Image and Vision Computing 32 (2014) 630–640

Contents lists available at ScienceDirect

Image and Vision Computing

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

Automatic measurement of ad preferences from facial responses gathered over the Internet☆

Daniel McDuff [a,1], Rana El Kaliouby [b], Thibaud Senechal [b], David Demirdjian [c], Rosalind Picard [a]

[a] MIT Media Lab, Cambridge 02139, USA
[b] Affectiva, Waltham 02452, USA
[c] CSAIL, MIT, Cambridge 02139, USA

ARTICLE INFO

Article history:
Received 2 June 2013
Received in revised form 29 October 2013
Accepted 9 January 2014
Available online 17 January 2014

Keywords:
Facial expressions
Preferences
Classification
Crowdsourcing

ABSTRACT

We present an automated method for classifying "liking" and "desire to view again" of online video ads based on 3268 facial responses to media collected over the Internet. The results demonstrate the possibility for an ecologically valid, unobtrusive, evaluation of commercial "liking" and "desire to view again", strong predictors of marketing success, based only on facial responses. The area under the curve for the best "liking" classifier was 0.82 when using a challenging leave-one-commercial-out testing regime (accuracy = 81%). We build on preliminary findings and show that improved smile detection can lead to a reduction in misclassifications. Comparison of the two smile detection algorithms showed that improved smile detection helps correctly classify responses recorded in challenging lighting conditions and those in which the expressions were subtle. Temporal discriminative approaches to classification performed most strongly showing that temporal information about an individual's response is important; it is not just how much a viewer smiles but when they smile. The technique could be employed in personalizing video content that is presented to people while they view videos over the Internet or in copy testing of ads to unobtrusively quantify ad effectiveness.

1. Introduction

The face has been shown to communicate discriminative valence information with zygomatic muscle (AU12) activity greater in ads with a positive emotional tone and corrugator muscle (AU4) activity greater in ads with a negative tone [3]. There is evidence that facial expressions can predict variables related to advertising success, with facial responses correlated with recall [8] and ad "zapping" [26]. The Facial Action Coding System (FACS) [6] is a catalog of 44 unique action units (AUs) that correspond to each independent movement of the face's 27 muscles. Computer vision systems can now reliably code many of these actions automatically [30]. In this paper we show that self-reported video advertisement liking and desire to view again can be accurately predicted from automatically detected spontaneous smile (AU12) responses captured in unconstrained settings over the Internet. Fig. 1 shows the framework we use to automatically classify media preferences.

Advertisement likability is a key measure of sales success in marketing [7,23]. Likability is described as having the dimensions of entertainment, energy, relevance, empathy, irritation and familiarity. However, these metrics are hard to quantify objectively and in many real-life applications self-report measures are impractical to capture (e.g. when people are watching TV). Advertisers wish to increase a viewer's desire to view an advertisement again; thus desire to view the ad is another measure of advertising effectiveness. Knowledge of likability and desire to view again are not only useful in advertisement copy-testing but could also be used to personalize the content viewers are shown when watching TV over the Internet using platforms such as Netflix or Hulu. In the case of humorous ads, smile activity is a good measure of positive advertisement attitude or liking, and this can be measured continuously and unobtrusively from video images [14].

Earlier work has shown that facial responses to content can be collected efficiently over the Internet, and that there are significant differences in the aggregate smile responses of groups that report liking a commercial compared to those that report disliking it [16]. A similar difference in the aggregate responses was observed between individuals who report a desire to watch the content again versus those that report no such desire. However, whether these aggregate trends allow accurate discrimination of liking versus disliking responses on an individual basis was not explored. The first published automated analysis on individual level prediction has shown that it is possible to accurately predict preferences [17]. This paper presents work extending and improving these initial results.

The dynamics of smile responses are rich and can be used to distinguish between different message judgments associated with them [2,9]

☆ This paper has been recommended for acceptance by Stan Sclaroff.
URL: djmcduff@media.mit.edu, http://media.mit.edu/~djmcduff (D. McDuff), kaliouby@affectiva.com (R. El Kaliouby), thibaud.senechal@affectiva.com (T. Senechal), demirdji@csail.mit.edu (D. Demirdjian), picard@media.mit.edu (R. Picard).
[1] Tel.: +1 617 253 8628.

0262-8856/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2014.01.004

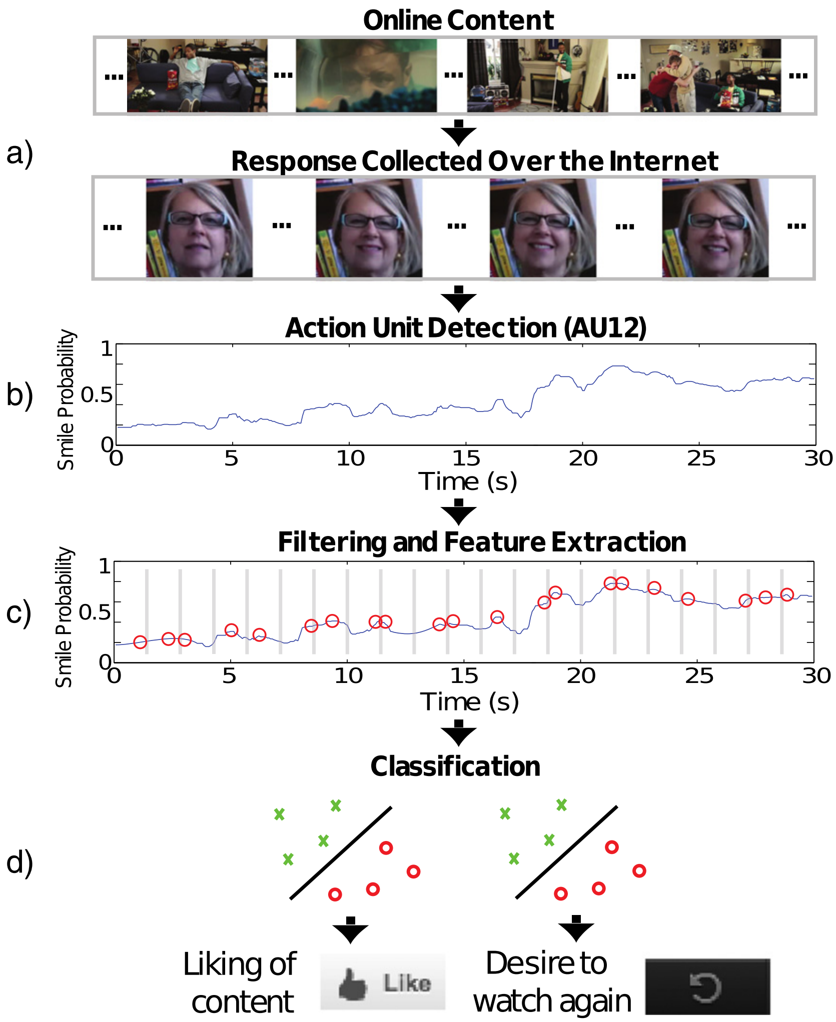


Fig. 1. Framework for classification of content liking and desire to view again based on automatically detected smile responses recorded over the web. a) Facial response to online media captured, b) smiles (AU12) detected automatically, c) temporal features extracted from smoothed smile track, d) features used to predict liking and desire to view again.

and whether they are posed or naturalistic [27]. Discriminative temporal models, in particular, Hidden Conditional Random Fields (HCRFs) and Latent Dynamic Random Fields (LDCRFs) have been shown to perform well in classification of noisy behavioral data [28,24]. In this work we test the power of both static and dynamic models to predict advertisement liking and desire to watch again from an ecologically valid, but challenging and noisy, dataset of automatically detected spontaneous smile responses. Many applications of facial expression recognition for predicting consumer preferences or highlights are tested on data collected within a lab setting which is unlike the environment in which the media is normally viewed [10,31]. In our work we use data collected in a setting much closer to that in which it is normally consumed.

