Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles

Mohammed E. Hoque, Daniel J. McDuff, and Rosalind W. Picard, *Member, IEEE*

Abstract—We create two experimental situations to elicit two affective states: frustration, and delight. In the first experiment, participants were asked to recall situations while expressing either delight or frustration, while the second experiment tried to elicit these states naturally through a frustrating experience and through a delightful video. There were two significant differences in the nature of the acted vs. natural occurences of expressions. First, the acted instances were much easier for the computer to classify. Second, in 90% of the acted cases, participants did not smile when frustrated, whereas in 90% of the natural cases, participants smied during the frustrating interaction, despite self-reporting significant frustration with the experience. As a follow up study, we develop an automated system to distinguish between naturally occurring spontaneous smiles under frustrating and delightful stimuli by exploring their temporal patterns given video of both. We extracted local and global features related to human smile dynamics. Next, we evaluated and compared two variants of Support Vector Machine (SVM), Hidden Markov Models (HMM), and Hidden-state Conditional Random Fields (HCRF) for binary classification. While human classification of the smile videos under frustrating stimuli was below chance, an accuracy of 92% distinguishing smiles under frustrating and delighted stimuli was obtained using a dynamic SVM classifier.

Index Terms— expressions classification, temporal patterns, natural dataset, natural vs. acted data, smile while frustrated.

1 Introduction

Automating the process of recognizing human facial expressions during natural interactions is a difficult computer vision and machine learning problem. Most of the previous exploratory studies have attempted to classify so-called “basic emotions” (anger, disgust, fear, happiness, sadness, and surprise) from images and videos ([2], [3] as reported in [1]). Basic emotion facial expressions are widely believed to be universally expressed, and their dynamics are typically much stronger than in spontaneous day-to-day facial expressions, which make them a natural place to start training expression recognition systems. Also, given that the majority of the available affective datasets contain basic emotions, it is desired to work on them towards developing a common benchmark. Through the use and analysis of basic emotions, there has been a trend to correlate certain Facial Action Coding Units (FACS) with affective states. In this work, we demonstrate that correlating certain FACS with affective states may contain surprising challenges while working with spontaneous affective data.

One of the major challenges in affect recognition is collecting datasets, which could be difficult, time consuming and expensive to construct. In the past, there have been efforts to collect spontaneous sequences of basic and natural emotions while the participants were acting, reacting or interacting. A few examples of such datasets include RU-FACS [4], SAL [5], Spaghetti [5], SEMAINE [6], Mind-Reading [7] and MMI [8]. Figure 1 demonstrates a graphical representation of each dataset in terms of whether it is acted vs. spontaneous and whether it contains basic vs. beyond basic emotions. Ideally, we would like to use a dataset that contains spontaneous natural emotion for affect analysis, and includes more than basic emotions, as shown in Figure 1.

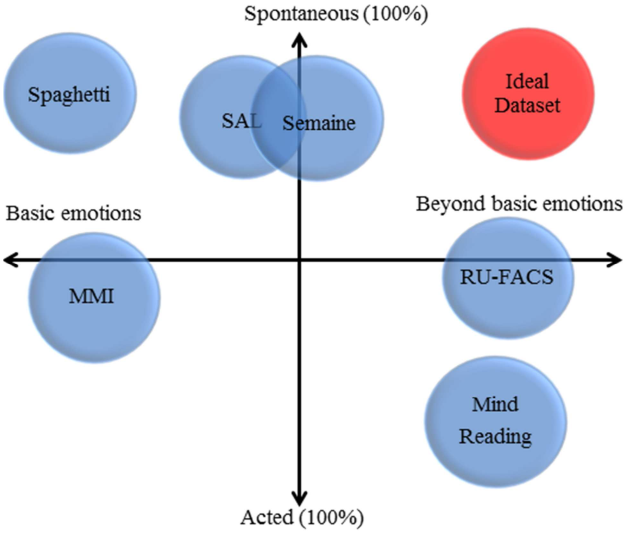


Figure 1: Comparison of existing datasets in terms of spontaneous vs. acted and basic vs. beyond basic. An ideal dataset would be spontaneous and contain a complete set of expressions.

MMI is a publicly available dataset where 87% of the data are acted for basic expressions, whereas the remain-

- *M. E. Hoque is with the MIT Media Lab, Cambridge, MA 02139. E-mail: mehoque@media.mit.edu.*
- *D. J. McDuff with the MIT Media Lab, Cambridge, MA 02139. E-mail: djmcduff@mit.edu.*
- *R. W. Picard is with the MIT Media Lab, Cambridge, MA 02139 and also with Affectiva, Inc, Waltham, MA 02452. E-mail: picard@media.mit.edu.*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

ing 13% are based on spontaneous basic expressions. Given the distribution of spontaneity vs. acted in MMI dataset, we position MMI dataset in Figure 1 more towards acted, than spontaneous.

In the RU-FACS database, participants were given a choice to lie about an opinion and receive $50 in return if they successfully convinced the interviewer. Otherwise, they would have to fill out a boring and time-consuming questionnaire. Therefore, participants were more inclined to lie, eliciting stronger emotions. Since the participants had to act to hide their true position, one could argue that the RU-FACS dataset is not fully spontaneous. Also, the RU-FACS dataset is not publicly available at this time. SAL and SEMAINE are publicly available datasets where participants are worked through a range of emotional states through an interface. The interface is controlled by an operator who acts out one of four basic emotions (happy, sad, angry, and neutral). The SEMAINE dataset contains 578 labeled annotations, and 26% of them are "basic", 37% of them are "epistemic", 31% of them are "interaction process analysis" and the rest are instances of "validity". In the Spaghetti dataset, participants were asked to insert their hand inside a box that contained a warm bowl of Spaghetti. Since the participants didn't know what was inside the box, they reacted strongly with disgust, surprise, fear or happiness. The Spaghetti dataset only contains 3 participants with a total of 1 minute and 35 seconds long data, but it is highly spontaneous. The "SAL" data consists of audio-visual recordings of human-computer conversations. The conversations are elicited through an interface called "Sensitive Artificial Listener". The interface contains four characters with four different personalities – Poppy (happy), Obadiah (sad), Spike (angry), and Prudence (pragmatic). Each character has a set of responses that match their personalities. It is hypothesized that as the participants interact with Poppy/Obadiah/Spike/Prudence, the participants get drawn into the affect that those characters display. The Mind reading dataset contains examples of more complex mental states, e.g., concentrating, thinking, confused, interested, agreement, and disagreement, and over a dozen others, but it has professional actors acting all the states.

