on the classification performance. To help the comparison with the past reports on the static approach, the same RBF-SVM was trained on the set of features proposed in this study. Table 7 reports results for McDuff's signals as well as for ours as described in Section 2. The features are not exact replicas of the ones used in Ref. [7], but are very similar ("valence" metric, which is actually derived from the activation of other classifiers like smile, was replaced by our own disgust classifier outputs, eye brow raise was replaced by our own surprise classifier). Also included are separate results for representations using only head pose information and representation using only facial expression information (based on smile, surprise and disgust dynamics). For our proposed model, performance was better when head and face dynamics were combined rather than used exclusively. This suggests that the packaging of nonverbal behavior, head pose and motion, independently contributes to predicting sales lift. For both LOO and 10-fold cross-validation, our combined representation produced much higher performance, while using McDuff's representation yielded about random chance performance. This finding emphasizes the importance of head pose information and session level analysis. The magnitude of the difference between the representations suggests that procedural differences (such as number of commercials viewed) play at most a minor role. Further research is needed to evaluate this matter. We also report the number of support vectors (#SV) kept after training as an indicator of generalization problems. For 147 samples in 10-fold cross validation scheme, the size of a training fold is about 132. An SVM model cannot generalize well if #SV is as large as the entire training fold. The results confirmed our assumption that low performance as reported in Ref. [7] is due to the fact that classification of high dimensional representations by non-linear SVM requires more data.

Table 6
Generalization performance of the proposed sales prediction model on ads from different regions. The validation scheme is LOLO so train fold does not contain samples from the region the test ads belong to. #low and #high denote the number of samples in the low and high performing classes, respectively.

Region	Acc.	ROC AUC	#low	#high
Australia	77.8%	0.753	18	9
France	66.7%	0.795	8	7
Germany	80.9%	0.889	9	12
Russia	68.2%	0.533	15	7
UK	72.7%	0.789	19	14
USA	72.4%	0.707	13	16
Average	73.1%	0.744		

Table 7
Impact of the different representations on the classification performance. The classifier is the same SVM with non-linear radial basis function kernel. This comparison also shows the complementary nature of head pose and facial expression information. SVM classifier achieves the highest performance with the Combined signals and the resulting models have lower complexity (fewer Support Vectors) than the model using McDuff's signals. The best trade off solution is denoted with bold letters.

Validation	Signal	ROC AUC	#SV
LOO	Head pose	0.685	127
	Facial expressions	0.623	107
	Combined signals	0.732	122
	McDuff's signals	0.503	130
10-Fold CV	Head pose	0.610 ± 0.021	90
	Facial expression	0.677 ± 0.023	96
	Combined signals	$\mathbf{0.701} \pm 0.021$	109
	McDuff's signals	0.580 ± 0.023	118

The ensemble model not only performed better on the combined signal than the SVM model of McDuff (0.747 ± 0.025 versus 0.701 ± 0.021), but it is markedly simpler (as indicated by the number of parameters in the two trained models). In turn it is expected to result in smaller generalization error on unseen data. Another advantage is that improvement by adding other behavioral signals increases model complexity in a well controlled way thus preserving generalization of the improved model.

4. Conclusion

One of the biggest challenges in today's market research is the exponential growth of the number of media contents to be analyzed since traditional survey based methods do not scale well. In addition, those methods fail to capture the important emotional aspects of the interaction between content and consumers.

We have created a feasible data acquisition system that allows for large scale behavioral data collection and analysis for practical market research. We have also trained a classification model that learned to distinguish ads with high and low sales performance. Although the size and structure of the training data are limited we managed to show that the learned model generalizes well over some factors not used in the modeling. These promising results may pave the way for a new generation of automated, cost-efficient, behavioral cue driven market research tools for analysis.

To further improve methodology, several limitations need to be addressed. Behavioral analysis is based on average responses assuming that individual differences are just random perturbations. However, it is more likely that these individual differences carry relevant information about the differences between the ads. Another limitation is that our model does not allow for more complex interactions between observations. Once more samples are available our method can be extended to include more features and it can also capture linear or non-linear interactions between features (generalized stepwise linear regression models can systematically check pairwise or higher order interactions between features). Finally, hybrid models that test conscious recollection and immediate behavioral–emotional responses must be developed to fully understand the impact of ads on consumer behavior.

Acknowledgment

This work was financially supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA — Automatic Sentiment Analysis in the Wild). The authors would like to thank the Ehrenberg-Bass Institute for Marketing Science for helping in defining the data collection experiment,

and MARS, Incorporated for providing the valuable sales lift data that made this research possible. The authors are especially grateful for the anonymous reviewers for their concerns and comments on transparency and simplicity.

