Crowdsourcing Facial Responses to Online Videos

Daniel McDuff, *Student Member, IEEE,* Rana El Kaliouby, *Member, IEEE,*
and Rosalind Picard, *Fellow, IEEE*

Abstract—We present results validating a novel framework for collecting and analyzing facial responses to media content over the Internet. This system allowed 3,268 trackable face videos to be collected and analyzed in under two months. We characterize the data and present analysis of the smile responses of viewers to three commercials. We compare statistics from this corpus to those from the Cohn-Kanade+ (CK+) and MMI databases and show that distributions of position, scale, pose, movement and luminance of the facial region are significantly different from those represented in these traditionally used datasets. Next we analyze the intensity and dynamics of smile responses, and show that there are significantly different facial responses from subgroups who report liking the commercials compared to those that report not liking the commercials. Similarly, we unveil significant differences between groups who were previously familiar with a commercial and those that were not and propose a link to virality. Finally, we present relationships between head movement and facial behavior that were observed within the data. The framework, data collected and analysis demonstrate an ecologically valid method for unobtrusive evaluation of facial responses to media content that is robust to challenging real-world conditions and requires no explicit recruitment or compensation of participants.

Index Terms—Crowdsourcing, Facial expressions, Non-verbal behavior, Advertising, Market research.

1 Introduction

The face is one of the richest sources of communicating affective and cognitive information [1]. In relation to advertising it as been shown that facial expressions exhibited in viewers while watching commercials can predict strength of recall of the commercial [2]. The face has been described as the window to the soul, to quote Cicero (circa 100 b.c.) *'Ut imago est animi voltus sic indices oculi'* (the face is a picture of the mind as the eyes are its interpreter). In this paper, we present results validating a first-in-the-world framework for collecting and analyzing ecologically valid facial responses to media content over the Internet.

The Internet provides the ability to crowdsource lots of useful information [3]. Previous work has shown that many people are willing to engage and share visual images from their webcams and these can be used for training automatic algorithms for learning [4]. Moreover, webcams are now ubiquitous and have become a standard component on many media devices, laptops and tablets. In 2010, the number of camera phones in use totaled 1.8 billion, which accounted for a third of all mobile phones[1]. In addition, about half of the videos shared on Facebook every day are personal videos recorded from a desktop or phone camera[2]. This provides a strong indication that people are willing to turn their webcams on and will readily do so if there is value in it for them

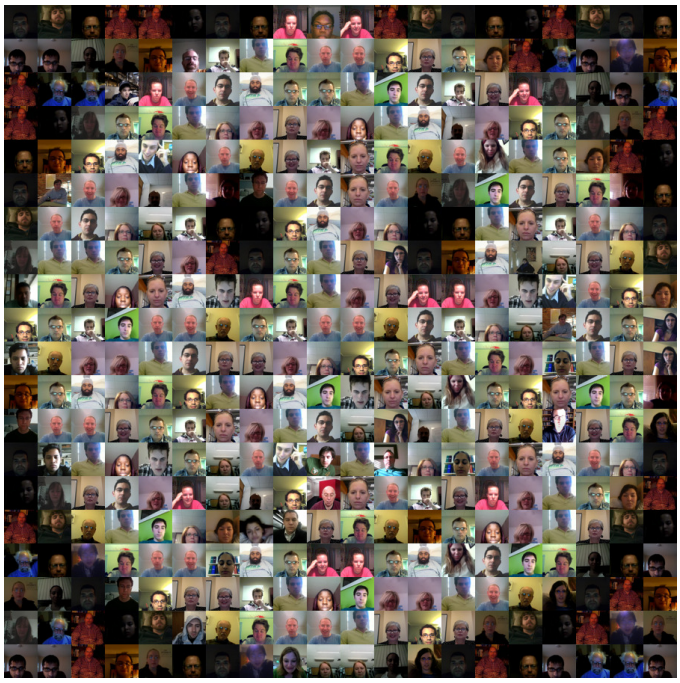

Fig. 1. A sample of frames from the 3,268 videos collected. There are significant variations in position, scale, pose, lighting and movement in the responses. These represent a subset of the public data.

- *D. McDuff*, Dr. El Kaliouby and Dr. Picard are with the Affective Computing Group at the Media Lab, Massachusetts Institute of Technology, Cambridge, US, 02139.*

  **E-mail: djmcduff@mit.edu*
- *Dr. El Kaliouby and Dr. Picard also hold positions at Affectiva, Inc..*

1. http://www.economist.com/node/15865270

2. http://gigaom.com/video/facebook-40-of-videos-are-webcam-uploads/

(e.g. social sharing). Inspired by these approaches, we undertook an experiment to see if people would opt in to have their video responses and facial expressions analyzed for no fee, how many would agree to this, and whether the quality would be adequate for analysis given no constraints over lighting, pose, movement, or system set-up.

Online and mobile video is the fastest growing medium in history[3]. In June 2011 alone, 178 million US Internet users (85.6% of the US Internet audience) watched online video content for an average of 16.8 hours per viewer. This audience engaged in more than 6.2 billion viewing sessions during the month, an all-time high (up from 172 million viewers and an average of 15 hours per viewer in Dec 2010)[4]. Moreover, video streaming made up 39% of all mobile data traffic worldwide in the first half of 2011, a growth rate of 77% over the second half of 2010[5]. Following this shift in consumer behavior, marketers are shifting advertising spend to online video, with over $1 billion in annual global video marketing spending anticipated to grow globally to $10 billion by 2015.

With so much content out there competing for the viewer's eyeballs, there is an increasing desire for content creators, marketers and advertisers to objectively measure engagement with content. In addition, brands are increasingly striving to build emotional connections with consumers, but marketers are struggling to objectively measure their success in achieving these aims. In advertising, the emotional response of viewers to advertisements has been shown to influence attitude to both the commercial and the brand [5] as well as the engagement of viewers [6]. However, most of these studies are performed in laboratory settings, which, in addition to being time-consuming and costly, leads to an unnatural setting for participants, which may heavily influence how they experience and engage with content. When considering facial behavior in a laboratory it has been shown that facial behavior in artificial settings differs from that exhibited in real world contexts [7]. This may be due to the subject being restricted in terms of motion, limited due to social constraints and may also be affected by the difference in context, an unfamiliar room, device or people [8]. As an example, Fridlund [9] shows that viewer's responses to videos are influenced by context, responses in social situations are heightened when compared to responses in non-social settings. Furthermore, this effect occurs even when the sociality of the situation is implicit. It is important for advertisers to be able to evaluate the affective impact of their commercials not just in laboratory settings. In this study the participants were aware that they were being recorded which might have caused a certain amount of behavior change; however, this is a significantly more comfortable and familiar setting than a laboratory and the effects are likely to be significantly reduced. In addition this much more closely reflects the real consumption environment for viewing Internet advertisements.

Traditionally consumer testing of video advertising, whether TV or Internet, has been conducted in laboratory settings [6], [10]. Self-report is the current standard measure of affect, where people are interviewed, asked to rate their feeling on a Likert scale or turn a dial to quantify their state. While convenient and inexpensive, self-report is problematic because it is subject to biasing from the interviewer, the context and other factors of little relevance to the stimulus being tested [11]. The act of introspection is challenging to perform in conjunction with another task and may in itself alter that state [12]. Unlike self-report, facial expressions are implicit and do not interrupt a person's experience. In addition, facial expressions are continuous and dynamic, allowing for a representation of how affect changes over time.

Smile detection is one of the most robust forms of automated facial analysis available. Whitehall et al. [13] present a smile classifier based on images collected over the Interest and demonstrate strong performance on this challenging dataset. Prior research has shown that the dynamics of smiles can be very informative in determining the underlying state [7], [14]. In this paper we focus on the automatic detection of smile intensity from video frames collected over the Internet.

Public datasets truly help accelerate research in an area, not just because they provide a benchmark, or a common language, through which researchers can communicate and compare their different algorithms in an objective manner, but also because compiling such a corpus is tedious work - requiring a lot of effort which many researchers may not have the resources to do. Computer-based machine learning and pattern analysis depends hugely on the number of training examples. To date much of the work automating the analysis of facial expressions and gestures has had to make do with limited datasets for training and testing. However, due to the limited representation of different cultural, age and gender demographics this often leads to over-fitting. Our framework allows for the efficient collection of large amounts of data from different demographics.