In this paper, we make the argument that while working with basic emotions has helped promote progress in expression recognition, it is also important to push the boundary of working with spontaneous naturalistic data congruent with realistic tasks. For example, tools and techniques derived to correlate FACS with basic emotions may work well with acted or other limited forms of data; however, the same techniques may not generalize well when applied to more challenging natural data. To further strengthen our hypothesis, let us provide an example. People diagnosed with Autism Spectrum Disorder (ASD) often have difficulty recognizing emotions [31][32], specially in natural contexts. Through therapy, they are taught to look for certain features to determine the occurrence of a particular emotion. Let's say according to their therapy, they were told that lip corner puller (AU 12) and cheek raiser (AU 6) would signal the emotion "delight". According to this rule, a person with ASD would label all the images in Figure 2 as "delight". But in reality, half of the images in Figure 2 were from participants who were in frustrating situations and self-reported to be strongly frustrated. If this rule were applied in real life, say to declare one's boss as "delighted" when she or he was actually frustrated, then this could jeopardize not only a single interaction, but potentially also the person's job. Better understanding is needed of spontaneous expressions and where expressions like smiles with AU 6 + AU 12 truly occur. To further stimulate the rest of the content of this paper, the readers are requested to look at Figure 2 and guess the images where the participants were frustrated and delighted. Answers are provided at the "Acknowledgements" section of this paper.

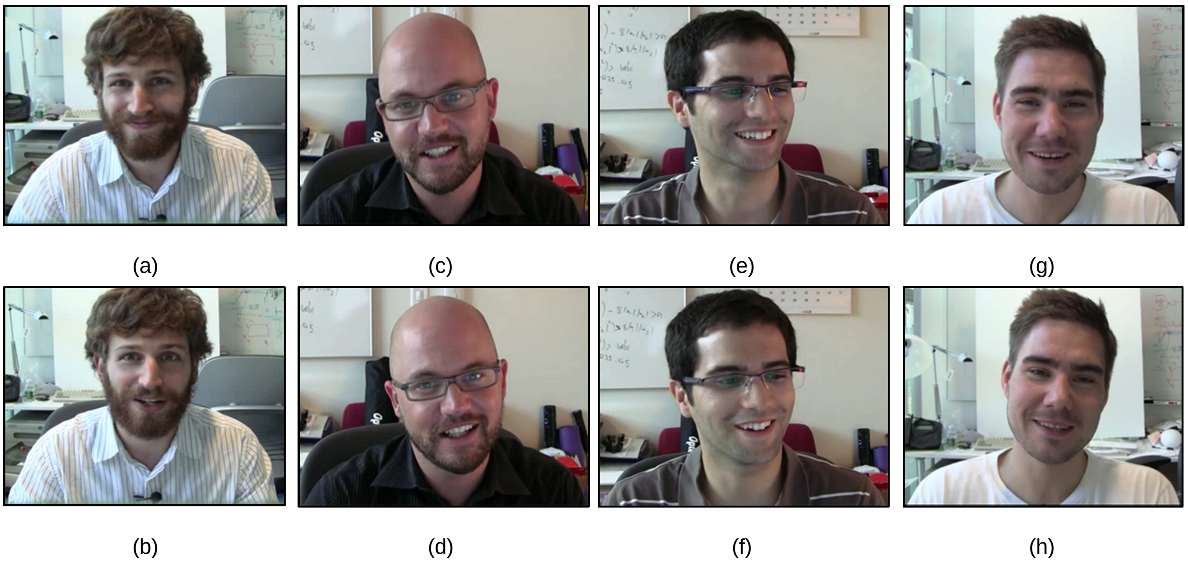


Figure 2. Four participants, each smiling while being in either a (i) frustrated or (ii) delighted state. Can you tell which smile is which state? Answers are provided in the Acknowledgements section. The images are printed with the written consent from the participants.

Contrary to popular belief [20], a lot of resarchers have argued that facial expressions serve as a mode of communication that may not necessarily reflect our emotional experience [21][22][23][24][26][27][28]. To further motivate this notion, let us provide a few scenarios:

a) Let's assume that one of your close colleagues just cracked a joke. However, you were not amused by it. What would you do? Politely smile? Or hold a neutral face?
b) Assume that you meet the same colleague in a memorial service for one of your relatives. Your colleague cracks a joke again on a different topic and you found it to be hilarious. Would you break out in laughter or just hold a neutral face given that you are in a memorial service?
c) Assume that you are interacting with your boss who happens to be monologuing without realizing it. You notice that you are running late for your next meeting, but your boss is still speaking, which adds to your frustration. What would you do in this context? Would you show the prototypical signs of frustration in your face to indicate that you are running late for your next meeting? Or would you rather provide subtle cues (e.g., look at your watch, appear busy) to indicate that you are interested to wrap up the conversation?

It is possible that in many social contexts, we encode information through our facial expressions which could be different from our experienced emotion. Do we do the same when we interact with computers? With today's technology, computers are oblivious to our mental states and we don't have any social pressure to hide our emotional experience when we interact with computers. However, it has been shown [19] that our 'interactions with computers, television, and new media *are fundamentally social and natural*, just like interactions in real life'. Therefore, it would be interesting to test the hypothesis of whether we remain socially aware with our facial expressions when we interact with tangible objects like computers. This is more likely to be possible if our experimental setup can capture spontaneous natural expressions.

In this study, we have set up two different experiments (acted-recalled and naturally-elicited) [30] to better understand two particular mental states such as frustration and delight, especially in context of human computer interaction. We observe that many participants smile, in contrary to showing protypical signs of frustration, under natural frustration. However, quite the opposite phenomenon was observed when the participants recalled a frustrating experience. This makes us wonder whether we implictely remain socially aware even when we interact with computers. It was interesting to see a lot of people smiling under frustration. Is there a difference when people smile under frustration as opposed to being genuinely delighted? How do the classifiers perform on recognizing mental states such as frustration and delight when acted, as well as when naturally elicited? What can we infer from the results about data collected through natural means as opposed to asking people to act? How do humans perform in correctly labeling smiles elicited under frustrated and delighted stimuli? Can we develop automated systems to distinguish between frustrated smiles and delighted smiles that perform better or as well as their human counterpart? This paper attempts to answer all these questions through a series of studies.

The remaining part of the paper is organized as follows: Section 2 describes the acted-data experiment. Section 3 describes the elicited-data experiment. Section 4 (analysis 1) reflects on recognition algorithms to distinguish among mental states such as frustration, delight and neutral when acted and elicited, and provides a general discussion on performance analysis and deeper insights on the problem. Section 5 (analysis 2) investigates the difference between frustrated smiles and delighted smiles and proposes an algorithm to distinguish between the two that performs better than its human counterpart.

2 Experiment 1: Acted data

The experiment took place in a well-lit empty room where participants were expected to interact with a computer program. The participants interacted with the computer program which consisted of a 2d image of an avatar (Figure 3). During the interaction, the avatar would ask a sequence of questions. The questions would appear in form of text on the interface (Figure 3). The participants would wear a headset and speak directly to the avatar to answer the questions. Additionally, there was a video camera to capture the face of the participant. The exact interaction between the avatar and the participant was as below:

Avatar: Hi There! I am Sam. I hope to be a real avatar someday. But today, I am just a 2d image who would like to interact with you. (Pause for 15 seconds)

Avatar: I hope you have signed the participant agreement form. If yes, please say your participant number. Otherwise, just state your name. (Avatar waits for the participant to speak and finish)

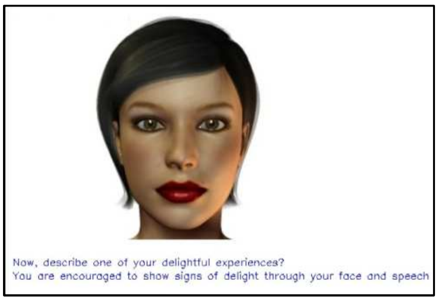
Avatar: Please briefly say a few sentences about why you are interested in this study? (Avatar waits for the participant to speak and finish)

Avatar: Now describe one of your most frustrating experiences. You are encouraged to show signs of frustration through your face and speech. (Avatar waits for the participant to speak and finish)

Avatar: Now describe one of your most delightful experiences. You are encouraged to show signs of delight through your face and speech. (Avatar waits for the participant to speak and finish).