The main contributions of this paper are: 1) to present results on classification of liking and desire to view again of Internet videos based on the facial responses analyzed over the Internet, 2) to identify conditions under which misclassifications (false positives, false negatives) occurred and 3) to show that improved smile detection can reduce the number of misclassified responses considerably. The remainder of the paper will discuss the data collection, feature extraction, modeling and results of the work.

2. Related work

Smile detection is one of the most robust forms of automated facial analysis available. Whitehall et al. [29] present a smile classifier based on images collected over the Internet and demonstrates strong performance on this dataset. A subset of the data was released as the MPL GENKI[2] dataset. Shan [22] demonstrates an accurate and faster smile detector on the MPL GENKI-4K dataset.

Joho et al. [10] showed that it is possible to predict personal highlights in video clips by analyzing facial activity. However, they also noted the considerable amount of individual variation in responses. These experiments were conducted in a laboratory setting and not in a natural context; our work demonstrates the possibility of extending this work to online content and real-world data. Zhao et al. [31] designed a video indexing and recommendation system based on automatically detected expressions of the six basic emotions (sadness, anger, fear, disgust, happiness, surprise). However, this was tested on only a small number of viewers ($n = 10$) in a lab setting.

Teixeira et al. [25] showed that inducing affect is important in engaging viewers in online advertising and is associated with reducing their frequency of “zapping” (skipping the advertisement). They demonstrated that joy, as measured by smile responses, was one of the states that increased viewer retention in the commercial. Again, these studies were performed in a laboratory setting rather than in the wild. Micu and Plummer [19] measured zygomatic major (AU12) activity using facial electromyography (EMG) while people watched TV ads. They

[2] http://mplab.ucsd.edu, The MPLab GENKI Database, GENKI-4K Subset.

showed that physiological measurements capture different information from self-reported responses. This aligns with [11] that shows facial expressions are related to self-reported experiences but the two measurements do not capture exactly the same information. Our real-world data supports these findings.

3. Data collection

Using a web-based framework similar to that described in [16] 3268 videos (2,615,800 frames) were collected of facial responses to three successful Super Bowl ads: 1. Doritos ("House sitting", 30 s), 2. Google ("Parisian Love", 53 s) and 3. Volkswagen ("The Force", 62 s). All three ads were somewhat amusing and were designed to elicit smile or laughter responses. In addition the VW ad was voted the most successful ad of the year 2011.[3] The responses were collected in natural settings via the Internet and the application was promoted on the Forbes website [1]. In total 6729 people opted-in and completed the experiment. For the automatic analysis here videos for which it was not possible to identify and track a face in at least 90% of the frames were disregarded; this left 4502 videos (67%). All videos were recorded with a resolution of 320 × 240 and a frame rate of 14 fps. Participants were aware from the permissions that their camera would be turned on and this may have had an impact on their facial response. In addition, we must consider the self-selection bias that may occur as the viewers were not recruited to reflect a specific demographic profile. However, it was observed that people responded naturally in a vast majority of cases. McDuff et al. [15] provide more detailed information about the data collection and quantify a number of characteristics of the videos recorded.

Following each ad, viewers could optionally answer three multiple choice questions: "Did you like the video?" (liking), "Have you seen it before?" (familiarity) and Would you watch this video again?" (rewatchability). Fig. 2 shows a screenshot of the questions asked. In this paper the relationship between the smile responses and the self-report responses for each question is examined. Since viewers were not obligated to complete the responses and the questions "timed out" once the smile response was computed, some participants only answered some of the questions and some none of the questions. On average each question was answered by 47.6% of viewers, which still provides almost 2400 labeled examples for each question and commercial combination.

4. Smile detection and dynamics

To compute the smile probability measure we used custom algorithms developed by Affectiva. In initial experiments version 1 (V1) of the smile detection algorithms was used. However, it was found that this did not capture the smile responses accurately in some cases, leading to misclassifications of liking [17]. As a result a second version (V2) was developed in order to improve performance. We describe the algorithms and compare the performance of the two smile detectors to quantify the impact of more accurate smile detection on the prediction of viewer preferences, showing improved preference prediction with the better smile detector V2. In each case using the algorithms a 1-dimensional smile track was computed for each video with length equal to the number of frames of the video. These smile tracks, and the corresponding self-report liking response labels, are used for the analysis and classification in the rest of the paper.

4.1. Smile detector V1

Detector V1 tracks a region around the mouth using the Nevenvision facial feature tracker[4] and computes Local Binary Pattern (LBP) [21] features within this region. An ensemble of bagged decision trees is used for classification. The classifier outputs a probability that the expression is a smile. A smile probability value (between 0 to 1) is calculated for every frame in which a face was tracked, yielding a one-dimensional smile track for each video. Fig. 3 (top) shows an example of one smile track with screenshots of six frames and demonstrates how the smile probability is positively correlated with the intensity of the expression.

The smile classifier was trained on examples from the CK+ and MPL[5] databases. The images from these datasets were labeled for the presence of a smile by three labelers with the majority label taken. We tested how well the smile classifier performs on crowdsourced face videos from a webcam where there is no control on the quality of the resulting face videos (these videos were taken from the study described here and are from the labeled public portion of the AM-FED dataset [18]). In total 52,294 frames were labeled with ground truth labels (presence of a smile). Three labelers labeled every frame of each video and the majority label was taken for each frame. The resulting ROC curve is shown in Fig. 4 (blue line); the area under the curve is 0.874. The resulting precision-recall curve is shown in Fig. 4 (blue line); the area under the curve is 0.73.

4.2. Smile detector V2

Detector V2 tracks the whole face region using the Nevenvision facial feature tracker and computes Histogram of Oriented Gradient (HOG) [5] features within this region. A support vector machine (SVM) with radial basis function kernel is used for classification. The signed distance of the sample from the classifier hyperplane is taken and normalized using a monotonic function that in the training phase rescaled points between [0–1]. Thus a smile value (between 0 and 1) is calculated for every frame in which a face was tracked, yielding a one-dimensional smile track for each video. Fig. 3 (bottom) shows an example of one smile track with screenshots of six frames and demonstrates how the smile detector output is positively correlated with the intensity of the expression.

The smile classifier was trained on images from crowdsourced face videos from a webcam where there is no control on the quality of the resulting face videos (these videos were from a similar but different study to the one described here). The images from the data were labeled for the presence of a smile by three labelers with the majority label taken. In total over 2,000,000 frames were randomly selected for ground truth labeling of smiles. These were taken from webcam videos recorded in 20 separate studies across Asia, Europe and America. The detector was tested on the same labeled frames as detector V1 from the AMFED dataset. The resulting ROC curve is shown in Fig. 4 (green line); the area under the curve is 0.91. The resulting precision-recall curve is shown in Fig. 4 (green line); the area under the curve is 0.80.

The results show that the smile detector V2 performs more robustly than detector V1 on the challenging webcam videos. The overall performance is strong and based on a very large number of images representing very diverse conditions.