The main contribution of this paper is to present the new corpus of data collected, and provide analysis and results from the first-in-the-world crowdsourcing of facial responses over the web. We show that despite significant individual differences in responses and challenging real world data we are able to distinguish significant patterns within the results and relate these to difference in self-report measures. We believe that this new method of collecting and analyzing facial video can provide unobtrusive evaluation of facial responses to media content without relying on self-report ratings. It also has the potential to truly accelerate research in automated understanding of facial expressions and

3. http://www.wpp.com/wpp/press

4. http://www.comscore.com/Press_Events/Press_Releases/2011/7/comScore_Releases_June_2011_U.S._Online_Video_Rankings)

5. http://www.pcworld.com/article/236606/survey_video_dominates_mobile_traffic.html

gestures by allowing the collection of huge corpuses of naturalistic and spontaneous data. We show that the method can also provide a mechanism to ask entirely new research questions, and to answer those questions with data that is ecologically valid. We present a massive dataset, collected via the Internet in 54 days, containing 3,268 videos captured in natural environments whilst the viewers were presented with public stimuli, one of three commercials. Figure 1 shows example frames from the 3,268 videos collected. These represent a subset of the public data.

In the remainder of the paper, we will; 1) describe the framework for data collection and analysis, 2) characterize the data collected in terms of viewer demographics and video qualities, 3) present results of facial response analysis, principally smiles, with different self-report and different stimuli, 4) and show the synchronous relationship between facial expressions and head gestures revealed by the data.

2 Related Work

2.1 Affective Computing Approaches

The Facial Action Coding System (FACS) [15] is a catalogue of 44 unique action units (AUs) that correspond to each independent movement of the face's 27 muscles. The action units combine to create thousands of unique and meaningful facial expressions. FACS enables the measurement and scoring of facial activity in an objective, reliable and quantitative way. However, FACS-coding requires extensive training and is a labor intensive task. It can take almost 100 hours of training to become a certified coder, and one to three hours to code a minute of video. Over the past 20 years, there has been significant progress building systems that unobtrusively capture facial expressions and head gestures [16]. This progress has been aided by improvements in facial alignment and feature tracking, which are now very efficient and can be performed online or offline in real-time.

State of the art face tracking and registration methods include Active Appearance Models (AAM) and Constrained Local Models (CLM). There are numerous features used in facial action unit detection. Geometric features, Gabor Wavelet coefficients, Local Binary Patterns and SIFT descriptors have all be demonstrated with success. Support Vector Machines (SVM) are the most commonly used classification method used in action unit detection. Different forms of boosting (e.g. AdaBoost) have also been effective at improving performance. A comprehensive review of methods for facial affect recognition methods can be found in [16].

2.2 Available Datasets

In the area of facial expression analysis, the Cohn-Kanade database, in its extended form named CK+ [17], played a key role in advancing the state of the art in this area. The CK+ database, contains 593 recordings of posed and non-posed sequences. The sequences are recorded under controlled conditions of light and head motion, and range between 9-60 frames per sequence. Each sequence represents a single facial expression that starts with a neutral frame and ends with a peak facial action. Transitions between expressions are not included. Several systems use the CK, or CK+, databases for training and/or testing. Since it was first published, a number of papers have appeared that were trained and/or tested on this data set including: Bartlett et al [18], Cohen et al. [19], Cohn et al. [20], Littlewort et al. [21] and Michel and El Kaliouby [22]. Since then, a few other databases have emerged, including: MMI [23], SEMAINE [24], RU-FACS [25], SAL [26], GENKI [13] and UNBC-McMaster Shoulder Pain Archive [27]. A survey of databases and affect recognition systems can be found in [16]. However, there is a need for mechanisms to quickly and efficiently collect numerous examples of natural and spontaneous responses. Lab-based studies pose many challenges including recruitment, scheduling and payment. Efforts have been made to collect significant amounts of spontaneous facial responses; however, the logistics of a laboratory based study typically limits the number of participants to under 100, e.g. 42 in [28]. By using the Internet we can make data collection efficient, asynchronous, less resource intensive, and get at least an order of magnitude more participants. Perhaps more importantly, we can begin to systematically explore the meaning of facial expressions and their relationship to memory and decision-making in an ecologically valid manner.

2.3 Market Research

Joho et al. [29] showed that it is possible to predict personal highlights in video clips by analyzing facial activity. However, they also note the considerable amount of individual variation in responses. These experiments were conducted in a laboratory setting and not in the wild; our work demonstrates the possibility of extending this work to online content and real-world data.

Teixeira et al. [6] show that inducing affect is important in engaging viewers in online video adverts and to reduce the frequency of "zapping" (skipping the advertisement). They demonstrated that joy was one of the states that stimulated viewer retention in the commercial and it is thus intuitive that smiles would be a significant indicator in evaluating this. Again, these studies were performed in a laboratory setting rather than in the wild.

Companies will frequently place their commercials on free video broadcasting websites such as YouTube with the hope that they will be shared by people. If a video circulates rapidly across the Internet it can be considered as being "viral". Berger and Milkman [30] investigated what makes online content viral and found that positive affect inducing content was more viral than negative affect inducing content and that virality was also driven by high physiological arousal. Although

their study focuses on written content is it reasonable to think that similar principles may apply to videos and that commercials that induce higher intensity positive responses would be more likely to go viral.

Ambler and Burne [31] and Mehta and Purvis [32] show that emotion plays an important role in the relationship between brand and advertising recall and that emotional content in well-executed commercials can boost recall. Haley [33] concluded that the concept of "likeability" of a commercial was the best predictor of sales effectiveness - a considerable measure of success in advertising. Smit et al. [34] found that commercials were perceived as less likeable over time. However, the predictive power of likeability was not diminished. In this study we focus primarily on smiles elicited during amusing commercials, which we hypothesize capture partially the complex concept of likeability. We also investigate the relationship between smile responses and familiarity, or ad wear out effects.

Poels [35] provides a survey of work on emotion measurement in advertising including evaluation of physiological, self-report and facial measurement techniques. We apply the latter two forms of measurement in this study as the research was conducted over the Internet and we have no direct contact with the participants.

2.4 Crowdsourcing

Crowdsourcing [36] aims to coordinate the effort and resources of large groups of people. Morris [37] presents the case for the use of crowdsourcing technology to serve applications in affective computing, which he calls "Affective Crowdsourcing". This involves both the use of the "crowd" to provide data for training algorithms, provide labels for existing data and to provide interventions that aim to improve well-being. In this paper, we leverage crowdsourcing concepts to elicit a large amount of data from a wide demographic, something that has not been possible through traditional research practices.

3 Crowdsourcing Platform

Figure 2 shows the web-based framework that was used to crowdsource the facial videos and the user experience. The website was promoted on Forbes.com for the first day that it was live. Visitors may have found it via this route, a search engine or a shared link. Visitors to the website opt-in to watch short videos while their facial expressions are being recorded and analyzed. Immediately following each video, visitors get to see where they smiled and with what intensity. They can compare their "smile track" to the aggregate smile track. On the client-side, all that is needed is a browser with Flash support and a webcam. The video from the webcam is streamed in real-time at 15 frames a second at a resolution of 320x240 to a server where automated facial expression analysis is performed, and the results are rendered back to the browser for display. There is no need to download or install anything on the client

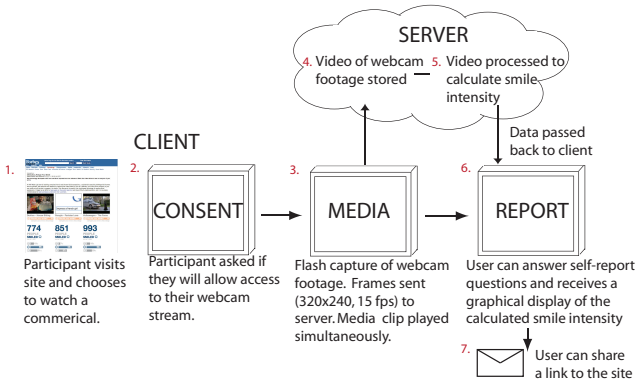

Fig. 2. Overview of what the user experience was like and Affectiva's (www.affectiva.com) web-based framework that was used to crowdsource the facial videos. From the viewer's perspective, all that is needed is a browser with Flash support and a webcam. The video from the webcam is streamed in real-time to a server where automated facial expression analysis is performed, and the results are rendered back to the browser for display. All the video processing was done on the server side.

side, making it very simple for people to participate. Furthermore, it is straightforward to easily set up and customize "experiments" to enable new research questions to be posed. For this experiment, we chose three successful Super Bowl commercials: 1. Doritos ("House sitting", 30 s), 2. Google ("Parisian Love", 53 s) and 3. Volkswagen ("The Force", 62 s). Large sums of money are spent on Super Bowl commercials and as such their effectiveness is of particular interest to advertisers. All three ads were somewhat amusing and were designed to elicit smile or laughter responses. Results showed that significant smiles were present in 71%, 65% and 80% of the responses to the respective ads.