2.1 Participants

The "Acted Data Experiment" consisted of 15 participants – 10 male and 5 female. All of them were employees at a major corporation and their age ranged from 25-40. From 15 participants, we gathered 45 clips of frustration, delight and neutral expressions (3 clips from each participant). The average duration per clip for delight and frustration was over 20 seconds, whereas the average duration for neutral was around 10 seconds. Participants wore Logitech ClearChat Comfort USB Headset to communicate with the avatar. The frontal face of the participant) was recorded using a Logitech 2 MP Portable Webcam C905. Logitech webcam software was used to connect the webcam with the PC providing 30 frames per second.



Now, describe one of your delightful experiences?
You are encouraged to show signs of delight through your face and speech

Figure 3. 2d image of the computer program used in the “Acted data experiment”

3 Experiment 2: Elicited data

For this study, 27 new participants were recruited. The participants not part of “Acted data experiment” and were blind to the hypothesis. The participants were told that they would have to evaluate the usability of a web form, and provide suggestions for improvement, if necessary. After the participant entered the room, the participant was told that s/he would have to fill out a web form. They were also instructed that based on how the task progressed; the participant may or may not be asked to speak to the camera to provide feedback on the form. The form contained 10 biographical questions (details in Table 1), including a field for date and current time without instructions on the format. The participants were instructed not to leave the experiment room until they nagivate to the confirmation screen of the form (screen 16 of Table 1). The exact sequence of interactions between the form and the participant is provided in Table 1.

Table 1. The Sequence of Screens for the natural experiment. The same sequence was maintained for all the participants [30]

Screen	Purpose	Message
1	Welcome screen	Click here to move on with this study.
2	Greetings to welcome the participant	Hi there! I hope you are doing well. Please click here to move forward with this experiment.
3	Elicit a neutral expression (Neutral)	Can you look at the camera and say a few sentences about why you are participating in this study? Please click here when done.
4	Elicit a neutral expression (Neutral)	Thank for your kind participation in this study. Before we move on, there is one more thing. Can you again look at the camera and say a few sentences about your regular activities in this department? Please click here when done.
5	Biographical form	Before you move on with this study, fill out the form below. 94.5% of the previous participants in this study were able to do this in less than 2 minutes.
6	ERROR	Error: You either did not enter the date or entered it in wrong format (correct format is: Month/Day/Year, Hour: Minute, AM/PM)
7	Biographical form	Before you move on with this study, fill out the form below. 94.5% of the previous participants in this study were able to do this in less than 2 minutes.
8	ERROR	Error: Your "About Me" section did not contain the minimum of 500 characters.
9	Biographical form	Before you move on with this study, fill out the form below. 94.5% of the previous participants in this study were able to do this in less than 2 minutes.
10	Confirmation	Your form has been submitted. Since you took a few trials to submit this form, please solve the following CAPTCHA to move forward.
11	ERROR	ERROR: Wrong values entered. Please solve this CAPTCHA to move forward.
12	ERROR	ERROR: Wrong values entered. Please solve this CAPTCHA to move forward.
13	Feedback (Frustration)	Since you are one of those participants who could not finish the form within 2 minutes, we want your feedback. Look at the camera and say a few things about why you could not finish the form within 2 minutes, unlike most of the participants.
14	Prepare for the next phase	Wonderful!! Thanks for your honest feedback. For the next phase of the experiment, you will be asked to share an experience from your past that you think is funny and delightful. To help you get started, I am sharing a click from youtube which hopefully will put you in the right mood. When ready, click here to move to the next screen and share the experience.
15	Share an experience (delight)	Now please look at the camera and share a funny experience from your past.
16	Thank you	Thank you! Your study has been completed!

All the text messages in Table 1 were converted into .wav files. As the participants navigated from one screen to another, the interface would read the text message out loud. The texts were converted into .wav files using ATT’s publicly available text to speech engine with a female American accented voice. Initially, the participants are asked two questions (screens 3 and 4 of Table 1), one after another. The purpose of those questions was to elicit expressions that were more likely to be neutral. The reason we opted for two consecutive questions is because during the pilot study we noticed that a lot of participants felt awkward looking at the camera for the first time. As a result, they either laughed out of embarrassment or provided a very brief answer, when asked, “Why are you participating in this study?” Adding a follow up question in the next screen helped them to loosen up, which resulted in a more neutral answer for the second question. We have seen this “first expression” effect dominate expressed emotions regardless of which emotion the stimuli were designed to elicit, and we encourage scientists to consider this when designing emotion elicitation experiments.

The biographical forms (screens 5, 7, 9 in Table 1) contained a timer that started counting the elapsed time. We intentionally put the timer in the middle of the screen in large font. Right mouse click and CTRL keys of the keyboard were disabled to prevent participants from copying content from one screen to another. The claim that 94.5% of the previous participants were able to finish this study in less than 2 minutes was a made up number to put more pressure on the participants. After three attempts to submit the form, the participants eventually reach screen 10 when, they are asked to solve a CAPTCHA to move forward. We used Google images (images.google.com) to select a few nearly impossible CAPATCHAs for this study. Therefore, regardless of whatever the participants typed, the interface kept on prompting error message asking participants to solve another CAPTCHA. After 3 trails, the participants would reach screen 13, where the interface would prompt them to provide feedback on what they had done wrong and why they were unable to finish the form in less than 2 minutes unlike most participants. In this phase of the study, we expected the participants to be somewhat frustrated and demonstrate signs of frustrations either through their face, speech or both.

In screen 14, participants begin the second phase of the study. In this phase, participants were given time to relax a bit and think of a funny experience that they would have to share momentarily. To help them transition into a relaxed state of mind, the interface shows them a funny YouTube video of a baby laughing uncontrollably. This particular video has more than 11 million views since 2006 and can be viewed through this link http://tinyurl.com/tac-affective. This video was picked because we felt that laughing is contagious and it may help to distract the participants from their frustrating experience of filling out the form. At the end of the experiment, majority of the participants mentioned even though they had watched the video before, they still found it funny and exhilarating. After the end of the interaction with the web form, we set up a post de-briefing session asking the participant to self-report how frustrated and delighted they were in a scale of 1-10, while they were filling out the form and watching the funny video. The entire interaction was recorded using a Canon 3.89 MP VIXIA HF M300 Camcorder and an Azden WMS-PRO Wireless Microphone. The Canon VIXIA HF M300 captured video in 30 frames per second.

