Crowdsourcing Affective Responses for Predicting Media Effectiveness

by

Daniel Jonathan McDuff

M.Eng., University of Cambridge (2009)
B.A., University of Cambridge (2008)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

© Massachusetts Institute of Technology 2014. All rights reserved.
Crowdsourcing Affective Responses for Predicting Media Effectiveness

by

Daniel Jonathan McDuff

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 2, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

Emotion is key to the effectiveness of media, whether it be in influencing memory, likability or persuasion. Stories and narratives, even if fictional, have the ability to induce a genuine emotional response. However, the understanding of the role of emotions in media and advertising effectiveness has been limited due to the difficulty in measuring emotions in real-life contexts. Video advertising is a ubiquitous form of a short story, usually 30-60 seconds in length, designed to influence, persuade, entertain and engage, in which media with emotional content is frequently used. The lack of understanding of the effects of emotion in advertising results in large amounts of wasted time, money and other resources; in this thesis I present several studies measuring responses to advertising.

Facial expressions, heart rate, respiration rate and heart rate variability can inform us about the emotional valence, arousal and engagement of a person. In this thesis I demonstrate how automatically-detected naturalistic and spontaneous facial responses and physiological responses can be used to predict the effectiveness of stories.

I present a framework for automatically measuring facial and physiological responses in addition to self-report and behavioral measures to content (e.g. video advertisements) over the Internet in order to understand the role of emotions in story effectiveness. Specifically, I will present analysis of the first large scale data of facial, physiological, behavioral and self-report responses to video content collected “in-the-wild” using the cloud. I have developed models for evaluating the effectiveness of media content (e.g. likability, persuasion and short-term sales impact) based on the automatically extracted features. This work shows success in predicting measures of story effectiveness that are useful in creation of content whether that be in copy-testing or content development.

Thesis Supervisor: Rosalind W. Picard
Title: Professor of Media Arts and Sciences,
Crowdsourcing Affective Responses for Predicting Media Effectiveness

by

Daniel Jonathan McDuff

The following people served as readers for this thesis:

Thesis Reader

Rosalind W. Picard
Professor of Media Arts and Sciences
MIT Media Lab, Cambridge

Thesis Reader

Jeffrey Cohn
Professor of Psychology
University of Pittsburgh, Pittsburgh

Thesis Reader

Ashish Kapoor
Research Scientist
Microsoft Research, Redmond

Thesis Reader

Thales Teixeira
Assistant Professor of Business Administration
Harvard Business School, Cambridge
Acknowledgements

I would like to express my sincere thanks to all the people that have helped me complete this degree. This work would not have been possible without many friends and colleagues. The following acknowledgements are by no means exhaustive, for which I apologize.

I would like to thank Roz Picard for being the best advisor I could have wished for. It has been a wonderful opportunity to be at the Media Lab for five years and to have unwavering support and encouragement in my academic life. I would like to thank my thesis committee: Ashish Kapoor, Thales Teixeira and Jeffrey Cohn, for all their time spent developing the research, traveling to meetings and reviewing thesis drafts. Their advice and contributions to this thesis have been, and will continue to be, invaluable.

I am grateful to Rana El Kaliouby who has been a great collaborator and helped to make this thesis work possible through collaboration with Affectiva. Rana’s dedication to a strong scientific grounding for the work has been really encouraging.

Ming-Zher Poh inspired me to apply to the Media Lab. I would like to thank him for being a great roommate, labmate, collaborator and most of all friend. I am grateful for all the dinners, movies and basketball games shared with Ming and Yukkee who are two of my favorite people; I wouldn’t be at this point if I hadn’t met them. The support from friends in the Graduate Christian Fellowship (GCF) - Eric, Victoria, Steph, Michelle, Po-Ru, Kunle, Sam, Adam, Keith, Emily, Marcus, Courtney and many more whom I don’t have space to mention here - has been invaluable. The friendships I have found at the Greater Boston Vineyard, in particular through the youth group, have helpful me to get to know my community and grow in ways I wasn’t expecting during graduate school, in particular I’d like to thank Evadne, Matt, Andrew and Nate.

Rob Morris and Javier Hernandez Rivera have been the most fantastic office-mates (even through those crazy days near deadlines). The projects we worked on together were some of the most fun and innovative. I’d like to thank the rest of the Affective Computing group: Akane Sano, Elliot Hedman, Micah Eckhardt, Yadid Ayzenberg, Ehsan Hoque, Yuta Kuboyama. Jackie Lee, Hyungil Ahn, Kristina Bonakowski, for helping me solve
problems and challenges along the way. The creativity of all my colleagues at the Media Lab has been a constant inspiration throughout my time. Thanks to Dustin Smith, Roarke Hortsmeyer, David Sengeh and Aithne Sheng-Pao to name just a few.

Ashish Kapoor and Mary Czerwinski gave me the wonderful opportunity to intern at Microsoft Research and have a memorable summer in Seattle. This was one of the most fruitful experiences of my PhD. I’d also like to thank my friends at Affectiva: Dan Bender, Evan Kodra, Thibaud Senechal, Jay Turcot, Seth Plough, Rich Sadowsky and Oliver Wilder-Smith. It is really exciting to see much of this technology having an impact.

I have had the pleasure of supervising some fantastic UROPs. Thanks to Abigail Klein, Miriam Greshenson, Fangheng Zhou, Jennifer Shi and Sarah Gontarek for all their hard work and willingness to explore interesting ideas. I would also like to thank all the people who took part in experiments whether in person or over the Internet. The findings in this research are only possible due to the thousands of individuals who took part - often with no compensation - in our studies.

I am indebted to Sina, Yusef, Lottie, Maria, Jimmy, Kate G., Kate B. and Ian for keeping me in touch with England and being so willing to see me at short notice. I hope that I can stay in touch and that these friendships continue to grow stronger.

Heather Beem has been the best friend I could have hoped to have met. I’d like to thank her for sharing life with me, helping me to appreciate things beyond MIT and for supporting me when life was tough - I hope the adventure continues.

I am deeply grateful to Joel, Abbie, Jake and Evie for letting me be a part of their lives - even when I was far away - in particular for the Skype calls that reminded me about what is most important in life. I’d like to thank Nannie, Rob, Shona, Aaron, Jerry, Keith, Julie, Jo, Thomas, Tim, Jo and Toby for the time we have enjoyed together and I hope that I can see more of you all now! My utmost thanks goes to my parents - they let me follow my desire to apply to MIT and supported me in all my decisions - I could never have done it without them. My final recognition and praise are to God, for from Him and through Him and for Him are all things.
In memory of Luke McDuff. You inspired me to study, read and try to be the best I could be. I will miss you always.
In this thesis I present analysis of large-scale emotion measurement via facial responses recorded over the Internet. Only images from videos with consent to be shared publicly are shown in this thesis and related publications.
## Contents

Abstract 3

1 Introduction 34

1.1 Motivation ........................................... 35
1.2 Status Quo ........................................... 37
1.3 Proposed Model ........................................ 38
1.4 Thesis Aims ........................................... 39
1.5 Thesis Outline ........................................ 40

2 Storytelling and Emotion 43

2.1 Advertising ........................................... 44
2.2 Political Ads and Debates ............................. 46
2.3 Movie Clips ........................................... 47

3 Emotion Models and Measurement 48

3.1 Models of Emotion ..................................... 48
3.2 Facial Expressions and Emotion ...................... 50
  3.2.1 Facial Action Coding (FACS) ...................... 51
3.3 Physiology and Emotion (Psychophysiology) ........ 53
3.4 Measurement of Emotions ............................. 54
  3.4.1 Self-Report ........................................ 55
  3.4.2 Facial Behavior ..................................... 58
4 Affective Computing

4.1 Automated Analysis of Facial Behavior
   4.1.1 History
   4.1.2 State of the Art
   4.1.3 Dataset for Training

4.2 Examples of Machine Learning Approaches

4.3 Affective Crowdsourcing

5 Overview of the Six Experiments

5.1 Experiments

5.2 Research Questions

6 Data Collection and Processing

6.1 Data Collection Over the Internet
   6.1.1 Methodology
   6.1.2 Characterizing Data

6.2 Automated Facial Expression Analysis
   6.2.1 Data Labeling
   6.2.2 The Automated System
   6.2.3 Facial Action Units
   6.2.4 Continuous Measures of Emotion
   6.2.5 Accuracy of the Classifiers
   6.2.6 Comparison with Dial Measures

6.3 Remote Measurement of Physiology
   6.3.1 Using a Webcam
   6.3.2 Using a Five Band Digital Camera
7 Preliminary Analysis of Facial Responses over the Internet 100
  7.1 Multiple Viewings and the Impact on Facial Responses ............... 100
  7.2 Intra-Person Variability in Facial Response .......................... 102
  7.3 Inter-Person Variability in Facial Response ........................... 103

8 Super Bowl: Predicting Viewer Preferences 107
  8.1 Aims and Motivation .................................................. 107
  8.2 Data ................................................................. 108
  8.3 Aggregate Responses ................................................. 110
  8.4 Predicting Preferences .............................................. 111
    8.4.1 Pre-Processing Facial Responses Data ......................... 113
    8.4.2 Classification of Responses ................................ 115
    8.4.3 Impact of Smile Detector Performance on Preference Prediction . 116
  8.5 Results/Conclusions ................................................ 119

9 Mars: Predicting Ad Likability and Purchase Intent 122
  9.1 Aims and Motivation .................................................. 122
  9.2 Data and Data Collection ........................................... 123
    9.2.1 Video ads ...................................................... 123
    9.2.2 Respondents ................................................... 125
    9.2.3 Methodology ................................................... 127
    9.2.4 Facial Coding .................................................. 129
    9.2.5 Expressiveness of Viewers ................................ 129
    9.2.6 Aggregate Characteristics ................................... 133
  9.3 Classification ....................................................... 133
    9.3.1 Calculating Aggregate Metrics ................................. 134
    9.3.2 Extracting Summary Features ................................ 134
    9.3.3 Computing Labels .............................................. 135
    9.3.4 Model .......................................................... 137
9.4 Results and Discussion .............................................. 137
  9.4.1 Ad Liking Prediction ........................................... 137
  9.4.2 Purchase Intent Prediction ................................. 139
9.5 Discussion and Conclusions .................................... 142

10 Mars: Predicting Sales ............................................. 145
  10.1 Aims and Motivation ........................................... 145
  10.2 Data and Data Collection .................................... 147
    10.2.1 Video ads ................................................. 147
    10.2.2 Respondents .............................................. 147
    10.2.3 Methodology ............................................. 148
    10.2.4 Sales Data ................................................ 148
  10.3 Features .......................................................... 149
    10.3.1 Calculating Aggregate Metrics ......................... 149
    10.3.2 Extracting Summary Features ......................... 149
    10.3.3 Computing Labels ....................................... 150
    10.3.4 Model ..................................................... 150
  10.4 Results .......................................................... 150
    10.4.1 Comparison of Sales Labels with Human Coding ...... 150
    10.4.2 Facial Metrics and Self-report Prediction ............ 151
  10.5 Case Studies .................................................... 152
    10.5.1 Case I: Facial responses correctly predict sales (not self-report) . 153
    10.5.2 Case II: Self-report correctly predicts sales (not facial response) . 153
  10.6 Discussion and Conclusions .................................. 154

11 Electoral Debate: Predicting Candidate Preferences ....... 156
  11.1 Aims and Motivation ........................................... 156
  11.2 Data ............................................................ 158
  11.3 Results and Discussion ...................................... 161
13.3.3 The Interactions ........................................ 183
13.3.4 Potential Applications .................................. 183
13.4 Social Impact and Participation ............................ 185

14 Conclusions ....................................................... 188
14.1 Thesis Contributions ......................................... 188
14.2 Other Relevant Contributions ............................... 190
  14.2.1 Exploiting Structure for Facial Action Recognition 190
  14.2.2 Acume ....................................................... 190
  14.2.3 AM-FED Dataset ........................................ 191
  14.2.4 AffectAura ............................................... 191
  14.2.5 CardioCam and the Affect Mirror ...................... 192
14.3 Future Work .................................................. 195
  14.3.1 Realtime Prediction of Ad Effectiveness ............... 195
  14.3.2 Affective Responses with Large Data ................... 195
  14.3.3 Affective Preference Systems ......................... 196

A Classifier Training Data ........................................ 197
  A.1 FACS Coding ............................................. 197
  A.2 Reliability of Labels ...................................... 198

B Survey Details .................................................... 200
  B.1 Mars ......................................................... 200

C FACS Codes ....................................................... 205

Bibliography ......................................................... 208
References .......................................................... 209
List of Figures

1-1 Thesis Model: Automatically measured facial and physiological responses capture dimensions of viewers’ affective responses. These measurements can predict media effectiveness including individual and corporate preferences, purchase intent and sales. ................................................................. 39

2-1 In this thesis I will be analyzing affective responses to video advertisements, political debates and Hollywood movies. ................................................................. 44

3-1 Action Unit Spectrograms (AUSpec) from the 66 participants watching clips from ‘The Lion King’ (left) - Sadness and ‘When Harry Met Sally’ - Amusement (right). The color intensity of each block reflects the percentage of viewers within the population who exhibited the corresponding action unit at that time. Deeper color corresponds to more viewers. The action unit numbers refer to FACS codes (see FACS code glossary in Appendix C). Greater instances of AU4 are present during the sadness clip and greater instances of AU12 are present during the amusement clip. Data from McDuff et al. (2010). ................................................................. 51

3-2 The Self-Assessment Manikin (SAM). Top) valence report, middle) arousal report, bottom) dominance report. ................................................................. 58

3-3 Feeltrace is a continuous emotion reporting tool that uses a circumplex. . . . 58
4-1 The three main components present in most facial expression or AU recognition systems: 1) a method for registering the face, 2) calculation of shape and/or appearance feature descriptors and 3) classification of AUs or expressions using the feature descriptors.

6-1 Overview of the user experience and web-based framework that is used to crowdsource the facial videos. The video from the webcam is streamed in real-time to a server where automated facial expression analysis is performed. All the video processing can be performed on the server side.

6-2 Example frames from the crowdsourced videos. The videos vary in illumination, image quality, pose, position and movement of the viewer, number of viewers, occlusions and background activity.

6-3 A) Histogram of head scales for the CK+ (top), MMI (center) and our webcam (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 our webcam 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 our webcam 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 our webcam (bottom) datasets. Examples of poses with pitch=-0.13 rads, yaw=-0.26 rads and roll=-0.19 rads are shown.

6-4 Screenshot of the video labeling tool ViDL used to label the video data. The onset and offset of each action unit was labeled by at least three labelers with the majority label being taken as ground truth.
6-5 Cropped examples of frames from ViDL with positive labels for action units that were coded in the groundtruth labeling stage. Smile and negative AU12 are labeled separately instead of labeling symmetrical AU12. The AU definitions can be found in Appendix C.

6-6 Locations of the 22 landmark points automatically labeled using the Neven-vision tracker.

6-7 Descriptions and examples of the action units (AU2, AU4, AU9/AU10) and dimensions continuous emotion (valence) which are captured by the facial coding technology.

6-8 Comparison between action unit classifier predictions (green) and manually coded action unit labels (blue and black dashed). Threshold of hand labels based on > 0.5 agreement between coders. Frames from the sequences are shown above. Top) Smile classification example, middle) AU2 classification example, bottom) AU4 classification example.

6-9 Time series plot of aggregate self-report dial responses (self-reported interest) (in Black) versus the smoothed facial metrics (in Red: smile metrics for Crime Drama and AU02 metrics for Sitcom). The non-smoothed facial data is shown in faded blue. Correlation between the smoothed metrics and dial data: -0.748 and 0.656 respectively. Self-reported interest is negatively correlated with smile metrics for the crime drama and positively correlated with AU02 metrics for the sitcom.

6-10 Algorithm for photoplethysmographic measurement from a webcam video and the extraction of heart rate, respiration rate and HRV from the resulting BVP signal.
6-11 Left) Five band camera light sensitivity profile. In addition to red, green and blue light sensors this camera can measure orange and cyan bands. Right) Five band camera sensor layout. The image shows the arrangement of the colors in a 4x4 pattern that repeats across the sensor. Each pixel on the sensor measures one color as determined by its position.

6-12 Comparison of remotely measured physiological parameters for 10 participants during rest and cognitive load conditions. a) Heart rate, b) respiration rate, c) HRV HF, d) HRV LF/HF ratio. Mean parameters are shown for: e) heart rate, f) respiration rate and g) HRV LF/HF. Error bars show one standard deviation either side of the mean. Respiration rate and LF/HF ratio are the most discriminative features between the two states, heart rate was the least discriminative.

6-13 Heart rate variability spectrograms (normalized amplitude) calculated using: Top) RGB camera signals, middle) GCO camera signals and bottom) contact finger sensor. Qualitatively the measurements from the GCO channels more closely match those from the contact sensor (reinforcing the quantitative comparisons in Table 6.2). Shown are sessions in which participants 5 and 10 were at rest and sessions in which participants 5 and 10 were under cognitive load. There are stronger low frequency components in the latter as we would expect less parasympathetic nervous system activity. The measurements for participant 5, under cognitive load, made using the camera suggest that the camera (using GCO bands) may have been more accurate than the finger sensors.

7-1 Aggregate (no. of viewers = 23) smile classifier output for a commercial on first (blue) and second (green) viewing. Correlation = 0.802, p<0.05. Shaded area shows the standard error.
7-2 Aggregate (no. of viewers = 23) smile classifier output for a commercial on first (blue) and second (green) viewing. Correlation = 0.430, p<0.05. Shaded area shows the standard error. ............................................ 102

7-3 Individual smile (blue) and disgust (red) responses to four video ads (a-d) and the baby clip (e) for four individuals (i-iv) - chosen to illustrate the variety of responses. Example frames from the first ad for two of the subjects are shown. ............................................ 104

7-4 Aggregate smile and disgust metrics with different population sizes, shaded area = 95% confidence bounds. Top) mean smile probability for Ad A, middle) mean disgust probability for Ad B, bottom) mean smile probability for baby video. ............................................ 106

8-1 Map showing the location of the 3268 viewers, based on their IP address. No viewers IP was located outside of the latitudes shown. ................. 108

8-2 Screenshots of the web interface in the Super Bowl study. a) Homepage, b) Webcam permissions, c) Self-report questions, d) Graphical representation of smile intensity. ................................. 109

8-3 The self-report questions with which the viewers were presented after watching the ads in the Super Bowl study. ................................. 109

8-4 Framework for classification of content liking and desire to view again based on automatically detected smile responses recorded over the web. Smile metrics are extracted from the recorded facial response. The smile metrics are filtered and temporal features extracted from 20 evenly spaced bins. Resulting features used for classification. ................................. 112

8-5 Funnel chart showing the participation at the early stage of the Super Bowl experiment. 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. ................................. 113
8-6 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.

8-7 Linear Discriminant Analysis (LDA) mapping of the individual 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.

8-8 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 ads 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.

8-9 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.

8-10 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.
8-11 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.

9-1 Screenshots of a subset of the 170 ads tested in this study. The ads were for products in one of the four categories: pet care, confectionery (chocolate, gum and candy), food (instant rice and pasta products) and cosmetics/toiletries. The ads were from different countries: France, Germany, the UK and the US.

9-2 The a) age (in years), b) gender and c) economic split (annual salary in $1000s) of the 1,223 viewers who took part in the Mars study.

9-3 Experimental procedure for the Mars experiment. The experiment was divided into two parts. 1) A pre-survey with baseline questions and 2-4) a main survey with ads and follow-up questions. During the main survey each viewer watched 10 ads and answered questions about liking and desire to share following each and then answered questions about purchase intent and brand likability at the end of the survey.

9-4 Example of the brand recall task. Participants were asked to enter the name of the brand represented from a forced choice list.

9-5 Example images from the brand recognition task. Brand names and logos were pixelated. Participants were asked to enter the name of the brand represented from a forced choice list.

9-6 Histograms of the number of frames with each expression probability. From top to bottom: eyebrow raise, smile, disgust, valence. In 82.8% of frames was there no detectable eyebrow raise, smile, disgust or positive/negative valence expression with magnitude above 0.1. Responses to ads in naturalistic settings are sparse but for all the ads there were expressive responses within the 70+ viewers.
Mean valence metrics for the 170 ads sorted by ascending mean valence.

1) Aggregate metric tracks calculated from all viewers who watched the ad. 2) Features extracted from each of the aggregate metric tracks: a) mean value, b) minimum value, c) maximum value, d) the gradient of the linear trend. 3) Summary features extracted from the facial expression metrics used to predict ad effectiveness.

Distribution of average: left) ad liking response and right) delta in purchase intent response for all ads. The median, minimum and maximum values are all shown. The report of liking was significantly (p<0.001) greater than neutral.

Scatter plot of ads with aggregate smile trend against aggregate disgust trend. Ads with higher mean ad liking score are shown in blue, ads with lower mean ad liking score are shown in red. The smile gradient is more discriminative than the disgust gradient.

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the ad liking models varying the SVM decision threshold. Black) the performance using face features, blue) the performance using face and context features combined. Unbroken lines) results for all ads, broken lines) results for only the amusing ads.

Aggregate ad response metrics that correctly and incorrectly classified by the ad likability model. True positives, true negatives, false positives and false negatives shown. Aggregate: eyebrow raise - black, smiles - green, disgust - red. High peak levels of positive expressions, high expressiveness and strong increasing trend in positive expressions predict high ad likability. Low expressiveness predicts low ad likability. Individual plot outlines indicate the product category the advertised product.
9-13 Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the purchase intent models varying the SVM decision threshold. Black) the performance using face features, blue) the performance using face and context features combined. Unbroken lines) results for all ads, broken lines) results for only the amusing ads. 

9-14 Aggregate ad response metrics that were correctly and incorrectly classified by the purchase intent model. True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) shown. Brand appearances within the ads are indicated by the broken gray lines. Notice how the peak in smile activity is preceded by a brand appearance in the TPs and not in the TNs. Aggregate: eyebrow raise - black, smiles - green, disgust - red. Individual plot outlines indicate the product category the advertised product. The results seem to generalize across product categories.

10-1 Flow diagram of the feature extraction for the sales prediction task. Facial metrics extracted from each video and aggregated. Temporal features extracted from the aggregate metrics.

10-2 Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the sales models (for non pet care ads) varying the SVM decision threshold. green) the performance using self-report features, red) the performance using face features, blue) the performance using self-report and face features combined.

10-3 a) Aggregate self-report and b) aggregate facial responses for the ad in case study I. The facial responses correctly predicted the sales measure for this ad and the self-report responses did not. Brand appearances within the ad are indicated by the broken gray lines.
10-4 a) Aggregate self-report and b) aggregate facial responses for the ad in case study II. The self-report responses correctly predicted the sales measure for this ad and the facial responses did not. Brand appearances within the ad are indicated by the broken gray lines. .......................... 155

11-1 Top) Images from the election debate between President Barack Obama and Governor Romney. Bottom) Images taken from a subset of facial responses data collected with permission to share publicly. ....................... 157

11-2 The number of participants who took part in the Electoral Debate study and watched at least one of the debate clips. ............................... 159

11-3 Questions asked before the participants watched the debate clips during the Electoral Debate study. Responses to all questions were required. ............... 159

11-4 Mean valence during the debate clips for those that reported a preference for Obama (blue) and Romney (red) after watching the clip. The shaded area represents the standard error range. Below the plots we show which candidate was speaking and which (or both) was on screen during the clip. The letters and dotted lines correspond to significant parts of the clips - transcripts of these parts of the clips can be found in Table 11.3. ............ 160

11-5 Top) Examples of smirks during two parts of the debate clips. Top left) Smirks that occurred during Clip 1, which were followed by smiles/positive valence. Top right) Smirks that were not followed by smiles/positive valence. Bottom) Plots of the aggregate smiles and number of smirks for debate Clip 1, and debate Clip 4. Regions in which a greater number of smirks occurred are highlighted in green. ......................... 162

11-6 Precision-recall (top) and ROC curves (bottom) for the voter preference prediction task. Green) Results for all eligible voters. Blue) Results for all eligible voters with no or an independent party affiliation. ............... 165
12-1 Arrangement of apparatus for multimodal data collection. EDA, BVP and respiration were measured with contact sensors. EDA was measured from both hands. A video of the viewer was recorded simultaneously. . . . . . . . 173

12-2 Examples of the physiological data collected during the multimodal experiment. Screenshots of the content and frames of the response are shown above. Notice the sharp pinching of the BVP amplitude as the EDA increases at 19 mins and again 40 seconds later. This characteristic was observed frequently - see Figure 12-4 for more examples. . . . . . . . . . . . 173

12-3 Bland-Altman plots showing the agreement between 30s epoch heart rate measurements obtained from participants watching media content using contact BVP and remote camera method (total of 1,607 pairs from 6 participants). The dotted lines represent the mean and 95% confidence intervals. Different colors represent different participants. . . . . . . . . . . . . . . . 174

12-4 BVP, respiration and EDA responses of six viewers watching eight emotion eliciting clips separated by neutral clips of a beach scene. Clips are the same for all viewers but the viewing order is not. The black lines indicate the transition between clips. Note: the scales for the BVP and EDA plots are not the same. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 176

13-1 The Affective Billboard. a) Schematic: A webcam and facial expression toolbox are used to detect the presence of a viewer and analyze their response to displayed content. A custom prediction engine evaluates viewers’ preferences based on their facial response and adapts the displayed content. b) Real-time viewer interaction with the Affective Billboard. . . . 183

13-2 Design of the user interface, a) scoreboard showing the performance of each ad based on the results of the liking prediction algorithm, b) results page showing an individual’s response to an ad and recommended ads, c) flow diagram of the interactions that can occur with the system. . . . . . . . 184
13-3 Digital advertising displays, such as the ones above, are common and facial analytics could be used intelligently and dynamically to display the content to make them more efficient. .................................................. 185

14-1 A screenshot of Acume, an open-source toolkit for exploring and visualizing behavioral data on multiple scales. .................................................. 192
14-2 A screenshot of the AffectAura emotional memory visualization tool. . . . . 193
14-3 The Affect Mirror. Top left) In vital signs mode measuring and displaying the user’s heart rate, top right) in life-logging mode with images and EDA measurements, bottom) a diagram of the composition of the affect mirror. . 194

A-1 Bar graph showing the mean in the Spearman-Brown reliability for each of the labels ................................................................. 199

B-1 Instructions... ................................................................. 201
# List of Tables


3.2 Summary of the autonomic measurement of emotions in media and marketing. SR = Self-report, EDA = Electrodermal Activity, EMG = Facial Electromyography, FC = Facial coding. ................................. 65

5.1 Summary of datasets collected in the five experiments, including the number of videos tests, the number of responses and labels. SR = Self-report, M=Memory tests, S=Sales. U = unknown. ................................. 76

6.1 Number of videos and frames used for training the action unit and expression classifiers and the area under the receiver operating characteristic (ROC) curve for testing. ................................. 89

6.2 Comparison of the correlations between the contact finger sensor measurements and camera measurements for all combinations of the camera color channel signals. For all correlations $p < 0.01$. On the right are the channel combinations ordered from lowest mean correlation to highest mean correlation. The GCO channel combination performed best. ................. 95

6.3 Classification accuracy for prediction of rest vs. cognitive load using different combinations of features. ................................. 98
8.1 Distribution of responses to self-report questions “Did you like the video?” and “Would you like to watch this video again?”.

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

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

9.1 Number of videos tested from each product category and emotion category (categorized using MTurk labelers). A majority of the ads were intentionally amusing.

9.2 Percentage of the 3,714,156 frames with expressions metrics within 10 evenly spaced classifier output bins centered on the values shown.

9.3 Area under the ROC and PR curves for the ad liking classifier: top) all ads (N=165), bottom) only amusing ads (N=75).

9.4 Confusion matrices for the optimal liking classifier: top) all ads (N=165), bottom) only amusing ads (N=75). Based on threshold of point closest to (0,1) on the ROC curve.

9.5 Area under the ROC and PR curves for the purchase intent classifier: top) all ads (N=163), middle) only non pet care product ads (N=111), bottom) only pet care product ads (N=52).