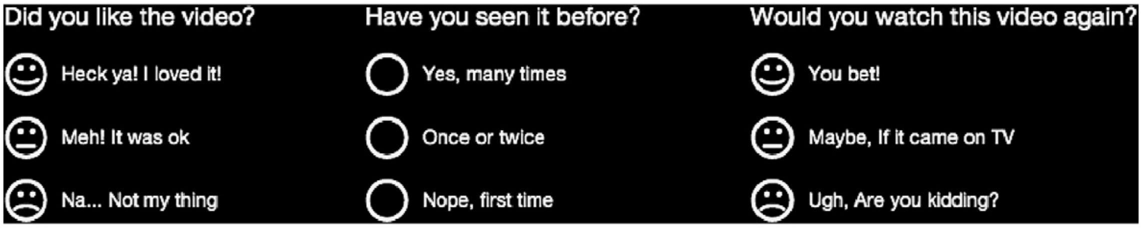
4.3. Appearance vs. geometric features for smile detection

It is worth mentioning the performance of a smile detector that uses geometric features from the facial landmark detector in order to provide some justification for the use of appearance based features in this work. Appearance-based features (e.g. HOG features) were initially chosen as they have been shown to perform more strongly in cases in which registration of the face is challenging [13]. With low resolution webcam videos this is often the case. Fig. 5 shows examples of smile tracks calculated using geometric features and HOG appearance features. Cropped images of the faces at intervals during each video are shown above the smile tracks. The geometric features were calculated from the landmark points identified by the tracker. The absolute distance

[3] http://www.adweek.com/

[4] Licensed from Google, Inc.

[5] http://mplab.ucsd.edu, The MPLab GENKI Database, GENKI-4K Subset.



Did you like the video?

- Heck ya! I loved it!
- Meh! It was ok
- Na... Not my thing

Have you seen it before?

- ○ Yes, many times
- ○ Once or twice
- ○ Nope, first time

Would you watch this video again?

- You bet!
- Maybe, If it came on TV
- Ugh, Are you kidding?

Fig. 2. The self-report questions that the viewers were presented with after watching the commercial.

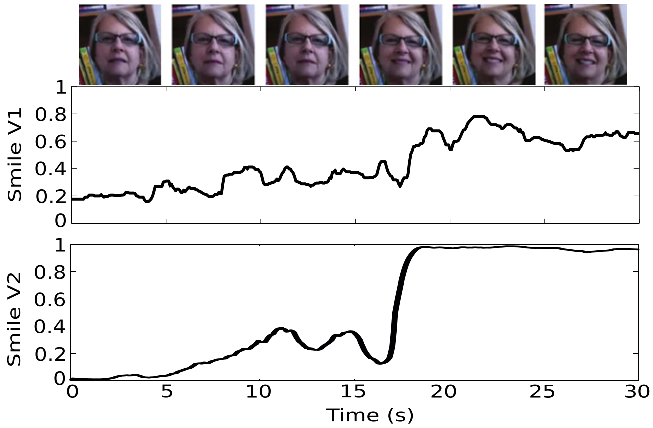


Fig. 3. A smile track with screenshots of the response, demonstrating how greater smile intensity is positively correlated with the output from the smile detectors. Top) Smile track using smile detector V1, bottom) smile track using smile detector V2.

between the lip corners is normalized by the distance between the eye corners (which are rigid points on the face). The resulting measurements are warped to a scale from 0 to 100 based on the minimum and maximum values observed in the data.

From Fig. 5 it is clear that the smile tracks calculated from geometric features and the appearance based features generally have similar shapes e.g. (a) and (b). However, there are other cases where subtle smiles are not detected using the geometric features as they do not result in large movements (c) or in which false positives occur due to noisy registration (d). We felt that due to the challenging registration issues with this data that appearance features represented a more robust option.

5. Feature extraction

Only smile tracks that satisfied the 90% trackable criteria (a face could be positively identified in greater, or equal to, 90% of frames) were used. First they were filtered with a low-pass FIR filter (with Hamming window) to smooth the signals. The 3 dB cut-off frequency of the low-pass filter was 0.75 Hz. Secondly, features were extracted from the smile tracks as shown in Fig. 6. The filtered tracks were divided evenly into 20 segments and the peak smile intensity for each segment calculated to form a feature vector of length 20. This removed any prior information held in the length of the content and will help promote generalizability of the resulting model. We present an experiment in

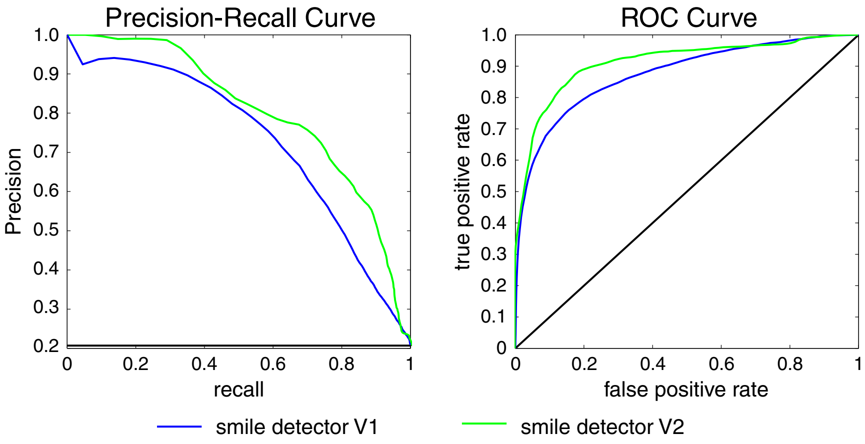


Fig. 4. Performance characteristics for the two smile detectors. Left) Precision-recall curves for detector V1 (blue) and detector V2 (green) tested on images from the AMFED dataset (no. of images = 52,294). Right) ROC curves for detector V1 (blue) and detector V2 (green) tested on the same images from the AMFED dataset.

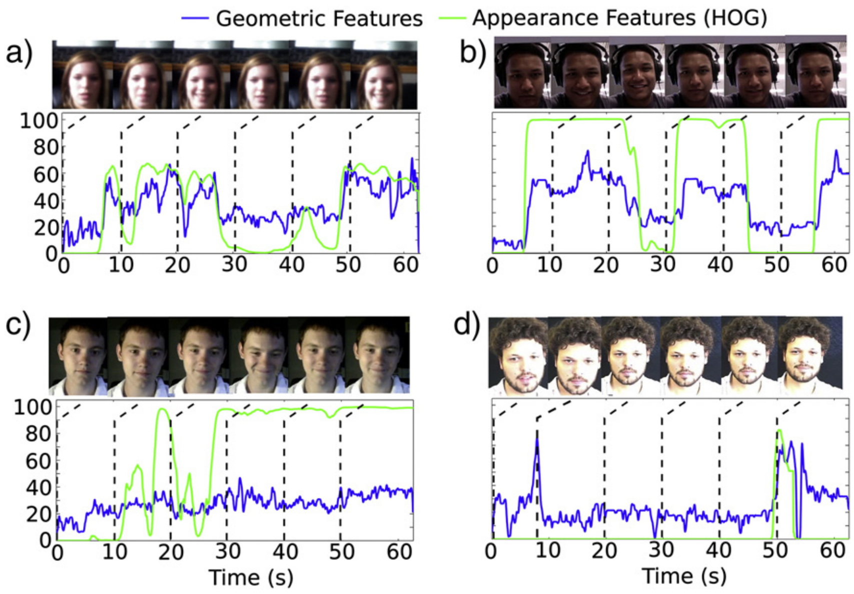


Fig. 5. Examples of smile tracks calculated using geometric features and HOG appearance features (smile detector V2) for four of the response videos. Cropped images of the faces at intervals during each video are shown above the graphs.

Section 7.4 to justify the choice of 20 bins. The videos have a frame rate of 14 fps and therefore the number of frames for each segment was 21 (Doritos), 37 (Google) and 43 (VW). Tests were run with more than 20 features but there was no significant change of performance of the classifiers.

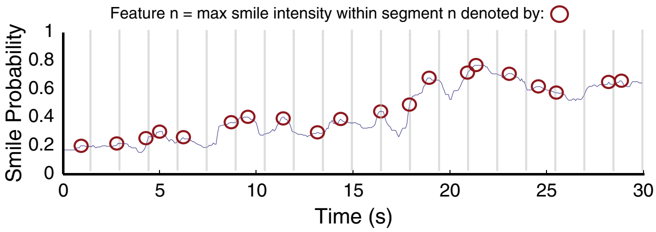


Fig. 6. Feature extraction involved selecting the maximum peaks from the filtered smile track within 20 evenly spaced temporal bins. This formed a feature vector of length 20.