The recorded footage was split into two different categories: 1) “Feedback” (contains both audio and video) 2) “Interaction” (contains only video, but no audio). Feedback dataset consisted of facial expressions and speech data of participants as they directly spoke to the camera with their feedback regarding the form and sharing a funny experience (e.g., screens 4, 13, and 15 of Table 1). Interaction dataset consisted of clips of participants when they were either filling out the form or watching the YouTube video (e.g., screens 5, 7, 9 and 14 of Table 1).

3.1 Participants and dataset

There were a total of 27 graduate students who participated in this study. Five of them were female and 22 male. All of them were blind to the hypothesis of this study. In post-experimental de-briefing, three participants informed us that they were able to figure out that the forms were intentionally designed to be buggy to provoke frustration from them. Since they were able to determine the objective of the study, we eliminated their data, resulting in 24 clips of frustration for the “feedback” dataset. Four of our participants were unable to remember a funny experience from their past during the experiment. Two of the participants told us in the de-briefing that they were so frustrated filling out the form that they were reluctant to share a delightful experience to the camera. As a result, from 27 participants, we ended up having 21 clips of delight for the “feedback” dataset. For neutral expressions, we only considered expressions from screen 4, as indicated in Table 1, and ignored the expressions elicited in screen 3. Therefore, we had 27 instances of neutral expressions for the “feedback” dataset. The average length of each clip in the “feedback” dataset for frustration and delight was a little over 30 seconds, and for neutral it was around 15 seconds.

4 Analysis 1: Acted vs. elicited Feedback

In Analyis 1, we take acted instances of frustration, delight and neutral from experiment 1 and naturally elicited instances of frustration, delight, and neutral from the “feedback” dataset of experiment 2. The goal was to allow for a comparison of recognition results on both acted and elicited data, where both facial expressions and speech were present. Below are the descriptions of the facial and speech features used for classification.

4.1 Face Analysis

We used Google’s facial feature tracker (formerly known as Nevenvision) [33] to track 22 feature points: 8 points surrounding the mouth region, 3 points for each eye, 2 points for each eye-brow, and 4 points for two nostrils, nose tip, and nose root. Points 23 and 24 shown in Figure 4 were extrapolated.

We calculated raw distances (in pixels) as well as their standard deviations across facial feature points. For example, distances and standard deviations be-

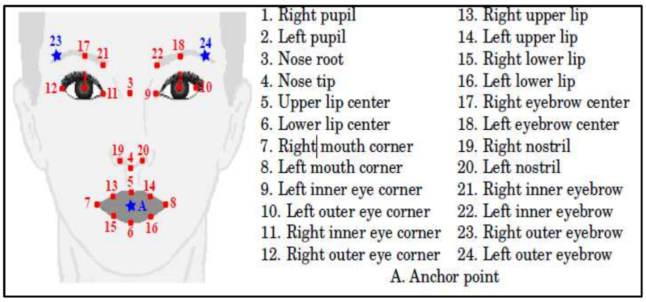


Figure 4. Extracted feature points of the face using Google Tracker

tween 12 and 11, 9 and 10, 2 and 18, 1 and 17, 11 and 21, 9 and 22, 7 and 8, 5 and 6 etc. were calculated.

The local distances among those points as well as their standard deviations were measured in every frame and used as features [9]. Additionally, we used Sophisticated Highspeed Object Recognition Engine (SHORE) [10] [17] API by Fraunhofer to detect the intensity of smiles. The SHORE API provides an agnostic score between 0-100 for smiles by analyzing the entire face including mouth widening, zygomaticus muscles, orbicularis oculi and other regions of the face in every frame. In this paper, the score is referred to as the *smile intensity*. All the features were tracked in every frame. The features extracted per clip were averaged to form a feature vector per clip. In the first experiment with acted data, while trying different techniques, averaging all the features across each clip yielded satisfactory results. Therefore, to allow for a valid comparison, in the second experiment with naturally elicited "feedback" data, we also averaged all the features across each clip. We have also investigated temporal patterns of the features per clip, which is reported in Section 5 of this paper.

4.2 Speech Analysis

We computed prosodic features related to segmental and supra-segmental information, which were believed to be correlates of emotion. Using *Praat* [4], an open source speech processing software package, we extracted features related to pitch (mean, standard deviation, maximum, minimum), perceptual loudness, pauses, rhythm and intensity, per clip.

4.3 Final Feature Set

There were 45 clips from experiment 1 and 72 clips from the "feedback" dataset from experiment 2. For each individual clip, we extracted audio and video features and concatenated them in a vector such that each clip's feature vector was as follows: V_{clip} ={ A_1,…A_n, F_1,…F_m}, where A_1,…A_n are n speech features, and F_1,…F_m are m facial features. In this study, n was equal to 15 and m was equal to 25; features are described below.

4.3 Results

We used five classifiers (BayesNet, SVM, Random-Forest, AdaBoost, and Multilayer Perceptron) from the WEKA toolbox [6], to compare the classification accuracy between the elicited face+voice data and the acted face+voice data. There were 45 instances of acted data and 72 instaces of naturally elicited feedback data. One sample was removed for each dataset and held out as the test sample. Leave-one-out K-fold cross validation (K=44 for acted, and K=71 for naturally elicited feedback) was applied. The model was trained on K-1 samples, while testing its parameters on the remaining sample, and repeating leaving a different one out each time. Through this iterative process, optimal parameters were chosen and then tested on the unseen test sample. This was repeated for all samples in the dataset yielding 45 test results for acted and 72 test results for feeback dataset for each classifier. Figure 5 shows all the classifiers performed significantly better with acted data compared to elicited data (using a leave-one-out test). The highest accuracy for acted data was 88.23% (chance for each category was 15 out of 45 or 33%) while the highest accuracy for naturally elicited feedback data was only 48.1% (chance for delight was 21 out of 72 or 29%, chance for neutral was 27 out of 72 or 38%, and chance for frustration was 24 out of 72 or 33%). The higher accuracy for the acted data held across the models with the average accuracy across all the classifiers for acted data around 82.34%, a value that dropped to 41.76% for the three-class classification of the elicited data.