On selecting a commercial to watch, visitors are asked to 1) grant access to their webcam for video recording and 2) to allow Affectiva and MIT to use the facial video for internal research. Further consent for the data to be shared with the research community at large is also sought, and only videos with consent to be shared publicly are shown in this paper. This data collection protocol was approved by the Massachusetts Institute of Technology Committee On the Use of Humans as Experimental Subjects (COUHES) prior to launching the site. A screenshot of the consent form is shown in Figure 3. If consent is granted, the commercial is played in the browser whilst simultaneously streaming the facial video to a server. In accordance with MIT COUHES, viewers could opt-out if they chose to at any point while watching the videos, in which case their facial video is immediately deleted from the server. If a viewer watches a video to the end, then his/her facial video data is stored along with the time at which the session was started, their IP address, the ID of the video they

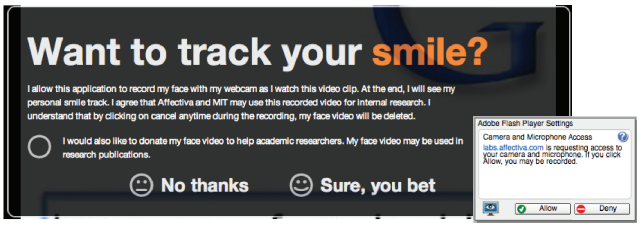

Fig. 3. The consent forms that the viewers were presented with before watching the commercial and before the webcam stream began.

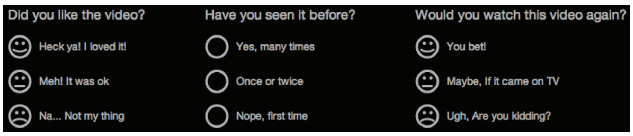

Fig. 4. The self-report questions the viewers were presented with after watching the commercial.

watched and self-reported responses (if any) to the self report questions. No other data is stored.

Following each commercial, the webcam is automatically stopped and a message clearly states that the "webcam has now been turned off". Viewers could then optionally answer three multiple choice questions: "Did you like the video?", "Have you seen it before?" and "Would you watch this video again?". A screenshot of the questions is shown in Figure 4. Finally, viewers were provided with a graphical representation of their smile intensity during the clip compared to other viewers who watched the same video; viewers were also given the option to tweet their result page or email it to a friend. All in all, it took under 5 seconds to turn around the facial analysis results once the video was completed so viewers perceived the results as instantaneous. Viewers were free to watch one, two or three videos and could watch a video as many times as they liked.

4 Data Collection

Using the framework described we collected 3,268 videos (2,615,800 frames) over a period of 54 days from 03/03/2011 to 04/25/2011. The application was promoted on the Forbes website [38]. Figure 6 shows the number of the trackable videos that were completed on each of the 54 days. We refer to the data collected as the Forbes dataset. We don't know how many viewed the site but the number of visitors who clicked a video was 16,366. Of these 7,562 (46.2%) had a webcam, had a computer that met the system requirements and opted-in to allow webcam access. A total of 5,268 (32.2%) completed the experiment. For the analysis here we disregard videos for which the Nevenvision tracker was unable to identify a face in at least 90% of frames; this left 3,268 videos (20.0%). Figure 7 shows the participation graphically. All videos were recorded with a resolution of 320x240 and a frame rate of 15 fps.

Fig. 5. The result of the smile analysis presented to the viewer comparing: their smile track (orange) with an aggregate track (gray).

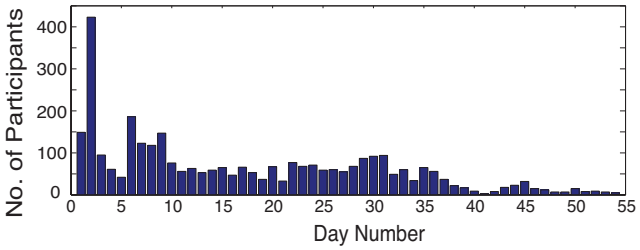

Fig. 6. Histogram of the number of viewers that successfully completed the study on each of the 54 consecutive days (from 3/3/2011) that it was live.

4.1 Demographics

We use IP information to provide statistics on the locations of viewers by finding the latitude and longitude corresponding to each address. Statistics for gender and facial hair were obtained by a labeler who watched the videos. IP addresses have been shown to be a reliable measure of location [39]. The IP address geo-location was performed using IPInfoDB[6]. We could not guarantee

6. http://www.ipinfodb.com/ip_location_api.php

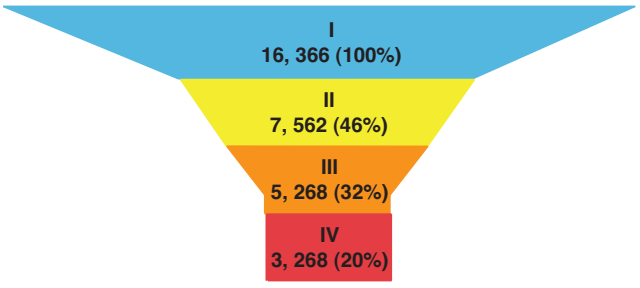

Fig. 7. Funnel chart showing the participation. I) 16,366 visitors clicked on a video, II) 7,562 opted-in to all webcam access, III) 5,268 completed watching the video and IV) 3,268 had identifiable faces in greater than 90% frames.

TABLE 1
Table showing the number of videos for each commercial broken down by continent and gender (no. of females shown in brackets).

	No. of viewers (female)		
Continent	Doritos	Google	VW
Africa	14 (4)	14 (8)	18 (8)
Asia	74 (22)	68 (20)	88 (24)
Europe	226 (75)	228 (65)	222 (61)
North America	681 (245)	730 (273)	714 (260)
South America	42 (13)	43 (15)	43 (12)
Oceania	23 (6)	21 (5)	19 (5)
Total	1,060 (365)	1,104 (386)	1,104 (370)

Fig. 8. Map showing the location of the 3268 viewers, based on their IP address. No viewers IP was located outside of the latitudes shown.

that the same viewer would watch all three of the commercials or that some may watch them more than once. As we do not have identifiable information from the viewers and we do not have the number of distinct viewers who took part, only a coarse calculation can be provided by the number of distinct IP addresses 1,495 (45.8%). This suggests that on average each location successfully completed the task for two viewings. Table 1 shows the number of viewers in each continent and in brackets the number of females. A majority of the viewers were located in North America and Europe. The geographic location of each of the viewers is shown in Figure 8.

Of the 3,268 videos, 1,121 (34.3%) featured females as the main subject. The age of viewers was restricted to those over the age of 13 or with a parent or legal guardian's consent. In 924 (28.3%) of the videos the viewer was wearing glasses. In 664 (20.3%) of the videos the viewer had some form of facial hair. Both glasses and facial hair are likely to introduce some degree of error in the feature tracking.

4.2 Characterizing the Face Videos

The framework presented allows the first large scale collection of natural and spontaneous responses to media content with no control over the placement of the camera, lighting conditions and movement of the viewers. As such there is significantly greater variability in characteristics when compared to data collected in the laboratory. We characterized the videos and compared them to two existing facial expression datasets collected in laboratory settings. We compare the statistics for these data collected with videos from the CK+ [17] and

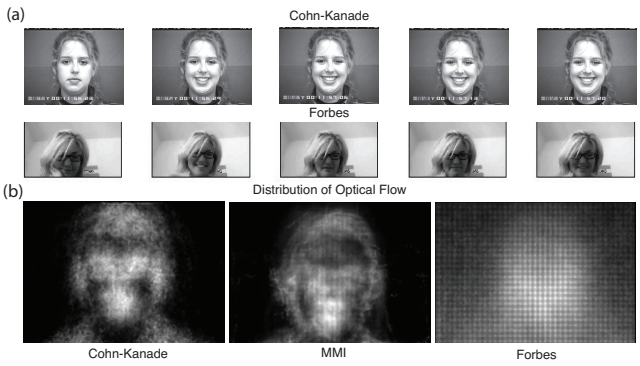

Fig. 10. (a) Examples of smiles in a video from the CK+ and Forbes dataset (no public images for MMI). (b) The optical flow across frontal face videos for each of the three databases, CK+, MMI and Forbes.

MMI [23] databases, data traditionally used for training and testing facial expression and affect recognition systems. For the analysis we took all 722 videos from the MMI database that featured participants filmed with a frontal pose (14,360 frames) and all 593 videos from the CK+ dataset (10,708 frames).