In order to reveal more about how facial responses to the content are related to the self-report responses we perform a logistic regression between each of the twenty smile track features and the self-report labels. We take all the smile track features from the first temporal bin and perform a logistic regression between these and the labels, then we take all the smile track features from the second temporal bin and so on. This will help reveal which parts of the temporal response show strongest correlation with the reported liking or disliking. Fig. 7 shows the inverse of the deviance for the models, as a measure of the goodness of fit, using each of the twenty features. It is clear that the smile features towards the end of the response (especially within the final 25% of each ad) are much more highly correlated with the liking self-report labels. This agrees with the previous work showing that post-hoc self-report responses are impacted more by experiences towards the end [19]. However, it should also be noted that the beginning of the ads tended to set-up the context and the end of the ads had the more obviously amusing content which would also contribute to this result.

6. Liking and desire to watch again

In this paper the responses to the questions “Did you like the video?” and “Would you watch this video again?” are considered. The answers available for the first question were: “Heck ya! I loved it!” (liking), “Meh! It was ok” (neutral) and “Na… not my thing” (disliking). The answers available for the second question were: “You bet!” (strong

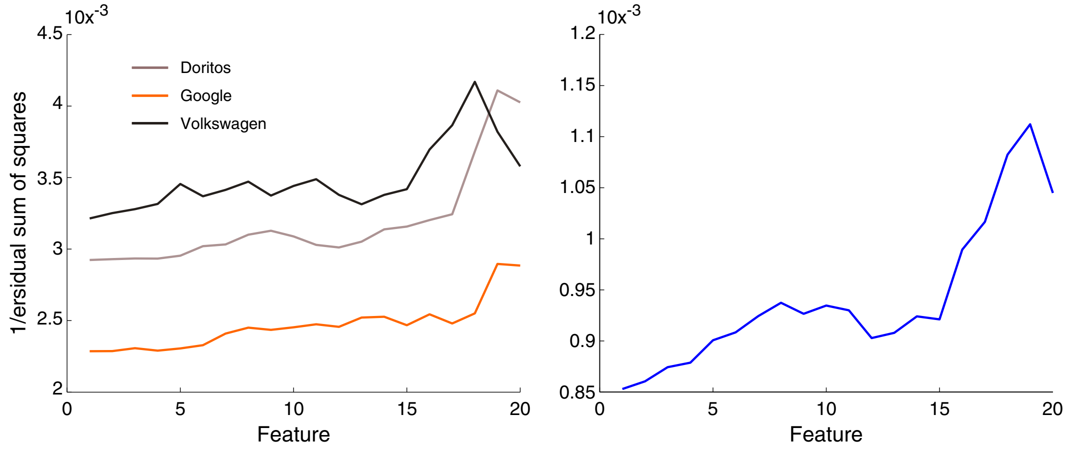


Fig. 7. Inverse of the residual sum of squares for a logistic regression between each of the features and the liking self-report labels. Left) For each of the ads separately, right) for all the ads combined.

Table 1
Number (and percentages) of "liking" and "desire to view again" classes for each of the ads.

Ad	Liking		Desire to view again	
	Like (+ve class)	Dislike (−ve class)	Strong (+ve class)	Weak (−ve class)
Doritos	280 (80%)	70 (20%)	91 (73%)	34 (27%)
Google	341 (79%)	92 (21%)	142 (63%)	84 (37%)
VW	711 (95%)	40 (5%)	381 (89%)	46 (11%)
Total	**1332 (86%)**	**212 (14%)**	**614 (79%)**	**164 (21%)**

desire), "Maybe, if it came on TV" (mild desire) and "Ugh, are you kidding?" (weak desire).

For this analysis we consider the binary cases in which the viewer would report liking vs. disliking the commercial and strong vs. weak desire to view the commercial again. Therefore the neutral and mild responses are not considered in the classification. This is reasonable as the people in these categories do not show a strong feeling towards the content and therefore showing them the content again, or not, will not represent a missed opportunity or negative consequence (misclassification for this group represents a low cost), whereas showing someone the content again who had a very weak desire to see it may be a waste of resources or have a negative impact on a viewer's perception of the brand. In addition, the majority of viewers in this data reported either a liking or disliking response and not a neutral response. Table 1 shows the number of samples, class priors, for each of the classes.

7. Experiments and results

7.1. Comparing models

We compare both static and temporal, generative and discriminative approaches in predicting reported liking and desire to view again based on smile features. For the classification we attempt to correctly classify examples of disliking and liking responses and strong and weak desire responses all of which satisfy the 90% trackable criteria.

Before performing the classification experiments we computed a 2D mapping of the data to get an intuition about the manifold responses. Fig. 8 shows a 2D mapping of the responses computed using Linear Discriminant Analysis (LDA). The reported "disliking" responses are shown in red and the reported "liking" in green. Examples of four of the smile responses are also shown. Gaussian distributions have been fitted to the data along the most discriminative axis. The distributions of the two classes are different with the disliking class characterized by a lower mean intensity and in particular a much lower response towards the end of the commercial. The global gradient, or trend, of the responses is also an important feature. However, it is clear that global features seem insufficient for accurate classification, for instance the mean of example b in Fig. 8 is greater than that of example c despite being a disliking response compared to a liking response. Looking at the temporal profiles examples c and d are the closest to one another.

We compared the following types of models in predicting reported liking and desire to view again based on smile features:

Naive (Class Priors): We provide a naive baseline for which test labels are predicted randomly but with a distribution that matches the training data class priors. Table 1 shows the class priors for each of the ads. The results shown for the class priors model are based on an average performance of over 200 sets of predictions.

Naive Bayes: A Naive Bayes (NB) classifier was used to provide a baseline performance. As the Naive Bayes classifier is non-parametric no validation step was performed.

Support Vector Machine: Support Vector Machines (SVM) are a static approach to classification and therefore do not explicitly model temporal dynamics. A Radial Basis Function (RBF) kernel was used. During validation the penalty parameter, C, and the RBF kernel parameter, γ, were each varied from 10^k with $k = -3, -2, \ldots, 3$. The SVM's were implemented using libSVM [4].

Hidden Markov Models: An HMM was trained for each of the two classes. This is a generative approach to modeling the data. During validation, the number of hidden states (3, 5 and 7) was varied.

Hidden-state Conditional Random Fields: HCRFs [28] and Latent Dynamic Condition Random Fields (LDCRFs) [20] are discriminative approaches to modeling temporal data. The CRF model and its