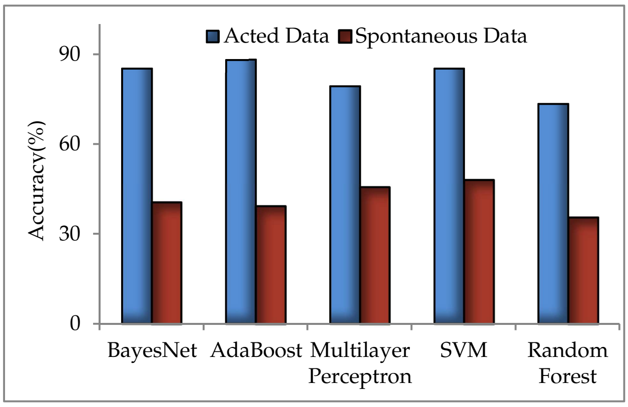
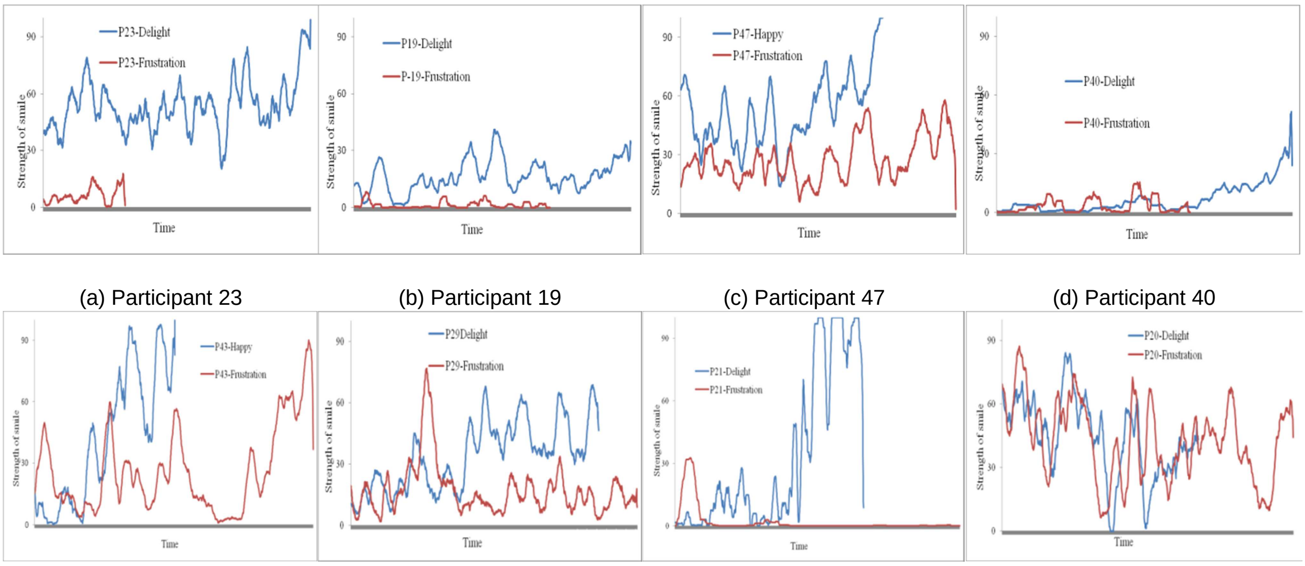


Figure 5. Classification accuracy for recognition of frustration, delight and neutral states using various classifiers with elicited and acted data. The accuracy is reported using the leave-one-out method.

Additional analysis on the feature vectors for participants from experiment 1 and experiment 2 revealed that in the acted data, close to 90% of the participants did not smile when they were encouraged to show frustration while recalling being frustrated. On the contrary, in the elicited data, close to 90% of the participants did smile when they were frustrated.



(a) Participant 23 (b) Participant 19 (c) Participant 47 (d) Participant 40

(e) Participant 43 (f) Participant 29 (g) Participant 21 (h) Participant 20

Figure 6: (a-h) Graphs of 8 participants whose patterns are representative of the rest of the participants. X axis is the time in seconds and y axis is the smile intensity/strength. (a, b, and c) are examples of participants who have distinct patterns of smile intensity when they are frustrated and delighted. (d, e, f, and g) provide examples of how the state of delight builds up in terms of smile intensity through time. f, g are examples of participants who initiated their frustration with a social smile. (h) is an example of one person who exhibited smiliar smile patterns regardless of whether delighted or frustrated.

The results shown in Figure 5 demonstrate significant differences in correctly classifying instances when the expressions are acted as opposed to being elicited. One possible explanation is that acted expressions seem to contain prototypical facial features, whereas elicited data may not contain similar facial attributes. That might be why recognizing unique features of expressions and feeding them in a classifier worked fairly well with acted data, but the performance degraded significantly when applied on elicited data. To further stimulate our findings, along with reporting the average, we also conducted an examination of subtle individual differences in terms of expressions. As part of post-analysis, we went through the analysis of each individual to get more insights on whether there are sub-categorical patterns among our participants. Specifically, we zoom into a narrow set of smiles to analyze the intrinsic dynamics of the expressions.

Analyzing each individual clip from feedback dataset of Experiment 2, for all the participants, revealed interesting findings. We noticed that almost all of the participants, despite self-reporting to be extremely frustrated, did not show the prototypical signs of frustration. In fact, in most cases, participants showed signatures of delight (e.g., smile) while providing their unpleasant feedback of filling out the form. One possible explanation is that all the participants were MIT colleagues and therefore, they refrained from being impolite given the dynamics of everyday social interaction. However, they were in a room alone during the study. Another possible reason for the greater smiling might be that the population in this study uses smiling to cope with frustration and to keep going. The participants in the second study, MIT graduate students, are all very accomplished and part of what might have helped them get where they are today is that they may have great coping abilities that perhaps use smiling to make them feel better when things go wrong. However, the participants in the first study, while none were students, were all also accomplished professional researchers at a top industrial research lab and one could argue that they would have similar excellent abilities for coping with frustration, and probably even more experience in doing so.

The occurrences of frequent smiling in elicited frustration may help explain why some people diagnosed with an Autism Spectrum Disorder (ASD) find it hard to make precise sense out of spontaneous facial expressions. If one is taught that smiles mean happiness then it would be easy to mistake smiles from a frustrated person as evidence that things are going great. Subsequently, walking up and smiling to share that person's "happiness" could be misconstrued as insensitivity or worse, and lead to numerous problems.

Almost all of our participants from experiment 2, hether frustrated or delighted, demonstrated signatures of smile (AU 12) during their interaction. This is problematic data for those who promote that smile is a strong disambiguating feature between delight and other affective states. To better understand this phenomenon, we analyzed and compared the smiling patterns of each participant when they were frustrated and delighted. Some of the interesting characterizing patterns are plotted in Figure 6. A small subset of the participants, as shown in Figure 6 (a, b, c), have clear separation of their smiles in terms of magnitude or intensity when they were frustrated and delighted. However, the pattern dissolves immediately when averaged with the rest of the participants. This phenomenon, once again, motivates the need to look at individual differences rather than reporting the average. In the context of delight, the intensity traces in Figure 6 (d, e, f, g) demonstrate that some participants gradually progressed into peaks in terms of smile. This finding is very insightful because now it supports the need to analyze the temporal dynamics of the smiles. Another interesting occurrence to observe, especially in Figure 6 (g) and Figure 6 (f), is that some people could initiate a frustrating conversation with a big social smile and then not smile much for the rest of the conversation. The prevalence of smiles when the participants were frustrated could likely be the social smile that people use to appear polite or even to cope with a bad situation by trying to "put a smile on".

Smiling under the condition of frustration or failure, even though surprising, is not a new phenomenon that we are reporting in this paper. Paul Ekman mentioned in [25] that people often smile when experiencing unpleasant emotions in the presence of others. It has been shown in earlier work [16] that pre-schoolers tended to demonstrate more true smiles in the sense of a "Duchenne" smile (Lip Corner Pull or AU 12, and cheek raised or AU 6) when they failed as opposed to when they succeeded. In this study, we observe that people seem to smile in unpleasant situations even when they interact with computers. Since it has been argued that interactions between people and computers are social and natural [19], it is possible that the participants under frustrating situations were trying to communicate their aggravation and acceptance of the situation, and trying to communicate that they were being put upon - to an imaginary interactant. This explanation does not come as a surprise since Fridlund [26] demonstrated that people who watched a pleasant videotape with friends smiled the same amount as people who watched a video with the belief that their friends were also watching the same in another room. In other words, it is possible for people to experience relevant social context even if they are alone, and enable it to guide the interaction patterns.