10.1 Number of videos tested from each product category and emotion category (categorized using MTurk labelers).

10.2 Confusion matrix for the human coding of sales performance.

10.3 Area under the ROC and PR curves for the sales classifier: top) all ads (N=163), middle) only non pet care product ads (N=111), bottom) only pet care product ads (N=52).
10.4 Confusion matrices for the best performing sales classifier: top) all ads (N=163), middle) only non pet care product ads (N=111), bottom) only per care product ads (N=52). Based on threshold of point closest to (0,1) on the ROC curve. ................................................................. 153

11.1 Demographic profile of the participants in the 611 facial responses videos collected. ................................................................. 158

11.2 Synopses and lengths of the five debate clips that viewers watched during the Electoral Debate study. The number of trackable face video responses collected is also shown on the right. ........................................ 160

11.3 Excerpts from the debate transcript which correspond to the marked points on Figure 11-4. The color of the dots refers to the speaker: blue = Obama, red = Romney. ................................................................. 160

11.4 Top) Confusion matrix for prediction of voter preference across all the eligible voters. Bottom) Confusion matrix for prediction of voter preference across the eligible voters with no or an independent party affiliation. Threshold determined as the case closest to the ROC (0,1) point. .......... 166

12.1 Order, descriptions and durations of the media clips viewed. Ads were placed in between the emotion eliciting clips that were viewed in the Multimodal studies. ................................................................. 169

A.1 Definitions of the labels for the dataset and the number of frames and videos in which each label was present (agreed by majority of labelers). Positive examples of each of the labels are shown in Figure 6-5. .......... 199
Chapter 1

Introduction

“You kids don’t know what you want! That’s why you’re still kids, because you’re stupid! Just tell me what’s wrong with the freakin’ show!”

Roger Meyers Jr. - The Simpsons

Thesis Statement

The Internet can be used to crowdsource responses to media content more quickly and efficiently than traditional methods of research. Rich affective information can be measured from ubiquitous sensors (e.g. webcams). Taking advantage of these qualities via an online framework and state-of-the-art computer vision methods we can drastically improve our understanding of the effectiveness of media content and storytelling and build models to predict self-report and behavioral outcomes. In addition, the data collected can reveal fundamental information about the way humans express their feelings.
1.1 Motivation

Emotions influence perception (Zadra & Clore, 2011; Phelps, Ling, & Carrasco, 2006; Scott et al., 1997), memory (Reisberg & Hertel, 2004), decision making (LeDoux, 2002; Damasio, 1994) and many other processes. Stories, even if fictional, have the ability to induce a genuine emotional response and this is something people experience many times a day. As such, emotions are key to the effectiveness of narratives and storytelling whether it be in influencing one’s memory of the content, liking of the characters or how much one is persuaded to change an opinion or behavior. These processes may not be consciously accessible or easy to verbalize.

Despite its power, the understanding of the role of emotions in storytelling and advertising effectiveness has been limited due to the difficulty of measuring emotions in real-life contexts. The epigraph above is a quotation from a scene in the TV show “The Simpsons” in which the creator of an amusing cartoon series, Roger Meyers Jr., is trying to use a focus group of children in order to understand how to make his cartoon show (“The Itchy and Scratchy Show”) more appealing after the ratings have dropped. The children have difficulty verbally expressing what they would like the show to feature; attempts to measure their attitudes toward the show are clearly lacking. The scene highlights the challenges presented by verbal and dial based self-report measures in such a context.

There are many examples of stories and narratives used around us. Advertisements - from traditional print, radio and TV adverts to Internet banner and online video adverts - are increasingly placing more emphasis on emotional content. Indeed, advertisers have long strived to design ads that elicit emotions in viewers and have struggled to measure the extent with which they have been successful in doing so. A case in point: in one ad for a popular chocolate bar, two construction workers are driving excavators and having fun playing a crazy game of soccer with oil barrels; in another, several Monks in a monastery have a chocolate bar then start ringing the bells while jumping up and down in a crazy way. Both ads are clearly aiming to entertain, but one of the two ads yielded a much greater positive impact on sales.
Online video is growing fast with viewing figures consistently being broken year after year. In the US in November 2013¹ over 189 million viewers watched videos online and the average viewer watched 19 hours of video online. In the US billions of dollars are spent on online video ads each year. A total of nearly 27 billion ads were viewed in November 2013 and this reached more than 50% of the US population. This is almost three times the number of viewings compared to the same month in 2012. Video ads accounted for 36.2% of all videos viewed. In addition to viewing videos more and more people are sharing video content with others. In 2013 72% of adult Internet users used video-sharing sites.² Internet TV sites like Hulu and Netflix frequently ask viewers about the relevance or their enjoyment of ads in an attempt to target them more effectively and measure performance. However, there remains truth in Ray and Batra’s (1982) statement: “an inadequate understanding of the role of affect in advertising has probably been the cause of more wasted advertising money than any other single reason.” This statement applies beyond advertising to many other forms of media and is due in part to the lack of understanding about how to measure emotion accurately in real-life contexts.

This thesis deals with evaluating the effectiveness of emotional content in storytelling and advertising beyond the laboratory environment and traditional measurement techniques using remotely measured facial and physiological responses and an online data collection framework. I analyze challenging ecologically valid data collected over the Internet in the same contexts in which the media would normally be consumed and build a framework and set of models for automatic prediction of effectiveness based on affective responses. In particular I will look at stories in the form of TV advertisements, movie clips, TV shows and political debates.

¹http://www.comscore.com
²http://pewinternet.org/
1.2 Status Quo

Traditionally, consumer testing of video advertising, whether by self-report, facial response or physiology, has been conducted in laboratory settings. Lab-based studies have many benefits such as allowing highly accurate measurement of physiological parameters or muscle movements. However, there are also a number of challenges. Subjects can be influenced by the presence of an experimenter and/or their comfort with the situation, factors that are unrelated to the stimulus of interest. These may impact the participant’s emotional experience (Wilhelm & Grossman, 2010) and the influence is difficult to quantify. In addition, running such studies is labor intensive and may not be cost effective. Conducting experiments outside a lab-based context, and without the use of expensive, obtrusive and uncomfortable sensors, can avoid such problems. However, this type of data collection does present a number of technical challenges that must be overcome. How can you measure emotions without physical contact with the viewer? How can you collect data reliably and efficiently?

Self-report is the current standard measure of affect, where people are typically interviewed, asked to rate their feeling on a Likert scale or turn a dial to quantify their state (affect dial approaches). While convenient and inexpensive, self-report is problematic because it is also subject to biasing from the context, increased cognitive load and other factors of little relevance to the stimulus being tested (Schwarz & Strack, 1999). Other drawbacks of self-report methods include the difficulty for people to access information about their emotional experiences and their willingness to report feelings even if they didn’t have them (Cornelius, 1996). For many the act of introspection is challenging to perform in conjunction with another task and may in itself alter that state (Lieberman et al., 2007). Survey methods are limited in resolution as questions can only be asked at discrete time points. Although affect dial approaches provide a higher resolution report of a subject’s response compared to a post-hoc survey, subjects are often required to view the stimuli twice in order to help the participant introspect on their emotional state. Self-reporting experiences can become laborious too. New ways of tapping into the emotional experiences
of a viewer/reader without constant interruptions could be very powerful.

Kassam (2010) performed one of the largest analyses of facial responses to media content. However, due to the time and expertise required to hand code Facial Action Coding System (FACS) annotations this was limited to 88 participants watching eight one to two minute videos clips (with coding performed at one second intervals). Kassam showed facial expressions are related to self-reported experiences but the two measurements do not capture exactly the same information. The results presented in this thesis support and extend these findings with real-world data collected outside the lab from a much greater number of clips and many orders of magnitude more participants. This thesis is the first to conduct facial analysis studies with large crowd-sourced online data. I show that facial, physiological and self-report measures of affective information can be used to predict viewer preferences and other measures of success and that they yield deeper insights than are available from conventional techniques. I will focus on facial expressions as the main remote measure of affect but I also consider physiological measures of emotion beyond the face in the form of heart and respiration rate and heart rate variability - which can all be measured remotely using methods I have developed.

1.3 Proposed Model

In this thesis I propose a model for predicting measures of storytelling impact and effectiveness from automatically measured affective responses. Specifically, I focus on data collected remotely over the Internet. The measured features include facial expressions, physiological responses and self-report responses. Figure 1-1 shows the model design. The remote measurements capture different dimensions of the affective response. These are modeled as discrete expressions such as smiles and disgust and continuous dimensions such as valence. I propose that these measures of affect (captured remotely) can predict success of content in a number of contexts. Firstly, that emotional responses to content are closely related to individual’s reported preferences of content they like or would like
Figure 1-1: Thesis Model: Automatically measured facial and physiological responses capture dimensions of viewers’ affective responses. These measurements can predict media effectiveness including individual and corporate preferences, purchase intent and sales.

to watch (b). Secondly, that emotional responses to content can predict behavioral and aggregated metrics of success (e.g. short-term sales lift) (d). Finally, I hypothesize that predicted effectiveness of content measured via cognitively evaluated self-report will not be as accurate as predicted effectiveness that takes into account automatic measures of emotional response (c) < (d). Further to this, the combination of self-reported measures and automatically measured emotional responses will be more predictive still (d) < (b) + (c) + (d).

1.4 Thesis Aims

The aims of this thesis, and as outlined in my thesis proposal, are as follows:

- To use a custom cloud-based framework for collecting a large corpus of response videos to online media content (advertisements, debate clips) with self-report responses and measures of success (sales). To make use of this framework to collect data from a diverse population to a broad range of content.
• To collect supplementary data in more controlled settings with gold-standard physiological measurements alongside facial coding. In order to improve and validate remote physiological measurement techniques and to demonstrate how these can be useful for media measurement.

• To design, train and evaluate a set of models for predicting key measures of story/advertisement effectiveness based on facial responses automatically extracted from the videos and self-report ratings collected online.

• To propose generalizable emotional profiles that describe an effective story/advertisement in order to practically inform the development of new content.

• To implement an intelligent real-time system (demo) that incorporates the findings into a fully automated intelligent classification of responses to advertisements on a digital public display - *The Affective Billboard*.

### 1.5 Thesis Outline

The remainder of this thesis will cover the following material:

**Chapters 2, 3 and 4** provide a background on the role of emotion in storytelling and an introduction to the related work on emotion and media measurement and affective computing on which this thesis is built. I will motivate the need for the development of technology for measuring emotional responses to media “in-the-wild.”

**Chapter 5** presents a summary of the experiments performed during my thesis work and the hypotheses they were designed to test.

**Chapter 6** describes the novel cloud-based framework and methodology used for data collection. I will also describe technical details about the automated analysis of facial ex-
pressions and physiology from video. Here I will present the details of existing systems I have employed in addition to systems I have designed during the course of my thesis work.

**Chapter 7** presents general observations from the analysis of crowdsourced affective responses that will be relevant in the following chapters (7-11). These results include: effects of familiarity, range of expressiveness and baseline differences. The large-scale data presented in this thesis not only allow us to better understand viewer responses to video content but also to examine more fundamental questions about the variability in facial behavior and expressiveness of viewers.

**Chapter 8** describes how affective measurements can be used to predict liking and desire to watch again preferences in advertising/media. In particular, I show that it is possible to predict individuals’ preferences (liking, desire to watch again) for online media based on smile responses. This was the pilot study that laid the groundwork for the experiments described in the following chapters.

**Chapter 9** describes how affective measurements to video advertising can be used to predict ad liking and purchase intent. I will analyze aggregate responses from over 12,000 videos and evaluate how automated copy-testing based on affective responses would perform.

**Chapter 10** describes how affective measurements can be used to predict short-term sales effectiveness of ads. I show that for intentionally humorous ads, those which induce increasing positive valence and low levels of negative expressions (such as disgust) are likely to be more successful.

**Chapter 11** shows how facial expressions can be used in contexts other than advertisements and short film clips. I demonstrate that it is possible to predict candidate preferences
in an election debate based on affective responses. In addition, I show that the cloud based framework presents a very efficient and fast way to collect facial responses to media content.

Chapter 12 presents multi-modal experiments that show the relationship between facial responses and physiological responses to ads and other media clips. In addition, I show further validation of remote physiological measurements using a webcam.

Chapter 13 is a generalized discussion of the results from all the experiments and the key findings. I will present a demo system that I have built to integrate the findings of this thesis into a real-time intelligent public display.

Chapter 14 summarizes the conclusions and contributions of this thesis and provides a discussion of future work. This chapter will also mention other contributions I have made that are beyond the main topic area of this thesis.
Chapter 2

Storytelling and Emotion

Emotion is key to the effectiveness of narratives and storytelling (Green, Strange, & Brock, 2002). Stories, even if fictional, have the ability to induce a genuine emotional response (Green, 2004; Kassam, 2010). However, there are nuances in the emotional response to narrative representations compared to everyday social dialogue (Parkinson & Manstead, 1993) and therefore context specific analysis is important. Imagery, music and dialogue can all contribute to inducing emotions within a viewer. The temporal nature of a viewer, listener or reader’s emotional response is important (Baumgartner, Sujan, & Padgett, 1997), with people reporting to prefer sequences of outcomes that improve over time (G. F. Loewenstein & Prelec, 1993) and rapid increases in improvement rather than slower increases (Hsee & Abelson, 1991). However, their memory may not capture all aspects of the temporal affective response. There is evidence that peak moments and final moments are weighted more heavily in post-hoc evaluations (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993).

Context plays a significant role in how emotion influences the effectiveness of a story. In this thesis I will focus mainly on emotional responses to TV advertisements - short stories which typically aim to persuade, inform and/or appeal to viewers. However, I will also present analysis of responses to election debates and movie clips.
Figure 2-1: In this thesis I will be analyzing affective responses to video advertisements, political debates and Hollywood movies.

2.1 Advertising

Marketing, and more specifically advertising, makes much use of narratives and stories. In advertising, the states of surprise and amusement are often intended emotions (Alden, Mukherjee, & Hoyer, 2000). In addition, when communicating a story or message confusion is of particular interest to the storyteller. Surprise, amusement and confusion could be modeled as discrete states or continuous measures of valence and arousal could be used to distinguish between them - depending on the model of emotion used.

The role of emotion in marketing and advertising effectiveness has been considered extensively since early work by Zajonc (1980) that argued emotion functions independently of cognition and can indeed override it. Since then, emotions in advertising have been shown to enhance the emotional connection consumers have with brands (Mehta & Purvis, 2006), viewer engagement (Teixeira, Wedel, & Pieters, 2010), ad recall (Hazlett & Hazlett, 1999; Ambler & Burne, 1999) and decisions to share (virality) (Berger & Milkman, 2011).

“Ad liking” was found to be the best predictor of sales success in the Advertising Research Foundation Copy validation Research Project (Haley, 1990). Biel (1990) and Gordon (2006) state that likability is the best predictor of sales effectiveness. Haley (1990)
concluded that the concept of “likability” of a commercial was the best predictor of sales effectiveness.

Memory of stories, advertisements and brands is an important measure for success when it is not possible to measure sales effects directly. By creating advertisements that surprise, that engage, and entertain, advertisers and storytellers hope to create memorable material. Explicit memory of advertising (recall and recognition) is one of the most frequently used metrics for measuring advertising success. Independent studies have demonstrated the sales validity of recall (Haley, 1990; Mehta & Purvis, 2006). Indeed, recall was found to be the second best predictor of advertising effectiveness (after ad liking) as measured by increased sales in the Advertising Research Foundation Copy validation Research Project (Haley, 1990). Ambler and Burne (1999) and Mehta and Purvis (2006) 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.

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 (2011) 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 it is 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.

Behavioral methods such as ad zapping or banner click-through rates are frequently used methods of measuring success. Teixeira et al. (2010) show that inducing affect is important in engaging viewers in online video adverts and in reducing the frequency of “zapping” (skipping the advertisement). They demonstrated that joy was one of the states that stimulated viewer retention in the commercial. Micu and Plummer (2010) measured zygomatic major activity using facial electromyography (EMG) whilst people watched TV ads. They showed that physiological measurements capture different information compared
to self-reported responses.

The previous work provides compelling evidence that the emotions elicited by a video ad are related to its success. This evidence generalizes across many different measures of success. Ultimately, purchase decisions are the truest measure of advertising effectiveness. When making decisions, our past and current emotion experiences bias our decision-making unconsciously, making emotions an important influencer on our decisions (G. Loewenstein & Lerner, 2003). Sales as a measure of the success of advertising and predicting sales success from affective responses will be one focus. However, the success of an advertisement varies from person to person and sales figures at this level are often not available. I will also consider other measures of success, in particular liking, desire to view again and stated purchase intent (PI).

2.2 Political Ads and Debates

Brader (2005) found that political campaigning - in particular TV advertising - achieves its goal in part by appealing to the emotions of the viewers and that different emotional states led to different self-report responses. In particular, whether an ad appeals to fear or enthusiasm can have a considerable effect on its persuasive power. The success of political media is typically measured by polling audiences. This may be performed via telephone interviews or focus groups. Luntz (1994) highlights the power in audience measurement. However, he also identifies that focus groups, in which people gather in a single room and report their feelings via a button- or dial-operated computer, can be the least financially profitable tool in political polling. Focus groups present many challenges, reporting emotions using a dial or slider can detract from the experience of interest and participants are typically limited to those within a small geographic area. I will show that analyzing facial responses to election debates clips can reveal information about viewers’ reported preferences of candidates and potentially represent a much more efficient method of polling audiences in a short space of time.
2.3 Movie Clips

Gross and Levenson (1995) compiled a widely used dataset of Hollywood film clips that have been validated to induce certain emotions in a majority of viewers. Well validated emotion eliciting material, whether it be images (P. J. Lang, Bradley, & Cuthbert, 1999), music (Strapparava, Mihalcea, & Battocchi, 2012) or movie clips (Gross & Levenson, 1995), are a very useful resource in emotion research (Coan & Allen, 2007). Using movie clips provides a well known and controlled set of stimuli by which we can understand the relationship between different modalities in different types of emotional response. In addition, we can compare these responses to responses to other media (TV ads, political debates, etc) which have not been validated to elicit a specific emotional state and may elicit a variety of emotions. I use the standard dataset presented in Gross and Levenson (1995) in order to collect facial and physiological responses to well known and widely used stimuli and provide reference point for the comparison of responses to media over the Internet.
Chapter 3

Emotion Models and Measurement

3.1 Models of Emotion

Emotion theorists have long debated the exact definition of an emotion and a number of models and taxonomies of emotion have been proposed. Three commonly used approaches are discrete or categorical models, continuous or dimensional models and cognitive-appraisal models. I will briefly summarize these main approaches; a deeper discussion can be found in Fox (2008).

The discrete categorization of emotion posits that there are “affect” programs that drive a set of core basic emotions (Silvan & McCarter, 1964). The model makes the assumption that there is high agreement in the way that people express and perceive expressions with respect to emotion. Most discrete theories of emotion list a set of prototypical (basic) states. A commonly used set of basic states is: happiness, sadness, fear, anger, disgust, and surprise. This set was first proposed by Ekman, Friesen and Ellsworth (1972). Ortoney and Turner (1990) present the different categories of basic emotions that have been proposed. These vary from two to 11 emotion labels. Other researchers (e.g. El Kaliouby and Robinson (2005)) have used categorizations of affective and cognitive mental states (such as agreeing, concentrating and disagreeing) rather than basic emotions. For some applications these may be more appropriate. However, there is much less evidence to suggest that
they are universally expressed.

Russell (1980) argues for a continuous model of affect rather than a discrete model. The most commonly used continuous model of the emotion space is the circumplex model. A circumplex is a circular two-dimensional space in which points close to one another are highly correlated. Valence (pleasantness) and arousal (activation) are the most frequently selected descriptions used for the two axes of the circumplex (J. A. Russell, 1980). A third dimension of dominance, or power, is in some cases added. The most appropriate principal axes of the emotion space have been debated (R. J. Larsen & Diener, 1992). A popular interpretation is to use main axes of “Positive Affect” (PA) and “Negative Affect” (NA) with each of these axes containing an activation component.

The third model of affect that is commonly used is the cognitive-appraisal model, in which the influences of emotions on judgements are considered. There have been a number of different presentations of this theory (K. R. Scherer, 1999). However, the central tenet to all is that emotions are elicited and differentiated based on the person’s evaluation of an event or object. In this case a person’s appraisal of a situation (such as viewing a video) will have a bearing on the emotion that they experience and people in different contexts watching the same video will not necessarily experience the same emotion.

Details of discrete and continuous scales used for eliciting reports of emotions are discussed later in the chapter. In this thesis I will utilize both discrete and dimensional representations of affect, classifying responses in categories such as disgust or amusement but also characterizing the response in terms of a continuous measure of valence. In certain cases it will be helpful to consider specific discrete emotion labels (in particular amusement and disgust) in order to communicate the conclusions drawn. Appraisal models of emotion are very interesting and powerful. However, I will not be using this type of model in my thesis. The following subsections will discuss theories of how emotions relate to facial expressions and physiological changes.
3.2 Facial Expressions and Emotion

Charles Darwin was one of the first to demonstrate universality in facial expressions in his book, “The Expression of the Emotions in Man and Animals” (Darwin, Ekman, & Prodger, 2002). His work was partially inspired by that of surgeon Sir Charles Bell and anatomist Duchenne de Bologne (who was studying the muscular anatomy of the face using electrical stimulation). Since then a number of other studies have demonstrated that facial actions communicate underlying emotional information and that some of these expressions are consistent across cultures (Silvan & McCarter, 1964; Ekman & Friesen, 1971; Ekman, 1993; R. Larsen & Fredrickson, 1999). In the 1960s and 1970s Tomkins and McCarter (1964), Izard (1971) and Ekman and Friesen (1971) presented a series of studies finding cross-cultural universality in expression and interpretation of expressions - beyond Darwin’s somewhat anecdotal evidence - that they suggested was innate. More recently, researchers have shown that universality of expression extends to infants not just adults (Camras, Oster, Campos, Miyake, & Bradshaw, 1992).

This idea of an innate and universal link between emotions and facial expression is not without its critics (J. Russell, 1994). To clarify before we begin, I am going to use facial expression to describe facial actions here rather than the term “facial display” or “facial behavior”. However, the use may not always imply an emotional input as might be suggested by an emotion-expression compared to a motive-communication view. This follows a reasoning similar to that of Schmidt and Cohn (2001) and will simplify the discussion somewhat.

Kassam’s analysis of facial expressions (Kassam, 2010) demonstrates that both facial expressions and self-report responses have significant variance. Results show that expression analysis provides unique insight into emotional experiences, different to information obtained via self-report questioning. Examples of the data collected are shown in the Affect Unit Spectrograms (AUSpec) (D. J. McDuff et al., 2010) shown in Figure 3-1. The color intensity of each block represents the percentage of people within the population of viewers who exhibited the corresponding action unit at that time. For the two clips, one
3.2.1 Facial Action Coding (FACS)

When using facial behavior as a measure of a person’s response it is important to clearly state how the behavior is quantified. There are two main approaches for coding of facial displays, “sign judgment” and “message judgment.” “Sign judgment” involves the labeling of facial muscle movements or actions, such as those defined in the Facial Action Coding Scheme (FACS) (Ekman & Friesen, 1977) taxonomy. “Message judgments” are labels of human perceptual judgment of the underlying state. In my work I will use facial expression classifiers trained on “sign judgments”, specific action units or combinations of action units, as they are objective and not open to contextual variation.

The Facial Action Coding System (FACS) (Ekman & Friesen, 1977) is the most widely used and descriptive system for coding facial actions. It is based on an initial system proposed by Hjortsjo (1969). A number of other coding systems also exist; EMFACS (Friesen
Ekman, 1984), MAX (C. E. Izard, 1979), and AFFEX (C. Izard, Dougherty, & Hembree, 1983) but these are coarser systems and are specifically focused on labeling emotions. Fasel and Luettin (2003) provide a summary of these approaches. Being the most comprehensive, descriptive and widely used system I will look mostly at FACS in this section.

FACS 2002 defines 27 action units (AU) - 9 upper face and 18 lower face, 14 head positions and movements, 9 eye positions and movements and 28 other descriptors, behaviors and visibility codes (J. Cohn, Ambadar, & Ekman, 2005). The action units can be further defined using five intensity ratings from A (minimum) to E (maximum). More than 7000 AU combinations have been observed (K. Scherer & Ekman, 1982) emphasizing the complexity of this channel of information. The system provides the most objective measurement of facial actions available as it is based on these movements and not on any interpretation of meaning derived from them. A key strength of FACS is that it has standardized documentation and training. Training for labelers is available in the form of self-taught instructional material (FACS 2002) (Hager, Ekman, & Friesen, 2002), an intensive taught course, and a certification examination.

The objective and quantitative nature of FACS has allowed significant advances in the understanding of facial expressions. Ekman and Rosenberg (Ekman & Rosenberg, 1997) show a number of examples that demonstrate how FACS was used to discover new behavior. For example, in the case of Chesney et al. (1990) a facial expression (glare) was discovered that would have been missed if a coarser coding scheme such as EMFACS were used. FACS analyses have been shown to be useful in the detection of pain (Prkachin, 1992), detection of depression (J. F. Cohn et al., 2009), classification of suicidal faces (Heller & Haynal, 1997) and analyzing mother-infant interactions (Messinger, Mahoor, Chow, & Cohn, 2009).

One of the main challenges presented by a comprehensive coding system such as FACS is the time and expense required to code expressions. It can take an experienced coder more than one hour to code a minute of video. As a result large-scale analysis of facial expressions can be infeasible in many cases. Thus, the development of automated algorithms for
coding offers obvious advantages. In addition, automated algorithms offer the advantage of being able to code observations that would be difficult by hand, such as the velocity of an action (Ambadar, Cohn, & Reed, 2009). Automated facial expressions recognition systems will be reviewed in Chapter 4.

3.3 Physiology and Emotion (Psychophysiology)

There are many physiological processes that are influenced by emotions. The physiological component of an emotional response is mostly concerned with the autonomic nervous system (ANS). The ANS has two branches: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS).

One of the most reliable measures of sympathetic arousal is electrodermal activity (EDA) (Boucsein, 1992), previously known as Galvanic Skin Response (GSR). EDA is typically characterized by measuring the conductance of the skin at bodily locations with high density of eccrine sweat glands. Some of the traditional locations are the fingers or the palm of the hand, but it has been successfully measured on other body locations such as foot or shoulders (van Dooren, de Vries, & Janssen, 2012). EDA is controlled bilaterally and therefore the responses on different sides of the body can contain different information (Boucsein, 1992). Recently wireless, wearable sensors have enabled measurement of data outside of laboratory settings (M.-Z. Poh, Swenson, & Picard, 2010).

Blood volume pulse (BVP) measurements can capture information both about the valence and arousal of an individual (Ekman, Levenson, & Friesen, 1983; Gomez, Stahel, & Danuser, 2004; Kushki, Fairley, Merja, King, & Chau, 2011). Healey (2000) shows examples of the relationship between BVP amplitude and EDA response in which the BVP amplitude “pinches” with increased EDA response.

Inter-beat-intervals (IBI) of the heart are controlled by the ANS. Heart rate can be measured as both a tonic value, the average frequency over a time window (typically between 5 and 30s) and as a phasic value which is equal to the inverse of the instantaneous heart
rate (1/IBI). Heart rate variability is common quantified by calculating the power spectral density (PSD) of the heart’s IBIs as this can provide information about the parasympathetic and sympathetic influences. The low frequency power (LF) and high frequency power (HF) are typically measured as the area under the PSD curve which correspond to the frequency ranges 0.04-0.15Hz and 0.15-0.4Hz respectively. The frequency components are calculated in normalized units (n.u.) to minimize the effect of changes in the total power. The LF component of the HRV spectrum is modulated by baroreflex activity and is influenced by both the sympathetic and parasympathetic parts of the nervous system (Akselrod et al., 1981). The HF component reflects parasympathetic influence on the heart. The ratio between the two (LF/HF) is often calculated as an estimate of sympathetic modulations but it is not a pure measure of sympathetic activity.