Figure 9 shows distributions of face scale, luminance and contrast of the facial regions, and head pose for the CK+, MMI and Forbes datasets. Figure 10 (b) shows the distribution of optical flow across the frames for the CK+, MMI and Forbes datasets. Face scale was calculated using the Nevenvision tracker with a scale of 1 representing an interocular distance of approximately 50 pixels. For Figure 9 (d) the measurements are with respect to a fully frontal face. Details of the calculations and in-depth analysis can be found in [40].

Our analyses show that there are marked differences between the position, scale and pose of participants in these natural interactions compared to those in datasets traditionally used for training expression and affect recognition systems, the MMI and CK+ datasets. In particular we found that scale of the face within the field of view of the camera and yaw of the head have significantly different distributions to those in traditional lab-based datasets in which these degrees-of-freedom are often constrained. The mean head scale is significantly lower for the Forbes set (mean=0.987) versus the MMI (mean=1.39) and CK+ sets (mean=1.22), $p<0.05$. There is greater deviation in the scales and yaw distributions for the Forbes set than both the MMI and CK+ sets.

Similarly, we identified a statistically significant difference between the average luminance within the facial region between the Forbes dataset and the CK+ and MMI sets. The luminance is significantly lower for the Forbes set (mean=84.3) versus the MMI (mean=128) and CK+ sets (mean=168), $p<0.05$, although the variance of the luminance and the distributions of contrast were not significantly different. These differences help define the range of performance needed for tracker and affect recognition systems. Further analysis of the datasets can

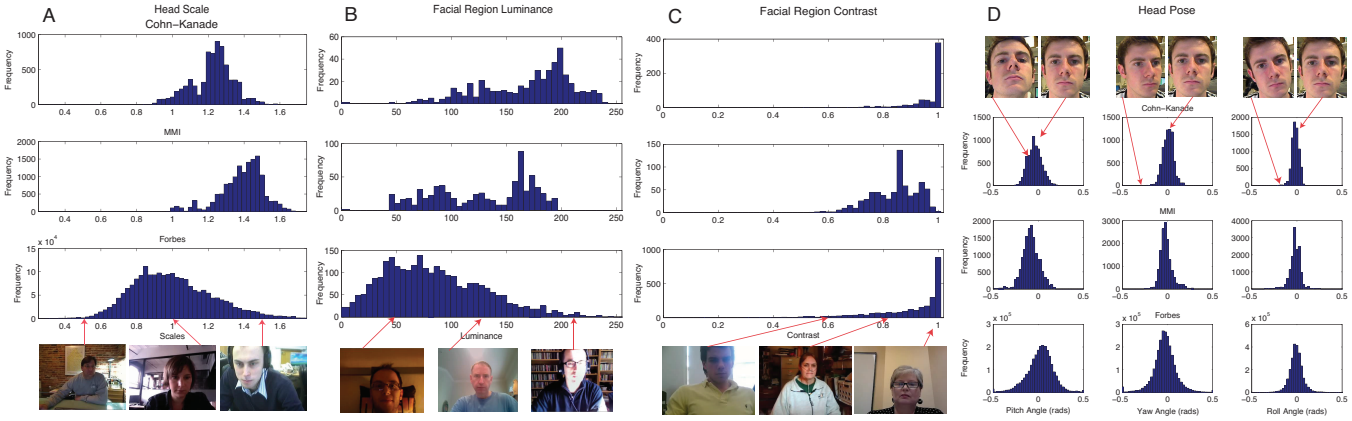

Fig. 9. A) Histogram of head scales for the CK+ (top), MMI (center) and Forbes (bottom) datasets. The head scale was calculated for every frame in which a head was tracked. Examples of head scales of 0.5, 1 and 1.5 are shown below. B) Histograms of the average luminance for the facial region for CK+, MMI and Forbes datasets. Examples are shown for luminance values of 50, 125 and 216. C) Histograms of the Michelson contrast for the facial region for CK+, MMI and Forbes datasets. Examples are shown for contrast values of 0.60, 0.82 and 1.0. D) Histograms showing the pose angles (relative to a fully frontal face) of the heads in the CK+ (top), MMI (center) and Forbes (bottom) datasets. Examples of poses with pitch=-0.13 rads, yaw=-0.26 rads and roll=-0.19 rads are shown.

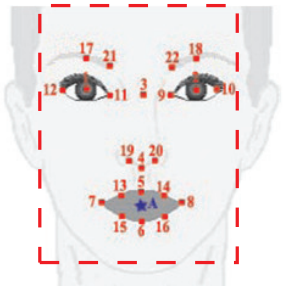

Fig. 11. Location of the 22 feature points tracked by the Nevenvision tracker, the red line highlights the facial region used for evaluating illumination.

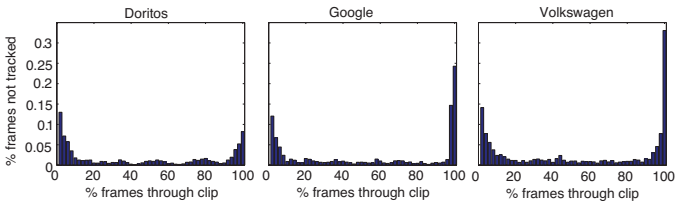

Fig. 12. Percentage of frames in which a face could not be tracked for the responses to each of the three commercials.

be found in [40].

5 Automated Facial Analysis

In this paper we focus primarily on the smile responses of the viewers to the video clips. We also investigate the correlation of smiles with head pose data.

5.1 Face Detection

The Nevenvision facial feature tracker[7] was used to automatically detect the face and track 22 facial feature points within each frame of the videos. The location of the facial landmarks is shown in Figure 11. Due to the low quality of the flash videos recorded (320x240) the Nevenvision tracker was deemed to provide better tracking performance than a AAM or CLM tracker.

The metrics and results need to be considered in the limitations of the facial feature tracker used. About three axes of pitch, yaw (turning) and roll (tilting), the limits are 32.6 (std=4.84), 33.4 (std=2.34) and 18.6 (std=3.75) degrees from the frontal position respectively (deviations reflect variability in performance in different lighting). These were computed in independent performance tests.

Figure 12 shows the percentage of frames in which a face could not be tracked for each of the three commercials. Tracking was most problematic at the beginning and end of the videos. This was when a majority of the movement occurred in the clips as reflected in Figure 13 which shows the average movement within the videos for each commercial. The increased level of movement and the auto-adjustment parameters of the webcams could explain why the tracking was more difficult at the beginning of the videos. The reasons for the increased difficultly in tracking at the end of the videos may reflect a shift in behavior of the viewers that signals disengagement with the commercial.

5.2 Head Pose

Three Euler angles for the pose of the head, pitch, yaw and roll were calculated. The head scale within the frame was also calculated using the feature tracker; this can be approximated as an inverse measurement of the face from the camera. These parameters were calculated for every frame in the responses in which a face was tracked.

7. Licensed from Google, Inc.

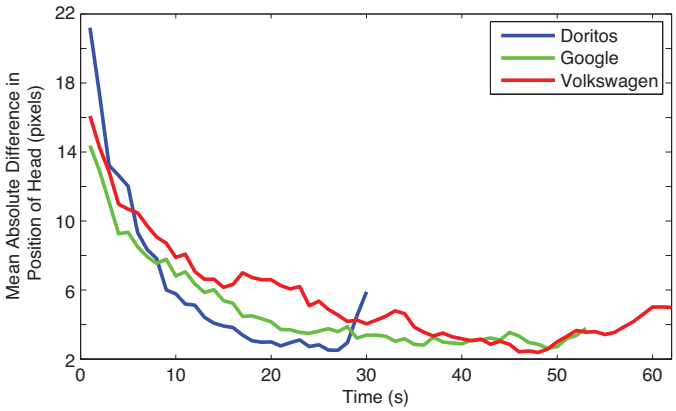

Fig. 13. Mean absolute difference of the position of the viewer's heads (pixels) for each second during the videos. The data is divided into responses to each of the stimuli.

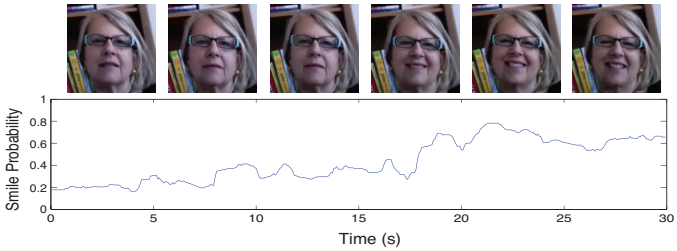

Fig. 14. A smile track with screenshots of the response, demonstrating how greater smile intensity is positively correlated with the probability output from the classifier.