Is it possible for smiles under delighted and frustrated stimuli to have different temporal patterns? Messinger et al. [18] demonstrate that in context of face to face interactions between adults and infants, contrasted types of smiles (e.g., Duchenne and non-Duchenne) can happen one after another in similar situations. But they usually occur in different temporal phases of a continuous emotional process. All these previous [29] findings further strengthened our observation that it might be useful to analyze the temporal patterns of smiles as they occued under delighted and frustrated stimuli as opposed to equating the presence of smiles with delight and absence of smiles with frustration.

5 Analysis 2: Elicited Interaction data

In analysis 2, we zoom into the naturally elicited "interaction" data from experiment 2 towards development of an algorithm utilizing temporal patterns to classify them into appropriate classes.

The "interaction" dataset contained instances of smiles under frustrated and delighted stimuli, as the participants were either filling out the forms or they were watching the YouTube video. Since the participants were asked to hold their natural posture to elicit natural interaction during the experiment, in the post data-analysis stage, we noticed a lot of participants moved out of the camera frame as a result of natural movement. This resulted in 14 sequences of smiles under delighted stimuli, and 20 sequences of smiles under frustrated stimuli. The examples of smiles under frustrated stimuli were 7.45 seconds (std: 3.64) and smiles under delighted stimuli were around 13.84 seconds (std: 9.94) long on average.

The system diagram of our algorithm to distinguish between the smiles under frustrated and delighted stimuli is provided in Figure 7. Figure 7 (a) refers to the video stream which was segmented into smaller sequences based on the rule described in Figure 8.

Each sequence is then run through a feature extraction algorithm to determine the probability of the participant smiling in each frame, as explained in Section 4.1. The resultant graph per sequence looks like Figure 7 (b), where the x-axis represents time, and the y-axis represents the intensity of the smile. We split each sequence into smaller segments (30 frames, 1 second) and extract local features (section 7.1) per segment, as shown in Figure 7 (c). The feature sets are then classified to distinguish between the smiles under frustrated and delighted stimuli.

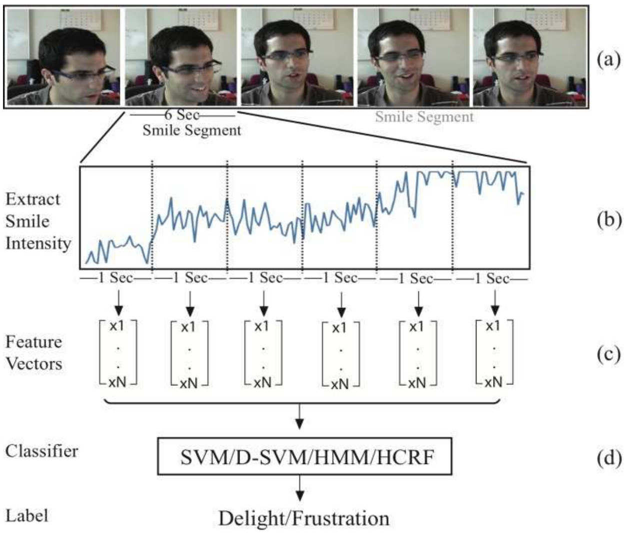


Figure 7: Methodology for smile classification. a) Segment smile sequence from clip, b) extract smile intensity from frames of smile segment, c) Form feature vectors from 1-second segments of smile intensity, d) classify input vector using SVM, HMM, or HCRF.

```
if (movement of the lip entails a smile)
        mark the beginning of clip
else if (lips retract to a neutral position)
        mark the end of a clip
```

Figure 8: Logic of clip extraction from a larger file

5.1 Features Extraction

Local features:

As mentioned in the previous section, each smile sequence gets broken into smaller segments. We measure global peak and the global gradient across the entire segment. From the smaller segments, we only extract local mean and local peak, as shown in Figure 9. Given all the extracted local and global features, we infer the following 4 features that compare each local segment with the entire segment.

1) Percentage of local frames above global mean.
2) Local mean: *mean value within the segment*
3) Gradient across the segment: *change in smile intensity per frame along with the x axis*
4) Peak comparison = $\frac{LocalPeak}{GlobalPeak}$

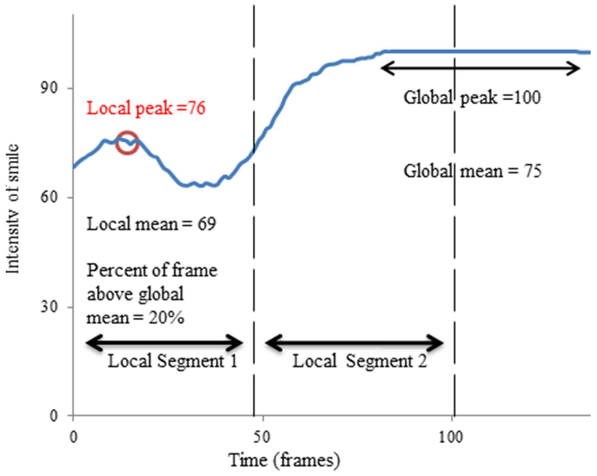


Figure 9: Description of the local and global features

5.2 Classification

The feature vectors and labels were used to train, validate, and test four models (SVM, D-SVM, HMM, HCRF) (details of data splits are in 5.3). These experiments were carried out in order to evaluate the performance of classifiers with different dependence assumptions and to compare the performance of static vs. dynamic and generative vs. discriminative classifiers. The SVMs were implemented using LIBSVM [11]. The HMMs were implemented using the HMM toolbox for MATLAB [12]. The HCRF classifiers were implemented using the HCRF toolbox [13].

Support Vector Machines

A Support Vector Machine classifier, a static discriminative approach to classification, was used as the first benchmark. Binary classifiers were trained with one class being delight and another one being frustration. A Radial Basis Function (RBF) kernel was used. During the validation the penalty parameter, C, and the RBF kernel parameter, γ, were each varied from 10^k with k = -3,...,3.

For the SVM, all the local features for the 1 second long segments were averaged over the entire sequence to form a 4-d vector. These inputs were used to train an SVM. The D-SVM was a pseudo-dynamic model in which the time samples were appended. As the video samples were of varying lengths, zeros were appended to the end to form input vectors of equal length, 128 (32 seconds *4 features/second). After subtracting the mean from the data matrix, Principal Component Analysis (PCA) was used to reduce the dimensions. The four largest principal components were used to build the model. This process was repeated for all iterations of the validation and training scheme.

Hidden Markov Model

HMMs are one of the most commonly used methods in modeling temporal data. We trained one HMM each for the delight and frustration classes. This is a dynamic generative approach to modeling the data. In testing, the class label associated with the highest likelihood HMM was assigned to the final frame of the sequence. During the validation the number of hidden states (1,...,5) was varied, with two states being the most frequently chosen as performing the best.