3.4 Measurement of Emotions

Affective responses are a significant component in the impact of media content and as such it is important to consider the ways in which affect can be quantified. Many techniques have been used to measure affective responses in advertising: self-report, autonomic, behavioral, physiological and brain imaging measures. Poels and Dewitte (2006) review these techniques. However, this review does not contain any examples of the use of automatic vision-based facial coding or remote measurement of physiology. In a review of the advertising and marketing literature that I performed I found seven examples of use of only self-report responses, four examples of use of physiological measures (e.g. cardiovascular signals or electrodermal activity (EDA) - all measured using wired sensors) and six examples of use of facial behavior (e.g. electromyogram (EMG) or manual facial coding). The following subsections will provide an overview of these techniques and the findings related to them. Tables 3.1 and 3.2 summarize the related work covering only self-report measures (Table 3.1) and autonomic measures (Table 3.2) of emotion respectively.

Whilst I will focus on findings related to media and advertising research a broader
summary of emotion elicitation methods and emotion assessment techniques can be found in (Coan & Allen, 2007).

### 3.4.1 Self-Report

Currently the most commonly used method for measuring emotion is self-report. This can take the form of a verbal self-report, visual analogue self-report or moment-to-moment rating (Poels & Dewitte, 2006). Self-report measures are relatively simple, cheap to implement and are quite reliable (Kassam, 2010). However, they require cognitive processing that can introduce a bias, referred to as “cognitive bias”, to the results. In addition, there are other potential pitfalls with self-report measurement. The act of introspection is challenging to perform in conjunction with another task and may in itself alter that state (Schwarz & Strack, 1999). Some subjects may be willing to report feelings even if they do not have them (Cornelius, 1996). Furthermore, emotion is a dynamic phenomenon that unfolds and changes over time, even during a 20-second advertisement. Self-report measures typically capture only occasional snapshots of these dynamics, heavily influenced by the end moments when the report is made (Baumgartner et al., 1997). If moment-to-moment measures are used it can be difficult to measure responses on more than two dimensions (e.g. valence and arousal/activation). Finally, self-report techniques can interrupt a viewer’s experience and there may be a non-negligible delay between a stimulus and a cognitive response. Finally, people may become tired of continuously reporting their feelings in different situations - imagine if you personally had to rate the emotion you felt during every ad you watched on TV using a set of extensive criteria. Automatic and unobtrusive measures of one’s experience could be much less tiresome. Furthermore, people may often experience more than one emotion during media viewing (e.g. surprise followed by amusement). We will see that continuous automatic measures of emotion can capture the evolution of feelings with higher fidelity than can be captured using a single self-report scale. I will now discuss the three forms of self-report in turn: verbal, visual analogue and moment-to-moment ratings.
Verbal Self-report

Verbal self-report is the most common and perhaps up until this point the most reliable measurement of how people feel (Kassam, 2010). As such it may currently be considered to hold out the best chance of success. In verbal self-report participants are asked to say or write their response to a question about the experience. Often five to seven point Likert scales with labels (anchors) at either end are used.

Emotion Scales: A number of discrete emotion scales have been developed and used for eliciting verbal reports of emotion. Plutchik’s “Emotional Profile Index” (Plutchik & Kellerman, 1974), Izard’s “Differential Emotion Scale” (C. E. Izard, 1993) and the Positive and Negative Affect Schedule (PANAS) (Watson, Clark, & Tellegen, 1988) are the most commonly used. The PANAS system consists of two ten-item mood scales and participants respond on a scale of one to five how strongly they felt each of the feelings or emotions. Taking a dimensional approach the “Pleasure, Arousal and Dominance (PAD)” space (Mehrabian & Russell, 1974) is designed to capture the full-spectrum of human emotions. Adjectives can be associated with the each of the PAD-dimensions in order to make interpretation of the axes more intuitive.

Ad Effectiveness Scales: Unsurprisingly ad-evoked feelings as measured via verbal self-report have a significant impact on attitude towards the ad, $A_{ad}$ (Pieters & de Klerk-Warmerdam, 1996). More interestingly, high-intensity pleasant feelings reported can significantly impact ad recall. It has been shown that this type of affective measurement is a good supplement to more cognitive responses which are elicited with more direct questions. Attitudes towards an ad also have a significant impact on brand attitudes (Batra & Ray, 1986). Ad recall and brand attitudes are both proxies for advertisement effectiveness. Ad content including emotional responses, measured using a verbal self-report system, influence viewing time (the inverse of zapping rate) (Olney, Holbrook, & Batra, 1991) another key measure of whether an ad is effective. I hypothesize that these relationships can be more clearly defined and explained if the self-report measures of affect are supplemented with facial and physiological measures.
Visual Analogue Self-report

Visual self-report methods require the participant to choose an image which most closely represents their experience. The most popular visual self-report measure is the Self Assessment Manikin (SAM) (P. Lang, 1980). It has been successfully applied in advertising research (Bradley & Lang, 1994; J. Morris, 1995). This usually has three dimensions of arousal, valence and dominance which are each represented by a series of figures. Morris et al. (2002) developed AdSam® in which 232 emotion words were scored on SAM. AdSam allows emotion words to be associated with emotional responses to an ad. Figure 3-2 shows an example of the pictorial representations of valence, arousal and dominance used in SAM. Using a visual analogue measure it was found that emotional responses to ads were related to behavioral intentions (purchase intent (PI) or store visit). However, brand attitude was not found to be a precursor to intention (J. D. Morris et al., 2002). Other visual self-report methods exist (Wood, 2007) although they have not been rigorously evaluated.

Moment-to-Moment Self-report

Moment-to-moment measurement can be captured using a dial or a pencil and paper method. Since reporting continuously interrupts the subject’s experience, the report is often made on the second viewing of a media clip. The clip is watched once and then watched again during which the dial is used to report. However, as we will show in this thesis (see Chapter 7) there is evidence that an individual’s expressed emotions can differ significantly on a second viewing of content compared to the first viewing and therefore by only recording the self-reported response on the second viewing there may be a bias introduced into the measurements. As with verbal self-report the “Pleasure, Arousal and Dominance (PAD)” (Mehrabian & Russell, 1974) dimensions can be used for eliciting moment-to-moment reports. Continuous emotion labeling is possible using interfaces such as FeelTrace (Cowie et al., 2000) which has the main axes of arousal and valence, see Figure 3-3. However, reporting emotions in a two dimensional space is not necessarily intuitive for a novice.
A popular one-dimensional measure of emotion applied in marketing research is the warmth monitor (Aaker, Stayman, & Hagerty, 1986). Warmth, as reported on this scale is typically correlated with arousal. Warmth may be more intuitive for a novice reporter to interpret than valence or arousal.

Aaker et al. (1986) found that a warm ad preceded by a non-warm ad was more effective (as measured using attitude towards the ad, $A_{ad}$, and purchase intent) than warm followed by warm. Baumgartner et al. (1997) asked viewers to report emotional valence via the “feeling monitor” - a linear computer scale with anchors “strong negative feelings” to “neutral” to “strong positive feelings.” They found that viewers’ assessments were dominated by peak emotional experiences and by the final moment of the experience. Viewers tended to prefer ads with high positive emotional peaks and particularly those that ended on a strong positive note with sharp increases in affect over time. I will show that large-scale analysis of facial responses to ads collected over the Internet support these findings.

### 3.4.2 Facial Behavior

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 emotions change over time. The two main approaches to measuring facial behavior
are via contact-based Electromyogram measurements or via observational methods, such as FACS coding.

**Electromyography (EMG) Measurement**

The measurement of muscle potentials on the face is non-trivial and the majority of research that makes use of EMG measurement is limited to measurement of two to three muscles. The most common muscles measured are the zygomaticus major, corrugator supercilii and orbicularis oculi. As it is only possible to measure a small number of muscles, facial EMG measurements primarily capture valence information and are not sufficient for detecting other dimensions of a response or specific discrete expressions of emotion. However, due to the fact that electrical potentials are measured they can capture activity that may not be observable to the eye.

Facial EMG measurement has been shown to correlate with human valence judgements when viewing pictures (P. Lang, Greenwald, Bradley, & Hamm, 2007), listening to radio commercials (Bolls, Lang, & Potter, 2001) and viewing video messages (Ravaja, Kallinen, Saari, & Keltikangas-Jarvinen, 2004). In particular, Orbicularis Oculi (cheek raiser/lid tightener) was higher during positive messages and corrugator (brow furrow) was significantly higher during negative news messages (Ravaja et al., 2004). Zygomatic Major was higher during positive messages (Bolls et al., 2001). Hazlett and Hazlett (1999) found that facial EMG was a more sensitive discriminator between commercials than self-report and that it was more closely related to recall. However, emotional valence from the measurements was not validated.

EMG can capture specific elements of emotional processes that other more traditional methods (e.g. self report) may miss (Cacioppo, Losch, Tassinary, & Petty, 1986). However, EMG analysis is obtrusive requiring electrodes to be connected to the subject’s face. This has several drawbacks. Firstly, EMG measurements require specialized hardware and thus are not possible to perform remotely, making data collection difficult to scale. Secondly, only a limited number of electrodes can be connected simultaneously, meaning that only a
subset of the facial muscles can be monitored at one time. Thirdly, it creates an unnatural experience which is considerably different to the experience of consuming content in real-life.

**Observational Measurement**

Derbaix (1995) is one of the first examples to use facial expressions to measure emotional responses to ads. Derbaix found that the contribution of affective responses on ad attitude and brand attitude were evident in verbal responses but not facial measures. The facial coders assigned basic emotion labels to the frames of the viewer’s response. Basic emotions may not be a suitable taxonomy for this task. This study is in a minority as we shall see that many other examples show utility in using facial expression measures to evaluate media success.

Joho et al. (2011) 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. (2013) 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 only tested on a small number of viewers (n=10) in a lab setting.

Teixeira et al. (2010) showed that inducing affect is important in engaging viewers in online video adverts and to reduce the frequency of “zapping” (skipping the advertisement). It was found that joy was one of the states that stimulated viewer retention in commercials. Although not directly related, it is intuitive that smiles might play a significant role in a joy response. In this study viewing occurred in a laboratory setting rather than in the wild. The work was extended by collecting viewer responses over the Internet (Teixeira, El Kaliouby, & Picard, 2014). Purchase intent of viewers was linked to emotional responses. However, no behavioral data (e.g. sharing or sales metrics) was linked to the results.
In this thesis I build on and extend the work of Teixeira et al. (2010, 2014) collecting real-world data over the Internet of viewers watching ads, self-reported references, desire to share and objective measures of advertising effectiveness including short-term sales impact. In addition, I consider a larger number of automatically coded emotion expressions in the analysis.

3.4.3 Autonomic Measures

There is a large amount of work in the field of psychophysiology about emotions and physiology that has informed studies specifically in the media measurement and marketing domains. As with facial behavior physiology is continuous and dynamic, allowing for a representation of how emotions change over time. In addition, physiological responses may be able to capture emotional responses that are not expressed observably. However, changes in physiology can be subtle and challenging to interpret.

Electrodermal Activity (EDA)

LaBarbera and Tucciarone (1995) argue strongly for the use of EDA in marketing research and present a number of studies to demonstrate the validity of physiological measurement in evaluating ad effectiveness. The strongest link found was between EDA and sales. Subsequent studies have found EDA to be a better predictor of memory (free-recall and brand recognition tests) than valence, as measured via EMG or self-report (Bolls et al., 2001).

Heart Rate and Respiration Rate

Tonic and phasic changes in heart rate can be used to capture arousal and attention of viewers. This was demonstrated by Lang (1990) in a study of viewers watching TV ads embedded between two sitcoms to simulate a more realistic viewing experience. Significant heart rate increases were observed for emotional ads compared to rational or mixed (emotion and rational) ads. EDA was not used in this case but could be used to help justify the link to arousal. Respiratory sinus arrhythmia (RSA) has been used as a measure of attention to
business news messages - not ads. RSA was found to increase with emotional over neutral content - specifically, happy and angry faces (Ravaja et al., 2004). Features from the EDA and BVP response of a computer game player were effective at detecting periods of frustration (Scheirer, Fernandez, Klein, & Picard, 2002). Kapoor et al. (2007) presented one of the first examples of multimodal fusion of physiological responses and facial behavior for the automatic prediction of frustration. There is a considerable body of work using multimodal fusion in the detection of affective states. A much more comprehensive review of the literature can be found in Gunes et al. (2008).

I focus primarily on facial behavior in this work. However, I will show that physiological responses (EDA and cardiopulmonary measurements) could play a vital role in media measurement, particularly when viewers are inexpressive. Furthermore, some of these parameters can be measured remotely with a standard webcam. I will show that HRV features and respiration rate detected remotely using a digital camera can distinguish accurately between subjects at rest and under cognitive load.

### 3.4.4 Brain Imaging

Neuromarketing is a growing area of research in the media and marketing industries (Lewis & Phil, 2004). Brain imaging tools are of keen interest in measuring human responses to media. Electroencephalography (EEG) captures surface level electrical signals mainly due to cortical activity. Functional magnetic resonance imaging (fMRI) is more suitable for imaging subcortical brain structures. However, both these methods require relatively expensive hardware, are obtrusive and are sensitive to motion. Indeed the measurements could be corrupted by facial expressions or head gestures.

There are interesting research questions that could be answered using brain imaging techniques and potential for uncovering information not revealed using other measurement techniques (Ariely & Berns, 2010). For instance, Ioannides et al. (Ioannides et al., 2000) found a difference in the brain activation during affective versus cognitive advertising stimuli. However, the measurement of these signals in natural situations is very challenging.
The participant is required to lie down in a noisy machine and it is possible that the feelings towards the ad are dominated by the feelings toward the machine. In this thesis I will focus only on self-report, facial and physiological responses, with an emphasis on collecting data during natural viewing experiences.
<table>
<thead>
<tr>
<th>Paper</th>
<th>Measures</th>
<th>Indep. Variables</th>
<th>Population</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaker et al. (1986)</td>
<td>MM</td>
<td>$A_{ad}$, Ad Recall, PI</td>
<td>67 undergraduate students</td>
<td>Warm ad preceded by a non-warm ad was more effective than warm followed by warm. ER → $A_{ad}$ → $A_{br}$ → PI</td>
</tr>
<tr>
<td>Batra and Ray (1986)</td>
<td>VB</td>
<td>$A_{ad}$, $A_{br}$, PI</td>
<td>102 graduate students</td>
<td>Demonstrated a chain of effects from emotion response to ads, to ad attitude and viewing behavior. Visual analogue scales are useful in advertising research.</td>
</tr>
<tr>
<td>Olney et al. (1991)</td>
<td>VB</td>
<td>Attention (Ad “zipping” and “zapping”), $A_{ad}$</td>
<td>102 graduate students</td>
<td>Pleasant feelings have a significant impact on ad recall. Unpleasant feelings and low-intensity pleasant feelings have a significant impact on $A_{ad}$.</td>
</tr>
<tr>
<td>Morris et al. (1995)</td>
<td>VA (SAM)</td>
<td>No others</td>
<td>Not specified</td>
<td>Viewers’ assessments dominated by peak emotional experience and final moment. Viewers prefer ads with high peaks, which end on a strong positive note and exhibit sharp increases in affect over time. Visual measure of emotion was more predictive of attitude and action (store visit) than cognitive measures.</td>
</tr>
<tr>
<td>Pieters and de Klerk-Warmerdam (1996)</td>
<td>VB</td>
<td>$A_{ad}$, brand recall</td>
<td>Not specified</td>
<td></td>
</tr>
<tr>
<td>Baumgartner et al. (1997)</td>
<td>MM</td>
<td>$A_{ad}$, $A_{br}$, brand recall</td>
<td>28-94 undergraduate business students</td>
<td></td>
</tr>
<tr>
<td>Morris et al. (2002)</td>
<td>VA (ad-SAM)</td>
<td>$A_{ad}$, brand interest, PI</td>
<td>Not specified</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paper</th>
<th>Measures</th>
<th>Indep. Variables</th>
<th>Population</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lang et al. (1990)</td>
<td>HR</td>
<td>Attention</td>
<td>14 students</td>
<td>Significant changes in heart rate observed with emotional ads compared to rational or mixed ads.</td>
</tr>
<tr>
<td>LaBarbera et al. (1995)</td>
<td>EDA</td>
<td>Sales</td>
<td></td>
<td>Strongest link between EDA in response to marketing communications and sales.</td>
</tr>
<tr>
<td>Hazlett and Hazlett (1999)</td>
<td>SR, EMG</td>
<td>Recall</td>
<td>49 working professionals and undergraduates</td>
<td>Facial EMG most closely related to recall and a more sensitive discriminator between commercials.</td>
</tr>
<tr>
<td>Bolls, Lang and Potter (2001)</td>
<td>EDA, EMG</td>
<td>Free-recall, brand recognition</td>
<td>41 undergraduates</td>
<td>EMG can be used to predict emotional tone of content but that arousal was a better predictor of memory than valence.</td>
</tr>
<tr>
<td>Teixeira et al. (2010)</td>
<td>FC</td>
<td>Ad “zapping”</td>
<td>58 students and professionals</td>
<td>Joy stimulated viewer retention in a commercial.</td>
</tr>
<tr>
<td>Teixeira et al. (2014)</td>
<td>FC</td>
<td>Purchase Intent</td>
<td></td>
<td>Entertainment and brand placement influence purchase intent.</td>
</tr>
</tbody>
</table>

Table 3.2: Summary of the autonomic measurement of emotions in media and marketing. SR = Self-report, EDA = Electrodermal Activity, EMG = Facial Electromyography, FC = Facial coding.
Chapter 4

Affective Computing

I will now introduce some of the computer vision and machine learning approaches that are relevant to the automated analysis of responses I performed in my thesis work. In particular I will look at related work in the automated detection of facial actions and expressions, the machine learning techniques for modeling human behavior and affect and finally the relatively new area of affective crowdsourcing.

4.1 Automated Analysis of Facial Behavior

4.1.1 History

The automated analysis of facial behavior has focused on both sign and message judgement methods. However, the majority of the work has been applied to facial action unit recognition. The first example of automated facial expression recognition was presented by Suwa et al. (1978). Essa and Pentland (1997) demonstrated how optical flow could be used to quantify facial muscle movements. Lyons et al. (1999) presented a method for classifying emotional expressions in single face images. Kapoor, Qi and Picard (2003) present one of the first demonstrations of fully automated detection of AUs from naturalistic data. Significant progress in automated FACS coding includes work by Tian et al. (2001), Bartlett et

Smile detection is one of the most robust forms of automated facial analysis available. Whitehill et al. (2009) present a smile classifier based on images collected over the Internet and showed strong performance on this dataset. A subset of the data was released as the MPL GENKI\(^1\) dataset. Shan (2012) reports an accurate and faster smile detector on the MPL GENKI-4K dataset.

Continuous affect detection from facial behavior has had a lot of attention recently with classification of arousal, valence and power from video sequences (Gunes & Pantic, 2010; Schuller et al., 2011).

### 4.1.2 State of the Art

Recently, several facial expression recognition challenges have been run in order to compare state-of-the-art approaches on a common dataset. The Facial Expression Recognition and Analysis Challenge (2011) was the first challenge to specifically compare performance on FACS detection (M. Valstar, Jiang, Mehu, Pantic, & Scherer, 2011). Most systems opted for static classification of AUs (10 of 12 teams using Support Vector Machines (SVM)) and basic emotions with only one modeling dynamics explicitly (M. F. Valstar, Mehu, Jiang, Pantic, & Scherer, 2012).

Almost all systems have three main components: a method for registering the face, calculation of feature descriptors and classification of AUs or expressions based on the feature descriptors. Figure 4-1 shows a high-level diagram of an automated facial expression recognition system. There is a large amount of literature on automated coding of facial actions and expressions. Comprehensive reviews of the work can be found in Zeng et al. (2009) and De la Torre and Cohn (2011).

\(^1\)http://mplab.ucsd.edu, The MPlab GENKI Database, GENKI-4K Subset
Figure 4-1: The three main components present in most facial expression or AU recognition systems: 1) a method for registering the face, 2) calculation of shape and/or appearance feature descriptors and 3) classification of AUs or expressions using the feature descriptors.
Face Tracking and Registration

Registration of the face involves aligning the face in images to minimize the impact of non-expression related variations. Better registration considerably helps with the detection of facial expressions from images and videos (P. Lucey, Lucey, & Cohn, 2010).

The Viola-Jones face detector is the most commonly used form of face detection algorithm (Viola & Jones, 2001). However, for most practical facial action detection algorithms a more precise registration method is needed and at least the eyes need to be located. In this work I use a landmark detector that uses Gabor filters for locating 22 landmarks on the face.

State of the art face registration methods include Active Appearance Models (AAM) (Cootes, Edwards, & Taylor, 2001) and Constrained Local Models (CLM) (Saragih, Lucey, & Cohn, 2009). These come from a class of models know as Parameterized Appearance Models (PAMs) (De la Torre & Cohn, 2011).

The aim of registration is to normalize the images as effectively as possible - removing rigid head motions. Normalization can be performed by using a two or three dimensional transformation. In 2D the image can be normalized using an affine model of six parameters (rotation and scaling).

Feature Descriptors

Once the face has been registered features are extracted for identifying which actions are present. Features generally fall into two categories - geometric features or appearance features.

Geometric features are calculated from the geometric relationships between permanent facial features, typically positions of the landmarks detected (e.g. the distance from the nose root to the middle of the eyebrow or the distance between the mouth corners). It can be hard to detect subtle facial expressions (e.g. AU14 - dimpler) from geometric features. Appearance features are computed from the pixel values within the facial ROI and aim to capture the texture or gradients of the image in that region (e.g. wrinkles and furrows of the
There are numerous appearance features used in facial action unit detection: Gabor Wavelet coefficients, Local Binary Patterns (LBP), Histogram of Oriented Gradient (HOG), and scale-invariant feature transformation (SIFT) descriptors have all been demonstrated with success. Other features include optical flow and Motion History Images (MHI).

In this work appearance features are primarily used as they tend to be more robust in cases where registration is not highly accurate, which can often be the case in low resolution webcam images.

### Classification

In early work Neural Networks were used to classify the presence of action units (Tian et al., 2001). However, SVMs 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. Dynamic models are useful for detecting the onset and offset of actions in addition to their presence. Valstar and Pantic (2007) present a variant of SVMs and Hidden Markov Models (HMM). Models which capture longer range dependencies such as Hidden Conditional Random Fields (HCRF) have also been shown to perform well (K.-Y. Chang, Liu, & Lai, 2009).

Detecting the intensity of action units or expressions is a challenging problem. FACS allows for action units to be coded on five different levels of intensity. However, these labels are extremely time consuming and costly to code by hand making the availability of training data limited. Mahoor et al. (2009) present a framework for automatic coding of AU intensity, although this is only applied to AU12 and AU6. Other work has found empirical evidence that the distance from the classification boundary in an SVM can capture intensity quite effectively (M. Bartlett et al., 2005).

### 4.1.3 Dataset for Training

As with many applications of machine learning, the data used for training models is important and can be very time consuming to collect. In the area of facial expression analysis,
the Cohn-Kanade database in its extended form named CK+ (P. Lucey, Cohn, et al., 2010), has played a key role in advancing the state of the art. 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 including: Bartlett et al (2003), Cohen et al. (2003), Cohn et al. (2004), Littlewort et al. (2004) and Michel and El Kaliouby (2003). Since then, a few other databases have emerged, including: MMI (Pantic, Valstar, Rademaker, & Maat, 2005), SEMAINE (Mckeown, Valstar, Cowie, & Pantic, 2010), RU-FACS (M. Bartlett et al., 2006), SAL (Douglas-Cowie et al., 2007), GENKI (Whitehill et al., 2009) and UNBC-McMaster Shoulder Pain Archive (P. Lucey, Cohn, Prkachin, Solomon, & Matthews, 2011). A survey of databases and affect recognition systems can be found in (Zeng et al., 2009; De la Torre & Cohn, 2011). 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 limit the number of participants to under 100, e.g. 88 in (Kassam, 2010). By using the Internet I can make data collection efficient, asynchronous, less resource intensive, and get at least an order of magnitude more participants. Perhaps more importantly, I can begin to systematically explore the meaning of facial expressions and their relationship to memory and decision-making in an ecologically valid manner. In this work I will show that facial responses from thousands of individuals around the world can be collected quickly. I will also describe the Affectiva-MIT Facial Expression Dataset (AM-FED) which I collected during my thesis work and which is now publicly available for researchers (McDuff et al., 2013).
4.2 Examples of Machine Learning Approaches

In machine learning and statistics there are many methods for modeling data and recognizing patterns. I will briefly review some relevant work here. This is by no means exhaustive but gives examples of some of the most commonly used modeling techniques.

Static Models

Naive Bayes (NB) models are a simple non-parametric method of classification. They are a useful baseline method for measuring performance. NB models capture the distribution of features in the different classes and assume their independence. Support Vector Machines (SVM) are commonly used in affective computing applications. They are generally effective, robust and can work efficiently in real-time applications. In this thesis SVMs are used in a number of cases. They provide a baseline by which to compare the performance of other models and can often perform as well as, or even better than, more complex models.

Temporal Models

Increasingly, the importance of considering temporal information and dynamics of facial expressions has been highlighted. Dynamics can be important in distinguishing between the underlying meaning behind an expression (Ambadar et al., 2009; Hoque, McDuff, & Picard, 2012). I will implement a method that considers temporal responses to commercials taking advantage of the rich moment-to-moment data I can collect using automated facial and physiological analysis.

Hidden Markov Models (HMM) are a commonly used generative approach to modeling temporal data and Hidden Conditional Random Fields (HCRFs) (Wang, Quattoni, Morency, Demirdjian, & Darrell, 2006) and Latent Dynamic Conditional Random Fields (LDCRFs) (Morency, Quattoni, & Darrell, 2007) are discriminative approaches to modeling temporal data. The CRF model and its 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) (Lafferty, McCallum, & Pereira, 2001). The dynamics of smiles are significant in distinguishing between their meanings (Ambadar et al., 2009); as such, I hypothesized a potential benefit in explicitly modeling the temporal dynamics of the responses. Song et al. (2011) describe a Gaussian temporal-smoothing kernel that improved performance without increasing the computational complexity of inference. This method takes a Gaussian-weighted average of observations within a moving window of size N.

**Hierarchical Models**

The modeling of emotion in marketing has received some considerable attention. The most commonly used models are hierarchical Bayesian models (Olney et al., 1991; Teixeira et al., 2010). These allow for consideration of many factors that influence the target variable. A hierarchical model was used by Olney et al. (1991) to investigate the relationship between emotional response, attitudes towards the ads and viewing behavior.

### 4.3 Affective Crowdsourcing

Crowdsourcing (Howe, 2008) aims to coordinate the effort and resources of large groups of people. Morris (2011) 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.

The widespread availability of webcams and video platforms such as YouTube has facilitated the collection of large amounts of rich image and video data. For instance, Taylor et al. (2011) and Spiro (2012) describe ways to use webcams to crowdsource posed data for training gesture recognition systems. Collecting naturalistic affective responses via the web raises some interesting issues. Opt-in participation is particularly important in cases where images from webcams are captured and stored. Another issue relates to data quality.
High bandwidth data such as real-time videos may be limited in resolution and frame-rate, which can present challenges for data analysis and feature detection.

In this thesis, I leverage crowdsourcing concepts to elicit a large amount of data from a wide demographic, something that has not been possible through traditional research practices. Remote data collection from crowds does raise issues of reliability and consistency. I will discuss how some of these issues can be overcome or their effects mediated in the discussion. More detail about the nuances of crowdsourcing affective data can be found in (R. Morris & McDuff, 2014).
Chapter 5

Overview of the Six Experiments

5.1 Experiments

I performed a number of experiments during this research. These experiments were designed in different ways in order to answer specific research questions and build upon lessons learnt during previous experiments. The details of the specific methodology, data and results are described in the following chapters. Table 5.1 summarizes the data collected during the experiments. The aims of the experiments were to investigate the relationship between automatically measured facial, physiological and self-report responses and sales. Four of the experiments involved ads, one involved videos of a political debate and one emotion eliciting movie and TV clips.