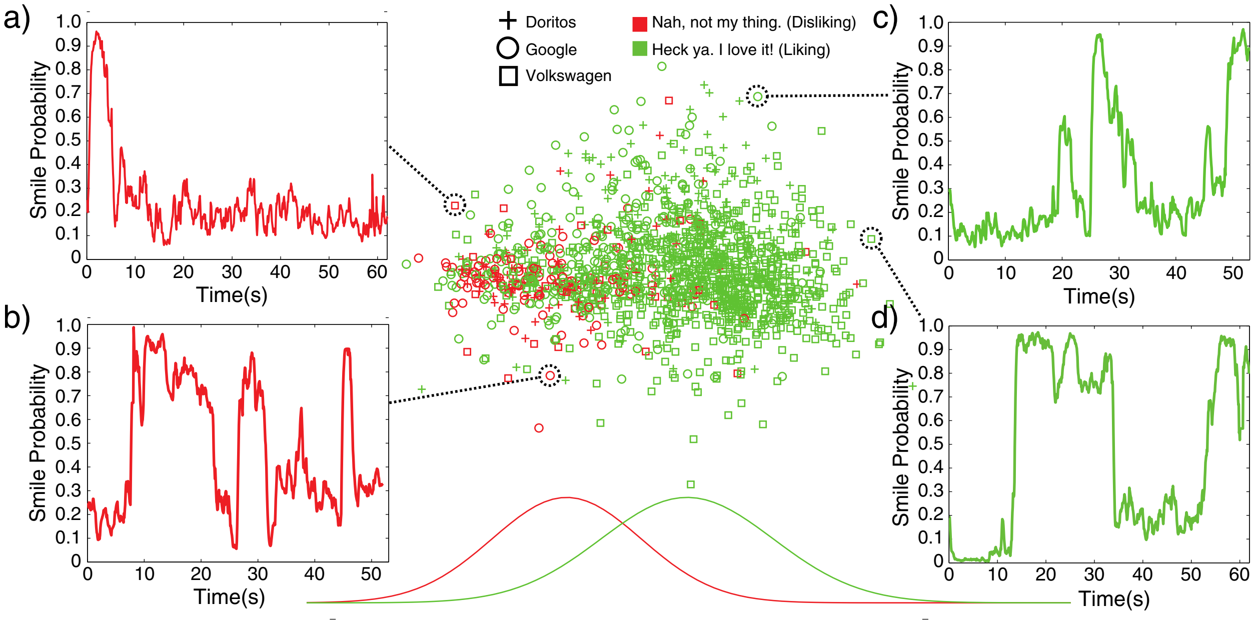


Fig. 8. Linear Discriminant Analysis (LDA) mapping of the smile tracks (calculated using smile detector V1) with labels of reported liking. Smile tracks for those that report disliking (red) and liking (green). Examples of four of the smile tracks are shown.

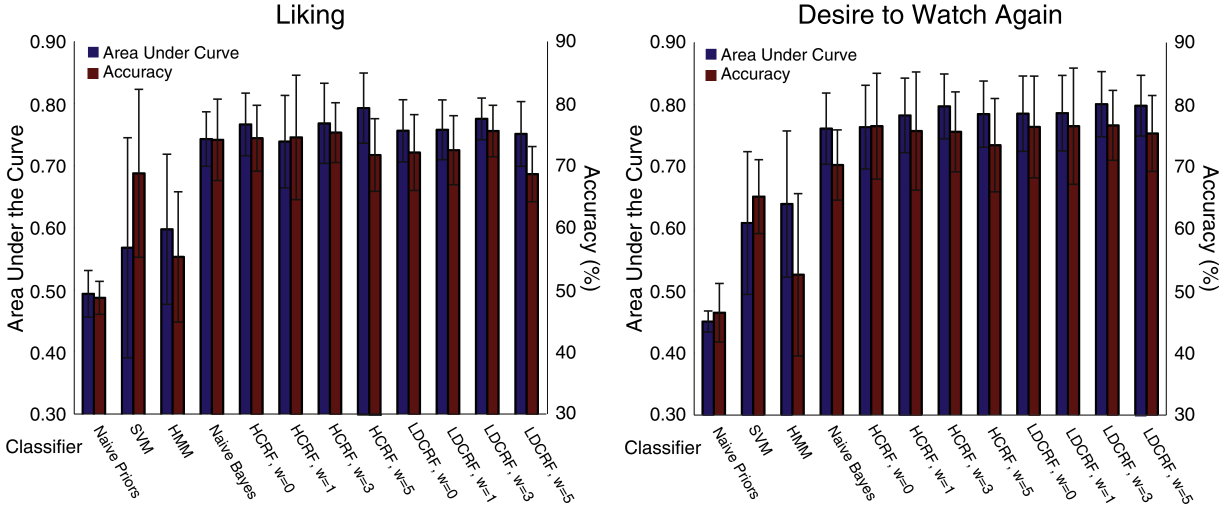


Fig. 9. Bar chart of area under the curve (AUC) and accuracy (%) for predicting liking (left) desire to view content again (right).

variants remove the independence assumption made in using Hidden Markov Models (HMMs) and also avoid the label-biasing problem of Maximum Entropy Markov Models (MEMMs) [12]. The dynamics of smiles are significant in distinguishing between their meanings [2,9]; as such, we hypothesized a potential benefit in explicitly modeling the temporal dynamics of the responses. Song et al. [24] describe a Gaussian temporal-smoothing kernel that improved performance without increasing the computational complexity of inference. This takes a Gaussian-weighted average of observations within a moving window of size N. The HCRF and LDCRF classifiers were tested with Gaussian temporal smoothing window sizes 0, 1, 3 and 5. We found that beyond a window size of 5 performance became saturated or began to fall off. During validation, the regularization factor (10^k with $k = -2, -1,\ldots, 2$) and number of hidden states (3, 5 and 7) was varied. The HCRF and LDCRF classifiers were implemented using the HCRF toolbox for MATLAB [20].

Validation: It is important that the prediction of success is independent of the particular ad and that the features generalize across new content. In order to test this we use a challenging leave-one-ad-out scheme. The data for one ad were removed for testing and the remaining data were used for training and validation. This was repeated for each of the three ads. For validation a leave-one-ad-out methodology was used. The training data set was split and data for one ad were withheld and validation performed to find the optimal model parameters, this was repeated for both ads in the training set. The area under the curve (AUC) was maximized in the validation stage to choose parameters.

Results: Fig. 9 shows the area under the ROC curve for each of the classifiers. The accuracy for the classifiers closest to the (0, 1) point on the ROC curve is also shown. For both liking and desire to view again the LDCRF classifier with Gaussian smoothing window, ω, of 3 shows the strongest performance with an area under the curve of 0.8. The HCRF classifier, $\omega = 3$, showed comparable performance and with the reduced complexity we use this from this point forward. More details on the model comparison can be found in [17].

Table 2
Prediction performance for liking and desire to watch again classifiers using features from smile detector V1 and smile detector V2.

	Liking		Desire to view again	
	Smile V1	Smile V2	Smile V1	Smile V2
ROC AUC	0.80	0.82	0.77	0.79
ROC PR	0.95	0.96	0.90	0.93
Accuracy	76%	81%	75%	73%

7.2. Impact of smile detector performance on preference prediction

Table 2 shows a comparison of the performances for liking and desire to watch prediction using features from the two versions of the smile classifier. An HCRF with Gaussian smoothing window size, $\omega = 3$ was used.

7.2.1. Using smile classifier V1 features

Using only temporal information about a person's smile response we can predict success with the area under the ROC curve for the liking and desire to watch again classifiers 0.8 and 0.78 respectively. These results were obtained using a challenging leave-one-ad-out training scheme to ensure generalizability across other amusing video ads. Table 3 (left) shows the confusion matrix for the HCRF liking classifier ($\omega = 3$) with the optimal decision threshold based on the point on the ROC curve closest to (0,1).

However, there are a number of misclassifications that occur. Fig. 10 (g–l) shows cases of false positive results. Fig. 10 (m–r) shows cases of false negative results. For comparison Fig. 10 (a–f) shows cases of true positive results. In a number of the true positive and false positive cases it is difficult to identify any difference in characteristics of the smile tracks for the positive and negative classes. For instance examples e and k seem to have high mean smile intensities and similar positive trends, similarly for examples a and g. For the false negative cases there are a number for which a very low smile probability was detected throughout the content (o and p in particular) but after which the participants reported loving the ad. In some cases (e.g. response p) the smile classifier correctly identified very little smile activity yet the participant reported liking the clip, it seems that the self-report response does not necessarily fit with the smile response that one might expect. Frames from response p can be seen in Fig. 10. From the smile response alone it is unlikely that we could achieve 100% accuracy in predicting liking or desire to view again. In other cases the misclassification is due to noise in the smile classifier prediction. In response i there was little smile activity yet the classifier predicted a relatively high smile probability. This was a dark video which may have been a cause of the error. The latter errors can potentially be removed by improving the

Table 3
Confusion matrices for the best performing liking classifier: left) using smile detector V1, right) using smile detector V2.