Hidden-state Conditional Random Field

In contrast to HMMs, Conditional Random Fields (CRFs) and CRF variants are discriminative approaches to modeling temporal data. The CRF model removes the independence assumption made in using HMMs and also avoids the label-biasing problem of Maximum Entropy Markov Models (MEMMs) [14]. The dynamics of smiles are significant in distinguishing between them [15]; as such, we hypothesized a potential benefit in removing the assumption that current features are solely dependent on the current valence label. During validation, the regularization factor (10^k with k = -3,...3) and number of hidden states (0,...,5) were varied, with a regularization factor of 10 and two states being the most frequently chosen as performing best.

5.3 Results

In this section, we present the performance of a static model (SVM), a pseudo-dynamic version of SVM, and two dynamic models (HMM, HCRF). We had 34 samples in the dataset. First, one sample was removed from the dataset and held out as the test sample. Leave-one-out K-fold cross validation (K=33,

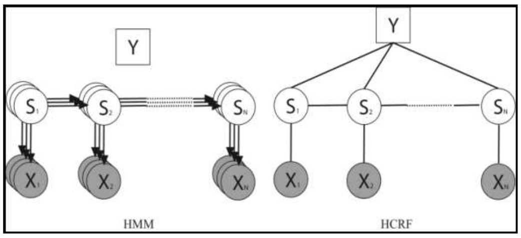


Figure 10: Structure of models. Xj represents the jth observation, Sj the jth hidden state and Y class label. The HMM requires a chain to be trained for each class label.

training the model on 32 samples, testing its parameters on the 33rd, and repeating leaving a different one out each time) was performed to find the optimum parameters. The best of these was tested on the test sample. This was repeated for all samples in the dataset (34), providing 34 test results for each model. The HMM models required no more than 30 iterations during training. The HCRF needed no more than 300 iterations in training. TABLE 2 provides a comparison of the performance of the models.

TABLE 2: PERFORMANCE STATISTICS FOR SVM, D-SVM, HMM, HCRF TOWARDS BINARY CLASSIFICATION

Model	SVM	D-SVM	HMM	HCRF	Human
Accuracy(%)	**85.30**	**92.30**	**82.40**	**79.40**	**68.98**
Sensitivity	0.89	0.91	0.83	0.82	0.81
Specificity	0.69	1.00	0.81	0.75	0.51
F-Score	0.82	0.92	0.75	0.72	0.68

In describing the following measures, we consider delight to be the positive class and frustration to be the negative class. Sensitivity measures the proportion of actual positives that are correctly identified, and specificity measures the proportion of negatives that are correctly identified. The F-score (the ratio of geometric mean and arithmetic mean of precision and recall) provides the coherence between the precision and recall values of the model and is a very good indicator of the reliability (higher F-score implies a better and more reliable model) of the predicted values.

$$F - score = 2 * \frac{precision * recall}{precision + recall}$$

In order to compare the machine performance with human performance, we asked 10 individuals, who were not part of this experiment and were oblivious of the experiment objective, to label the 34 video clips (without sound) "for frustrated smiles and for delighted smiles." The labelers were instructed to watch each clip and predict whether the participant in the clip was in happy and frustrated state of mind. During the labeling process, the average accuracy among 10 labelers towards labeling the delighted smiles was 84% (chance was 14 out of 34 or 41%), and the accuracy for the frustrated smiles was 54% (chance was 20 out of 34 or 59%), with an overall accuracy across both categories of 69%.

A detailed performance comparison between humans and the classifiers to recognize smiles under frustrated and delighted stimuli is provided in Figure 11. Figure 12 demonstrates visual sequences of smiles under delighted stimuli (Figure 12 A [I-III]) and frustrated stimuli (Figure 12 B [I-III]). Careful observation does reveal the fact there is a stronger smile signature in the frustrated smile compared to the delighted smile, which may explain why most people got it wrong. However, all of our classifiers (except for HCRF for the instance of delight) were able to classify the instances, shown in Figure 12, correctly. This demonstrates that our algorithm not only properly utilizes the signatures of smile (e.g., lip corner pull, cheek raiser etc), but also the pattern in which they appear in time.

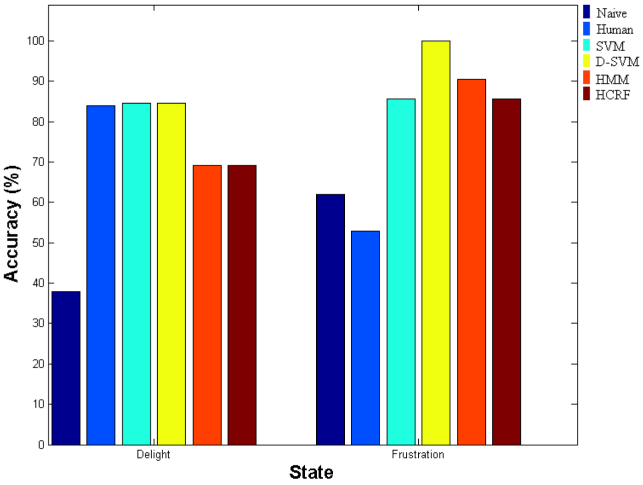


Figure 11: Bar chart comparing the performance of the human and computer labeling of 34 delighted and frustrated smile sequences.

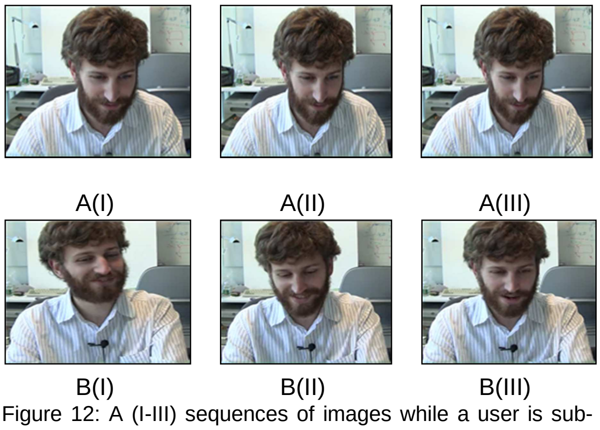


Figure 12: A (I-III) sequences of images while a user is subjected to a delightful stimuli. B (I-III) sequences of images while a user is subjected to a frustrating stimuli. Only 5 out of 10 of our human labelers were able to label the video sequence containing images A (I-III) as a delighted smile, and only 1 out of 10 of our human labelers was able to label the video sequence containing images B (I-III) as a frustrated smile. However, all of our classifiers (except for HCRF for the instance of delight) were able to classify the instances.

6 Discussion

We demonstrate in this work that it is useful to explore how the patterns of smile evolve through time, even over many seconds (smiles under frustrated stimuli averaged 7.5 sec. and smiles under delighted stimuli averaged 13.8 sec.). The average smile intensity per clip under delighted stimuli was 76.26% (std: 17.8) and for frustrated stimuli, it was 47.38% (std: 28.9). While a smile of similar intensity may occur in positive and in negative situations, its dynamic patterns may help to disambiguate the underlying state.