In the experiments that involved webcam recordings over the Internet, the participants were asked to: 1) grant access to their webcam for video recording; and 2) allow Affectiva and MIT to use the facial video for internal research, on opting in to the study. Further consent for the data to be shared with the research community at large was also sought, and only videos with consent to be shared publicly are shown in this thesis and related publications. In total over 20,000 webcam videos were recorded during the experiments represented in the largest analysis of this kind in the world.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Media Clips</th>
<th>Categories</th>
<th>Description</th>
<th>Videos</th>
<th>Ages</th>
<th>Locations</th>
<th>Meta-data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super Bowl</td>
<td>Three ads (three brands)</td>
<td>Superbowl ads</td>
<td>First in the world experiment to collect naturalistic facial responses to online videos, participants were not compensated.</td>
<td>6,729</td>
<td>U</td>
<td>Global</td>
<td>✔️</td>
</tr>
<tr>
<td>Cereal</td>
<td>16 ads (three brands)</td>
<td>Breakfast Cereal</td>
<td>Preliminary experiment to investigate the ability to predict sales from facial responses to commercials collected online.</td>
<td>100x16</td>
<td>U</td>
<td>US</td>
<td>✔️</td>
</tr>
<tr>
<td>Mars</td>
<td>170 ads</td>
<td>Snack foods, pet care, chewing gum</td>
<td>Large-scale study to analyze the relationship between emotional responses and sales.</td>
<td>100x170</td>
<td>18-65+</td>
<td>UK, DE, FR, US</td>
<td>✔️</td>
</tr>
<tr>
<td>Electoral Debate</td>
<td>5 debate clips</td>
<td>Election debate clips</td>
<td>Measuring reactions to five electoral debate clips from the 2012 US presidential election campaign.</td>
<td>610</td>
<td>U</td>
<td>Global</td>
<td>✔️</td>
</tr>
<tr>
<td>Multimodal</td>
<td>8 media clips and 5 ads (5 brands)</td>
<td>Media clips and cosmetics/cleaning product ads</td>
<td>Multimodal - physiological and facial - data in response to emotion eliciting clip.</td>
<td>12x10</td>
<td>18-30</td>
<td>US</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>8 media clips</td>
<td>Standard emotion eliciting media clips</td>
<td>(Gross &amp; Levenson, 1995)</td>
<td>8x10</td>
<td>18-30</td>
<td>US</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of datasets collected in the five experiments, including the number of videos tests, the number of responses and labels. SR = Self-report, M=Memory tests, S=Sales. U = unknown.
5.2 Research Questions

The experiments described were designed to collect data to answer several research questions.

Q1. Can affective information (both physiology and facial expressions) can be measured remotely using low-cost cameras?

Facial expressions and physiological signals (heart rate, respiration rate and heart rate variability) can be measured using computer vision algorithms. Although it is not always as precise as the information that can be measured using contact sensors we can capture high resolution temporal information about a viewer’s affective response.

Q2. Can the Internet can be used as an efficient method of sourcing large amounts of ecologically valid responses to online media?

Ecologically valid naturalistic and spontaneous facial responses to online media/stories can be collected efficiently and quickly using the Internet. Webcams and ubiquitous sensors can be used to capture data for measuring affective responses. People are willing to engage with the technology - in some cases even without compensation.

Q3. Is there congruency between facial expressions and self-report? (i.e. Affective responses to media content can be used to predict self-reported feelings toward content - in particular liking.)

There is congruency between affective responses and self-reported enjoyment, desire to watch again and desire to share. Automatically measured affective responses can be used to reliably predict these.
Q4. Does a link between facial expressions and short-term sales impact of ads exist?

Aggregate affective responses to ads can predict whether an ad is likely to perform strongly or weakly with respect to increasing purchase intent or even short-term sales. The high temporal resolution of facial coding will lead to insights into the emotional profile of a “successful” ad.

Q5. Are facial expressions and self-report complementary? (i.e. Facial responses combined with self-report measures will more accurately predict effectiveness of advertisements than self-report measures alone.)

The combination of automatically measured spontaneous affective responses and results from post hoc self-report questioning (such as purchase intent, enjoyment, brand appeal) will allow for more accurate predictions of effectiveness than the self-report questioning alone. The spontaneous visceral affective responses yield subtle information that is not available via post hoc questioning.
Chapter 6

Data Collection and Processing

In this chapter I will present details of the data collection framework used in the online experiments and technical information about the automated facial expression and physiological measurements used for measuring viewer responses from video sequences.

Although many of the experiments had similar components the exact methodology used for each was slightly different as the needs and technology evolved over time - I will explain details about data collection and analysis that are specific to each experiment in Chapters 8 - 11, while this chapter will describe elements common to all the experiments.

6.1 Data Collection Over the Internet

Computer-based machine learning and pattern analysis depend hugely on the number of training examples (Shotton et al., 2013). To date much of the work of automating the analysis of facial expressions and gestures has had to make do with limited datasets for training and testing. However, due to the considerable individual differences in responses, for example due to culture and gender variations (Hall, Carter, & Horgan, 2000; Matsumoto, 1990), this often leads to over-fitting. The diversity and complexity of naturalistic and spontaneous emotional expressions requires large amounts of data for learning to be effective.

The techniques found in crowdsourcing can be used effectively in the collection of
affective data (R. Morris, 2011; R. Morris & McDuff, 2014). The Internet allows for distributed collection of data very efficiently and has been used successfully to collect examples of gestures (Taylor et al., 2011; Spiro, 2012) and naturalistic facial responses (D. McDuff, Kaliouby, & Picard, 2012). In this thesis I mainly focus on spontaneous and naturalistic responses to media collected over the Internet. Using the Internet and remote measures of affect has several benefits:

- We can collect data much more efficiently and cost effectively than is possible by performing data collection in a laboratory environment.

- We can reach a much more diverse population and collect a significant number of samples from each demographic group.

- The environment in which people’s reactions are recorded more closely reflects the real setting of consumption.

In this section I present the framework for collecting facial responses over the Internet. The online framework was used for a majority of my data collection and is the first example of large-scale sourcing of facial expression responses over the Internet. In this thesis I will analyze over 20,000 videos collected in this new way.

This type of data presents several challenges. Firstly, as the data is collected outside our control it is impossible to know all the details of the context in which a person completes a task such as this. For instance, their location and who, if anyone, they are with are not strictly controlled. Secondly, the position of the webcam and the directionality and luminance of the lighting are not controlled. Finally, as a result of the informed consent, people are aware that their camera will be on and this knowledge may well influence their response. However, this is also true for lab studies. These issues will be addressed in detail in the discussion. I will now describe the framework and characterize the type of data that can be collected.
6.1.1 Methodology

Figure 14-2 shows the web-based framework and the user experience used to crowdsource the facial videos. For each experiment the structure of the components and the self-report questions were tailored in order to collect data to answer each of the research questions. The questions include the relationship between affective responses and content liking, desire to view content again, purchase intent and sales.

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 14 frames a second at a resolution of 320x240 to a server where automated facial expression analysis is performed. There is no need to download or install anything on the client 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.

6.1.2 Characterizing Data

The quality of the data collected using the crowdsourcing framework cannot be controlled as easily as with data collected in a laboratory setting. The data shows a large array of
qualities in lighting, video quality, pose, position and movement of the viewer(s), number of viewers, occlusions and background activity. Examples of frames from the data are shown in Figure 6-2. In some of the experiments the participants were given instruction to center themselves within the videos frame, to ensure they were well lit and to remove caps and chewing gum. However, in other experiments no instruction was given.

With great diversity in the data it is helpful to characterize the videos in some way compared to other datasets of similar behaviors. The data were compared to existing public datasets used for training and testing many AU and expression detectors. A detailed comparison can be found in (D. J. McDuff, El Kaliouby, & Picard, 2011; D. McDuff, Kaliouby, & Picard, 2012). I showed that distributions of position, scale, pose, movement and luminance of the facial region are significantly different from those represented in these traditionally used datasets (CK+ and MMI). Figure 6-3 shows a histogram of head scales, histograms of the average luminance for the facial region, histograms of the Michelson contrast for the facial region and histograms showing the pose angles for the CK+ (top), MMI (center) and our webcam (bottom) datasets. Across five out of six of these measures there was significantly higher variance in the webcam data.
Figure 6-3: A) Histogram of head scales for the CK+ (top), MMI (center) and our webcam (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 our webcam 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 our webcam 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 our webcam (bottom) datasets. Examples of poses with pitch=-0.13 rads, yaw=-0.26 rads and roll=-0.19 rads are shown.

6.2 Automated Facial Expression Analysis

6.2.1 Data Labeling

The performance of action unit and expression classifiers is closely tied to the number of samples that are available for training. The action unit and expression classifiers used for analysis in this thesis were trained on naturalistic and spontaneous examples of expressions collected over the Internet in addition to posed examples from publicly available datasets.

For labeling a web-based, distributed video labeling system (ViDL) was used which is specifically designed for labeling affective data (Eckhardt & Picard, 2009). An Affectiva-developed version of ViDL was used for the labeling task. ViDL allows the onset and offset of action units within a face video to be labeled. Figure 6-4 shows a screenshot of the ViDL interface.

Frames were labeled with ground truth labels (presence of action units). Three labelers labeled every frame of each video and the majority label for each action unit was taken for
Figure 6-4: Screenshot of the video labeling tool ViDL used to label the video data. The onset and offset of each action unit was labeled by at least three labelers with the majority label being taken as ground truth.

each frame. Appendix A shows analysis of the Spearman-Brown reliability of these labels broken down by action unit for 168,359 frames.

6.2.2 The Automated System

Several action unit classifiers were used at different stages of the analysis with the most accurate classifiers available being used at each point. Below we describe the details of the action units classifiers used. In the experiments described in Chapters 8 - 11 I will

Figure 6-5: Cropped examples of frames from ViDL with positive labels for action units that were coded in the groundtruth labeling stage. Smile and negative AU12 are labeled separately instead of labeling symmetrical AU12. The AU definitions can be found in Appendix C.
specify which of these action unit classifier metrics were used as inputs into each model. In general, the same methodology was used for detecting all action unit and expressions. The action unit classifiers were developed by Affectiva.

**Face Tracking and Registration**

The Nevenvision facial feature tracker\(^1\) was used to automatically detect the face and identify 22 facial feature points within each frame of the videos. The location of the facial landmarks is shown in Figure 6-6. The Nevenvision tracker performed well on the low quality flash videos recorded (320x240). Other methods of localizing points on the face such as Active Appearance Models (AAM) or Constrained Local Models (CLM) were not evaluated extensively.

**Feature Descriptors**

For each AU, a region of interest (ROI) around the appropriate part of the face is located using the landmark points. The image is cropped to the ROI and histogram of orientated

\(^1\)Licensed from Google, Inc.
gradients (HOG) (Dalal & Triggs, 2005) features computed for the region.

**Classification**

A Support Vector Machine (SVM) with RBF kernel is used for classification. The classification of each action unit is treated independently. These being all static, classifications were performed on a frame-by-frame basis without knowledge of features from previous or subsequent frames. In some cases the signed distance of the sample from the classifier hyperplane was taken and normalized using a monotonic function that in the training phase rescaled points between [0, 1].

### 6.2.3 Facial Action Units

In this analysis we use the Facial Action Coding Scheme (FACS) to characterize movements of the face in addition to several facial expressions. Using the method of registration, feature extraction and classification described above the following classifiers were trained:

**Outer eyebrow raiser: AU02**

Outer eyebrow raiser (*frontalis - pars lateralis*) action pulls the outer portion of the eyebrows upwards. This is often combined with inner eyebrow raiser (*frontalis - pars medialis*) in which the inner portion of the eyebrow is raised.

**Eyebrow lowerer: AU04**

Eyebrow lowerer (*corrugator supercillii*), or brow furrow, is the action of pulling the inner eyebrows down and toward one another. This often causes vertical wrinkles running between the eyebrows.
Nose wrinkler: AU09

Nose wrinkler (levator labii superioris alaquae nasi) is commonly seen in a disgust expression. In many cases the action leads to much more subtle appearance changes than other action units such as AU12. In general we use a classifier for AU09 and AU10 which fires when either or both are present.

Upper Lip Raiser: AU10

Upper lip raiser (levator labii superioris, caput infraorbitalis) is also commonly seen in a disgust expression. In general we use on classifier for AU09 and AU10 which fires when either or both are present.

Lip Corner Depressor: AU15

Lip corner depressor (depressor anguli oris) is commonly seen in a sadness expression. The lip corners are pulled downward. AU15 was only used in the definition of negative valence, I did not use an independent AU15 classifier in this work.

Smile

Smile probabilities were calculated distinct from AU12. A smile was characterized as the presence of AU12 without the presence of AU4 or AU9. Where AU12 is lip corner pull (Zygomaticus major) is the action that pulls the lip corners apart. This classifier detects either symmetric smiles only.

Smirk

In this work a smirk is defined as presence of an asymmetric AU12. More information about the smirk detector used can be found in (Sénéchal, Turcot, & el Kaliouby, 2013).
6.2.4 Continuous Measures of Emotion

Valence

Valence is defined as the positive/negative affect of the expression. The valence classifier was trained using a three label system: negative valence (-1), neutral (0) and positive valence (1). The training data was given ground truth labels using the following criteria:

if (smile present) \{valence = +1\}
else if (AU04 or AU09 or AU15 present) \{valence = -1\}
else \{valence = 0\}

Expressiveness

The expressiveness measure is defined as a linear combination (mean) of all the actions described above (AU2, AU4, AU9, AU10, smile and smirk).
Table 6.1: Number of videos and frames used for training the action unit and expression classifiers and the area under the receiver operating characteristic (ROC) curve for testing.

<table>
<thead>
<tr>
<th></th>
<th>Videos</th>
<th>Frames</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU02</td>
<td>93</td>
<td>15,700</td>
<td>0.85</td>
</tr>
<tr>
<td>AU04</td>
<td>308</td>
<td>16,000</td>
<td>0.72</td>
</tr>
<tr>
<td>AU09/AU10</td>
<td>254</td>
<td>58,000</td>
<td>0.84</td>
</tr>
<tr>
<td>Smile</td>
<td>868</td>
<td>114,000</td>
<td>0.97</td>
</tr>
<tr>
<td>Smirk</td>
<td>201</td>
<td>5,100</td>
<td>0.88</td>
</tr>
<tr>
<td>Valence</td>
<td>500</td>
<td>65,000</td>
<td>0.90</td>
</tr>
</tbody>
</table>

6.2.5 Accuracy of the Classifiers

Here I will characterize the performances of the classifiers on images similar to those collected in the experiments described in Chapter 5 (webcam videos of people responding to online media). Videos of facial responses were collected over the Internet and labeled on a frame-by-frame basis as described above.

The classifiers were tested against these images. The performance is quantified as the area under the receiver operating characteristic curve (ROC) when applied to testing data that was not used for training or validation. Table 6.1 shows the number of videos and frames used for training the classifiers and the area under the ROC curve for each (1.0 = perfect performance).

Figure 6-8 shows examples of the comparison between the human labelers coding and the machine coding of smiles, AU2 and AU4 sequences.

6.2.6 Comparison with Dial Measures

Measurement of facial behavior provides rich temporal information about a person’s affective responses. This is similar to the type of data that might be collected through an affect dial report, where someone is asked to turn a dial during an experience to quantify their affective response. A comparison was performed between the information gained from automated analysis of facial responses and dial reports (Kodra, Senechal, McDuff, & el Kaliouby, 2013).
Figure 6-8: Comparison between action unit classifier predictions (green) and manually coded action unit labels (blue and black dashed). Threshold of hand labels based on > 0.5 agreement between coders. Frames from the sequences are shown above. Top) Smile classification example, middle) AU2 classification example, bottom) AU4 classification example.
Figure 6-9: Time series plot of aggregate self-report dial responses (self-reported interest) (in Black) versus the smoothed facial metrics (in Red: smile metrics for Crime Drama and AU02 metrics for Sitcom). The non-smoothed facial data is shown in faded blue. Correlation between the smoothed metrics and dial data: -0.748 and 0.656 respectively. Self-reported interest is negatively correlated with smile metrics for the crime drama and positively correlated with AU02 metrics for the sitcom.

Participants were recruited to watch a sitcom (5 males, 5 females) and a crime drama (6 males, 5 females). Whilst watching the shows they were asked to report their interest in the show. Figure 6-9 shows time series plots of the measured facial metrics (raw and smoothed) and the dial measurements. For the sitcom, smile responses show high correlation with the dial measures and for the crime drama AU02 shows high inverse correlation with dial measures. The results suggested that automated facial expression analysis can serve as an accurate proxy for the self-report dial method of measurement and potentially yield deep insights beyond it. Furthermore, it may be able to achieve such results with a fraction of the sample size and with several logistical benefits. For instance, in this experiment it was found that facial expression analysis can require only 5% of the sample size (~10 rather than 200 participants) necessary for dial studies to yield similar results. In addition, the probability outputs from the expression classifiers seem to correlate strongly with the dial measures suggesting that they might be a good measure of the underlying state.

6.3 Remote Measurement of Physiology

During my research I have developed methods for remotely measuring physiological signals from the human face using digital cameras. Heart rate, respiration rate and heart rate
variability features can be calculated using a non-contact method described in (M. Poh, McDuff, & Picard, 2010, 2011). Figure 6-10 shows graphically how our algorithm can be used to extract the blood volume pulse (BVP) and subsequently HR information from the color channels in a video containing a face. This approach has been tested with a standard webcam and a novel five-band digital single-lens reflex (DSLR) camera.

6.3.1 Using a Webcam

The following is a description of the method for extracting the physiological parameters. A face tracker (such as the Nevenvision tracker or OpenCV face detector) is used to identify the facial ROI within each frame of the recorded videos and then a spatial average of the R, G and B channels is taken from the image ROI. For a time window of 30 seconds these form three time series. The signals are detrended using Tarvainen et al.’s method (Tarvainen, Ranta-aho, & Karjalainen, 2002), lambda = 100. The resulting signals are normalized by subtracting the mean and dividing by the standard deviation. The resulting zero mean, and unit standard deviation, signals are inputs into an Independent Component Analysis (ICA). In this step we compute the most non-Gaussian mixtures of the source signals. The output signals are bandpass filtered (128-point Hamming window, 0.7 - 4 Hz). The power spectra density for the resulting filtered signals are calculated using the Lomb periodogram. Of the three output “source” signals we assume that one contains a stronger BVP signal. The channel is selected automatically by finding the spectrum that contains the greatest amplitude peak. The computed frequency of the heart rate is the frequency corresponding to the maximum peak.

The estimated BVP signal was interpolated with a cubic spline function at a sampling frequency of 256Hz. Peak detection was performed using a custom algorithm with a moving time window of length 0.25s. To avoid artifacts (such as motion or ectopic beats) which can impact the HRV analysis, the resulting IBIs were filtered using the non causal of variable threshold (NC-VT) algorithm (Vila et al., 1997) with a tolerance of 30%. Finally, inter-beat intervals were filtered using a low pass filter with cut-off frequency 0.4Hz. We
construct the HRV spectrograms by calculating the power spectral density from the IBIs for sequential windows. For each window the power spectral density (PSD) of the inter-beat intervals was calculated using the Lomb periodogram. In this analysis we use a moving window of one minute, the sessions were two minutes in length and the step size was one second.

The method for remote measurement of BVP has been validated against contact sensors and proven to be accurate. On 12 subjects the correlations between heart rate measures from the algorithm and a gold-standard contact finger PPG measurement was 1.00, the correlation between respiration rate measure was 0.94 and the correlation between heart rate variability LF/HF measure was 0.88. See Poh et al. (2011) for more details.

6.3.2 Using a Five Band Digital Camera

In subsequent work I have shown that the RGB channels of a standard webcam are not necessarily optimal. Using a five band digital camera that can capture orange and cyan color bands in addition to the red, green and blue colors I found that a combination of the orange, green and cyan bands outperformed the red, green and blue combination.

The camera used to collect the video sequences for analysis was a digital single-lens
reflex (DSLR) camera. The lens used was a standard Zuiko 50mm lens. The camera’s sensor has the capability of capturing five color bands including the typical three frequency band sensors (red, green and blue (RGB)) and also cyan and orange frequency band sensors (RGBCO). Figure 6-11 shows the sensitivities for the five band camera and the five band camera sensor layout. The image shows the arrangement of the colors in a 4x4 pattern that repeats across the sensor. Each pixel on the sensor measures one color as determined by its position. Further details about the sensor and demosaicking can be found in (Monno, Tanaka, & Okutomi, 2012).

Using a similar procedure to that described above I calculated the HR, RR and HRV components using all combinations of color channels. The correlations between the physiological parameters measured using a contact BVP sensor and the camera method are shown in Table 6.2.

Predicting Cognitive Load

I performed an experiment to test whether the camera measurements were accurate enough to detect differences in a person’s physiological state at rest and under cognitive load. Our
Table 6.2: Comparison of the correlations between the contact finger sensor measurements and camera measurements for all combinations of the camera color channel signals. For all correlations $p < 0.01$. On the right are the channel combinations ordered from lowest mean correlation to highest mean correlation. The GCO channel combination performed best.

<table>
<thead>
<tr>
<th></th>
<th>HR</th>
<th>RR</th>
<th>LF</th>
<th>HF</th>
<th>LF/HF</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.99</td>
<td>0.95</td>
<td>0.60</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>G</td>
<td>0.99</td>
<td>0.91</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>B</td>
<td>0.99</td>
<td>0.93</td>
<td>0.68</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>C</td>
<td>0.85</td>
<td>0.44</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>O</td>
<td>0.83</td>
<td>-0.02</td>
<td>0.43</td>
<td>0.43</td>
<td>0.34</td>
</tr>
<tr>
<td>RG</td>
<td>0.97</td>
<td>0.66</td>
<td>0.72</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>RB</td>
<td>0.95</td>
<td>0.89</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>RC</td>
<td>0.99</td>
<td>0.67</td>
<td>0.69</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>RO</td>
<td>1.00</td>
<td>0.93</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>GB</td>
<td>0.89</td>
<td>0.75</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>GC</td>
<td>0.99</td>
<td>0.83</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>GO</td>
<td>1.00</td>
<td>0.98</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>BC</td>
<td>0.99</td>
<td>0.68</td>
<td>0.61</td>
<td>0.61</td>
<td>0.65</td>
</tr>
<tr>
<td>BO</td>
<td>1.00</td>
<td>0.92</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>CO</td>
<td>0.99</td>
<td>0.67</td>
<td>0.40</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td>RGB</td>
<td>0.85</td>
<td>0.67</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>RGC</td>
<td>0.99</td>
<td>0.75</td>
<td>0.67</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>RGO</td>
<td>1.00</td>
<td>0.92</td>
<td>0.83</td>
<td>0.83</td>
<td>0.86</td>
</tr>
<tr>
<td>RBC</td>
<td>0.99</td>
<td>0.69</td>
<td>0.71</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>RBO</td>
<td>1.00</td>
<td>0.92</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>RCO</td>
<td>1.00</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>GBC</td>
<td>0.99</td>
<td>0.77</td>
<td>0.80</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>GBO</td>
<td>1.00</td>
<td>0.93</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>GCO</td>
<td>1.00</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>BCO</td>
<td>0.99</td>
<td>0.84</td>
<td>0.69</td>
<td>0.69</td>
<td>0.77</td>
</tr>
<tr>
<td>RGBC</td>
<td>0.99</td>
<td>0.89</td>
<td>0.72</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>RGBO</td>
<td>1.00</td>
<td>0.81</td>
<td>0.79</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>RGCO</td>
<td>1.00</td>
<td>0.90</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>RBCO</td>
<td>1.00</td>
<td>0.90</td>
<td>0.81</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td>GBCO</td>
<td>1.00</td>
<td>0.72</td>
<td>0.83</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td>RGBCO</td>
<td>1.00</td>
<td>0.74</td>
<td>0.81</td>
<td>0.81</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Lowest $\tau$

Highest $\tau$
experiments featured 10 healthy participants of both genders (seven females), different ages (18-30) and skin color. During the experiment participants were seated approximately 3m from the camera and asked to face the camera while the videos were recorded. This study was approved by the Institutional Review Board of the Massachusetts Institute of Technology. All experiments were conducted indoors and with a varying amount of ambient light (a combination of sunlight and indoor illumination). Participants were seated and the data were recorded on a laptop (Toshiba running Windows 7).

**Measurements at rest:** In the first experiment participants were asked to sit still, look toward the camera and relax. The video and physiological recordings were captured for two minutes. For one of the participants in the rest condition the contact finger BVP measurements were noisy due to motion artifacts; this session was not used for the validation of the remote measurements. Although the camera method is also susceptible to motion artifacts this highlights some of the challenges associated with contact measurements.

**Measurements under Cognitive Load:** In the second experiment participants were asked to perform a mental arithmetic test (MAT). Starting with the number 4000 they were required to subtract 7, then subtract 7 again, and so on, as quickly as possible. The video and physiological recordings were captured for two minutes. The participants started the task immediately after the recordings were started. In order to increase the cognitive load induced, prior to starting the task we told the participants that they were competing against the other people to reach the lowest number after two minutes.

**Results:** Figure 6-12 (a-d) shows the values of heart rate, respiration rate, HRV LF and HRV LF/HF ratio for each participant (blue = rest, red = session). Figure 6-12 (e-g) shows the mean HR, RR and HRV LF/HF ratio across all participants with error bars showing one standard deviation either side of the mean.

Using the physiological parameters measured with the camera as features, I built and
tested a classifier for predicting whether an individual was under cognitive load (mental arithmetic task) or at rest during the video. I tested a Naive Bayes model and a support vector machine (SVM) with linear kernel for the classification. SVMs have been shown to perform well on many classification problems. The input features were the mean heart rate, mean respiration rate, normalized HRV LF power, normalized HRV HF power and HRV LF/HF power ratio for each session. The cost, $C$, parameter was set to 0.1. I performed a person-independent testing by withholding the data for one participant in the test set and using all the remaining data for training. I repeated this 10 times, once for each participant.

The prediction accuracy of the model for classifying rest or cognitive load using a linear SVM was 85%, this is a two-class case with balanced class sizes and therefore a
naive prediction would be 50%. Table 6.3 shows the accuracy using the HR, RR and HRV features alone and in combination.

**HRV Spectrograms**

I calculated the HRV spectrograms from the two minute sessions using a one minute sliding window with one second increments. Figure 6-13 shows a comparison of spectrograms recovered from the three band (RGB) and three band (GCO) recordings next to those from the contact finger measurements. On the left are examples from sessions in which the participants were at rest and on the right are examples in which the participants were under cognitive load.

The spectrograms calculated from the five band observations are closer to those of the finger measurements. The remotely measured spectrogram in example P5 of Fig. 6-13 actually appears to be more accurate than that from the finger sensor - perhaps due to motion artifacts as a result of the fingers moving. It is clear there is less high frequency power in the HRV spectra for those individuals under cognitive load: this is what we would expect due to less parasympathetic activity.
Figure 6-13: Heart rate variability spectrograms (normalized amplitude) calculated using: Top) RGB camera signals, middle) GCO camera signals and bottom) contact finger sensor. Qualitatively the measurements from the GCO channels more closely match those from the contact sensor (reinforcing the quantitative comparisons in Table 6.2). Shown are sessions in which participants 5 and 10 were at rest and sessions in which participants 5 and 10 were under cognitive load. There are stronger low frequency components in the latter as we would expect less parasympathetic nervous system activity. The measurements for participant 5, under cognitive load, made using the camera suggest that the camera (using GCO bands) may have been more accurate than the finger sensors.
Chapter 7

Preliminary Analysis of Facial Responses over the Internet

This chapter presents general observations from the analysis of crowdsourced facial responses that will be relevant in the following chapters. I show that facial responses can reveal that responses to ads are impacted by familiarity. I characterize the variability within a person’s responses and how affective responses to video ads compare to other clips.