Liking	Actual + ve (liking)	Actual − ve (disliking)
Predict + ve	1027	53
Predict − ve	305	149
Predict + ve	1078	56
Predict − ve	236	142

performance of the smile prediction algorithm; however, the former errors raise much more interesting questions about self-report's accuracy as a reflection of feeling towards content and about the circumstances under which viewers express their feelings as facial expressions. As the data collection was unconstrained a third possibility is that people participating may have been playing with the system and intentionally recording false data. To help understand these observations further we compare the results using an improved version of the smile detector that was developed after detector V1.

7.2.2. Using smile classifier V2 features

Smile detector V2 was trained on more example images and showed higher accuracy. The results of preference prediction using features from the smile classifier V2 are also more accurate. Using the HCRF classifier, $\omega = 3$, the area under the ROC curve for the liking classifier was 0.82 compared to 0.80 with smile classifier V1. The accuracy was 81% compared to 76% with smile classifier V1. There are 82 (22%) fewer misclassifications. Table 3 (right) shows the confusion matrix for the HCRF liking classifier ($\omega = 3$) with the optimal decision threshold based on the point on the ROC curve closest to (0, 1). Fig. 11 shows some examples which were misclassified using the smile classifier V1 and were correctly classified using the smile classifier V2. Cases in which noisy smile tracks from V1 led to an incorrect "liking" prediction (Fig. 11 a and b) were corrected using V2 of the detector. Cases in which subtle smiles were missed by V1 and led to an incorrect "disliking" prediction (Fig. 11 c and d) were corrected too as V2 detected more subtle smiles.

Fig. 12 (g–l) shows cases of false positive results. Fig. 12 (m–r) shows cases of false negative results. For comparison Fig. 12 (a–f) shows cases of true positive results. With the improvement in performance we uncover more cases where similar smile patterns are seen across both the liking and disliking categories reinforcing the understanding that self-reported experiences are not perfectly correlated with facial behavior. Different individuals may cognitively evaluate an experience differently even if they display very similar smile activity during the experience. This results in false positives such as g and h. In some cases it was observed that people were not watching the ads alone. In these cases smiles were sometimes the result of social interactions between the viewers and not as a direct response to the content being viewed. This is another reason why the self-report rating of the ad may not seem coherent with the observed response.

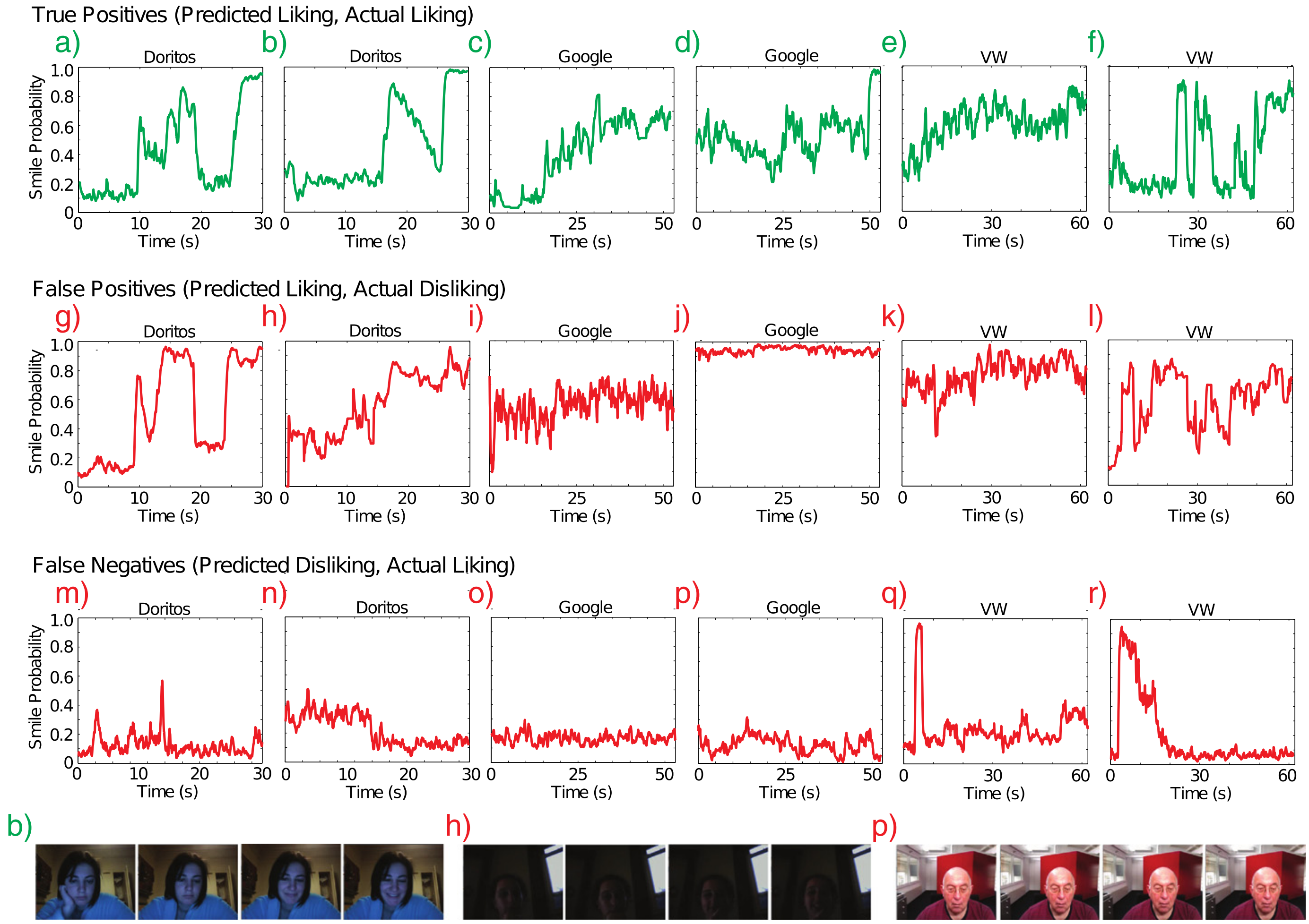


Fig. 10. Examples of true positives (top), false positive (center) and false negatives (bottom) using features from smile detector V1. Most of the false negative examples show responses with very low smile intensity despite the viewer reporting liking the commercial. Shown below are frames from examples of TP, FP and FN videos.

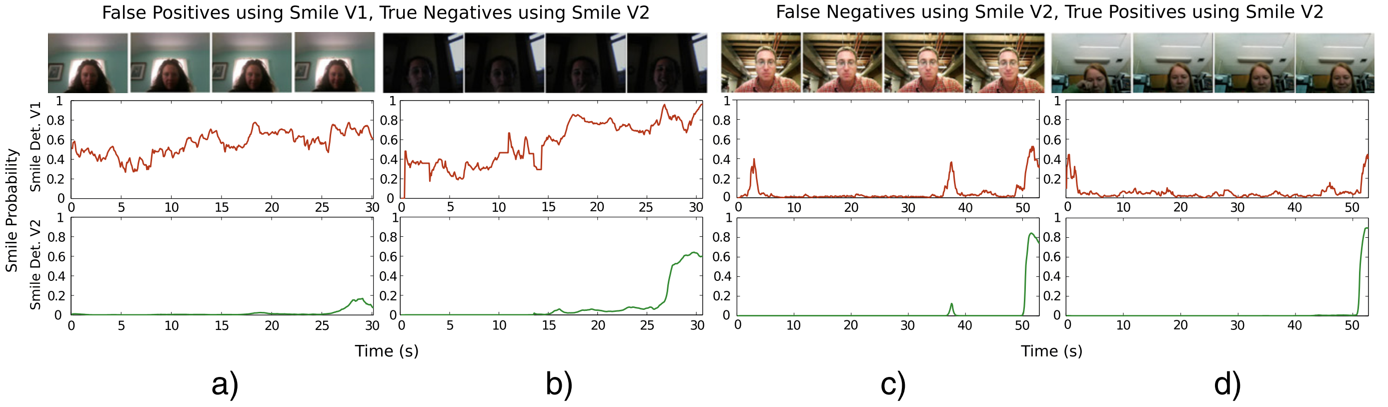


Fig. 11. Examples of video misclassified using features from smile detector V1 but correctly classified using smile detector V2. a and b) False positive examples — noisy smile tracks using V1 due to challenging lighting, c and d) false negative examples — missed smiles due to subtle expressions.