5.3 Smile Detection and Dynamics

To compute the smile probability measure we used a custom algorithm developed by Affectiva. This tracks a region around the mouth using the facial feature tracker and computes Local Binary Pattern (LBP) [41] features within this region. The segmented face images were rescaled to 120x120 pixels, with the region around the mouth 32x64. An affine warp was performed on the bounded face region to account for in-planar head movement. An ensemble of bagged decision trees is used for classification. SVMs and Gabor Energy filters have been shown to perform well on smile detection [13] but we found that the bagged decision tree classifier using LBP features has better performance. 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. Figure 14 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. We refer to the classifier output as the smile intensity from this point on. Figure 15 shows examples of 20 randomly selected smile tracks, for each of the three self-report liking classes, from responses to the Doritos ad.

The smile classifier was trained on examples from the

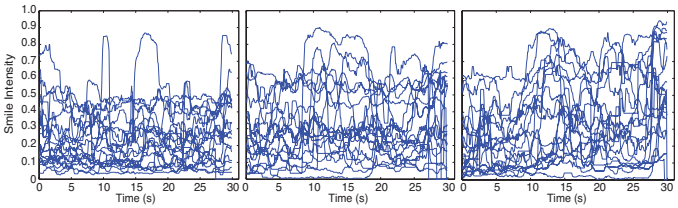

Fig. 15. Examples of 20 smile tracks from each of the self-report classes, "Na...Not my thing" (left), "Meh. It was ok" (middle) and "Heck ya! I loved it" (right).

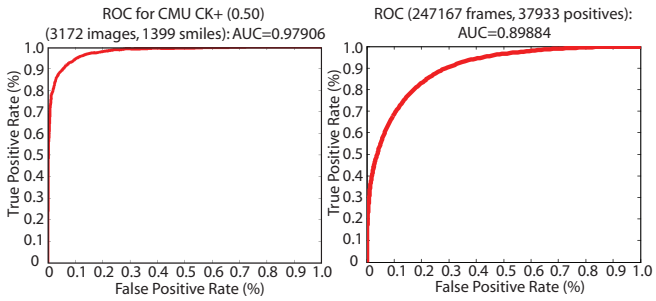

Fig. 16. ROC curves for the smile detection algorithm. ROC curve tested on CK+ database (left), ROC curve for training on CK+ and MPL and testing on webcam data (right).

CK+ and MPL-GENKI[8] databases. All frames were were labeled for smile vs. non-smile by coders. We tested the classifier on 3,172 frames from the CK+ database (the test videos were not included in the training set). The resulting ROC curve is shown in Figure 16 (left), and the area under the curve is 0.979. We also tested how well the smile classifier performs on crowdsourced face videos from a webcam where there was no control on the quality of the resulting face videos (these videos were from a similar but different study to the one described here). A set of 247,167 frames were randomly selected for ground truth labeling. Three labelers labeled each video and the majority label was taken. The resulting ROC curve is shown in Figure 16 (right); the area under the curve is 0.899. The performance of the smile classifier degrades with the uncontrolled videos compared to the CK+; however it is still very accurate. The training data contained examples with a large variation in head position and pose. We believe these results are comparable to those in [13].

The dynamics, speed of onset and offset, for smiles has been shown to be informative [7], [14]. We investigated the dynamics by calculating the gradient of the smile responses. Figure 17 shows histograms of these data for the three commercials. We found that there were not significant differences in the speed of change of the smile intensity in the responses to each commercial. However, this is not necessarily surprising as a large majority of smile responses are due to amusement and we expect

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

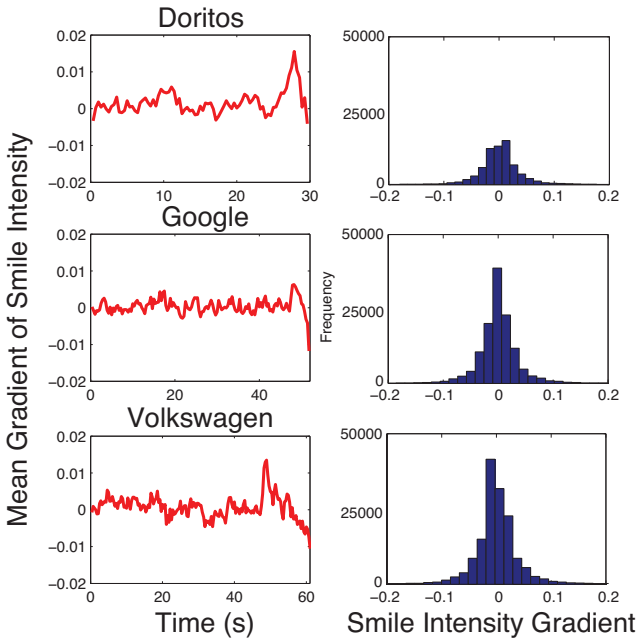

Fig. 17. Mean gradient of response trajectories for the three commercials (left column) and the distribution of response gradients (right column).

there to not be significant differences in the dynamics of specifically amused smiles.

5.4 Facial Expressions and Head Gestures

Previous work has shown that facial expressions and head gestures are highly correlated. Ambadar et al. [14] show that smiles associated with different interpretations occur with different head gestures. We investigated the relationship between head pitch, yaw and roll with smile intensities in the data collected in this study.

Due to the significant differences in the position and pose of the viewers within the videos, see Figure 9, the pose tracks were normalized for each video by subtracting the mean. Figure 18 shows the mean smile track and the mean normalized pose tracks for the three commercials.

The results show a strong association between the pitch and the most intense smile response at the climax of each clip. More specifically the results suggest that generally a pitch elevation is associated with strong intensity smile responses. These results seem congruent with Ambadar's [14] results that suggest that head gestures are linked closely to facial expressions. In Ambadar's work an embarrassed smile was shown to be linked with a downward head motion. The association shown here with predominantly amused smiles is quite different. This finding is an example of how crowdsourcing large amounts of data can contribute to the theory of emotions and their expression.

6 Results and Discussion

Following each commercial, 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). We examined the relationship between the smile responses and the self-report responses for each question. Since viewers were not obligated to complete the responses and the questions "timed out" once the smile response was computed, we do not have responses from all viewers to all the questions. The number of responses for each commercial and question combination are shown in Table 2. On average each question was answered by 45.6% of viewers, which still provides almost 500 examples for each question and commercial combination.

TABLE 2
Number of responses for each commercial to the three self-report questions (participants were not required to answer questions.)

Report	Commercial		
	Doritos	Google	VW
Did you like the video?	498 (47%)	622 (56%)	650 (59%)
Have you seen it before?	430 (41%)	588 (53%)	655 (59%)
Would you watch this video again?	248 (23%)	351 (32%)	443 (40%)

6.1 Liking

Figures 19 - 21 show the mean smile intensities, with standard error (SE) bars, for each of the three ads broken down by self-report of liking. SE is calculated as:

$$SE = \frac{\sigma}{\sqrt{n}} \quad (1)$$

Where σ is the standard deviation of the samples and n is the number of samples (viewers). The vertical lines on the plots indicate the timings of the scenes within the commercials. Below each graph are shown histograms of the timings of the maximum and minimum smile peaks for each of the three self-report classes.

There is a time period at the start of the clips during which the distributions of smile intensities are very similar for each self-report class. This period lasts for 8 secs (27% of the clip length) for the Doritos commercial, 16 secs (30% of the clip length) for Google and 5 secs (8% of the clip length) for the Volkswagen commercial. Table 3 shows the percentage of frames for which the mean smile tracks in each of the self report categories were statistically different ignoring this initial period during which the distributions are very similar. The mean p value for these frames is shown in brackets.

6.2 Familiarity

Figures 22 show the mean smile intensities, with standard error bars, for each of the three ads broken down by self-report of familiarity. The only significant difference in response is for the Volkswagen ad, where viewers watching for the first time show lower mean smile intensity compared to those who have watched it before.