7.1 Multiple Viewings and the Impact on Facial Responses

In the Cereal experiment participants were exposed to the same content twice, once at the start of the survey and then again following questions at the end. This allows us to explore the impact of multiple viewings on the affective response of viewers. In some cases the mean facial metrics during the first viewing were highly correlated with the facial metrics during the second viewing (e.g. Figure 7-1, n=23. \( \rho = 0.802 \)). The peaks are closely aligned and the smile responses lead to the same conclusions about the points of greatest positive response.

However, in other cases mean facial metrics during the first viewing were only weakly correlated with the facial metrics during the second viewing (e.g. Figure 7-2, n=23). In this
Figure 7-1: Aggregate (no. of viewers = 23) smile classifier output for a commercial on first (blue) and second (green) viewing. Correlation = 0.802, p<0.05. Shaded area shows the standard error.

case the peaks within the aggregate responses are very different and correlation of the mean lines is much lower (\( \rho = 0.430 \)). At points during the clip, such as at 18s, the responses were significantly different.

In traditional methods of moment-to-moment affective measurement viewers are asked to watch the content once and then on a second viewing report their affect using a slider or dial. However, as we see here it may be the case that their affective response is not the same on the first and second viewings, therefore requiring the viewer to watch the content twice - and recording their report only the second time can lead to different conclusions compared to the first-time capture. In addition, perhaps the changes in response might give us information about the effectiveness of content and these changes would not be captured if it is not possible to measure the response on the first viewing.
In this thesis the focus is mainly on predicting the effectiveness of video advertisements, many of which are intentionally humorous, from facial responses. We can only hope to predict outcomes from facial responses accurately if viewers are expressive during content. Here I will compare the responses of a number of individuals to humorous ads and to a popular YouTube clip featuring a baby making funny facial expressions and laughing.\(^1\).

Figure 7-3 shows the smile (blue) and disgust (red) responses, measured using the automated classifiers, to four video ads and the baby clip for four individuals (i-iv). Responses to some of the ads are not as strong as responses to the baby clip but there are measurable expressions and individual responses to certain ads are very strong. Example frames from the first ad for three of the subjects are shown.

One individual (Figure 7-3 (i)) shows elements of disgust and smile responses to the various ads and a strong smile response to the baby clip. Another (Figure 7-3 (ii)) shows very strong smile responses but very little evidence of disgust expressions - smiling a lot to

---

\(^1\)http://www.youtube.com/watch?v=xBSYD0dQCAw - clip has over 10,000,000 views on YouTube.
all the clips.

Some individuals do not show much expression to any of the clips (Figure 7-3 (iii)). Figure 7-3 (iv) shows an individual’s responses with very little expression to the ads but much stronger response to the baby clip - the baby clip appears much more humorous to this viewer than the ads.

These examples show that different viewers can respond quite differently to the same videos and that viewers do show detectable facial expressions during video ads.

7.3 Inter-Person Variability in Facial Response

It is helpful to characterize the variance with the responses measured across a group of viewers and to analyze how many participants are required to be able to identify significant dynamic changes within the mean response. This will be particularly important when we consider aggregated responses to ads and what they can help us predict.

In order to evaluate how the signals vary with the number of responses I show how the mean and 95% confidence bounds (standard error) for the aggregate metrics vary with the number of participants. Figure 7-4 (top row) shows the aggregate smile metrics for an ad with 10, 20, 30, 50 and 70 randomly selected viewers respectively. We see that aggregate trends converge and that the trends with 50 and 70 viewers are correlated ($\rho = 0.82, p << 0.05$). Perhaps more importantly the form of the trends is similar with peaks within the trends in similar places. However, we must be aware that there still remains some variance within the trends even with 70 viewers.

Smiles occur frequently in responses to humorous ads like this but do the trend lines for more infrequently occurring expressions converge in a similar way? Figure 7-4 (middle row) shows the aggregate disgust trends and standard error (95% confidence bounds) for different size samples (no. of viewers = 10, 20, 30, 50 and 70). Once again the aggregate trends converge and the trends with 50 and 70 viewers are similar.

To investigate how the mean trends change with a much greater numbers of viewers I
Figure 7-3: Individual smile (blue) and disgust (red) responses to four video ads (a-d) and the baby clip (e) for four individuals (i-iv) - chosen to illustrate the variety of responses. Example frames from the first ad for two of the subjects are shown.
had over 1,000 viewers watch the YouTube baby clip described above. Figure 7-4 (bottom row) shows the aggregate smile trends and standard error (95% confidence bounds) for different size samples (no. of viewers = 10, 20, 50, 100, 1,000).

We can see both qualitatively and quantitatively that the trend lines with a sample of 100 viewers are similar to that with 1,000 viewers suggesting that the order of magnitude more viewers may only incrementally improve our resulting inferences. However, with only 10 or 20 viewers the variance is extremely high and so it is likely that we really need to sample at least 50 viewers to get an accurate representation of a wider population. These results bring into question the validity of inferences that can be made using a small population and emphasize the benefits of using a data collection framework that means we can efficiently scale the number of observations.

Ideally we would be able to collect as high a number of samples as possible. However, in reality we are limited by cost and the time taken to recruit subjects. Therefore, in the aggregate analysis I perform in Chapters 8 - 11 I use data from between 50-100 viewers (the exact number varies depending on the number of videos in which viewers had their webcam in a suitable position) which represents a good trade-off between cost, time and information content.
Figure 7-4: Aggregate smile and disgust metrics with different population sizes, shaded area = 95% confidence bounds. Top) mean smile probability for Ad A, middle) mean disgust probability for Ad B, bottom) mean smile probability for baby video.
Chapter 8

Super Bowl: Predicting Viewer Preferences

8.1 Aims and Motivation

This experiment acted as a proof-of-concept test for the crowdsourcing framework and as a basis for comparing facial responses to self-reported preferences. Knowledge of likability and desire to view again are useful in advertisement copy-testing and could also be used to personalize the content viewers are shown when watching TV over the Internet using platforms such as Netflix or Hulu.

The aims of this experiment were:

- To evaluate if it was possible to efficiently collect ecologically valid facial responses to videos online.

- To determine whether facial responses were correlated with reported preferences, liking and desire to watch again.

- To test whether we could automatically an individual viewer’s preferences from their facial responses.
8.2 Data

The experiment was the first in the world data collection of natural and spontaneous facial responses to videos online. 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.

Respondents

The responses were collected in natural settings via the Internet and the application was promoted on the Forbes website (http://www.forbes.com/2011/02/28/detect-smile-webcam-affectiva-mit-media-lab.html). In total 6,729 sessions were completed by people who opted-in to the experiment. The respondents in this experiments were not paid. There was no control over the people who took part, the lighting conditions or the camera position.
Methodology

A screenshot of the consent form is shown in Figure 8-2. If consent is granted, the commercial is played in the browser whilst simultaneously streaming the facial video to a server. In accordance with the Massachusetts Institute of Technology Committee On the Use of Humans as Experimental Subjects (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 watched and self-reported responses (if any) to the self report questions. No other data is stored. This data collection protocol was approved by MIT COUHES prior to launching the site.

Following each commercial, the webcam is automatically stopped and a message clearly states that the “webcam has now been turned off”. 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). Figure 8-3 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 2,400 labeled examples for each question and commercial combination.

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.

### 8.3 Aggregate Responses

Figure 8-8 shows 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:

\[
SE = \frac{\sigma}{\sqrt{n}} \tag{8.1}
\]

Where \( \sigma \) 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 8.1: Distribution of responses to self-report questions “Did you like the video?” and “Would you like to watch this video again?”.

<table>
<thead>
<tr>
<th>“Did you like the video?”</th>
<th>“Would you like to watch this video again?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nah</td>
<td>Ugh</td>
</tr>
<tr>
<td>66</td>
<td>13</td>
</tr>
<tr>
<td>Meh</td>
<td>49</td>
</tr>
<tr>
<td>Heck ya</td>
<td>3</td>
</tr>
</tbody>
</table>

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 8.1 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?”.

### 8.4 Predicting Preferences

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 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 the content again to someone 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.

Figure 8-4 shows the framework we use to automatically classify media preferences from the facial responses. In this section I will describe the system and the results.
Figure 8-4: Framework for classification of content liking and desire to view again based on automatically detected smile responses recorded over the web. Smile metrics are extracted from the recorded facial response. The smile metrics are filtered and temporal features extracted from 20 evenly spaced bins. Resulting features used for classification.
8.4.1 Pre-Processing Facial Responses Data

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. Later a second version (V2) was developed. We compared 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. 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.

I tested how well the smile classifiers performed 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 (McDuff et al., 2013)). 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 for V1 is shown in Figure 11-6 (blue line); the area under the curve is 0.874. The resulting precision-recall curve for V1 is shown in Figure 8-6 (blue line); the area under the curve is 0.73. The

---

Figure 8-5: Funnel chart showing the participation at the early stage of the Super Bowl experiment. 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.
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.

The smile tracks that satisfied the 90% trackable criteria were used. First they were filtered with a low pass linear-phase FIR filter (with Hamming window) to smooth the signals. The high frequency 3dB cut-off of the low-pass filter was 0.75 Hz (filter order = 128). Secondly, features were extracted from the smile tracks as shown in Figure 8-4. 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. The videos have a frame rate of 14fps 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 in results.

Before performing the classification experiments we computed a 2D mapping of the resulting ROC curve for V2 is shown in Figure 8-6 (green line); the area under the curve is 0.91. The resulting precision-recall curve for V2 is shown in Figure 8-6 (green line); the area under the curve is 0.80.

Figure 8-6: 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.
data to get an intuition about the manifold of responses. Figure 8-7 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 Figure 8-7 is greater than that of example c despite it being a disliking response compared to a liking response. Looking at the temporal profiles examples c and d are the closest to one another.

8.4.2 Classification of Responses

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
Figure 8-8: 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 ads 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.

weak desire responses all of which satisfy the 90% trackable criteria.

**Validation** It is important that the prediction of success is independent of the particular commercial and that the features generalize across new content. In order to test this we use a challenging leave-one-commercial-out scheme. The data for one commercial was removed for testing and the remaining data was used for training and validation. This was repeated for each of the three ads. For validation a leave-one-commercial-out methodology was used. The training data set was split and data for one commercial 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.

**8.4.3 Impact of Smile Detector Performance on Preference Prediction**

Table 8.2 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
Table 8.2: Prediction performance for liking and desire to watch again classifiers using features from smile detector V1 and smile detector V2.

<table>
<thead>
<tr>
<th></th>
<th>Liking</th>
<th></th>
<th>Desire to View Again</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smile V1</td>
<td>Smile V2</td>
<td>Smile V1</td>
<td>Smile V2</td>
</tr>
<tr>
<td>ROC AUC</td>
<td>0.80</td>
<td>0.82</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>ROC PR</td>
<td>0.95</td>
<td>0.96</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>Accuracy</td>
<td>76%</td>
<td>81%</td>
<td>75%</td>
<td>73%</td>
</tr>
</tbody>
</table>

smoothing window size, $\omega = 3$ was used.

**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-commercial out training scheme to ensure generalizability across other amusing video ads. Table 8.3 (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).

Table 8.3: Confusion matrices for the best performing liking classifier: left) using smile detector V1, right) using smile detector V2.

<table>
<thead>
<tr>
<th>Liking</th>
<th>Actual +ve (Liking)</th>
<th>Actual -ve (Disliking)</th>
<th>Liking</th>
<th>Actual +ve (Liking)</th>
<th>Actual -ve (Disliking)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict +ve</td>
<td>1027</td>
<td>53</td>
<td>Predict +ve</td>
<td>1078</td>
<td>56</td>
</tr>
<tr>
<td>Predict -ve</td>
<td>305</td>
<td>149</td>
<td>Predict -ve</td>
<td>236</td>
<td>142</td>
</tr>
</tbody>
</table>

While the overall performance is good, there are a number of misclassifications that occur. Figure 8-9 (g-l) shows cases of false positive results. Figure 8-9 (m-r) shows cases of false negative results. For comparison Figure 8-9 (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 expression 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 Figure 8-9.
From 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 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 people 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 V1 of the system.

**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 8.3 (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). Figure 8-11 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 (Figure 8-
11 a and b) were corrected using V2 of the detector. Cases in which subtle smiles were
missed by V1 and this led to an incorrect “disliking” prediction (Figure 8-11 c and d) were
corrected too as V2 detected more subtle smiles.

Figure 8-11 (g-l) shows cases of false positive results. Figure 8-11 (m-r) shows cases of
false negative results. For comparison Figure 8-11 (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 a 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 sometimes not be coherent with the observed response.

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 Figure 8-11 k. It may be that detection of other facial actions could help improve the prediction accuracy as we know that smile may occur in both positive and negative situations (Hoque et al., 2012). Further details can be found in McDuff et al. (2013).

8.5 Results/Conclusions

This is the first Internet-based collection of facial responses and shows that ecologically valid data can be collected efficiently using an Internet framework. The facial responses were correlated with reported preferences of liking and desire to view again. I present an automated method for classifying “liking” and “desire to view again” based on 3,268 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. I 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 re-
Figure 8-9: 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.

Figure 8-10: 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.
Figure 8-11: 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.

spectively when using a challenging leave-one-commercial-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 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. 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.

The results show that strong increases in positive expressions of emotion and high peaks in positive expressions of emotion are indicative of a liking response. These results are congruent with the work of Kahneman et al. (Varey & Kahneman, 1992; Fredrickson & Kahneman, 1993). Baumgartner et al. (1997) reached a similar conclusion in the advertising space: “our findings suggest a way of designing advertisements that will utilize ad time most effectively and maximize overall positive emotion toward and liking for the advertisement”. My results show that the intuition stated here is confirmed in a real-world setting.
Chapter 9

Mars: Predicting Ad Likability and Purchase Intent

9.1 Aims and Motivation

Copy testing of ads typically involves asking panelists their feelings about ads and evaluating the likelihood of good performance based on the aggregated responses. The surveys are usually concerned with self-reported variables such as ad liking, familiarity, brand likability and purchase intent. In this chapter I will investigate the relationship between facial responses and aggregated measures of ad performance to test whether automated affect analysis could be used to screen ads rather than relying on self-reported responses. For instance, can we predict accurately whether one ad is likely to boost purchase intent by a greater amount than another using only the automatically measured affective response.

In this work I will consider a majority of ads which are intentionally humorous and it is therefore reasonable to think that the results from Chapter 8 will hold and that stronger positive emotional responses to online ads is more likely to lead to their being more successful. I will compare the performance a) for all the ads together; and b) for just the intentionally humorous ads:

The aims:
Figure 9-1: Screenshots of a subset of the 170 ads tested in this study. The ads were for products in one of the four categories: pet care, confectionery (chocolate, gum and candy), food (instant rice and pasta products) and cosmetics/toiletries. The ads were from different countries: France, Germany, the UK and the US.

- To collect a much larger data set of affective responses to ads.

- To model the relationship between facial responses and aggregate self-report metrics of ad liking and purchase intent.

- To identify affective profiles that are more likely to be associated with high ad liking and purchase intent.

9.2 Data and Data Collection

9.2.1 Video ads

The following experiment was designed to extend the understanding of the relationship between short term sales increase due to an advertisement and the measured affective response to the advertisement. In addition, substitute measures of success (self-reported brand attitude (liking), self-reported purchase intent and memory (brand and ad recognition) were measured. In this chapter and the following chapter I will show analysis of the same data.

I tested 170 video ads from four countries (30 from France, 50 from Germany, 60 from the UK and 30 from the US). The videos were all originally aired between 2001 - 2012. The mean length of the video content was 28s (std = 8.6s). The 170 copies under testing
Table 9.1: Number of videos tested from each product category and emotion category (categorized using MTurk labelers). A majority of the ads were intentionally amusing.

<table>
<thead>
<tr>
<th>Emotion Category</th>
<th>Petcare</th>
<th>Confec.</th>
<th>Food</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amusement</td>
<td>14</td>
<td>46</td>
<td>7</td>
<td>8</td>
<td>75</td>
</tr>
<tr>
<td>Heart-warming</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Cute</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Exciting</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Inspiring</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Sentimental</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>No Majority</td>
<td>11</td>
<td>17</td>
<td>3</td>
<td>6</td>
<td>37</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>75</td>
<td>19</td>
<td>23</td>
<td>170</td>
</tr>
</tbody>
</table>

were divided into 17 independent sets (batches) of 10 ads. Each participant watched the ads from one batch in a randomized order. The batching process was performed in order to minimize boredom (having individuals watch many more than 10 ads may have caused them to become bored towards the end). Randomization was performed to reduce the primacy, recency and carry-over effects. Screenshots of some of the ads can be seen in Figure 9-1.

Product Categories

The videos tested were mostly ads for products in one of the three following categories: pet care, confectionery (chocolate, gum and candy) and food (instant rice and pasta products). Of the 170 ads, 23 from other categories were also tested (mainly cosmetics and toiletries). Importantly, these are all products that might be bought frequently by category users and don’t represent a longterm purchasing decision (such as a new car might). Table 9.1 shows the number of videos from each product category. Each batch consisted of ads from only one region but from a range of product categories and a range of emotion categories.

Crowdsourced Ad Labels

The ads were not all designed to elicit the same emotions or to communicate the same messages. Two different ads could be designed to create very different feelings within the viewers. I used the crowdsourcing platform Amazon’s Mechanical Turk (MTurk) to
crowdsource labels for the videos.

At least three coders were recruited to watch each video and answer the following question: “CHOOSE the words that best describe the type of FEELINGS that you think this video was designed to induce”. Labelers were able to select one of more answers from the following list:
Sentimental, Inspiring, Exciting, Romantic, Heart-warming, Amusing, Cute.

The majority label was taken for the videos. If there was no majority label it was given a label of “No Majority”. Table 9.1 shows the number of videos from each emotion category. A majority of the ads were labeled as amusing.

9.2.2 Respondents

One hundred respondents were recruited to view each batch of copies. Participants were recruited from the four countries (UK, US, France and Germany). Recruitment was such that age groups, gender and economic status of the participants was as balanced as possible and also helped mitigate the effects of self-selection biases. In addition, in all cases at least 70% of the viewers who watched each ad were a user of the product category being advertised. The respondents were compensated with approximately $8.10 for participating in the study (~ $0.8 per ad). Figure 9-2 shows the distribution of viewers across gender, age and economic status. It took an average of 36 minutes for participants to complete the survey.

Not all the participants I contacted had a functioning webcam or were willing to let their responses be recorded. In either case they were not allowed to continue with the survey. Of the participants that started the survey 48% reported having a working webcam. Of these 48% of participants, 49% stated they were happy to have their facial responses recorded. These statistics show that a significantly larger number of people (perhaps up to 400% of the required population) need to be contacted in order to collect a dataset. However, contacts are very inexpensive and so this generally should not cause many issues.
Figure 9-2: The a) age (in years), b) gender and c) economic split (annual salary in $1000s) of the 1,223 viewers who took part in the Mars study.

In total 1,223 people successfully completed the survey. Each participant watched 10 ads giving a total of 12,230 facial response videos. Each ad was watched by an average of 72 viewers. Once a participant had taken the survey they were excluded from taking the survey again, even with a different set of ads.

At the end of the survey participants were asked: 1) How COMFORTABLE did you feel during the study? Of the viewers 88% reported “very comfortable” to “neutral”, 3% reported “very uncomfortable”. 2) Did you behave differently than you would have if you were watching these ads NOT as part of a study? Of the viewers 75% reported “no differently”, 21% reported “a little differently” and 4% reported “very differently”. These statistics along with observation of the recorded videos suggest that the responses of the viewers was in general natural. However, it is very difficult to conduct such a study with full consent from the participants that they will be recorded and not have some chance that their behavior will be different from how they might respond if they were not being recorded. I strongly believe that these data are more naturalistic than they would be if collected in a lab-based setting which is perhaps even more likely to cause unnatural behavior.
9.2.3 Methodology

The facial video recording framework was integrated into a survey with self-report questions and memory tests in a structure shown in Figure 9-3.

Self-Report Ad Liking, Brand Liking, Purchase Intent and Sharing

A pre-survey was designed in order to measure the self-reported brand liking and purchase intent of the participants prior to watching the commercials in the study. In order to minimize bias caused by the pre-survey the participants waited three days before completing the main survey. During the pre-survey participants were asked to answer the following questions about each of the brands in the study. If they were not familiar with a brand then they could indicate so rather than entering a rating.

Q. How **LIKABLE** do you find each of the following brands? Please rate all brands

<table>
<thead>
<tr>
<th>Very dislikable</th>
<th>Neutral</th>
<th>Very likable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

Q. Next time you are buying `[product category]` how likely are you **TO PURCHASE** products from each of these brands?

<table>
<thead>
<tr>
<th>Not likely to purchase</th>
<th>Neutral</th>
<th>Likely to purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

The main survey consisted of the participant viewing a series of ads followed by questions about them. The sequence of ads and questions in the main survey is shown in Figure 9-3. During the experiment the participants watched 10 ads in a randomized sequence, in order to minimize primacy, recency and carry-over effects. Following each ad viewers were asked four questions about the ad that they had just watched:

Q. If you watched this ad on a website such as YouTube how likely would you be to **SHARE** it with someone else?

<table>
<thead>
<tr>
<th>Very unlikely</th>
<th>Neutral</th>
<th>Very likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>
Q. How likely would you be to **Mention this ad** to someone else?

<table>
<thead>
<tr>
<th>Very unlikely</th>
<th>Neutral</th>
<th>Very likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
</tbody>
</table>

Q. How much did you **Like** the **AD** that you just watched?

<table>
<thead>
<tr>
<th>Not at all</th>
<th>Neutral</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
</tbody>
</table>

Q. Had you seen the ad you just watched before this study?

| 1. No, never | 2. Once or twice | 3. More than twice |

After watching all the ads, viewing the intermediate media, and answering the brand and ad recognition questions described below the viewers were asked the following questions about each brand:

Q. How **Likable** do you find each of the following brands? Please rate all brands

<table>
<thead>
<tr>
<th>Very dislikable</th>
<th>Neutral</th>
<th>Very likable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
</tbody>
</table>

Q. Next time you are buying *[product category]* how likely are you **To Purchase** products from each of these brands?

<table>
<thead>
<tr>
<th>Not likely to purchase</th>
<th>Neutral</th>
<th>Likely to purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
</tbody>
</table>

**Brand and Ad Recognition**

In order to test advertisement recognition the participants were shown a series of descriptions of the ads that they had just watched and they were asked to enter the name of the brand that featured in the ad from a forced choice list of names. An example is shown in Figure 9-4. The question asked was:

Q. Enter the brand name that corresponds to the ad being described. If you don’t know the name of the brand leave the field blank.
In order to test brand recognition the participants were shown a series of images taken from the commercials. Brand names and logos were pixelated out. The participants were asked to name the correct brand from a forced choice list of names:

Q. Enter the brand name represented in the image. If you don’t know the name of the brand leave the field blank.

*Comfort and Behavior Self-report*

As mentioned above the participants were asked about their comfort and behavior following the survey.

**Q. How** COMFORTABLE **did you feel during the study?**

<table>
<thead>
<tr>
<th>Very uncomfortable</th>
<th>Neutral</th>
<th>Very comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

**Q. Did you behave differently than you would have if you were watching these ads NOT as part of a study?**


**9.2.4 Facial Coding**

For the analysis in this chapter I used classifiers for eyebrow raises, smiles, disgust expressions and positive and negative valence expressions (for details of the facial coding classifiers see Chapter 6). I selected these facial expressions as they were deemed highly relevant to the context of advertising and viewer responses.

**9.2.5 Expressiveness of Viewers**

To characterize the expressiveness of viewers I analyzed the metrics across all videos. Frames for which the expression classifier output is smaller than 0.1 are classed as no ex-
Figure 9-3: Experimental procedure for the Mars experiment. The experiment was divided into two parts. 1) A pre-survey with baseline questions and 2-4) a main survey with ads and follow-up questions. During the main survey each viewer watched 10 ads and answered questions about liking and desire to share following each and then answered questions about purchase intent and brand likability at the end of the survey.

53. Enter the brand name that corresponds to the ad being described. If you don't know the name of the brand leave the field blank.

"A mother is preparing dinner. The other family members drop what they are doing to run to the kitchen."

Figure 9-4: Example of the brand recall task. Participants were asked to enter the name of the brand represented from a forced choice list.

Figure 9-5: Example images from the brand recognition task. Brand names and logos were pixelated. Participants were asked to enter the name of the brand represented from a forced choice list.
pression present. In 82.8% of the 3,714,156 frames in which a face was detected there was
no detected eyebrow raise, smile, disgust expression or non-neutral valence expression.

Figure 9-6 shows histograms of the number of frames with each expression (smiles,
disgust and positive and negative valence) probability. Examples of frames from select
buckets are shown. A vast majority of the frames did not have an eyebrow raise, smile,
expression of disgust or positive or negative valence detected as present. Table 9.2 shows
the percentage of frames which feature expressions of each metric value, only 6% of frames
had an eyebrow raise > 0.1, 7.9% of frames had a smile > 0.1 and 5.5% of frames an
expression of disgust > 0.1.

Table 9.2: Percentage of the 3,714,156 frames with expressions metrics within 10 evenly
spaced classifier output bins centered on the values shown.

<table>
<thead>
<tr>
<th>Bin</th>
<th>Eyebrow R.</th>
<th>Smile</th>
<th>Disgust</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>94.0</td>
<td>92.1</td>
<td>94.5</td>
<td>81.9</td>
</tr>
<tr>
<td>0.15</td>
<td>1.04</td>
<td>1.67</td>
<td>1.63</td>
<td>4.22</td>
</tr>
<tr>
<td>0.25</td>
<td>0.59</td>
<td>0.89</td>
<td>0.80</td>
<td>2.29</td>
</tr>
<tr>
<td>0.35</td>
<td>0.43</td>
<td>0.61</td>
<td>0.53</td>
<td>1.58</td>
</tr>
<tr>
<td>0.45</td>
<td>0.35</td>
<td>0.50</td>
<td>0.40</td>
<td>1.37</td>
</tr>
<tr>
<td>0.55</td>
<td>2.04</td>
<td>1.11</td>
<td>0.84</td>
<td>3.44</td>
</tr>
<tr>
<td>0.65</td>
<td>0.82</td>
<td>0.79</td>
<td>0.54</td>
<td>2.19</td>
</tr>
<tr>
<td>0.75</td>
<td>0.33</td>
<td>0.68</td>
<td>0.38</td>
<td>1.15</td>
</tr>
<tr>
<td>0.85</td>
<td>0.12</td>
<td>0.66</td>
<td>0.24</td>
<td>0.88</td>
</tr>
<tr>
<td>0.95</td>
<td>0.03</td>
<td>0.98</td>
<td>0.14</td>
<td>0.87</td>
</tr>
</tbody>
</table>

In only 54.5% of face videos were there any detected expressions greater than 0.1 at
any point. In 36.9% of the face videos were there any detected expressions greater than 0.5
at any point. However, with an average of over 70 viewers for each ad I found detectable
responses - greater than 0.5 - in at least one viewer for all the ads. In addition, note that there
are much larger numbers of detected positive valence expressions than detected negative
valence expressions. Considering that most ads probably aim to induce positive affect this
is to be expected.
Figure 9-6: Histograms of the number of frames with each expression probability. From top to bottom: eyebrow raise, smile, disgust, valence. In 82.8% of frames was there no detectable eyebrow raise, smile, disgust or positive/negative valence expression with magnitude above 0.1. Responses to ads in naturalistic settings are sparse but for all the ads there were expressive responses within the 70+ viewers.
9.2.6 Aggregate Characteristics

Figure 9-7 shows the mean valence metrics for the different ads tested (order by increasing mean positive valence). Interestingly, a large number of ads had negative valence mean expression metrics. These results, and those above, show that although responses are sparse, different people respond to the ads differently and the ads elicited a range of expressions (from strong positive valence to negative valence). A few ads elicited no aggregate positive or negative valence.

In the remaining part of the chapter I focus on the prediction of aggregate level results (how effective is an ad across all viewers) rather than individual level results (how effective is an ad for a specific individual). Examples of individual level prediction can be found in Chapter 8.