We also see that a large number of the false positive and false negatives occur when the viewers are relatively inexpressive (l, m, n, r). Some very subtle facial behavior is still missed by the classifiers, such as in Fig. 12 k. It may be that detecting other facial actions could help improve the prediction accuracy as we know that smiles may occur in both positive and negative situations [9].

Looking at the log likelihoods given to each of the test samples by the HCRF classifier we can gain further insight into the structure of liking responses learnt. Fig. 13 shows a histogram of the log likelihoods of the test samples with three examples of the responses. The optimal decision threshold based on the point on the ROC curve closest to (0, 1) is also shown. Examples which fall furthest from the decision boundary seem reasonable with the disliking example showing high amounts of smiling at the beginning but which disappear quickly and the liking example showing high amounts of smiling at significant parts of the ad. Closer to the decision boundary the responses are less coherent.

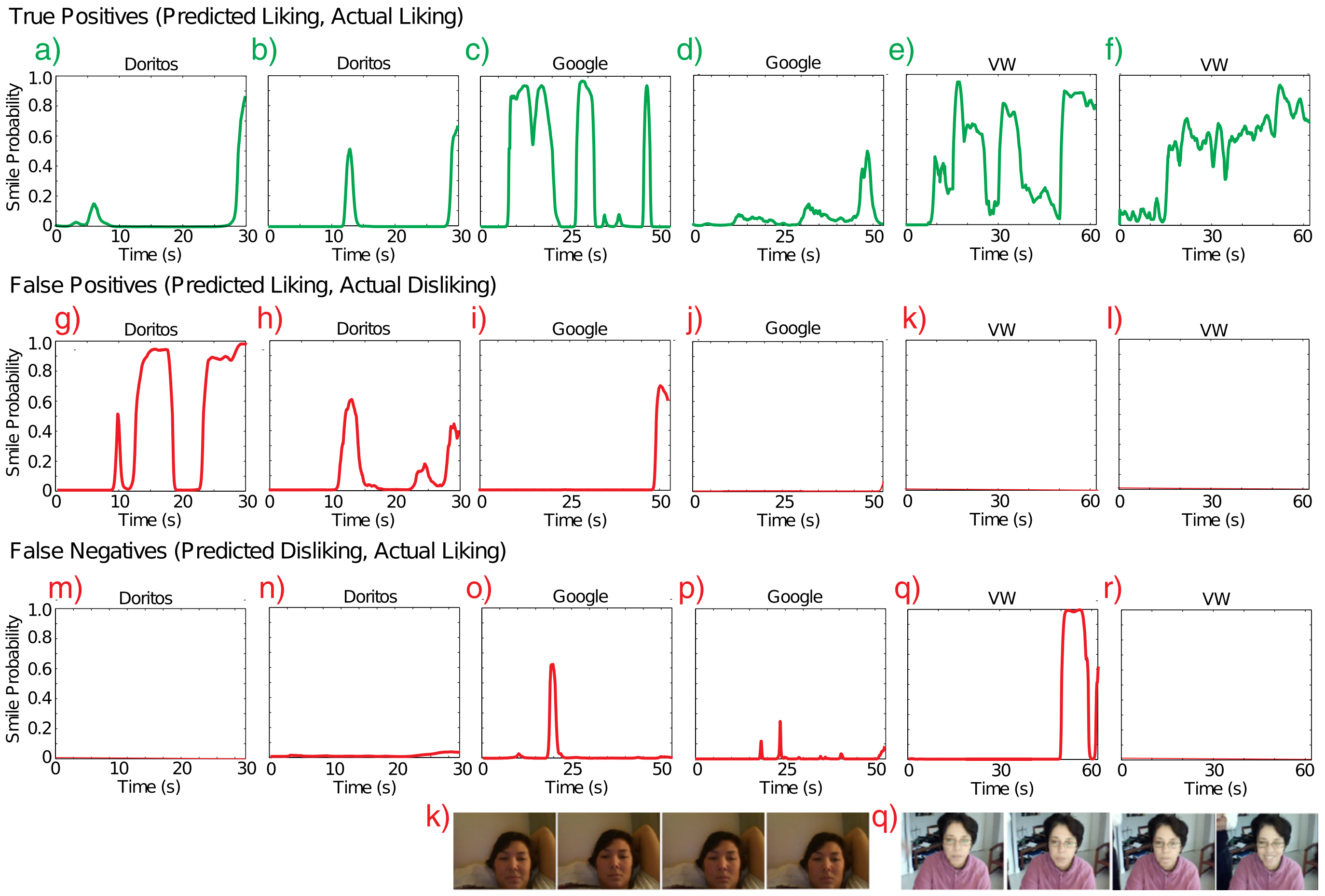


Fig. 12. Examples of true positives (top), false positive (center) and false negatives (bottom) using features from smile detector V2. Most of the false negative examples show responses with very low smile intensity despite the viewer reporting liking the commercial. Shown below are frames from examples of FP and FN videos.

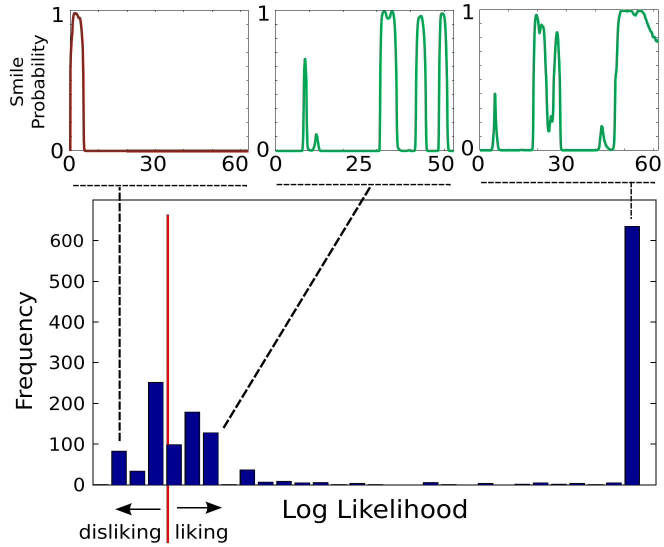


Fig. 13. Histogram of log likelihood outputs from the HCRF classifier for the test samples. The decision threshold based on the point on the ROC curve closest to (0, 1) is shown in red. Examples of a response from three of the buckets are shown.

7.3. Impact of testing scheme

In the experiments above we used a challenging leave-one-ad-out testing scheme. This means that no responses to the ad in the testing set were in the training and validation sets. Here we compare these results to the case in which there are examples of responses to the ad in the testing set also in the training and validation sets (these responses are not the same but responses from different people to the same ad). Table 4 shows the best performance in both cases. The leave-one-ad-out testing scheme does lead to lower prediction performance area under the ROC 0.79 compared to 0.84 for the desire to view again prediction task. However, the results are not greatly decreased in part due to the fact that all the ads are intentionally humorous and have similar structures but also due to the fact that the HCRF shows good generalization performance.