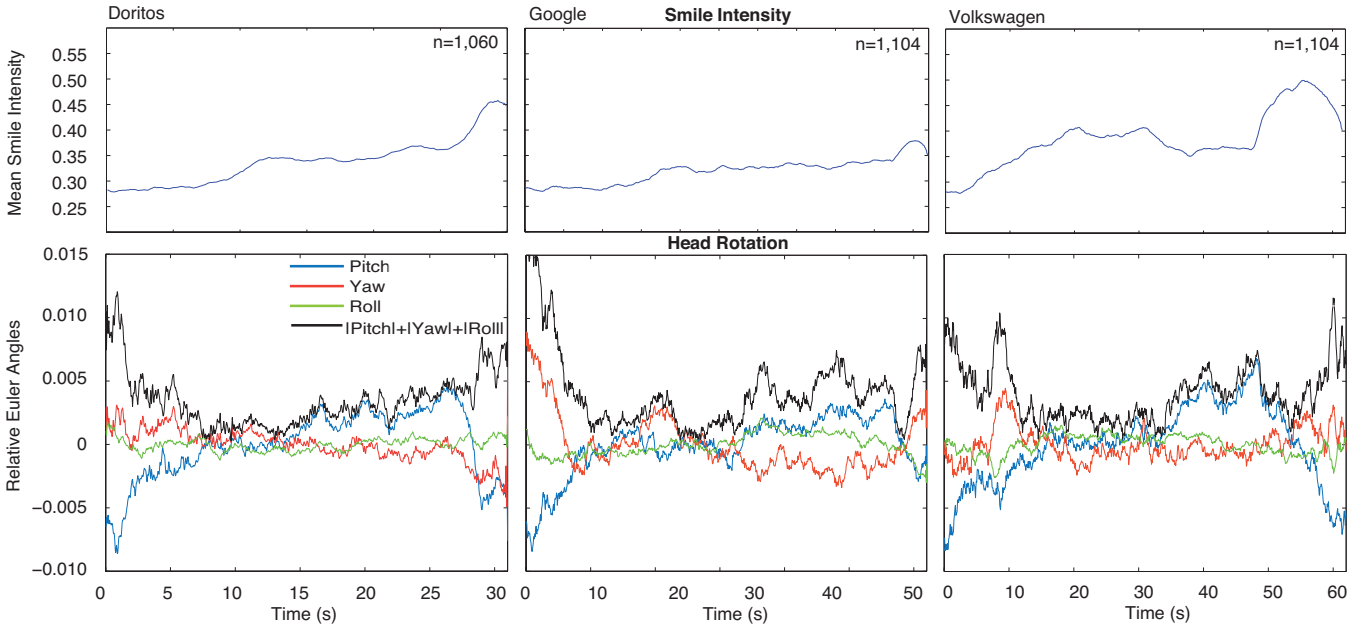

Fig. 18. The mean smile intensity track (top) and mean relative head pose tracks in radians (pitch, yaw and roll) (bottom) of viewers watching the three commercials. The pitch changes dramatically with increased smile responses.

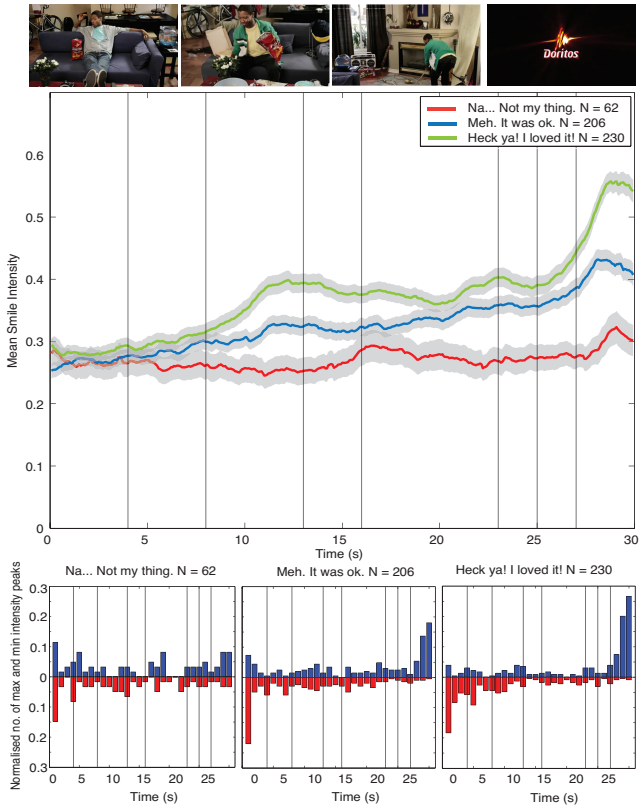

Fig. 19. There are significant differences in the smile responses between people that reported liking the ads more than others. The mean smile intensity and standard error whilst watching the Doritos ad for the three self-report classes (top). Histograms of the maximum (blue) and minimum (red) smile intensity peak locations whilst watching the Doritos ad for the three self-report classes.

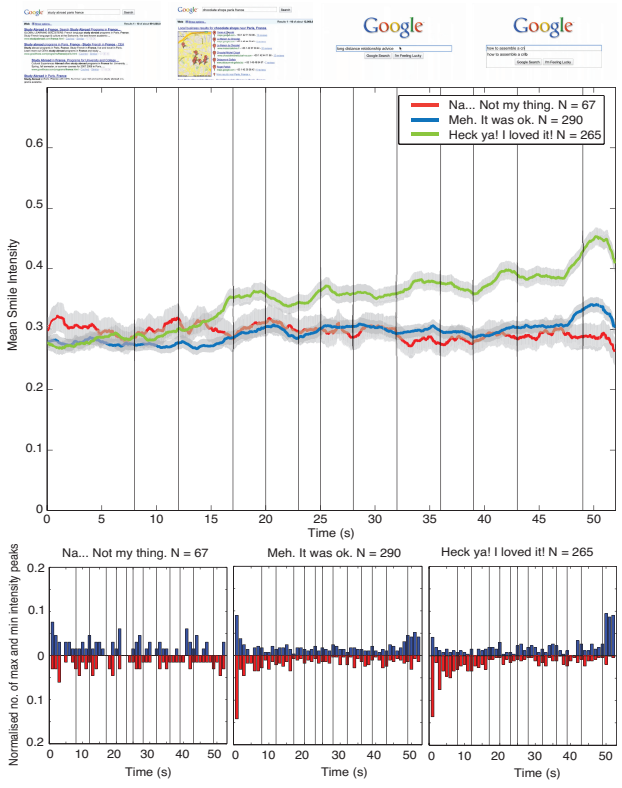

Fig. 20. There are significant differences in the smile responses between people that reported liking the ads more than others. The mean smile intensity and standard error whilst watching the Google ad for the three self-report classes (top). Histograms of the maximum (blue) and minimum (red) smile intensity peak locations whilst watching the Google ad for the three self-report classes.

TABLE 3
Percentage of the clip for which smile tracks were statistically different ($p<0.05$), ignoring the initial period in which all tracks were similar, 8s (Doritos), 16s (Google), 5s (VW). (Mean p value for these frames in.)

	Doritos	Google	Volkswagen
Na...not my thing vs. Meh! It was ok	64.8% (0.011)	0.8% (0.046)	2.8% (0.0382)
Meh! It was ok vs. Heck ya! I loved it!	60.3% (0.0088)	100% (0.0013)	100% (<0.0001)
Na...not my thing vs. Heck ya! I loved it!	99.4% (0.0024)	70.0% (0.0081)	95.2% (0.0014)

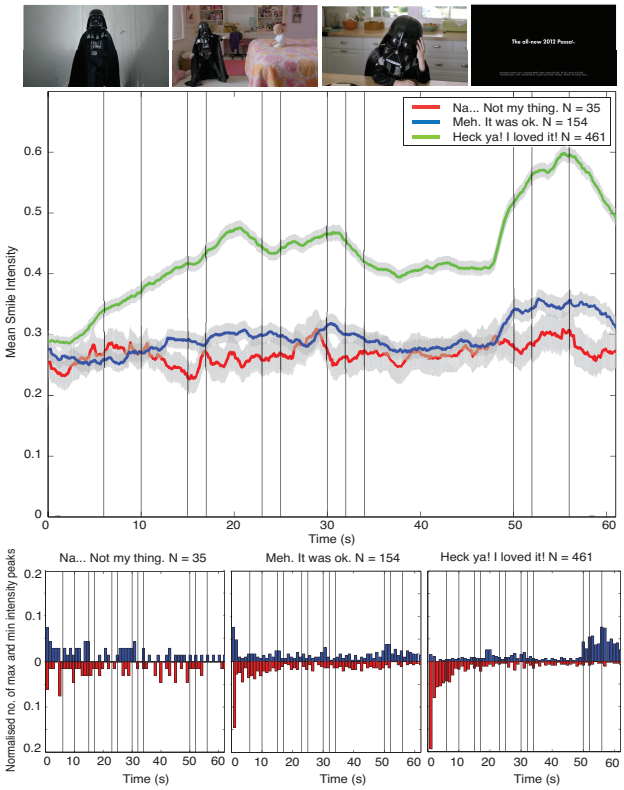

Fig. 21. There are significant differences in the smile responses between people that reported liking the ads more than others. The mean smile intensity and standard error whilst watching the Volkswagen ad for the three self-report classes (top). Histograms of the maximum (blue) and minimum (red) smile intensity peak locations whilst watching the Volkswagen ad for the three self-report classes.