9.3 Classification

To test the predictive performance of the facial responses I build and test classifiers for predicting ad effectiveness performance (effectiveness based on the self-reported ad liking and brand PI responses) directly from the measured facial response metrics and contextual information (product category and country). Below I explain how I calculate the features, the labels and how I validate, train and test the models. Figure 9-8 shows a flow diagram
9.3.1 Calculating Aggregate Metrics

I calculate aggregate expression metrics for each ad from the individual facial responses to that ad - Figure 9-8 (step 1). These were calculated as the mean metric intensity across all viewers to the ad (ignoring frames in which no face was detected). I compute mean tracks for the eyebrow raise, smile, disgust and valence classifiers.

9.3.2 Extracting Summary Features

Facial Metric Features: I extract summary features from the aggregate facial expression metrics. The summary features extracted from each summary metric were: mean, maximum, minimum and gradient. Figure 9-8 (step 2) shows how the features were extracted from an aggregate metric track. These four features for the four facial metrics classifiers led to a feature vector of length 16 for each ad.

Contextual Features: I use the product category and the country the ad is from as contextual features. These are coded as a binary matrix with columns that correspond to each of the five categories and columns that correspond to each of the four countries.
Figure 9-9: Distribution of average: left) ad liking response and right) delta in purchase intent response for all ads. The median, minimum and maximum values are all shown. The report of liking was significantly (p<0.001) greater than neutral.

9.3.3 Computing Labels

The labels I use are taken from the viewers’ self-report responses to the questions answered during the survey. In both cases I divide the data into two groups to create a binary classification problem. As this is the first time that prediction of ad effectiveness from crowdsourced responses has been attempted it helps to start with a simpler classification paradigm rather than one of multi-class classification or regression.

**Liking Score:** To compute the liking score for each commercial I calculate the average ad liking reported by each of the viewers, in response to the question “*How much did you LIKE the AD that you just watched?*” I divide the ads into two categories - those with average liking greater than the median score and those with average liking equal to, or lower than, the median. Since I separate the ads using the median value the classes are inherently balanced in size. Five of the ads did not have complete labels, therefore there are 165 liking examples.

**Purchase Intent Score:** To compute the purchase intent score for each commercial I calculate the mean delta in purchase intent reported by each of the viewers, in response to the question “*Next time you are buying [product category] how likely are you TO PURCHASE products from each of these brands?*” which was asked in the pre-survey and at the end of the main survey. I divide the ads into two categories - those with average purchase intent
Figure 9-10: Scatter plot of ads with aggregate smile trend against aggregate disgust trend. Ads with higher mean ad liking score are shown in blue, ads with lower mean ad liking score are shown in red. The smile gradient is more discriminative than the disgust gradient.

delta greater than the median and those with average purchase intent delta equal, or lower than, the median. Seven of the ads did not have complete labels, therefore there are 163 purchase intent examples.

Figure 9-9 shows the distribution of labels for the liking score and purchase intent score. The average reported liking was significantly (p<0.001) greater than “neutral”. As explained above, in both cases I normalize by the median average rating and split the data into two classes.
9.3.4 Model

For this analysis I test the performance of an SVM model. A Radial Basis Function (RBF) kernel was used. During validation the penalty parameter, $C$, and the RBF kernel parameter, $\gamma$, were each varied from $10^k$ with $k=-3, -2, ..., 3$. The SVMs were implemented using libSVM (C.-C. Chang & Lin, 2011). The choice of parameters during the validation state was made by maximizing the geometric mean of the area under the receiver operating characteristic (ROC) and precision-recall (PR) curves (when varying the SVM decision threshold).

In order to test a model that can generalize I use a leave-one-ad-out training and testing scheme. As such, data for one ad is taken out of the dataset and the remaining data is used for validation and training (cross validation performed on the set and the median parameters selected). This process is repeated $N$ times, where $N$ is the number of ads.

9.4 Results and Discussion

9.4.1 Ad Liking Prediction

Figure 9-11 shows the receiver operating characteristic (ROC) and precision-recall (PR) curves for the model for predicting ad liking score varying the SVM decision threshold in both cases. Table 9.3 shows the area under the curve (AUC) for the receiver operating characteristic (ROC) and precision-recall (PR) curves for the ad liking score prediction models. I compare the performance using just the face features and a combination of the face and contextual features. Table 9.4 shows the confusion matrix for the SVM classifier (with the optimal decision threshold - closest point to (0,1) on the ROC curve) with face and context features for the ad liking score prediction. During the validation process the median parameters $C$ and $\gamma$ were 10 and 0.1 respectively.

It is reasonable to think that ads with different aims (i.e. comical ads that aim to amuse vs. charity cause related ads that aim to evoke sympathy) would result in a different re-
relationship between viewer’s expressed responses and ad effectiveness. As a result I also performed the same analysis for just the ads labeled as intentionally amusing by the MTurk labelers. The AUC for the ROC and PR curves are shown in Table 9.3 and confusion matrix in Table 9.4. One can see that the amusing ad models performs slightly better, with greater ROC AUC and PR AUC in three of four cases. For the amusing ad model only 18 of 75 ads are misclassified (76% accuracy).

Table 9.3: Area under the ROC and PR curves for the ad liking classifier: top) all ads (N=165), bottom) only amusing ads (N=75).

<table>
<thead>
<tr>
<th>Ads</th>
<th>Features</th>
<th>ROC AUC</th>
<th>PR AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.779</td>
<td>0.762</td>
</tr>
<tr>
<td></td>
<td>Face &amp; Context</td>
<td>0.840</td>
<td>0.828</td>
</tr>
<tr>
<td>Amusing</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.790</td>
<td>0.798</td>
</tr>
<tr>
<td></td>
<td>Face &amp; Context</td>
<td>0.850</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Table 9.4: Confusion matrices for the optimal liking classifier: top) all ads (N=165), bottom) only amusing ads (N=75). Based on threshold of point closest to (0,1) on the ROC curve.

<table>
<thead>
<tr>
<th>Ads</th>
<th>Actual +ve (High Liking)</th>
<th>Actual –ve (Low Liking)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>63</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>59</td>
</tr>
<tr>
<td>Amusing</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>31</td>
</tr>
</tbody>
</table>

Figure 9-12 shows examples of the true positives, true negatives, false positives and false negatives from the best performing classifier. The emotion profile of ads that generate high ad liking is a strong gradient and high peak in positive expressions (valence and smiles). Emotion profiles of ads that do not generate high ad liking are either low across all metrics (i.e. very few people show the expressions I detect) or feature a greater amount of negative (disgust) expressions than positive expressions (smiles). These results are congruent with previous work (Varey & Kahneman, 1992; Fredrickson & Kahneman, 1993)
Figure 9-11: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the ad liking models varying the SVM decision threshold. Black) the performance using face features, blue) the performance using face and context features combined. Unbroken lines) results for all ads, broken lines) results for only the amusing ads.

showing that peak and final emotions experienced are disproportionately weighted when people recall their feelings during the experience.

There are some cases that break these trends. For example the responses to one ad (shown in Figure 9-12(n)) showed large amounts of smiling and very low levels of disgust but the average liking score was below the median response. This ad had a liking score of 3.36 which is very close to the class boundary of 3.44 which explains why it would easily be misclassified. In other cases examples in the positive class achieve similarly low facial responses (e.g. Figure 9-12(p and v)) and this explains why they would be misclassified.

9.4.2 Purchase Intent Prediction

Figure 9-13 shows the ROC and PR curves for the purchase intent score prediction model. Again, I compare the performance using just the face features and a combination of the face and contextual features. In addition, I show results with all ads and just with the amusing
Figure 9-12: Aggregate ad response metrics that correctly and incorrectly classified by the ad likability model. True positives, true negatives, false positives and false negatives shown. Aggregate: eyebrow raise - black, smiles - green, disgust - red. High peak levels of positive expressions, high expressiveness and strong increasing trend in positive expressions predict high ad likability. Low expressiveness predicts low ad likability. Individual plot outlines indicate the product category the advertised product.

ads. Table 9.6 shows the confusion matrix for the best performing SVM classifier for the purchase intent score prediction. Only 18 of the 74 ads (accuracy = 76%, F1-score = 0.757) were misclassified when considering just the amusing ads. During the validation process the median SVM parameters C and $\gamma$ were 1.0 and 1.0 respectively.

Figure 9-14 shows examples of the true positives, true negatives, false positives and false negatives from the best performing PI model. I also plot the timing of the appearances of the brand within each ad (broken grey line). The prediction performance was lower for the PI model than for the liking model, suggesting that the relationship between
Table 9.5: Area under the ROC and PR curves for the purchase intent classifier: top) all ads (N=163), bottom) only amusing ads (N=74).

<table>
<thead>
<tr>
<th>Ads</th>
<th>Features</th>
<th>ROC AUC</th>
<th>PR AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.755</td>
<td>0.804</td>
</tr>
<tr>
<td></td>
<td>Face &amp; Context</td>
<td>0.739</td>
<td>0.741</td>
</tr>
<tr>
<td>Amusing</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.647</td>
<td>0.696</td>
</tr>
<tr>
<td></td>
<td>Face &amp; Context</td>
<td>0.781</td>
<td>0.811</td>
</tr>
</tbody>
</table>

Table 9.6: Confusion matrices for the best performing purchase intent classifier: top) all ads (N=170), bottom) only amusing ads (N=75). Based on threshold of point closest to (0,1) on the ROC curve.

<table>
<thead>
<tr>
<th>Ads</th>
<th>Actual +ve (High Liking)</th>
<th>Actual –ve (Low Liking)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Predict +ve</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Predict -ve</td>
<td>29</td>
</tr>
<tr>
<td>Amusing</td>
<td>Predict +ve</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Predict -ve</td>
<td>9</td>
</tr>
</tbody>
</table>

facial responses and changes in PI is more complex; this is to be expected. purely eliciting more and more positive expressions is not as successful at driving purchase intent as it is for driving ad liking. However, notice that for all the true positives and false negatives in Figure 9-14 the peak in aggregate smiling is preceded by a brand appearance, whereas this is not the case for any of the true negatives. These results support the work of Teixeira et al. (2014) showing that emotion elicited by ads is more effective if associated with a brand. Our results suggest that brand appearances immediately prior to the peak positive emotion is a driver for increasing purchase intent. Furthermore, Figure 9-14 (g) shows a false positive that appears to exhibit the features of a good response (i.e. a brand appears preceding the peak positive response) but I also see that the peak in disgust responses is also shortly after the brand appearance. This suggests that negative emotions may become associated with the brand and outweigh the effects of the positive responses. This is something that would not have been identified had I only considered smile responses as was the case in the analysis performed by Teixeira et al. (2014).
9.5 Discussion and Conclusions

I present the largest ever analysis of facial responses to online ads. Using an online framework and state-of-the-art facial expression analysis I capture and code 12,230 facial responses to 170 ads from four countries (France, Germany, UK, US). In total over three million frames were analyzed. This analysis would not have been possible with traditional laboratory data collection methods and manual coding of the frames of facial video.

I measured eyebrow raises, smiles, disgust and positive and negative valence expressions of the viewers on a frame-by-frame basis and mapped these to two key measures of advertising effectiveness, ad liking and changes in brand purchase intent. I note that facial responses to the ads (viewed in natural settings) were sparse. In only 17.2% of the frames was there a detectable eyebrow raise, smile, disgust or positive or negative valence expression. Almost 50% of the facial response videos had no detectable behavior. However, aggregate metrics reveal that there were detectable responses from subsets of the viewers.
Figure 9-14: Aggregate ad response metrics that were correctly and incorrectly classified by the purchase intent model. True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) shown. Brand appearances within the ads are indicated by the broken gray lines. Notice how the peak in smile activity is preceded by a brand appearance in the TPs and not in the TNs. Aggregate: eyebrow raise - black, smiles - green, disgust - red. Individual plot outlines indicate the product category the advertised product. The results seem to generalize across product categories.

to all the ads and this yields rich temporal affective information.

I built and tested a model for predicting ad liking based on the emotional responses of the viewers. The model performs accurately (ROC AUC = 0.850). A strong positive trend in expressed valence and high peak positive expressions suggest an ad will have a high liking score. This supports previous work looking at individual responses. I built and tested a model for predicting changes in purchase intent based on the automatically coded facial responses (ROC AUC = 0.781). Performance in predicting both effectiveness measures was good. In addition, one can gain insight into the structure of effective ads, such
as where the brand should appear. My results suggest that brand appearances immediately prior to the peak positive emotion is a driver for increasing purchase intent.
Chapter 10

Mars: Predicting Sales

“Half the money I spend on advertising is wasted; the trouble is I don’t know which half.”

John Wannamaker

“There is more money wasted in advertising by underspending than by overspending ... underspending in advertising is like buying a ticket halfway to Europe. You’ve spent your money but you never get there.”

Morris Hite

10.1 Aims and Motivation

It is now widely accepted that human emotions play a significant role in driving our decisions, from the type of content we watch to the products and services we buy (LeDoux, 2002; Damasio, 1994).

An objective measure of success for advertising content, whether it be for consumer products, feature films or services, is the impact the advertisement has on sales. In particular the sponsor of the ad wants to see the sales increase during the time the media was
Figure 10-1: Flow diagram of the feature extraction for the sales prediction task. Facial metrics extracted from each video and aggregated. Temporal features extracted from the aggregate metrics.

aired and for a period after the content was aired. This chapter presents the first large-scale examination of the relationship between short term sales and automatically measured affective responses.

Today, the de facto approach to predicting sales effectiveness of advertising is through traditional surveys. Respondents watch an ad and then answer a number of questions that capture aspects of the viewers’ cognitive assessment of the ad, such as awareness of the brand, persuasion of a message, and purchase intent (how likely they are to purchase the product as a result of viewing the ad). These responses are then correlated to sales data. This approach models the cognitive aspects of decision-making but misses people’s visceral emotional responses, which play an equally important role in influencing a person’s memory for a product and purchase decisions.

I show results from a large number of viewers responding to 163 ads using the automated and highly scalable framework presented in my thesis work. This analysis considers data from the Mars experiment (see Chapter 5). In this chapter I consider sales figures that have been calculated accounting for other factors (such as: promotions, seasonality of media, media cost and media weighting). The sales figures are on an ad level and not a viewer level. For seven of the 170 ads described in Chapter 9 I did not have complete data and therefore I have 163 independent samples of data. The aims in this Chapter are:

- To understand the relationship between aggregate facial responses to video ads and
Table 10.1: Number of videos tested from each product category and emotion category (categorized using MTurk labelers).

<table>
<thead>
<tr>
<th>Emotion Category</th>
<th>Petcare</th>
<th>Confec.</th>
<th>Food</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amusement</td>
<td>14</td>
<td>45</td>
<td>7</td>
<td>8</td>
<td>74</td>
</tr>
<tr>
<td>Heart-warming</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Cute</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Exciting</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Inspiring</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Sentimental</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>No Majority</td>
<td>11</td>
<td>17</td>
<td>3</td>
<td>6</td>
<td>37</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>74</td>
<td>19</td>
<td>18</td>
<td>163</td>
</tr>
</tbody>
</table>

their short-term sales success.

- To identify which facial expressions (smiles, disgust, positive or negative valence) are most strongly linked to this success and for which categories of ads (e.g. food or pet care product categories).

- To compare the prediction of sales success using facial responses and self-report feelings.

### 10.2 Data and Data Collection

#### 10.2.1 Video ads

For this sales analysis I use the same set of data as described in the previous chapter. However, five of the ads did not have sales data associated with them and two ads did not have purchase intent data. These are the seven ads that were excluded. In total 163 ads from four countries (France, Germany, UK and US) were tested. Table 10.1 shows the number of videos from each product and emotion category.

#### 10.2.2 Respondents

One hundred respondents were recruited to view each batch (set of 10 videos) of copies. Participants were recruited from the four countries (UK, US, France and Germany). Re-
cruitment was such that age groups, gender and economic status of the participants was as balanced as possible and also helped mitigate the effects of self-selection biases. In addition, in all cases at least 70% of the viewers who watched each ad were a user of the product category being advertised. The respondents were compensated with approximately $8.10 for participating in the study. It took an average of 36 minutes for participants to complete the survey.

10.2.3 Methodology

The facial video recording framework was integrated into a survey with self-report questions and memory tests in a structure shown in Figure 9-3. During the experiment the participants watched 10 ads in a randomized sequence, in order to minimize primacy, recency and carry-over effects. For details of the survey structure see Chapter 9.

10.2.4 Sales Data

Sales figures for the products advertised were collected. These measures were calculated using single source household level panels. The sales figures are calculated as the relative probability of households buying the product given exposure against probability of households buying the product given no exposure (Equation 10.1).

\[ \frac{p(\text{buying} | \text{exposure})}{p(\text{buying} | \text{no exposure})} \]  

(10.1)

The measurements for each of these ads were taken for at least two flights and known biases removed (e.g. promotions, TV viewings, media weight, seasonality of media and media cost). Measurements were also calculated relative to the market average in order to account for economic conditions.

This assumes that the behavior during the short-term time period is homogeneous. Thus, single source data collection was performed outside seasonal periods (e.g. Christmas Day and St. Valentine’s Day). In addition, data captured during promotions was filtered
out. Although this process tries to take into account many influences on sales index that are not related to the viewing response the calculation is not perfect. This noise within the labeling processing is unavoidable given the data I have available and may well contribute to inaccuracies in the prediction process.

10.3 Features

As in the previous chapter I compare the prediction performance using different combinations of features. In particular I compare the performance using self-report and facial metric features. The facial metric features are calculated in the same way as in Chapter 9.

10.3.1 Calculating Aggregate Metrics

I calculate aggregate expression metrics for each ad from the individual facial responses to that ad - Figure 9-8 (step 1). These were calculated as the mean metric intensity across all viewers to the ad (ignoring frames in which no face was detected). I compute mean tracks for the eyebrow raise, smile, disgust and valence classifiers (as in Chapter 9).

10.3.2 Extracting Summary Features

Facial Metric Features: I extracted summary features from the aggregate facial expression metrics. The summary features extracted from each summary metric were: mean, maximum, minimum and gradient. Figure 9-8 (step 2) shows how the features were extracted from an aggregate metric track. These four features for the four facial metrics classifiers led to a feature vector of length 16 for each ad.

Self-Report Features: The ad liking and purchase intent score for each ad used in the previous chapter were also used as features. The ad liking score is the mean liking score across all viewers on a five point scale. The purchase intent score is the mean delta in purchase intent report from the pre-survey to the final survey across all viewers. The brand
likability score (not used in the previous chapter) is the mean delta in brand likability report from the pre-survey to the final survey across all viewers. Each of these features results in one number per ad. Therefore the length of the self-report feature vector is three.

10.3.3 Computing Labels

I divide the ads into two categories - those with higher sales score and those with lower sales score. This gives two classes which were almost equally balanced (54% in class one and 46% in class two).

10.3.4 Model

For this analysis I test the performance of an SVM model. A Radial Basis Function (RBF) kernel was used. During validation the penalty parameter, C, and the RBF kernel parameter, $\gamma$, were each varied from $10^k$ with $k=-3,-2,...,3$. The SVMs were implemented using libSVM (C.-C. Chang & Lin, 2011). The choice of parameters during the validation state was made by maximizing the geometric mean of the area under the receiver operating characteristic (ROC) and precision-recall (PR) curves (when varying the SVM decision threshold).

In order to test a model that can generalize I use a leave-one-ad-out training and testing scheme. As such, data for one ad is taken out of the dataset and the remaining data is used for validation and training (cross validation performed on the set and the median parameters selected). This process is repeated N times, where N is the number of ads.

10.4 Results

10.4.1 Comparison of Sales Labels with Human Coding

In order to test how well the algorithm compared to the performance of human coders I used Amazon’s Mechanical Turk (MTurk) to collect viewers’ ratings of how well they felt
the ad would perform. Viewers were asked: “Do you think that this ad strongly, moderately or weakly increased the short-term sales of the product advertised?”. A minimum of three coders labeled each video.

The accuracy of the raters was 30% for the three class classification and 55% for the two class (weakly and moderately vs. strongly) classification. Table 10.2 shows the confusion matrix for the two class classification problem. The human labels were almost identical to a random guess (54%). The human labelers were most likely to pick “moderately” in response to the question - which explains the bias toward the negative class.

Table 10.2: Confusion matrix for the human coding of sales performance.

<table>
<thead>
<tr>
<th>Predict +ve</th>
<th>Actual +ve (Good Sales)</th>
<th>Actual -ve (Poor Sales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict +ve</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>Predict -ve</td>
<td>52</td>
<td>65</td>
</tr>
</tbody>
</table>

10.4.2 Facial Metrics and Self-report Prediction

In order to address the fifth hypothesis I compare the predictive performance of the self-reported responses, the facial expression metrics and a combination of the two modalities.

Figure 10-2 shows the receiver operating characteristic (ROC) and precision-recall (PR) curves for the model for predicting sales score varying the SVM decision threshold in both cases. Table 10.3 shows the area under the curve (AUC) for the receiver operating characteristic (ROC) and precision-recall (PR) curves for the sales score prediction model. I compare the performance using the face and context features and a combination of the face, context and self-report features. Table 10.4 shows the confusion matrix for the SVM classifier (with the optimal decision threshold - closest point to (0,1) on the ROC curve) with facial and self-report features for the sales score prediction. During the validation process the median parameters C and \( \gamma \) were both 0.1.
Figure 10-2: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the sales models (for non pet care ads) varying the SVM decision threshold. green) the performance using self-report features, red) the performance using face features, blue) the performance using self-report and face features combined.

Table 10.3: Area under the ROC and PR curves for the sales classifier: top) all ads (N=163), middle) only non pet care product ads (N=111), bottom) only pet care product ads (N=52).

<table>
<thead>
<tr>
<th>Ads</th>
<th>Features</th>
<th>ROC AUC</th>
<th>PR AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Self-report</td>
<td>0.517</td>
<td>0.456</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.514</td>
<td>0.513</td>
</tr>
<tr>
<td></td>
<td>Face, Self-report</td>
<td>0.518</td>
<td>0.44</td>
</tr>
<tr>
<td>Non pet care</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Self-report</td>
<td>0.644</td>
<td>0.571</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.645</td>
<td>0.618</td>
</tr>
<tr>
<td></td>
<td><strong>Face, Self-report</strong></td>
<td><strong>0.739</strong></td>
<td><strong>0.690</strong></td>
</tr>
<tr>
<td>Pet care</td>
<td>Naive</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Self-report</td>
<td>0.547</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td>Face</td>
<td>0.550</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td>Face, Self-report</td>
<td>0.569</td>
<td>0.566</td>
</tr>
</tbody>
</table>

10.5 Case Studies

In order to understand the importance of facial emotion dynamics, I will look at two case studies. Examples include a case where facial coding was a more accurate predictor of
Table 10.4: Confusion matrices for the best performing sales classifier: top) all ads (N=163), middle) only non pet care product ads (N=111), bottom) only per care product ads (N=52). Based on threshold of point closest to (0,1) on the ROC curve.

<table>
<thead>
<tr>
<th>Ads</th>
<th>Actual +ve (Good Sales)</th>
<th>Actual -ve (Poor Sales)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predict +ve</td>
<td>Predict -ve</td>
</tr>
<tr>
<td>All</td>
<td>55</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>37</td>
</tr>
<tr>
<td>Non pet care</td>
<td>38</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>38</td>
</tr>
<tr>
<td>Pet care</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>23</td>
</tr>
</tbody>
</table>

sales success and a case where the combination of self-report and facial coding was more informative than either alone.

10.5.1 Case I: Facial responses correctly predict sales (not self-report)

The first case study features an ad for a chocolate brand (confectionery product category) that was tested in France and performed badly according to the sales measure. The facial responses correctly predicted the lack of success using my model. However, the self-report responses incorrectly predicted the sales effectiveness. Figure 10-3 shows the aggregate: a) self-report and b) facial responses measured for this ad. The peak disgust measure is higher than the peak smile measure which may explain why the classifier using facial features correctly classified it as performing badly. There are positive increases in both self-reported brand likability and brand PI which may explain why the classifier using self-report features misclassified it.

10.5.2 Case II: Self-report correctly predicts sales (not facial response)

The second case study features an ad for a chocolate brand that was tested in France and performed well according to the sales measure. The self-report responses correctly predicted the success using our model. However, the facial responses incorrectly predicted the sales effectiveness. Figure 10-3 shows the aggregate: a) self-report and b) facial responses
measured for this ad. The peak disgust measure is higher than the peak smile measure which may explain why the classifier using facial features misclassified it as performing badly. There are positive increases in both brand likability and brand PI which may explain why the classifier using self-report features correctly classified it. Interestingly, when looking at the brand appearances it is notable that the peak disgust response is not near a brand appearance, whereas the peak smile response is. This example may not have been misclassified if I had taken this into account in the model. As identified in Chapter 9 the location of brand appearance relative to the facial expressions induced seems important when evaluating ad effectiveness.

### 10.6 Discussion and Conclusions

Predicting product sales is the holy grail of advertising and it is a very challenging problem - not least because isolating one effect (e.g. TV advertising) on sales is very hard. Here I present the first attempt to predict sales effectiveness of TV ads using spontaneous facial emotion responses. I show that there is indeed predictive power in affective responses. For all the products not in the pet care category, using a combination of facial and self-report features, it was possible to discriminate between poor and good sales performance for 76
out of 111 ads. This exceeded the performance when using just self-report or facial features alone. The performance is also much better than human judgements of sales effectiveness in this task. However, when also considering the pet care ads the performance is poor; it was not much better than chance. This shows that a universal model is unlikely to perform well and that the relationship for some product categories may be harder to model than others. This is also what we expected to find.

Although there is still much research to be done to understand the complex relationship between emotional responses and sales effectiveness of an ad, I believe that a crowdsourcing approach to the collection of data will help researchers understand this relationship.

The case studies described above show examples in which the aggregate self-report measures do not seem to match the expressed facial responses and that perhaps in some cases the measurement of visceral facial expressions may lead to insights that cannot be uncovered using self-report alone. In particular, increasing amusement throughout an ad was a predictor of success (as it was for ad liking and purchase intent in Chapter 9), while the presence of negative emotions was inversely correlated to sales. Temporal information about a response is hard to obtain from self-reported measures, unless a dial is used, which detracts from the viewing experience. Meanwhile, I measured rich temporal information from faces without disturbing the viewing experience.
Chapter 11

Electoral Debate: Predicting Candidate Preferences

11.1 Aims and Motivation

I have shown that viewer preferences to TV ads can be predicted from facial responses. Ad liking, brand purchase intent and short-term sales can all be predicted from facial behavior recorded in a natural setting over the Internet. In this chapter I aim to show that preferences towards political candidates can be predicted based on facial response to sections of a televised political debate. During the 2012 US presidential election campaign there were two live televised debates between the candidates President Barack Obama and Governor Mitt Romney (a third debate between the Vice-Presidential candidates was also held). I performed this experiment immediately following the second of the Presidential debates.

Political debates cover emotive issues that impact people’s lives. The policies the candidates present and the way in which they present them can have a significant bearing on their public perception and potentially on the outcome of the election. In the 2012 US Presidential election campaign the third live televised debate lasted one and a half hours and covered foreign and domestic policy.

Advertising is a huge component of political campaigns. In their 2012 US Presidential
election campaigns the Democratic and Republican parties spent almost $1,000,000,000 on TV advertising\(^1\). A number of studies have focused on responses to political advertisements in order to measure their effectiveness and evaluate the emotional responses of viewers. These studies have generally focused on self-reported emotions. Political debates cover many of the same campaign themes as political advertising and the techniques presented here could generalize to the evaluation of political ads, providing further evidence of the efficacy of automated facial coding in ad testing. Kraus (1999) provides detailed history on televised presidential debates which are viewed as a key element in modern campaigns.

The aims of this experiment were:

- To identify salient segments of the debate video content based on facial emotion responses.
- To evaluate a method for candidate preference prediction based on facial responses
- To identify significant facial behaviors that provide evidence of the predictive power

\(^1\)http://www.washingtonpost.com
Table 11.1: Demographic profile of the participants in the 611 facial responses videos collected.