References

[1] statista.com, U.S. Advertising Industry — Statistics & Facts, 2015. [Online]. http://www.statista.com/topics/979/advertising-in-the-us/.
[2] J. Liaukonyte, T. Teixeira, K.C. Wilbur, Television advertising and online shopping, Market. Sci. 34 (3) (2015) 311–330. http://dx.doi.org/10.1287/mksc.2014.0899.
[3] R.B. Zajonc, Feeling and thinking: preferences need no inferences, Am. Psychol. 35 (2) (1980) 151–175. http://dx.doi.org/10.1037/0003-066x.35.2.151.
[4] A.R. Damasio, Descartes' Error: Emotion, Reason, and the Human Brain, 1st ed., Harper Perennial. 1995.
[5] K.R. Scherer, Feelings integrate the central representation of appraisal-driven response organization in emotion, in: A.S.R. Manstead, N. Frijda, A. Fischer (Eds.), Feelings and Emotions, Cambridge University Press. 2004, pp. 136–157. cambridge Books Online. http://dx.doi.org/10.1017/CBO9780511806582.009.
[6] D. McDuff, R. el Kaliouby, J.F. Cohn, R.W. Picard, Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads, IEEE Trans. Affect. Comput. 6 (3) (2015) 223–235. http://dx.doi.org/10.1109/TAFFC.2014.2384198.
[7] D.J. McDuff, Crowdsourcing Affective Responses for Predicting Media Effectiveness, Massachusetts Institute of Technology Cambridge, MA, USA, 2014. (Ph.D. thesis).
[8] D. McDuff, R.E. Kaliouby, E. Kodra, L. Larguinet, Do emotions in advertising drive sales? Proceedings of ESOMAR Congress, 2013.
[9] Z. Hammal, J.F. Cohn, C. Heike, M.L. Speltz, What can head and facial movements convey about positive and negative affect? 2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21–24, 2015, 2015. pp. 281–287. http://dx.doi.org/10.1109/ACII.2015.7344584.
[10] H. Dibeklioğlu, Z. Hammal, Y. Yang, J.F. Cohn, Multimodal detection of depression in clinical interviews, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15,, ACM, New York, NY, USA, 2015, pp. 307–310. http://dx.doi.org/10.1145/2818346.2820776.
[11] D. Keltner, The signs of appeasement: evidence for the distinct displays of embarrassment, amusement, and shame, J. Pers. Soc. Psychol. (1995) 441–454.
[12] Z. Ambadar, J.F. Cohn, L.I. Reed, All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous, J. Nonverbal Behav. 33 (1) (2008) 17–34. http://dx.doi.org/10.1007/s10919-008-0059-5.
[13] J. Rottenberg, R.D. Ray, J.J. Gross, Emotion Elicitation Using Films, Oxford University Press, London, 2007. Ch. 2.
[14] S. Dolnicar, B. Grn, F. Leisch, Quick, simple and reliable: forced binary survey questions, Int. J. Mark. Res. 53 (2) (2011) 231–252.
[15] M. Vriens, M. Wedel, Z. Sandor, Split-questionnaire designs: a new tool in survey design and panel management, Mark. Res. 13 (1) (2001) 14–19.
[16] J. Orozco, B. Martinez, M. Pantic, Empirical analysis of cascade deformable models for multi-view face detection, Image Vis. Comput. 42 (2015) 47–61. http://dx.doi.org/10.1016/j.imavis.2015.07.002.
[17] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[18] M.G. Calvo, A. Gutiérrez-García, A. Fernández-Martín, L. Nummenmaa, Recognition of facial expressions of emotion is related to their frequency in everyday life, J. Nonverbal Behav. 38 (4) (2014) 549–567. http://dx.doi.org/10.1007/s10919-014-0191-3.
[19] M. LaFrance, M.A. Hecht, The social context of nonverbal behavior. Studies in emotion and social interaction, Ch. Option or Obligation to Smile: The Effects of Power and Gender on Facial Expression, Editions de la Maison des Sciences de l'Homme & Cambridge University Press. 1999, pp. 45–70.
[20] P.O. Glauner, Deep Convolutional Neural Networks for Smile Recognition, Imperial College London, London, UK, 2015. Master's thesis.
[21] M. Slaney, A. Stolcke, D. Hakkani-Tür, The relation of eye gaze and face pose: potential impact on speech recognition, Proceedings of the 16th International Conference on Multimodal Interaction, 2014. pp. 144–147.
[22] A. Doshi, M.M. Trivedi, Head and eye gaze dynamics during visual attention shifts in complex environments, J. Vis. 12 (2) (2012) 9. http://dx.doi.org/10.1167/12.2.9.
[23] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (1) (2009) 1–39. http://dx.doi.org/10.1007/s10462-009-9124-7.
[24] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. 11 (1999) 169–198.
[25] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc., New York, NY, USA, 1995.
[26] C. Cortes, V.N. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297. http://dx.doi.org/10.1007/BF00994018.
[27] L.A. Jeni, J.F. Cohn, F. De La Torre, Facing imbalanced data—recommendations for the use of performance metrics, Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII '13,, IEEE Computer Society, Washington, DC, USA, 2013, pp. 245–251. http://dx.doi.org/10.1109/ACII.2013.47.
[28] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.
[29] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966.
[30] P.I. Good, J.W. Hardin, Common Errors in Statistics (and How to Avoid Them), John Wiley & Sons. 2012.
[31] A. Airola, T. Pahikkala, W. Waegeman, B.D. Baets, T. Salakoski, A comparison of AUC estimators in small-sample studies., in: S. Dzeroski, P. Geurts, J. Rousu (Eds.), MLSB, Vol. 8 of JMLR Proceedings, JMLR.org, 2010. pp. 3–13.