6.3 Rewatchability

The self report responses to the question "Did you like the video?" and the question "Would you like to watch this video again?" were related. Table 4 shows the distribution of responses to the questions. The table has a strong diagonal. The smile responses categorized by responses to the question "Would you like to watch this video again?" were similar to the responses categorized by responses to the question "Did you like the video?".

TABLE 4
Distribution of responses to self-report questions "Did you like the video?" and "Would you like to watch this video again?".

	"Would you like to watch this video again?"		
"Did you like the video?"	Ugh	Maybe	You bet
Nah	66	13	0
Meh	49	258	18
Heck ya	3	151	420

6.4 Discussion

Figure 15 shows that there is considerable variability in the responses, due to individual differences in responses. There are consistent trends showing that on aggregate self-reported liking correlates highly with increased smile intensity, particularly in the case of the Doritos and Volkswagen commercials. As expected the largest difference across the three commercials was between the "Na...not my thing" responses and "Heck ya! I loved it!" responses, smile intensity being significantly different in over 88% of the frames across the three commercials. This supports our intuition and suggests that smile responses to this class of commercial, intentionally amusing commercials, can be used as an effective measure of predicting viewers self-reported liking without having to actually ask them. The smile tracks also provide a much finer level of resolution and avoid the cognitive load associated with self-report measures [35]. For instance, without a time consuming self-report questionnaire it would have not been possible to identify if the participants liked each part of the commercial equally as much or responded more strongly during certain scenes. However, in the behavioral response, such as Figure 14, we can identify when the peaks occur. This analysis allows us to unveil interesting timing information about the responses of people to the commercials.

However, it is possible that the distinction viewers made between the labels "Na...not my thing" and "Meh! It was ok" was not strong enough as for two of the three commercials those that report the commercial as "ok" showed statistically similar results to those that report it as "not my thing". Likert scales could be used to replace these labels in future studies. The difference in smile intensity for the three different classes does not occur immediately but there is a time period at the start of the clips during which the distributions are very similar, up to 16 seconds for the Google ad. Suggesting it takes time for liking or disliking of an ad to become apparent.

Considering the position of the maximum and minimum smile intensity peaks within the responses we can see that there is greater coherence in the responses (more consistently showing greatest smile intensity at the same points) for those that report the commercials were not their thing when compared to the groups who reported liking the commercials. Assuming that one of the advertiser's intentions was to create a commercial

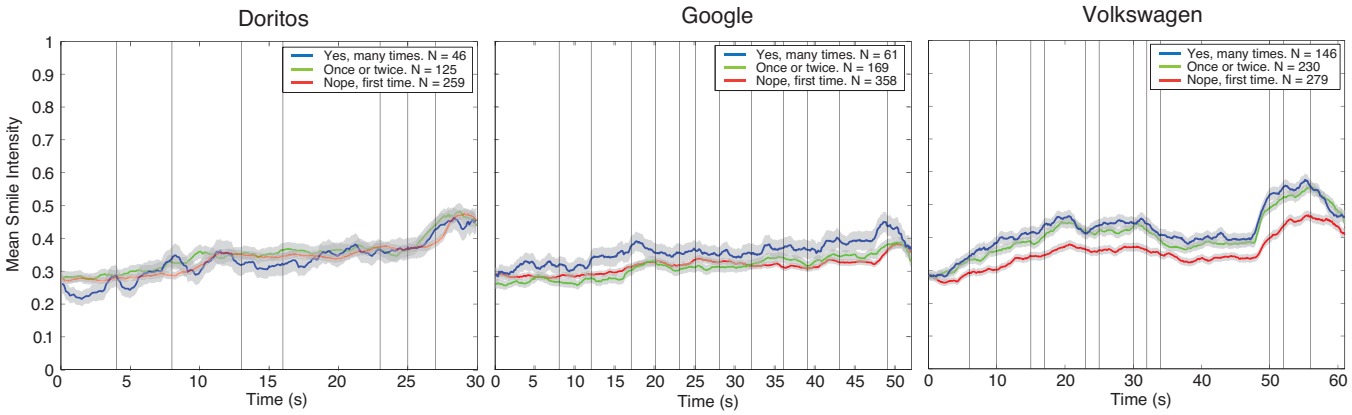

Fig. 22. Graph showing the mean smile intensity and standard error whilst watching the three ads for the three familiarity classes responding to "Have you seen it before?", Doritos (left), Google (middle), Volkswagen (right).

that consumers like, these results suggest that the latter group "got" the intended message of the commercial.

With regard to familiarity, the only significantly different trend exists between those that were seeing the Volkswagen commercial for the first time and those that were not. The mean smile intensity for those that were watching it for the first time was lower. The results suggest that the affective impact of the advertisement increased upon multiple viewing. Interestingly, the Volkswagen commercial was the only one of the three tested here that went viral on the Internet and has subsequently received awards[9]. Inducing positive and consistently strong responses in viewers is one of the potential reasons for the advertisement going viral: our data supports the conclusions of Berger and Milkman [30]. There are a number of other explanations for this difference, for instance that those that chose to watch the commercial for a second or third time may have chosen it because they liked it. However, since there is not a significant difference in the other two videos it would suggest that there is something unique about the VW ad. More generally the results show that on repeated viewings a strong response is still obtained and perhaps more importantly the dynamics of the smile responses are still present. Biel [42] identifies that one of the main reasons why "likeability" might be such a key indicator of a commercials success is that if it is well liked people may be willing to watch it again. In our study we ask participants after the clip whether they would like to see the clip again and these reports were highly correlated with the strength of their self-reported liking.

7 Conclusions and Future Work

We have presented results from the first crowdsourced collection of natural and spontaneous facial responses over the web. The framework allows very efficient collection of examples of natural and spontaneous responses from a large and varied population. In less than two months we collected 5,268 videos from around the world, of which 3,268 were trackable in over 90% of the frames. These responses are aligned with stimuli that were simultaneously presented to the participants. The method did not require payment or recruitment of the viewers but rather used popular media to motivate opt-in participation. The method has allowed us to ask new research questions in a scalable and ecologically valid way. It has further potential to allow exploration of cross-cultural differences in emotion expression as well as non-verbal behavior in atypical populations.

Our system analyzed the responses of people to three intentionally amusing commercials. We have shown that automated analysis of facial responses yields results coherent with self-reports but also provides greater time and intensity resolution. There are significant differences in the intensity and dynamics of smile responses between those that report not liking a particular commercial and those that report liking it. One of the commercials showed significantly increased smile intensity in the responses of those that were not watching it for the first time. This was also the only commercial to "go viral".

In addition to learning about responses to different media from a wide demographic this framework and the data allowed us to learn fundamental relationships between head gestures and facial expressions. We found a relationship between head motions and smile responses, namely that smiles in this context were associated with a backward and upward tilted motion of the head. Findings based on large sets of data have the potential to make contributions to the theory of emotion.

Our results demonstrate that facial responses are potentially a viable alternative to eliciting self-report from viewers. They have the benefit of offering greater temporal resolution, can be measured simultaneously with content and do not place a burden on the participant to fill out tedious questionnaires. In addition, it is arguable that the cognitive load imposed by a self-report system actually means that facial behavior is more accurate a measure, or at least less likely to be cognitively biased.

Our analyses have shown that there are marked differences between the position, scale and pose of partici-

9. http://www.adweek.com/

pants in these natural interactions compared to those in datasets traditionally used for training expression and affect recognition systems, the MMI and CK+ datasets. In particular we showed that position along the vertical axis of the frame, scale of the face within the field of view of the camera and yaw of the head had significantly different distributions to those in traditional lab-based datasets in which these degrees-of-freedom are often constrained. Similarly, we identified a statistically significant difference between the average luminance within the facial region between the Forbes dataset and the CK+ and MMI sets, although the variance of the luminance and the distributions of contrast were not significantly different. These results show that significantly more examples that accurately represent the full extent of these ranges should be included in data used for training and testing systems that might be used in the wild.

Although, these data demonstrate that the dynamic range of viewer position, pose, movement and illumination are greater than those represented in existing datasets we have shown that we were able to collect thousands of trackable videos via the crowdsourcing platform and that these data reveal very interesting trends in the facial responses across a large demographic. This presents a lot of promise for obtaining data for training future algorithms.