<table>
<thead>
<tr>
<th>Ages (years)</th>
<th>18-25: 168 (27%), 26-35: 165 (27%), 36-45: 138 (23%), 46-55: 79 (13%), 56-65: 39 (6%), 65+: 22 (4%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male: 407 (67%), Female: 204 (33%)</td>
</tr>
<tr>
<td>Location</td>
<td>United States: 378 (62%), Outside US: 233 (38%) Australia (11), Bulgaria (1), Canada (10), China (6), Czech Rep. (5), Denmark (6), Egypt (11), Estonia (4), Germany (13), Honduras (11), Italy (8), Rep. of Korea (1), Netherlands (13), Norway (5), Sweden (3), Switzerland (5), UK (113).</td>
</tr>
<tr>
<td>Eligibility to Vote</td>
<td>Yes: 415 (68%), No: 196 (32%)</td>
</tr>
<tr>
<td>Party Affiliation</td>
<td>Democratic: 226 (37%), Republican: 51 (8%), Independent: 150 (25%), None: 184 (30%)</td>
</tr>
</tbody>
</table>

of facial responses.

11.2 Data

Respondents

Using the online framework I was able to collect over 60% of the trackable video responses (374 videos from 98 people) within one day of the live debate and over 80% (501 from 135 people) within three days. This represents a very efficient method of measuring responses - necessary when responses to the material may be time sensitive as with topical debates. No participants were compensated for taking the survey making it also cost effective.

People in over 19 countries took part with 62% in the US. The demographics are summarized in Table 11.1. Of the viewers 37% declared a Democratic party affiliation. As the survey was open to the public the participants were not limited to being in the US or to being eligible to vote in the election. For the modeling of voter preferences I disregard those participants who said they were ineligible to vote.

Methodology

All videos were recorded with a resolution of 320x240 and a frame rate of 14 fps - collected using the online framework. In order to perform the experiment for this work a web-based survey was created which required participants to answer multiple choice questions related
Figure 11-2: The number of participants who took part in the Electoral Debate study and watched at least one of the debate clips.

Figure 11-3: Questions asked before the participants watched the debate clips during the Electoral Debate study. Responses to all questions were required.

to their party affiliation, candidate preferences, familiarity with the debate and demographic profile. Following this they watched five clips from the debates and their facial responses were recorded using the framework described above. Figure 11-3 shows a screenshot of the questions asked before the viewers watched the clips. Following each clip viewers were asked the following question: “After watching this clip, which candidate do you prefer?” Viewers were required to pick either “President Barack Obama” or “Governor Mitt Romney.”
Table 11.2: Synopses and lengths of the five debate clips that viewers watched during the Electoral Debate study. The number of trackable face video responses collected is also shown on the right.

<table>
<thead>
<tr>
<th>Clip</th>
<th>Duration</th>
<th>Description</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55s</td>
<td><strong>Criticism of U.S. Navy:</strong> President Obama responds to Governor Romney’s criticism that the US Navy has fewer ships than it has since 1916.</td>
<td>154</td>
</tr>
<tr>
<td>2</td>
<td>53s</td>
<td><strong>Tour of the Middle East:</strong> Governor Romney comments on President Obama’s tour of the Middle East early in his presidency.</td>
<td>119</td>
</tr>
<tr>
<td>3</td>
<td>60s</td>
<td><strong>Bin Laden Assassination:</strong> President Obama speaks about the personal impact of killing Osama bin Laden.</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>43s</td>
<td><strong>Trade War with China:</strong> Governor Romney responds to a question about whether he would start a trade war with China.</td>
<td>111</td>
</tr>
<tr>
<td>5</td>
<td>68s</td>
<td><strong>Fate of Detroit:</strong> The candidates spar over the fate of Detroit and the auto industry.</td>
<td>107</td>
</tr>
</tbody>
</table>

Figure 11-4: Mean valence during the debate clips for those that reported a preference for Obama (blue) and Romney (red) after watching the clip. The shaded area represents the standard error range. Below the plots we show which candidate was speaking and which (or both) was on screen during the clip. The letters and dotted lines correspond to significant parts of the clips - transcripts of these parts of the clips can be found in Table 11.3.

Table 11.3: Excerpts from the debate transcript which correspond to the marked points on Figure 11-4. The color of the dots refers to the speaker: blue = Obama, red = Romney.

<table>
<thead>
<tr>
<th>Ref</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>The nature of our military’s changed. We have these things called aircraft carriers where planes land on them.</td>
</tr>
<tr>
<td>b</td>
<td>We have these ships that go underwater, nuclear submarines</td>
</tr>
<tr>
<td>c</td>
<td>One of the challenges we’ve had with Iran is that [people] felt that the administration was not as strong as it needed to be. I think they saw weakness where they had expected to find American strength.</td>
</tr>
<tr>
<td>d</td>
<td>Then when there were dissidents in the streets of Tehran, the Green Revolution, holding signs saying, is America with us, the president was silent. I think they noticed that as well.</td>
</tr>
<tr>
<td>e</td>
<td>Obama Interrupts: &quot;Governor Romney, that’s not what you said.&quot;</td>
</tr>
<tr>
<td>f</td>
<td>Under no circumstances would I do anything other than to help this industry get on its feet. And the idea that has been suggested that I would liquidate the industry, of course not. Of course not.</td>
</tr>
</tbody>
</table>
11.3 Results and Discussion

11.4 Insights from Facial Responses

In the first section of this analysis we look at the differences in aggregate responses between those that reported a preference for Obama or Romney following each clip. Figure 11-4 shows the mean valence measured for those that reported preference for Obama (blue) and Romney (red) after watching the clips. Immediately the facial responses tell us detailed information about the parts of the clips that had greatest emotional response. Both candidates appear to have “scored points” during the debates, in particular, Obama during clip one and Romney during clip five. The greatest difference in measured facial activity between the two groups (those who preferred Obama versus those who preferred Romney) occurred when Obama made a joke about why the Navy was not investing in more ships - quotes a and b in Table 11.3. Another highlight occurs when the candidates discuss the decline of the car industry in Detroit. Obama interrupts Romney (e) but it seems that Romney’s response is sufficient to cause a positive emotional response in the audience.

The next section of the analysis that we performed was to look at the moments during the debate clips that cause the greatest amount of facial behavior - as detected by our expression detection algorithms described above. Figure 11-5 shows examples in the responses to clip 1 and clip 4 during which the viewers smirked. However, in the two clips the relationship between the smirks and smiles was very different. During the first clip the smirks were followed in most cases by a symmetric AU12 or smile. However, during clip 4 the smirks were followed by more varied responses including AU1+2, AU25 and AU4 - generally suggesting a more negative valence than smiles (D. J. McDuff et al., 2010). Interesting combinations and temporal patterns of facial expressions is something that large-scale data collection of this type can help reveal. However, we must be aware that the coding used can impact whether or not we might be likely to discover new patterns - coarser coding will make this type of discovery less likely.
Figure 11-5: Top) Examples of smirks during two parts of the debate clips. Top left) Smirks that occurred during Clip 1, which were followed by smiles/positive valence. Top right) Smirks that were not followed by smiles/positive valence. Bottom) Plots of the aggregate smiles and number of smirks for debate Clip 1, and debate Clip 4. Regions in which a greater number of smirks occurred are highlighted in green.
11.5 Predicting Voter Preference

I present a method for predicting voters’ candidate preference from responses to the debate clips. A person-independent test was performed and an SVM used for classification. The modeling was performed by removing data from 10 participants from the data for a particular debate clip and then performing a leave-one-out validation on the remaining data. The classifier was then trained using the validated parameters and tested on the 10 withheld participant data samples. As the class sizes were unbalanced, we over-sampled the testing set in each case to reflect an equal distribution of Obama and Romney labels (therefore giving balanced classes). Thus, five of the 10 withheld testing data points were from the Obama class and five were from the Romney class.

A particularly interesting subset of the voting population are those that identify themselves as independent. Campaigns for the Democratic and Republican candidates often focus large amounts of effort on winning the vote of these people. We show the results for this population in particular.

11.5.1 Features

In order to train and test the preference classifier we extracted features from the raw metrics. The features were calculated from the valence metric as this is a combination of both positive and negative expressions. We divided each response into 10 evenly spaced temporal bins and took the maximum valence peak within each bin, a similar method to that used in Chapter 8 to predict liking preference. As additional features we took the area under the smirk track and the smile track. This gave a final feature vector of length 12. The features extracted were normalized using a z-transform (to result in zero mean and unit standard deviation).
11.5.2 Model

Support Vector Machines (SVM) are a static approach to classification and therefore do not explicitly model temporal dynamics. However, as described above, the features extracted from the data captured the dynamics. A Radial Basis Function (RBF) kernel was used. The model parameters were validated using a leave-one-out procedure on the training set. During validation the penalty parameter, $C$, and the RBF kernel parameter, $\gamma$, were each varied from $10^k$ with $k=-2, 1, ..., 2$. The SVM’s were implemented using libSVM (C.-C. Chang & Lin, 2011). The median parameters selected during the model validation were $\gamma = 0.1$ and $C = 10$.

11.5.3 Voter Preferences

Table 11.4 (top) shows the confusion matrix for the prediction of candidate preferences from the facial valence information. The accuracy of the model was 73.8%. The model represents a significant improvement over a naive model for which the accuracy would be 50%. The Precision-recall and ROC curves (green lines) are shown in Figure 11-6, the area under the ROC curve was 0.818.

11.5.4 Independent Voter Preferences

Table 11.4 (bottom) shows the confusion matrix for the prediction of candidate preferences from the facial valence information of self-reported independent voters. The Precision-recall and ROC curves (blue lines) are shown in Figure 11-6, the area under the ROC curve was 0.733. The prediction accuracy is still strong (75%) and the area under the precision-recall and ROC curves are slightly greater, at 0.841 and 0.801 respectively. This is encouraging as the preferences of independent voters are perhaps of more interest than those with a declared Democrat or Republican affiliation.
Figure 11-6: Precision-recall (top) and ROC curves (bottom) for the voter preference prediction task. Green) Results for all eligible voters. Blue) Results for all eligible voters with no or an independent party affiliation.
Table 11.4: Top) Confusion matrix for prediction of voter preference across all the eligible voters. Bottom) Confusion matrix for prediction of voter preference across the eligible voters with no or an independent party affiliation. Threshold determined as the case closest to the ROC (0,1) point.

<table>
<thead>
<tr>
<th>Pred. Outcome</th>
<th>Self-report Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obama</td>
</tr>
<tr>
<td>O'</td>
<td>39.9%</td>
</tr>
<tr>
<td>R'</td>
<td>10.1%</td>
</tr>
<tr>
<td>Total</td>
<td>50%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pred. Outcome</th>
<th>Self-report Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obama</td>
</tr>
<tr>
<td>O'</td>
<td>40.5%</td>
</tr>
<tr>
<td>R'</td>
<td>9.5%</td>
</tr>
<tr>
<td>Total</td>
<td>50%</td>
</tr>
</tbody>
</table>

### 11.6 Results/Conclusions

I used an Internet based framework to collect 611 naturalistic and spontaneous facial responses to five video clips from the third presidential debate during the 2012 American presidential election campaign. Using the Internet as a channel allows for data to be collected very efficiently, which is especially important when considering live topical events such as election debates. We were able to collect over 60% of these video responses (374 videos) within 1 day of the live debate and over 80% (501) within three days.

The results show that different responses can be detected from viewers with different political preferences and that similar expressions at significant moments can have very different meanings. In particular, I used algorithms to automatically identify asymmetric AU12 (smirks) at particular moments, the interpretation of which is highly dependent on
the subsequent expressions. A model for predicting candidate preference based on automatically measured facial responses is presented and we show that its accuracy is significantly greater than a naive model. The ROC AUC for the model is 0.779 for voters with a democrat or republican affiliation and 0.801 for just the independent voters with neither a democrat nor a republican affiliation.

This work presents much potential for the application of emotional measurement to political ads, debates and other media. The other work presented in this thesis has shown the power of facial responses to advertising in predicting effectiveness. This could be extended to specifically political advertising and the nuances associated with that domain. Future work will also address the temporal sequences in facial behavior in more depth.
Chapter 12

Complementary Aspects of Facial Expressions and Physiology

The majority of my analysis has focused on analysis of facial responses to media content. However, there are other modalities which contain useful information about an affective response. In particular, physiological changes can help us capture responses even if the viewer does not express anything on his or her face. As I showed in Chapter 6, during my PhD I have collaborated in developing new algorithms that allow us to capture heart rate, respiration and HRV from video sequences of the human face.

12.1 Aims and Motivation

The aim of this study was to collect multimodal data of responses to media clips (physiological and facial responses) in a more controlled setting. This enables the validation of the remote sensor measurements of physiology and also the investigation of other modalities. I conducted a smaller lab-based study in order to investigate these questions.

The main aims:

- Validate remote BVP measurement from the face in a media viewing setting.
Table 12.1: Order, descriptions and durations of the media clips viewed. Ads were placed in between the emotion eliciting clips that were viewed in the Multimodal studies.

<table>
<thead>
<tr>
<th>Clip</th>
<th>Emotion</th>
<th>Description</th>
<th>Dur. (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neutral</td>
<td>Beach scene</td>
<td>240</td>
</tr>
<tr>
<td>2</td>
<td>Sadness</td>
<td>A couple’s dog is put down</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advertisement</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Disgust</td>
<td>A graphic description of taxidermy</td>
<td>311</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advertisement</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>Amusement</td>
<td>Disney cartoon clip</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>Neutral</td>
<td>Scenery</td>
<td>140</td>
</tr>
<tr>
<td>6</td>
<td>Amusement</td>
<td>Stand-up comedy</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advertisement</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>Anger</td>
<td>Glenn Beck clip</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advertisement</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>Fear</td>
<td>Horror movie trailer</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advertisement</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Total</strong></td>
<td><strong>1345</strong></td>
</tr>
</tbody>
</table>

Experiment 2

<table>
<thead>
<tr>
<th>Clip</th>
<th>Emotion</th>
<th>Description</th>
<th>Dur. (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neutral</td>
<td>Beach Scene</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>Sadness</td>
<td>Lion King</td>
<td>131</td>
</tr>
<tr>
<td>3</td>
<td>Disgust</td>
<td>Pink Flamingos</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>Amusement</td>
<td>When Harry Met Sally</td>
<td>154</td>
</tr>
<tr>
<td>5</td>
<td>Amusement</td>
<td>Robin Williams Live</td>
<td>81</td>
</tr>
<tr>
<td>6</td>
<td>Surprise</td>
<td>Sea of Love</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>Surprise</td>
<td>Capricorn One</td>
<td>47</td>
</tr>
<tr>
<td>8</td>
<td>Fear</td>
<td>Silence of the Lambs</td>
<td>204</td>
</tr>
<tr>
<td>9</td>
<td>Anger</td>
<td>My Bodyguard</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Total</strong></td>
<td><strong>950</strong></td>
</tr>
</tbody>
</table>

- Investigate relationships between different modalities (BVP, EDA, respiration, face).

12.2 Data

12.2.1 Experiment One

Respondents

10 (5 females) participants watched eight emotion eliciting media clips. For two of the participants the video recordings were corrupted and only the contact physiological measures were used. They were compensated $10 for taking part in the study.
Methodology

Each viewer watched eight media clips and five advertisements. The running order of the content is shown in Table 12.1. The average length of the clips was 149 s (st.dev 86.9) and the average length of the advertisements was 30.2 s. The clips were chosen as they were content that might be shown on TV (clips from a drama show, documentary, Disney film, etc.) but also because they elicited specific emotions. There were two neutral and amusement clips, sad, disgust, anger and fear clips. All participants watched all the clips. In total the experiment took 22.5 minutes.

12.2.2 Experiment Two

Respondents

7 (4 females) participants were recruited to watch eight emotion eliciting movie clips. They were compensated $10 for taking part in the study.

Methodology

The movie clips were previously identified as strong elicitors of the six basic emotions (Gross & Levenson, 1995): Robin Williams Live (amusement, 81s), When Harry Met Sally (amusement, 154s), My Bodyguard (anger, 246 s), Pink Flamingos (disgust, 55s), Silence of the Lambs (fear, 204), Capricorn One (surprise, 47s), Sea of Love (surprise, 12s), The Lion King (sadness, 131s). Kassam (Kassam, 2010) also validated the emotion induced by each clip. The viewers were asked to watch the movie clips (11 to 204 seconds) and were simultaneously filmed. The average length of the advertisements was 30.2 s. All participants watched all the clips, in total the experiment took 23.8 minutes. The participants watched the emotion eliciting clips in a random order and a 1 minute neutral beach clip appeared between each.

In both experiments all the viewers wore headphones in order to minimize the likelihood
of any distractions. The viewers’ responses were recorded with a Microsoft LifeCam at 30 frames per second and 840 x 526 resolution. The recordings were cropped to be synchronized with the stimuli and the physiological recordings.

As we are interested in comparing contact sensor measurements with remote sensor measurements in this work, people were recruited to participate in a lab. A diagram of the experimental setup is shown in Figure 12-1. Blood volume pulse, respiration rate and electrodermal activity measurements were collected using a pulse oximeter, chest strap and finger tip electrodes respectively.

**Contact Measurements:**

A FlexComp Pro (Infiniti, Inc.) was used for the contact physiological measures. The measurements were collected simultaneously as the viewer watched the media. All the contact measurements were recorded at a sampling rate of 32Hz. Examples of segments of the signals can be seen in Figure 12-2.

*Heart Rate:* The blood volume pulse (BVP) was used to measure heart rate. This was measured from the index finger of one hand. The heart rate was calculated from the BVP signal using a moving 30s window. For each window the power spectrum of the BVP signal was calculated and the frequency of the maximum peak taken as the heart beat frequency.

*Respiration Rate:* A chest strap was used to measure respiration rate. This was attached around the chest and under the arms of the viewers. The respiration rate was calculated from the chest strap measurements using a moving 60s window. For each window the Lomb periodogram was calculated and the frequency of the maximum peak taken as the respiration frequency.

*Electrodermal Activity:* Electrodes attached to the third and fourth fingers of both hands were used to measure EDA. In order to get a strong signal we applied conductive gel to the
electrodes prior to each session.

**Remote Measurements:**

*Heart Rate:* The remote measurement of the BVP was performed using the method described in Chapter 6. The heart rate was calculated from the BVP over a moving 30s window.

*Facial Expressions:* These were coded on a frame-by-frame basis using the CERT toolbox. This provides probability estimates of the presence of each action based on the image features in the frame. The face is detected using the OpenCV face tracker and 10 facial features identified. These features are used to register the face and warp it to a canonical model. A filter bank of 72 Gabor filters extracted from the segmented facial region of the image forms the features vector. Separate support vector machine classifiers (SVM) are trained for each AU. Full details of the implementation can be found in (Littlewort et al., 2011).

### 12.3 Results

#### 12.3.1 Remote Measurements of the BVP vs. Contact Measurements

Here we compare the heart rate measurements calculated from the camera and those calculated from the contact sensor. Using the approach described in Chapter 6 we extract the BVP signal from the video images. We compare the HR measurements calculated using the video BVP to the HR measurements calculated using the contact BVP. For this analysis the HR is calculated as the dominant frequency within the FFT of the BVP signal. I use windows of 30s with no overlap.

I use the data from the second study for this validation. For one of the subjects the contact BVP sensor measurements were highly noisy - possibly due to motion of the fingers during the experiment and thus we do not use this data for the comparison. Figure 12-3
Figure 12-1: Arrangement of apparatus for multimodal data collection. EDA, BVP and respiration were measured with contact sensors. EDA was measured from both hands. A video of the viewer was recorded simultaneously.

Figure 12-2: Examples of the physiological data collected during the multimodal experiment. Screenshots of the content and frames of the response are shown above. Notice the sharp pinching of the BVP amplitude as the EDA increases at 19 mins and again 40 seconds later. This characteristic was observed frequently - see Figure 12-4 for more examples.
shows a Bland-Altman plot comparing the contact and remote measurements of heart rate for the remaining six subjects in the second study. The intervals on the plot show 95% confidence intervals at an error of 8.8 bpm which compares well with the data from (M. Poh et al., 2010) despite this being an uncontrolled setting. As shown in Chapter 6 a five-band camera could be used instead of a webcam and is likely to deliver more accurate results.

### 12.3.2 Arousal and BVP Amplitude

Previous work has shown examples of similarities between the EDA responses and the amplitude of the BVP (Healey, 2000). Figure 12-4 shows the BVP (measured from the contact sensor), EDA and facial responses of the participants. The BVP upper and lower
envelopes are highlighted. The instantaneous amplitude of the BVP is calculated as the vertical distance between the upper and lower envelopes.

Pinching of the BVP seems strongly related to large changes in EDA. In almost all cases a significant rise in EDA is associated with a pinching of the BVP. However, the reverse relationship is not as strong. Notice in particular the prolonged pinching of the BVP that occurs across all of the viewers during the anger inducing clip. This is one of the most consistent features across the data.

12.3.3 Facial Expressions and Physiological Responses

The amusement clips do indeed seem to induce strong smiles and expressions of joy (as detected by the CERT classifiers). Similarly, we detect negative expressions during the anger and fear clips in particular. However, there are many cases (see data for Subject 6 in Figure 12-4) in which there are few detected expressions but a number of skin conductance response (SCR) events and changes in BVP amplitude. It may be that we are not detecting the facial expressions; however, when qualitatively reviewing the video response a neutral face was observed through most of the video. This suggests that the EDA and cardio-pulmonary responses may be more informative for this individual than facial responses.

12.4 Discussion

The results presented show that there are clear trends within the physiological responses of individuals and the affect of the content they are watching. These are complex relationships that are often not as easy to interpret as facial responses. However, the data show consistent responses across individuals (e.g. suppression of BVP amplitude during the anger inducing clip). There are cases (Subjects 4 and 6) in which there is not much facial behavior but there are changes in physiological responses. In real applications there may be many cases where facial behavior is not detectable and in these cases physiological signals could be extremely useful.
Figure 12-4: BVP, respiration and EDA responses of six viewers watching eight emotion eliciting clips separated by neutral clips of a beach scene. Clips are the same for all viewers but the viewing order is not. The black lines indicate the transition between clips. Note: the scales for the BVP and EDA plots are not the same.
Chapter 13

Discussion

13.1 Meta-Analysis of Results

This chapter contains a meta-analysis of the results presented in this thesis. Many of the findings in the different experiments lead to similar conclusions and it is helpful to draw generalizable conclusions. The main focus of the discussion is how remotely measured affective responses can be used to predict media effectiveness metrics. A network television broadcaster can charge up to $300,000 for a 30 second ad during primetime. Thus, improving the efficiency and accuracy of effectiveness measurements is extremely valuable.

13.1.1 Preference Prediction

I have presented a number of comparisons between facial responses and self-reported preferences in this thesis: liking (Chapter 8), desire to share again (Chapter 8) and candidate preferences (Chapter 11). The results show that indeed there is a congruency between the facial response observed and how people report their experiences.

It is indeed the case that there seem to be consistencies and informative responses within groups of individuals that generalize. In all the cases studies, occurrences of positive valence expressions (characterized by presence of smiles and absence of AU4/AU9/AU10) were predictors of increased preference toward the ad or candidate in question.
However, it should be noted that the responses to ads in particular were sparse. For instance, 50% of responses did not show detectable smiles at any point even during content that was intentionally designed to be amusing. Therefore, if one were to design an ad recommendation system based on responses this would be an important consideration. How do you classify a response when there is no visible expression? Does one give it a label? In these cases the physiological responses could be particularly useful.

### 13.1.2 Purchase Intent Prediction

Performance in predicting self-report purchase intent from aggregate facial responses was good. In addition, I show that insight can be gained into the structure of effective ads, such as where the brand should appear. Driving purchase intent is not as simple as driving ad liking - where more smiling is linked to higher reported liking. My results suggest that brand appearances immediately prior to the peak positive emotion is a driver for increasing purchase intent. This is potentially because the positive emotion then becomes associated with the brand.

The prediction of purchase intent was slightly weaker than self-reported ad liking. This is reasonable as the decision to purchase is not solely, or even perhaps mainly, influenced by the emotions they experience during an ad. Whereas an individual’s evaluation of their liking of an ad is directly related to the emotions experience while watching the ad.

### 13.1.3 Sales Prediction

I have presented the first analysis of automatically measured facial responses to ads and short-term sales measures. Although there is still much research to be done to understand the complex relationship between emotional responses and sales effectiveness of an ad; I believe that a crowdsourcing approach to the collection of data will help us. Spontaneous facial expressions are important, alongside self-report metrics, to understanding the overall impact of an ad. Furthermore, I have shown that these can be measured automatically and
do not interrupt the viewer’s experience. I have also shown that building models for specific product categories is likely to improve the accuracy of sales performance predictions.

13.1.4 Physiological Responses

Chapter 12 showed examples of dynamic physiological responses as people watched media content (including ads and movie clips). In addition physiological changes were observed in cases where no facial expressions were observed (as measured by the automated algorithms). As described above physiological responses could add valuable affective information when people are inexpressive. As I have shown in Chapter 6 we can capture the blood volume pulse remotely using a webcam or digital camera. One of the strongest relationships observed in the data was the pinching of the BVP amplitude during large increases in EDA. Prolonged pinching occurred during the anger clips.

13.2 Research Questions

I will now revisit the research questions which were presented in Chapter 5:

Q1. Can affective information (both physiology and facial expressions) can be measured remotely using low-cost cameras?

The crowdsourced facial responses to videos collected online provided rich expressions. Although not everyone displayed facial actions there were detectable expressions in many of the videos. By using information from the facial metrics captured from these videos I was able to show a strong relationship between viewers’ reported preferences and their facial responses.

Heart rate, respiration rate and heart rate variability can be captured accurately using a webcam or DSLR camera. I have shown that a novel five-band DSLR sensor can capture more accurate measurements than an RGB camera. Using HRV and respiration features it was possible to distinguish accurately between a restful state and one of cognitive stress,
suggesting that remotely measured physiology could be used to capture viewer engagement in media content.

Multimodal data collection showed that blood volume pulse amplitude and changes in EDA were closely related as viewers watched media content. More specifically, pinching of the BVP amplitude was observed during increases in EDA.

**Q2. Can the Internet can be used as an efficient method of sourcing large amounts of ecologically valid responses to online media?**

I was able to source over 20,000 videos of facial responses to online media using the online framework. Videos were captured very quickly. In the debate study described in Chapter 11 60% of the trackable video responses (374 videos from 98 people) were collected within one day of the live debate and over 80% (501 from 135 people) within three days (with no compensation for participants).

Collecting this data using traditional lab-based approaches would not have been possible. Controlled lab-based studies are extremely important in understanding the relationship between emotions and media; however, these should be supplemented with data collected in more naturalistic settings. Large-scale data collection can be used to greatly improve AU detection systems.

**Q3. Is there congruency between facial expressions and self-report? (i.e. Affective responses to media content can be used to predict self-reported feelings toward content - in particular liking.)**

The results presented in this thesis support and extend previous findings (Kassam, 2010; Micu & Plummer, 2010; D. J. McDuff et al., 2010) that facial responses during media content are linked to subjective reports of the experience. In addition the results support the conclusion that the final and highest peaks in the affective response are most closely linked to subjective reports (Baumgartner et al., 1997). This thesis also extends these works by presenting models designed and evaluated for automatically predicting media preferences.
from webcam-collected spontaneous and naturalistic facial responses.

**Q4. Does a link between facial expressions and short-term sales impact of ads exist?**

I carried out the largest study of all time examining the relationship between facial emotional responses to ads and sales effectiveness. In the study I tested 170 ads from food, confectionery and pet care product categories. The results showed a link between facial responses and sales effectiveness. However, this link was weaker than those between the facial responses and self-reported preferences. For the food and confectionery product categories the prediction performance was the best (76 of 111 ads were correctly classified). The cost of producing a high quality ad can be \( \sim \$10 \text{ million} \) and the cost of screening an ad during primetime on a US network can be \( \sim \$300,000 \). With such large amounts of money at stake during an advertising campaign any insights that can be gained about the potential success of an ad are extremely valuable.