7.4. Impact of number of temporal bins

As described in Section 5 features were extracted from the smile tracks using temporal bins. For this analysis we chose 20 temporal bins. However, potentially a less computationally complex model that had fewer bins could have been used. We compared the performance of prediction with different numbers of temporal bins (5, 10, 20 and 30 bins). The number of bins does not have a very large impact on the performance of the overall system in this case as only three ads were tested and all had similar structures. Even with 5 bins relevant dynamics of the smile track are captured. In cases where ads have more distinct structures then features with finer temporal resolution may well be necessary.

Table 4
Prediction performance for liking and desire to watch again classifiers using an ad-independent (leave-one-ad-out) testing scheme and using a testing scheme in which different individuals' responses to the ad under testing are included in the training and validation sets.

	Liking		Desire to view again	
	Ad-ind. test	Not ad-ind. test	Ad-ind. test	Not ad-ind. test
ROC AUC	0.82	0.83	0.79	0.84

8. Conclusions

We present an automated method for classifying “liking” and “desire to view again” based on 3268 facial responses to media collected over the Internet. The results demonstrate the possibility for an ecologically valid, unobtrusive, evaluation of liking and desire to view again for advertisements, strong predictors of marketing success, based only on facial responses. We build on preliminary findings and show improvement in accuracy predicting viewer preferences. The accuracy and area under the curve for the best “liking” classifier were 81% and 0.82 respectively when using a challenging leave-one-ad-out testing regime. We built on preliminary findings and show that improved smile detection can lead to a 22% reduction in misclassifications. Comparison of the two smile detection algorithms showed that improved smile detection helps correctly classify responses recorded in challenging lighting conditions and those in which the expressions were subtle. With the improved smile classification most of the misclassifications occurred in the cases where people did not smile or where there were differences in reported liking despite very similar facial responses during the content.

HCRFs and LDCRFs (discriminative approaches to classification that model temporal structure) performed most strongly showing that temporal information about an individual's response is important. It is not just how much a viewer smiles but when they smile. Logistic regression analysis shows that the smile activity in the final 25% of the ads is the most strongly related to the liking reported after the ad.

This work considers only smiles which are the most commonly occurring action in this dataset [18] due to the humorous nature of the content. However, other action units occur and these will be considered in future work in particular to help distinguish preferences for those that do not smile. Future work will take advantage of a greater number of action units and gestures (such as eyebrow raise (AU1 + 2), brow lowerer (AU4) and nose wrinkler (AU9). The ground-truth considered in this study was self-reported and future work will consider behavioral measures of success (such as sharing) and sales. Finally, considering responses to a greater number of ads would give more confidence that the results generalize beyond intentionally humorous commercials.

Acknowledgments

Richard Sadowsky, Oliver Wilder-Smith, Jay Turcot, Zhihong Zeng and Affectiva provided generous support in providing access to the crowdsourcing platform. Brian Staats provided a great front end design for the site. Jon Bruner and Forbes kindly promoted the work on their site. Google provided use of their facial feature tracker. This work was funded by the Media Lab Consortium Members.

References

[1] Web address http://www.forbes.com/2011/02/28/detect-smile-webcam-affectiva-mit-media-lab.html.
[2] Z. Ambadar, J. Cohn, L. Reed, All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous, J. Nonverbal Behav. 33 (1) (2009) 17–34.
[3] P. Bolls, A. Lang, R. Potter, The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements, Commun. Res. 28 (5) (2001) 627.
[4] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (27) (2011) 1–27(27).
[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005, CVPR 2005. IEEE Computer Society Conference on, vol. 1, IEEE, 2005, pp. 886–893.
[6] P. Ekman, W. Friesen, Facial Action Coding System. Test, 1977.
[7] R.I. Haley, A.L. Baldinger, The ARF copy research validity project, J. Advert. Res. 40 (06) (2000) 114–135.

[8] R. Hazlett, S. Hazlett, Emotional response to television commercials: facial emg vs. self-report, J. Advert. Res. 39 (1999) 7–24.

[9] M. Hoque, D. McDuff, R. Picard, Exploring temporal patterns in classifying frustrated and delighted smiles, IEEE Trans. Affect. Comput. 3 (3) (july-september 2012) 323–334.

[10] H. Joho, J. Staiano, N. Sebe, J. Jose, Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents, Multimedia Tools and Applications2011. 1–19.

[11] K.S. Kassam, Assessment of Emotional Experience through Facial Expression, (PhD thesis) Harvard University, 2010.

[12] J. Lafferty, A. McCallum, F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, 2001.

[13] P. Lucey, S. Lucey, J.F. Cohn, Registration invariant representations for expression detection, Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on, IEEE, 2010, pp. 255–261.

[14] D. McDuff, R. El Kaliouby, K. Kassam, R. Picard, Affect valence inference from facial action unit spectrograms, Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 17–24.

[15] D. McDuff, R. el Kaliouby, R. Picard, Crowdsourced data collection of facial responses, Proceedings of the 13th International Conference on Multimodal Interfaces, ACM, 2011, pp. 11–18.

[16] D. McDuff, R. El Kaliouby, R. Picard, Crowdsourcing facial responses to online videos, IEEE Trans. Affect. Comput. 3 (4) (2012) 456–468.

[17] D. McDuff, R. El Kaliouby, R.W. Picard, Predicting online media effectiveness based on smile responses gathered over the Internet, Automatic Face & Gesture Recognition, 2013 IEEE International Conference on, IEEE, 2013.

[18] D. McDuff, R. El Kaliouby, S. Thibaud, A. May, J. Cohn, R. Picard, Affectiva-MIT facial expression dataset (AM-FED): naturalistic and spontaneous facial expressions collected in-the-wild, Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Computer Society Conference on, IEEE, 2013.

[19] A.C. MICU, J.T. Plummer, Measurable emotions: how television ads really work, J. Advert. Res. 50 (2) (2010) 137–153.

[20] L. Morency, A. Quattoni, T. Darrell, Latent-dynamic discriminative models for continuous gesture recognition, Computer Vision and Pattern Recognition, 2007, CVPR'07. IEEE Conference onIEEE, 2007, pp. 1–8.

[21] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[22] C. Shan, Smile detection by boosting pixel differences, IEEE Trans. Image Process. 21 (1) (2012) 431–436.

[23] E. Smit, L. Van Meurs, P. Neijens, Effects of advertising likeability: a 10-year perspective, J. Advert. Res. 46 (1) (2006) 73.

[24] Y. Song, D. Demirdjian, R. Davis, Multi-signal gesture recognition using temporal smoothing hidden conditional random fields, 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE, 2011, pp. 388–393.

[25] T. Teixeira, M. Wedel, R. Pieters, Emotion-induced engagement in Internet video ads, J. Mark. Res. (2010).

[26] T. Teixeira, M. Wedel, R. Pieters, Moment-to-moment optimal branding in TV commercials: preventing avoidance by pulsing, Mark. Sci. 29 (5) (2010) 783–804.

[27] M. Valstar, H. Gunes, M. Pantic, How to distinguish posed from spontaneous smiles using geometric features, Proceedings of the 9th International Conference on Multimodal Interfaces, ACM, 2007, pp. 38–45.

[28] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, T. Darrell, Hidden conditional random fields for gesture recognition, Computer Vision and Pattern Recognition, 2006, IEEE Computer Society Conference on, vol. 2, IEEE, 2006, pp. 1521–1527.

[29] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, J. Movellan, Toward practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 2106–2111.

[30] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.

[31] S. Zhao, H. Yao, X. Sun, Video classification and recommendation based on affective analysis of viewers, Neurocomputing (2013).