The dataset we have collected here (3,268 videos) represents a rich corpus of spontaneous facial responses. In this paper, we have reported on smile responses but we do not distinguish between different types of smiles. Recent research has shown examples of differences between types of smiles in different contexts [7]. It would be interesting to further examine the combinations of facial actions, in particular investigating the number of smiles that occur with the "Duchenne" marker, AU6. Furthermore, we have observed several occurrences of asymmetric smiles, which while part of the FACS system has been largely unstudied. We would like to investigate the automated detection of asymmetric facial expressions and explore the underlying meaning of this. Also, as described earlier, we only included in the analysis the facial videos which were tracked over 90% of time (about 62% of the viewers). We would like to explore several approaches to manipulating the lighting and contrast to potentially improve the face detection results. Finally, while we focus our analysis here on smiles and head pose, we would like to examine this dataset for other facial expressions, such as brow lowerer (AU4) as an indicator of confusion and outer eyebrow raise (AU2) as an indicator of surprise.

This paper validated the framework for crowd sourcing emotional responses, which beyond this experiment, enables numerous research questions to be tackled around automated facial expression recognition as well as understanding the meaning of facial expressions in different contexts. While we limited the content to amusing commercials in order to induce a significant number of smile responses, moving forward we would like to test the framework with content that elicits a wider range of affective responses, for instance disgust, sadness or confusion. We could also systematically explore the relationship between emotion responses and memory, testing various aspects of ad recall. We are interested in exploring the relationship between the intensity and dynamics of the emotional responses with the virality of content. In summary, We are excited about the future work that this platform enables toward characterizing behavior in natural spontaneous online contexts.

ACKNOWLEDGMENTS

Richard Sadowsky, Oliver Wilder-Smith, Zhihong Zeng, Jay Turcot and Affectiva provided access to and support with the crowdsourcing platform. Brian Staats provided front end design for the site. Jon Bruner and Forbes promoted the work on their front page and blog. Google provided use of their facial feature tracker. Procter and Gamble provided funding support for McDuff. This work was funded in part by the Media Lab Things That Think Consortium.

REFERENCES

[1] P. Ekman, W. Freisen, and S. Ancoli, "Facial signs of emotional experience." *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1125, 1980.

[2] R. Hazlett and S. Hazlett, "Emotional response to television commercials: Facial emg vs. self-report," *Journal of Advertising Research*, vol. 39, pp. 7–24, 1999.

[3] A. Quinn and B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011, pp. 1403–1412.

[4] G. Taylor, I. Spiro, C. Bregler, and R. Fergus, "Learning Invariance through Imitation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[5] R. Batra and M. Ray, "Affective responses mediating acceptance of advertising," *Journal of consumer research*, pp. 234–249, 1986.

[6] T. Teixeira, M. Wedel, and R. Pieters, "Emotion-induced engagement in internet video ads," *Journal of Marketing Research*, vol. 49, no. 2, pp. 144–159, 2010.

[7] M. E. Hoque and R. Picard, "Acted vs. natural frustration and delight: many people smile in natural frustration," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011.

[8] H. Gunes, M. Piccardi, and M. Pantic, "From the lab to the real world: Affect recognition using multiple cues and modalities," *Affective computing: focus on emotion expression, synthesis, and recognition*, pp. 185–218, 2008.

[9] A. Fridlund, "Sociality of solitary smiling: Potentiation by an implicit audience." *Journal of Personality and Social Psychology*, vol. 60, no. 2, p. 229, 1991.

[10] T. Teixeira, M. Wedel, and R. Pieters, "Moment-to-moment optimal branding in tv commercials: Preventing avoidance by pulsing," *Marketing Science*, vol. 29, no. 5, pp. 783–804, 2010.

[11] N. Schwarz and F. Strack, "Reports of subjective well-being: Judgmental processes and their methodological implications," *Well-being: The foundations of hedonic psychology*, pp. 61–84, 1999.

[12] M. Lieberman, N. Eisenberger, M. Crockett, S. Tom, J. Pfeifer, and B. Way, "Putting feelings into words," *Psychological Science*, vol. 18, no. 5, p. 421, 2007.

[13] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, pp. 2106–2111, 2009.

[14] Z. Ambadar, J. Cohn, and L. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *Journal of nonverbal behavior*, vol. 33, no. 1, pp. 17–34, 2009.
[15] P. Ekman and W. Friesen, "Facial action coding system," 1977.
[16] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.
[17] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.
[18] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction." *Computer Vision and Pattern Recognition Workshop*, vol. 5, p. 53, 2003.
[19] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 160–187, 2003.
[20] J. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 610–616.
[21] G. Littlewort, M. Bartlett, I. Fasel, J. Chenu, and J. Movellan, "Analysis of machine learning methods for real-time recognition of facial expressions from video," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
[22] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 258–264.
[23] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, p. 5.
[24] G. Mckeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *Proceedings of IEEE Int'l Conf. Multimedia, Expo*, July 2010, pp. 1079–1084.
[25] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.
[26] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," *Affective computing and intelligent interaction*, pp. 488–500, 2007.
[27] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 57–64.
[28] D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard, "Affect valence inference from facial action unit spectrograms," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2011, pp. 17–24.
[29] H. Joho, J. Staiano, N. Sebe, and J. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, pp. 1–19, 2011.
[30] J. Berger and K. Milkman, "What makes online content viral?" *Unpublished manuscript, University of Pennsylvania, Philadelphia*, 2011.
[31] T. Ambler and T. Burne, "The impact of affect on memory of advertising," *Journal of Advertising Research*, vol. 39, pp. 25–34, 1999.
[32] A. Mehta and S. Purvis, "Reconsidering recall and emotion in advertising," *Journal of Advertising Research*, vol. 46, no. 1, p. 49, 2006.
[33] R. Haley, "The arf copy research validity project: Final report," in *Transcript Proceedings of the Seventh Annual ARF Copy Research Workshop*, 1990.
[34] E. Smit, L. Van Meurs, and P. Neijens, "Effects of advertising likeability: A 10-year perspective," *Journal of Advertising Research*, vol. 46, no. 1, p. 73, 2006.
[35] K. Poels and S. Dewitte, "How to capture the heart? reviewing 20 years of emotion measurement in advertising," *Journal of Advertising Research*, vol. 46, no. 1, p. 18, 2006.
[36] J. Howe, *Crowdsourcing: How the power of the crowd is driving the future of business*. Century, 2008.
[37] R. Morris, "Crowdsourcing workshop: The emergence of affective crowdsourcing," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, 2011.
[38] Web address of data collection site:. [Online]. Available: http://www.forbes.com/2011/02/28/detect-smile-webcam-affectiva-mit-media-lab.html
[39] D. Svantesson, "Geo-location technologies and other means of placing borders on the 'borderless' internet," *J. Marshall J. Computer & Info. L.*, vol. 23, pp. 101–845, 2004.
[40] D. McDuff, R. El Kaliouby, and R. Picard, "Crowdsourced data collection of facial responses," in *Proceedings of the 13th international conference on Multimodal Interaction*. ACM, 2011.
[41] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.
[42] A. Biel, "Love the ad. buy the product?" *Admap, September*, 1990.

Daniel McDuff is a PhD candidate in the Affective Computing group at the MIT Media Lab. McDuff received his bachelors degree, with first-class honors, and masters degree in engineering from Cambridge University. Prior to joining the Media Lab, he worked for the Defense Science and Technology Laboratory (DSTL) in the United Kingdom. He is interested in using computer vision and machine learning to enable the automated recognition of affect. He is also interested in technology for remote measurement of physiology. Contact him at djmcduff@media.mit.edu.

Rana El Kaliouby Rana El Kaliouby is a Research Scientist at MIT Media Lab, inventing technologies that sense and have a common-sense understanding of people's affective and cognitive experiences. El Kaliouby holds a B.Sc and M.Sc in Computer Science from the American University in Cairo and a Ph.D. in Computer Science from the Computer Laboratory, University of Cambridge. She is also co-founder, and chief technology officer at Affectiva, Inc. Contact her at kaliouby@media.mit.edu.

Rosalind W. Picard is a fellow of the IEEE and member of the IEEE Computer Society, is Professor of Media Arts and Sciences at the MIT Media Lab, founder and director of the Affective Computing Group, and leader of a new Autism and Communication Technology Initiative at the Massachusetts Institute of Technology. She is also co-founder, chairman, and chief scientist of Affectiva, Inc. Her current research interests focus on the development of technology to help people comfortably and respectfully measure and communicate affective information, as well as on the development of models of affect that improve decision-making and learning. Picard has an Sc.D. in Electrical Engineering and Computer Science from the MIT. Contact her at picard@media.mit.edu.