**Q5. Are facial expressions and self-report complementary? (i.e. Facial responses combined with self-report measures will more accurately predict effectiveness of advertisements than self-report measures alone.)**

The results suggest that people’s self-report does not always match their facial response - even when it is accurately detected. Spontaneous emotional responses can be influenced by many factors (imagery, characters, brands, music) and therefore there is unlikely to be a one-to-one mapping between facial responses and self-report responses to specific questions. This result also agrees with the conclusions in (Kassam, 2010). Within my data I found that a combination of facial responses and self-report responses did lead to better prediction of sales effectiveness than either facial responses or self-report responses alone. This confirms that the two measurements are complementary.
13.3 Demo

One of the stated aims of my work was to create a practical application that demonstrates key findings from my analysis of the data collected. The following section describes the purpose, functionality and potential applications of the demo.

13.3.1 The system

The application of this demo is an intelligent advertising display. Improved targeting of advertisements presents benefits for consumers and a more effective form of marketing for advertisers. However, traditional public display advertising is static regardless of the viewers’ reactions.

The display uses facial analytics to measure the effectiveness of content and promotes ads that are predicted to be performing more effectively. The system consists of a digital display projected either from the front or rear or an LCD screen, a webcam for capturing video of the viewer and a computer for processing the data in real-time. Figure 13-1a shows a schematic of the system. Figure 13-1b shows the final system.

13.3.2 The Interface

Figure 13-2 shows examples of the billboard interface. Whilst no-one is close to the screen or facing it a scoreboard is displayed (Figure 13-2a) showing the number of viewers that have watched each ad and the percentage that were predicted to have liked the ad. Once a viewer approaches the screen and their face is detected, then an ad will begin playing after five seconds. During the ad the smile features are used to continuously update the liking prediction.

Following the ad I display a summary graph that shows the smile response over time alongside frames from the ad (Figure 13-2b). Such a screen may not be necessary in a real system but is included for demo purposes.
13.3.3 The Interactions

The following interactions are possible with our system. These are summarized in the flow diagram (Figure 13-2c).

1. If the viewer turns away or moves away from the screen for more than five seconds the interface will revert to the scoreboard screen.

2. If the preference prediction is that the viewer does not like the ad another ad will be suggested that was liked by a similar demographic (e.g. gender and age) that disliked the current ad: if (prob. liking > 0.25) then continue playing ad; else, suggest another ad end

3. If the preference engine predicts that the viewer is neutral or likes the ad, then it will continue to play. The following ad cued will be one liked by a majority of others that liked the viewed ad.

![Figure 13-1: The Affective Billboard. a) Schematic: A webcam and facial expression tool-box are used to detect the presence of a viewer and analyze their response to displayed content. A custom prediction engine evaluates viewers’ preferences based on their facial response and adapts the displayed content. b) Real-time viewer interaction with the Affective Billboard.](image)

13.3.4 Potential Applications

The demo application presented here takes the form of a digital display. The immediate application of our system is in public display advertising. The Affective Billboard could
Figure 13-2: Design of the user interface, a) scoreboard showing the performance of each ad based on the results of the liking prediction algorithm, b) results page showing an individual’s response to an ad and recommended ads, c) flow diagram of the interactions that can occur with the system.

track the response of viewers in real-time and target content on specific displays based on the performance in that area and to that audience. However, the technology could also be used for customizing content through personal entertainment systems not just ads presented on public displays. Sites like Netflix and Hulu might not have to ask for viewer’s preferences as often if they can capture information automatically using a webcam. Figure 13-3 shows examples of digital displays within stores or in public places (such as public transport stations) in which the technology could be particular effective. Digital displays and the use of augmented reality are likely to continue to make advertising more adaptive and
personalized. Therefore, incorporating affective measurement seems a natural extension.

Figure 13-3: Digital advertising displays, such as the ones above, are common and facial analytics could be used intelligently and dynamically to display the content to make them more efficient.

13.4 Social Impact and Participation

The use of the Internet to capture images or videos of viewers and the measurement of emotions to make communicating messages (e.g. in advertising) more effectively has serious social implications. I will devote discussion to these here as the work presented can shed light on the nature of these implications.

Using the Internet to Capture Videos of Viewers

The use of the Internet and viewer’s webcams to capture videos of them whilst they watch media raises privacy issues. Firstly, I would like to emphasize that ALL the data collected for this thesis involved people consenting at least twice to allow their webcam to be used. If multiple videos were recorded during a single survey they were asked for consent before the webcam was turned on each time.

Although participation does not in itself justify use of this technology and its benefit for society, it is helpful to look at the number of people who opted-in during the experiments. As an example, of approximately 20,000 people who were contacted during experiments, 45% indicated that they had a functioning webcam attached to their computer, and of these

185
53% indicated that it could be used for recording videos during the survey. These numbers show that a considerable percentage of people will engage with the technology as part of a market research survey. In addition, when asked following the survey if they felt uncomfortable 88% reported “very comfortable” to “neutral” and only 3% reported “very uncomfortable.” Across the four countries that were included in the survey, people in the United Kingdom were most likely to indicate that their webcam could be used (59%) and people in United States were least likely to do so (45%).

Further to this I have shown that it was possible to collect hundreds of videos in response to election debates without having to give viewers compensation. With a sufficiently interesting application, people may engage with this technology without having to be persuaded by financial incentives.

**Measuring Emotions to Communicate Messages More Effectively**

To many people advertising and political theatre can appear intentionally manipulative. Some would argue that subliminal effects can be very powerful (Eriksen, 1958). The intention in this thesis is not to design algorithms that allow people to create more manipulative content without concern for the interests of the viewers. Our work does not show how to do that. Rather the aims are to make an industry that spends $100,000,000,000s more efficient and to provide viewers with more relevant, entertaining and in some cases less offensive content. This technology gives viewers a non-verbal voice, allowing them to influence the content providers.

**Final Remarks**

The immediate application of technology presented is in copy testing of ads and media content. However, the technology could also be used for customizing content through personal entertainment systems not just ads presented on public displays. Sites like Netflix and Hulu might not have to ask for viewer’s preferences as often if they can capture information automatically using a webcam. The demo described above can track the response of viewers.
in real time and target content on specific displays based on the facial expressions observed from the viewers.

Personalization of content can be beneficial to consumers, whether it be in search results, display advertising or TV show recommendations. However, addressing privacy in the design of ubiquitous systems (Langheinrich, 2001), especially with technology that involves cameras in public places, is a very important and sensitive issue. Previous work has shown that people are willing to engage with systems that use cameras to measure facial behavior in public spaces and online at home or in the office (Hernandez, Hoque, Drevo, & Picard, 2012). However, there is also ongoing research into how these types of systems in public places could also have opt-out capabilities (Harvey, 2012). The choice for viewers to opt-out is important and should be available. In our system no video or identifying information about the viewer is stored. Only summary metrics about the facial response, time and video id are recorded.

I believe that this technology can be extremely powerful and have a positive impact on people’s lives. However, strict rules governing the use of cameras for these types of applications need to be in place and access to cameras on devices such as phones, tablets and computers should only be possible with explicit and, if necessary, repeated consent.
Chapter 14

Conclusions

14.1 Thesis Contributions

This thesis makes the following novel contributions to the fields of affective computing, marketing, computer vision and psychology:

- I present the first use of a cloud-based framework for collecting facial responses over the Internet. Using this approach I have collected the largest corpus of response videos to online media content (advertisements, debate clips) with ground truth success (sharing, likability, persuasion and sales). The experiments performed show that naturalistic and spontaneous facial responses can be collected efficiently over the Internet (Chapters 8-11).

- I have shown that data collected in uncontrolled settings has significantly more variability in lighting, pose and position of the viewer than datasets collected in lab-based settings (Chapter 6).

- I have shown, across a very large dataset, that many viewers are inexpressive to ads much of the time (no expressions detected in over 80% of the frames). However, with a sample size of ~50 people watching an ad we can obtain rich aggregate level information about the facial emotion response (Chapter 9). In addition, facial ex-
pression analysis can yield similar results to dial studies with an order of magnitude smaller sample size (Chapter 6).

• By measuring facial responses on first and repeat viewing of ads I have shown that responses are impacted by familiarity. This is important as in many studies as viewers are often asked to view an ad and then only report their feelings towards it on the second viewing (Chapter 7).

• I have designed and evaluated models for automatic unobtrusive evaluation of individual “liking” and “desire to view again” of online video content from webcam facial responses. Smiling in the final 25% of the ads tested was the strongest predictor of ad “liking”. This approach could be used for targeting content on platforms such as Netflix (Chapter 8).

• During the 2012 presidential election campaign I captured facial responses to candidate debate clips. I built and validated a model for predicting candidate preferences from the automatically measured facial responses (Chapter 11).

• Validated and improved remote physiological measurements. Experiments have shown that remotely measured physiological signals can be used to distinguish between a person at rest or under cognitive stress using a digital camera. I have shown that using a combination of cyan, green and orange camera color channels significantly out performs the traditional red, green and blue channel combination when recovering the blood volume pulse (Chapter 6, 12).

• I have presented the first analysis of the relationship between automatically measured affective responses to advertisements and short term sales increases. By modeling the relationship between facial responses, self-report responses and sales I have shown: 1) that temporal facial responses are linked to sales success, and 2) that self-report responses and facial responses combined improve the model further (Chapter 10).
I have implemented a system that incorporates the findings of the main experiments into a fully automated classification of facial responses to a story/advertisement. The predicted label will be the effect of the story in changing likability and purchase intent (Chapter 13).

14.2 Other Relevant Contributions

During my Ph.D. program I worked on several other projects in the field of affective computing, data visualization and crowdsourcing which are beyond the scope of this thesis. The contributions that these projects made are described briefly.

14.2.1 Exploiting Structure for Facial Action Recognition

I worked on a novel framework for facial action unit recognition. One of the key observations behind this work is sparsity: the fact that out of a possible 45 (and more) facial action units and gross behaviors, only very few are active at any given moment of time. Second, the method builds upon the fact that there is a strong statistical co-occurrence structure of the facial action units. The work presents a novel Bayesian graphical model that encodes these two natural aspects of facial action units. We derive an efficient inference scheme and show how such sparsity and co-occurrence can be automatically learned from data. Experiments on standard benchmark data show improved performance in facial AU recognition from videos and images. More details can be found in:


14.2.2 Acume

I designed an open-source tool that enables navigation of and interaction with dynamic face and gesture data across large groups of people, making it easy to see when multi-
ple facial actions co-occur, and how these patterns compare and cluster across groups of participants (D. McDuff, El Kaliouby, Kassam, & Picard, 2011). For more information see:


14.2.3 AM-FED Dataset

I helped to compile and release the Affectiva-MIT Facial Expression Dataset (AM-FED) which is the largest dataset of AU coded facial videos recording "in-the-wild." These videos are examples of naturalistic and spontaneous responses to online media (McDuff et al., 2013). For more information see:


14.2.4 AffectAura

I developed an emotional prosthetic, called AffectAura, that allows users to reflect on their emotional states over long periods of time (D. McDuff, Karlson, Kapoor, Roseway, & Czerwinski, 2012a, 2012b). I helped design a multimodal sensor set-up for continuous logging of audio, visual, physiological and contextual data, a classification scheme for predicting user affective state and an interface for user reflection. The system continuously predicts a user’s valence, arousal and engagement, and correlates this with information on events,
communications and data interactions. We evaluated the interface through a user study consisting of six users and over 240 hours of data, and demonstrated the utility of such a reflection tool. For more information see:


Figure 14-1: A screenshot of Acume, an open-source toolkit for exploring and visualizing behavioral data on multiple scales.

### 14.2.5 CardioCam and the Affect Mirror

I collaborated on the development of CardioCam a remote method for measuring vital signs (heart rate, respiration rate, heart rate variability). This led to the development of an novel interface for remote measurement, and reflection on, physiological signals (M.-Z. Poh, McDuff, & Picard, 2011). The affect mirror is a new approach to pervasive health monitoring. The gesturally controlled digital mirror interface provides a seamless way to
Figure 14-2: A screenshot of the AffectAura emotional memory visualization tool.

display and interact health information (Hernandez, McDuff, Fletcher, & Picard, 2013). There are two modes shown in Figure 14-3:

**Vital Signs**

The vital signs mode uses a camera mounted behind the two way mirror to capture images of the user. The face is segmented and the heart rate calculated using the remote method described in Section 6. The calculated heart rate is then displayed back to the user using a digital display mounted behind the two way screen.

**Life-Logging**

The life-logging mode can display photographs and time aligned EDA data on the mirror. The user can navigate through the data by moving their hand from right to left in front of the mirror. For more information see:
Figure 14-3: The Affect Mirror. Top left) In vital signs mode measuring and displaying the user’s heart rate, top right) in life-logging mode with images and EDA measurements, bottom) a diagram of the composition of the affect mirror.


14.3 Future Work

14.3.1 Realtime Prediction of Ad Effectiveness

This has a number of implications for the future of advertising and suggests that there is value in making the measurement of facial responses a routine part of copy testing. A real-time system that implements the algorithms shown in this thesis could be used for:

1. Ad optimization and real-time tracking: We can do live prediction of the performance of ads from the current emotional responses of viewers. This could impact the decision of when, or whether, to air the ad, as well as media spend.

2. From animatics to finished film: This technology could be employed powerfully in the creation or editing of advertising content in order to improve the likelihood of success of ads, even at an early mock-up stage.

14.3.2 Affective Responses with Large Data

This thesis is one of the first presentations of large scale affective responses. The data can help us ask and answer many new questions about the nature of individual differences, cross cultural differences, the influence of social factors, effects of familiarity and context. Online video consumption is only likely to increase in coming years as people consume more TV on mobile devices and not on regular TV sets. Data collected using this medium could continue to drive research forward. In the coming years the field of affective computing could benefit hugely from other analysis of large scale data. I believe that this is one of the main research challenges to come.
14.3.3 Affective Preference Systems

The results presented have demonstrated that predicting preferences based on non-contact measurements of facial expressions and physiology is possible, even in relatively unconstrained conditions. Many people benefit from systems that tailor content (articles, TV shows, movies, music, etc.) to their preferences. The technology and models presented in this thesis make it possible to have a completely automated content preference prediction system for a site such as Netflix - as long as viewers are comfortable with their webcam being on for some of the time. Content recommendation has thus far mainly relied on previous clicking and viewing activity and has not captured the rich affective information that is available.
Appendix A

Classifier Training Data

Public datasets help accelerate the progress of research by providing researchers with a benchmark resource. During my thesis work I helped to compile and release the Affectiva-MIT Facial Expression Dataset (AM-FED) which is the largest dataset of AU coded facial videos recording “in-the-wild.” Here I characterize the reliability of the AM-FED data which is similar to the automated facial coding classifier training data.

A.1 FACS Coding

Each of the videos were independently labeled, frame-by-frame, by at least three FACS trained coders chosen from a pool of 16 coders (labeling stage). All 16 coders had undergone FACS training and three were FACS certified. The labels were subsequently labeled by another independent FACS trained individual (QA stage) and discrepancies within the coding reviewed (relabeling stage). For labeling we used a web-based, distributed video labeling system (ViDL) which is specifically designed for labeling affective data (Eckhardt & Picard, 2009). A version of ViDL developed by [name to be adding in final version] was used for the labeling task. Figure 6-4 shows a screenshot of the ViDL interface. The labelers were working independently for the labeling. The coders labeled for presence (binary labels) of AU2, AU4, AU5, AU9, AU12 (unilateral and bilateral), AU14 (unilateral
and bilateral), AU15, AU17, AU18 and AU26. Smiles are labeled and are distinct from the labels for AU12 as AU12 may occur in an expression that would not necessarily be given the label of smile (e.g. a grimace). The expressiveness label describes the presence of any non-neutral facial expression. The trackerfail label indicates a frame in which the automated Nevenvision facial feature tracker (licensed from Google, Inc.), for which the detected points are provided with the dataset, were not accurately tracking the correct locations on the face. This gives a total of 168,359 FACS coded frames. Definitions of the labels and the number of frames in which they were labeled present by a majority of the labelers are shown in Table A.1. Although AU9 and AU15 were coded for, there were only one or two examples identified by a majority of the coders. Therefore we do not evaluate the reliability of AU9 and AU15. In the smile and action unit classification section of this paper, we assume a label is present if over 50% of the labelers agree it is present and assume that a label is not present if 100% of the labelers agree it is not present. We do not use the frames that do not satisfy these criteria for the classification task.

A.2 Reliability of Labels

A minimum of three coders labeled each frame of the data and agreement between the coders was not necessarily 100%. The labels provided in the archive give the breakdown of all the labelers judgements. I present the reliability of the FACS coding in Figure A-1. The reliability of the labels was calculated on a subset of the labeled data in order to evaluate it. The details of the results are reported in McDuff et al. (2013).
Table A.1: Definitions of the labels for the dataset and the number of frames and videos in which each label was present (agreed by majority of labelers). Positive examples of each of the labels are shown in Figure 6-5

<table>
<thead>
<tr>
<th>Label</th>
<th>Definition</th>
<th>Frames Present</th>
<th>Videos Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Gender of the viewer</td>
<td></td>
<td>242</td>
</tr>
<tr>
<td>AU2</td>
<td>Outer eyebrow raise</td>
<td>2,587</td>
<td>50</td>
</tr>
<tr>
<td>AU4</td>
<td>Brow lowerer</td>
<td>2,274</td>
<td>22</td>
</tr>
<tr>
<td>AU5</td>
<td>Upper lid raiser</td>
<td>991</td>
<td>11</td>
</tr>
<tr>
<td>AU9</td>
<td>Nose wrinkle</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>AU10</td>
<td>Upper lip raiser</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>AU14</td>
<td>Symmetrical dimpler</td>
<td>1,161</td>
<td>27</td>
</tr>
<tr>
<td>AU15</td>
<td>Lip corner depressor</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AU17</td>
<td>Chin raiser</td>
<td>1,500</td>
<td>27</td>
</tr>
<tr>
<td>AU18</td>
<td>Lip pucker</td>
<td>89</td>
<td>7</td>
</tr>
<tr>
<td>AU26</td>
<td>Jaw drop</td>
<td>476</td>
<td>6</td>
</tr>
<tr>
<td>AU57</td>
<td>Head is forward</td>
<td>253</td>
<td>22</td>
</tr>
<tr>
<td>AU58</td>
<td>Head is backward</td>
<td>336</td>
<td>37</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>Non-neutral face (may contain AUs that are not labeled)</td>
<td>68,028</td>
<td>208</td>
</tr>
<tr>
<td>Smile</td>
<td>Smile (distinct from AU12)</td>
<td>37,623</td>
<td>180</td>
</tr>
<tr>
<td>Trackerfail</td>
<td>Frames in which the track failed to accurately find the correct points on the face</td>
<td>18,060</td>
<td>76</td>
</tr>
<tr>
<td>Unilateral left AU12</td>
<td>Left asymmetric AU12</td>
<td>467</td>
<td>6</td>
</tr>
<tr>
<td>Unilateral right AU12</td>
<td>Right asymmetric AU12</td>
<td>2,330</td>
<td>14</td>
</tr>
<tr>
<td>Unilateral left AU14</td>
<td>Left asymmetric dimpler</td>
<td>226</td>
<td>8</td>
</tr>
<tr>
<td>Unilateral right AU14</td>
<td>Right asymmetric dimpler</td>
<td>105</td>
<td>4</td>
</tr>
<tr>
<td>Negative AU12</td>
<td>AU12 and AU4 together - distinct from AU12 in smile</td>
<td>62</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure A-1: Bar graph showing the mean in the Spearman-Brown reliability for each of the labels
Appendix B

Survey Details

B.1 Mars

Screening Questions

Q. Do you have a webcam that is functioning correctly and that is attached to the computer you are using?
A1. Yes, I have a functioning webcam attached to my computer.
A2. No, I do not have a functioning webcam to use for this study. As a result I understand I will be dropped from the survey.

Q. Do you have a working webcam and agree to have your reactions recorded?
A1. Yes, permit Affdex to anonymously process my facial expressions captured via my webcam as I watch video ads. I realize my facial image may be used by Affectiva, Affectiva’s partner and MIT, exclusively, for analysis purposes in a manner that safeguards my personal privacy.
A2. No, I’d rather not have my face recorded at this time, or I do not have a working webcam. I realize as a consequence that I will be dropped from this survey.

Q. Do you own a cat and/or dog?
Figure B-1: Instructions...

A1. I own a cat(s)
A2. I own a dog(s)
A3. I own a cat(s) and dog(s)
A4. I own neither a cat nor a dog

Q. Do you eat rice?
A1. Yes
A2. No

Q. Do you eat pasta/rice sauces?
A1. Yes
A2. No

Please read the following instructions carefully. **In order to receive payment you MUST follow these instructions:**
Pre-survey

Q. How LIKABLE do you find each of the following brands? Please rate all brands

<table>
<thead>
<tr>
<th>Very dislikable</th>
<th>Neutral</th>
<th>Very likable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

Q. Next time you are buying [product category] how likely are you TO PURCHASE products from each of these brands?

<table>
<thead>
<tr>
<th>Not likely to purchase</th>
<th>Neutral</th>
<th>Likely to purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

Survey

Q. If you watched this ad on a website such as YouTube how likely would you be to SHARE it with someone else?

<table>
<thead>
<tr>
<th>Very unlikely</th>
<th>Neutral</th>
<th>Very likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

Q. How likely would you be to MENTION this ad to someone else?

<table>
<thead>
<tr>
<th>Very unlikely</th>
<th>Neutral</th>
<th>Very likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

Q. How much did you LIKE the AD that you just watched?

<table>
<thead>
<tr>
<th>Not at all</th>
<th>Neutral</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>

Q. Had you seen the ad you just watched before this study?

1. No, never   2. Once or twice   3. More than twice

Q. How LIKABLE do you find each of the following brands? Please rate all brands

<table>
<thead>
<tr>
<th>Very dislikable</th>
<th>Neutral</th>
<th>Very likable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>5.</td>
<td></td>
</tr>
</tbody>
</table>
Q. Next time you are buying [product category] how likely are you TO PURCHASE products from each of these brands?

<table>
<thead>
<tr>
<th>Not likely to purchase</th>
<th>Neutral</th>
<th>Likely to purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.</td>
<td>3. 4. 5.</td>
</tr>
</tbody>
</table>

Brand and Ad Recognition

Q. Enter the brand name that corresponds to the ad being described. If you don’t know the name of the brand leave the field blank.

Q. Enter the brand name represented in the image. If you don’t know the name of the brand leave the field blank.

Q. How often do you PURCHASE the following products?

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Quite Often</th>
<th>Very Often</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snack Foods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chewing gum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pet Foods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pasta/Rice Sauces</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instant Rice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q. Which of the following [product category] would you consider for your NEXT PURCHASE OCCASION? Click all that apply.

Comfort and Behavior Self-report

Q. How COMFORTABLE did you feel during the study?

<table>
<thead>
<tr>
<th></th>
<th>Very uncomfortable</th>
<th>Neutral</th>
<th>Very comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3. 4. 5.</td>
</tr>
</tbody>
</table>

Q. Did you behave differently than you would have if you were watching these ads NOT as part of a study?

Q. Your responses are very helpful to us and the data can be used to understand more about human emotions. Are you happy for your face video to be shared with other researchers (outside Affectiva and MIT) to help improve understanding of facial expressions? (No other identifying information will be shared.)

A1. Yes, I would like to.

A2. No, I would prefer not.
## Appendix C

### FACS Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU1</td>
<td>Inner Eyebrow Raise</td>
</tr>
<tr>
<td>AU2</td>
<td>Outer Eyebrow Raise</td>
</tr>
<tr>
<td>AU4</td>
<td>Eyebrow Lowerer</td>
</tr>
<tr>
<td>AU5</td>
<td>Upper Lid Raiser</td>
</tr>
<tr>
<td>AU6</td>
<td>Cheek Raiser</td>
</tr>
<tr>
<td>AU7</td>
<td>Lid Tightener</td>
</tr>
<tr>
<td>AU9</td>
<td>Nose Wrinkler</td>
</tr>
<tr>
<td>AU10</td>
<td>Upper Lip Raiser</td>
</tr>
<tr>
<td>AU11</td>
<td>Nasolabial Deepener</td>
</tr>
<tr>
<td>AU12</td>
<td>Lip Corner Puller</td>
</tr>
<tr>
<td>AU14</td>
<td>Dimpler</td>
</tr>
<tr>
<td>AU15</td>
<td>Lip Corner Depressor</td>
</tr>
<tr>
<td>AU16</td>
<td>Lower Lip Depressor</td>
</tr>
<tr>
<td>AU17</td>
<td>Chin Raiser</td>
</tr>
<tr>
<td>AU18</td>
<td>Lip Puckerer</td>
</tr>
<tr>
<td>AU20</td>
<td>Lip Stretcher</td>
</tr>
<tr>
<td>AU22</td>
<td>Lip Funneler</td>
</tr>
<tr>
<td>AU23</td>
<td>Lip Tightener</td>
</tr>
<tr>
<td>AU24</td>
<td>Lips Pressor</td>
</tr>
<tr>
<td>AU25</td>
<td>Lips Part</td>
</tr>
<tr>
<td>AU26</td>
<td>Jaw Drop</td>
</tr>
<tr>
<td>AU27</td>
<td>Mouth Stretch</td>
</tr>
</tbody>
</table>
Glossary

AAM  Active Appearance Model
ANS  Autonomic Nervous System
AU   Action Unit
AUC  Area Under Curve
BVP  Blood Volume Pulse
CK+  Cohn-Kanade+ Dataset
CERT Computer Expression Recognition Toolbox
CLM  Constrained Local Model
CRF  Conditional Random Field
EDA  Electrodermal Activity
EEG  Electroencephalography
EMFACS Emotional Facial Action Coding System
EMG  Electromyography
FACS  Facial Action Coding System
FIR  Finite Impulse Response
fMRI  Functional Magnetic Resonance Imaging
GSR  Galvanic Skin Response
HCRF  Hidden Conditional Random Field
HOG  Histogram of Oriented Gradients
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>Heart Rate</td>
</tr>
<tr>
<td>HRV</td>
<td>Heart Rate Variability</td>
</tr>
<tr>
<td>IBI</td>
<td>Interbeat Interval</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>LBP</td>
<td>Local Binary Patterns</td>
</tr>
<tr>
<td>LDCRF</td>
<td>Latent Dynamic Conditional Random Field</td>
</tr>
<tr>
<td>MEG</td>
<td>Magnetoencephalography</td>
</tr>
<tr>
<td>NB</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
</tr>
<tr>
<td>PAD</td>
<td>Pleasure, Arousal and Dominance</td>
</tr>
<tr>
<td>PAM</td>
<td>Parameterized Appearance Models</td>
</tr>
<tr>
<td>PANAS</td>
<td>Positive and Negative Affect Schedule</td>
</tr>
<tr>
<td>PI</td>
<td>Purchase Intent</td>
</tr>
<tr>
<td>PNS</td>
<td>Parasympathetic Nervous System</td>
</tr>
<tr>
<td>PPG</td>
<td>Photoplethysmography</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>ROI</td>
<td>Region Of Interest</td>
</tr>
<tr>
<td>RR</td>
<td>Respiration Rate</td>
</tr>
<tr>
<td>RSA</td>
<td>Respiratory Sinus Arrhythmia</td>
</tr>
<tr>
<td>SAM</td>
<td>Self Assessment Manikin</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale-Invariant Feature Transformation</td>
</tr>
<tr>
<td>SNS</td>
<td>Sympathetic Nervous System</td>
</tr>
<tr>
<td>SR</td>
<td>Self-Report</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>SVR</td>
<td>Support Vector Regression</td>
</tr>
</tbody>
</table>
References


Ambadar, Z., Cohn, J., & Reed, L. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior, 33*(1), 17–34.


Green, M. (2004). Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse Processes, 38*(2), 247–266.

Green, M., Strange, J., & Brock, T. (2002). *Narrative impact: Social and cognitive foun-


unbc-mcmaster shoulder pain expression archive database. In IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (AFGR), 2011 (pp. 57–64).


Sénéchal, T., Turcot, J., & el Kaliouby, R. (2013). Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience. In *10th IEEE In-