Smiles are not only a universal, but also a multifaceted expression. We smile to express rapport, polite disagreement, delight, favor, sarcasm and empathy. Being able to automatically recognize and differentiate the different types of smiles could fundamentally change the way we interact with machines today. Moreover, it is very important that a machine discern the difference between a frustrated customer and a delighted one and not just assume that a smile means the customer is happy.

Analysis on the feedback datasets collected from experiment 1 and experiment 2 revealed that in the acted data, close to 90% of the participants did not smile when they were frustrated. On the contrary, in the naturally elicted feedback dataset of experiment 2, close to 90% of the participants did smile when they were frustrated. We were surprised to see a lot of participants smile despite self-reporting to be frustrated. This further motivated us to develop algorithms, described as part of analysis 2, to distinguish between the spontaneous naturalistic examples of smiles under delighted and frustrated stimuli. To do this, we have automated the process of extracting temporal facial features in real time that are believed to be correlates of smiles. Among the 4 classifiers, the most robust classification was achieved using D-SVM with an accuracy of 92% and F1 score 0.92. It is a little surprising that D-SVM outperformed HMM and HCRF for our dataset, especially when HMM and HCRF have been shown to perform well modeling temporal data. However, with the addition of more classes and training samples, the best model might change. All the classification models that we have used in this paper could be implemented as part of a real-time system. Also, it is worth noting that given the limited set of smiling instances, we used leave-one-out method as opposed to k-one-out, where k>1. Leave-one-out methods could provide optimistic results on unseen data. However, with the availability of more data, the system could scale to recognize a wide variety of smiles.

In our dataset, the gradient across the entire smiling instance was the most important feature towards distinguishing between delighted smiles. While this is an important finding, it needs to be further validated across larger dataset and individuals.

Our immediate extension of this work would be to explore other facial and speech features for characterizing individual sub-categorical patterns. Continued work in this direction will hopefully help us to redesign and reshape existing one-size-fits-all expression recognition algorithms.

How good are we at differentiating the patterns of delighted smiles and frustrated smiles if we can only look visually at videos of facial expressions? Our results, as plotted in Figure 11, show human ability to identify spontaneous frustrated smiles by looking at the facial cues is below chance, whereas we perform comparatively better in identifying the spontaneous delightful smiles. Therefore, one may question if we can build systems that perform better than the human counterpart disambiguating between naturally occurring smiles under delighted and frustrated stimuli by only analyzing facial expressions. Our results demonstrate that our automated system offers comparable or stronger performance in recognizing spontaneous delighted smiles. However, the system performs significantly better by correctly labeling all the spontaneous smiles under frustrated stimuli compared to the below-chance human performance.

It is interesting to note that even though it is possible for people to smile under frustration, we usually have a pre-defined mindset of not assciating smiles with frustration. This mindset was reflected in our study through the human's inability to label the frustrated smiles correctly, as well as the human posers who posed frustration without smiles. Presumably, they would have actually smiled if they had been frustrated.

One would wonder, and rightly so, why would a machine perform better than the humans in recognizing instances of spontaneous frustrated smiles? One possible explanation is that humans usually rely on additional information that they sense using other modalities (e.g., prosody, spoken words, context) to disambiguate among different kind of smiles. Unavailability of such information could reduce a person's ability to understand emotions. Machines, however, could utilize the local intrinsic structures of the temporal patterns in the context of the entire sequence discovering unique patterns that are typically not seen by humans. Another possible explanation is that we have used a skewed number of samples (62% instances of frustrated smiles, and 38% instances of delighted smile) in our training process. Therefore, the classifier is more likely to do better in categories where it has seen more examples. However, humans have seen examples of these smiles throughout their life in everyday interactions, so this does not explain why they are not better still.

7 Conclusion

In this study, we have set up experiments to collect acted and elicited expressions of frustration and delight, and run analysis with multiple methods to automatically classify them. We observe that even using the simplest approach of averaging features over video clips, we get an average of 82.3% accuracy on all the classifiers on the acted data, a value that dropped to

41.8% for the same three-class classification using elicted data. Additionally, in 90% of the acted cases, participants did not smile when frustrated, whereas in 90% of the elicted cases, participants smiled during the frustrating interaction despite self-reporting significant frustration with the experience.

We proposed, implemented and evaluated an automated system that can correctly extract and classify sequences containing smiles under delighted and frustrated stimuli gathered from experiment 2. As part of validation, we trained a variety of static and dynamic models, both generative and discriminative. The models were evaluated with K-fold validation and testing schemes. The best classifier distinguished between the patterns of spontaneous smiles under delighted and frustrated stimuli with 92% accuracy. Moreover, individually the classifier was able to identify all the instances of smiles under frustrated stimuli correctly, compared to below chance performance (53%) of humans. Meanwhile, the performance of recognizing delighted smiles was comparable between humans and machines.

We successfully demonstrate through our work that carefully designed experiments to elicit spontaneous natural expressions can show that sometimes surprising things occur – like 90% of frustrated participants smiling. These data can then be used to develop automated systems that recognize spontaneous expressions with accuracy higher than the human counterpart. We hope that our work will motivate the field to move beyond the trend of working with "six basic emotions", move beyond teaching people that "smiles mean happy" and continue to develop methods to interpret challenging spontaneous data that contain complex patterns of expression.

Acknowledgments

Figure 1 (a), (d), (f), (h) are taken from instances of frustration; (b), (c), (e), (g) are from instances of delight. We conducted an independent survey of 12 labelers of these, and all scored at or below chance (4 out of 8, or 50%) in labeling images from Figure 2.

The authors would like to acknowledge the participants for their time helping with this study and agreeing to have their data shared with other researchers. We also acknowledge the generous support of Media Lab consortium sponsors and of Procter and Gamble for this work.

References

[1] H. Gunes and M. Pantic, Automatic, Dimensional and Continuous Emotion Recognition, *International Journal of Synthetic Emotion*, Vol. 1, No. 1, pp. 68-99., 2010.

[2] D. Keltner, and P. Ekman, "Facial expression of emotion", In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* , pp. 236-249). New York: Guilford Press, 2000.

[3] P. N. Juslin, and K. R. Scherer, "Vocal expression of affect", *In Journal Harrigan*, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research*, pp. 65-135. Oxford, UK: Oxford University Press. 2005.

[4] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, and J.R. Movellan, Automatic recognition of facial actions in spontaneous expressions. *Journal of Mutlimedia*, pp. 1–14, Oct. 2006.

[5] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488–501, Jan., 2007.

[6] G. McKeown, M.F. Valstar, R. Cowie and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions", *IEEE Int'l Conf. Multimedia & Expo (ICME'10)*, Singapore, pp. 1079-1084, July 2010.

[7] Baron-Cohen, S., et al., "*Mind Reading: The Interactive Guide to Emotions".* 2004, Jessica Kingsley Publishers: London.

[8] M. Pantic, M.F. Valstar, R. Rademaker and L. Maat, "Web-based database for facial expression analysis", *IEEE Int'l Conf. on Multimedia and Expo (ICME '05)*, Amsterdam, The Netherlands, pp. 317-321, July 2005.

[9] M. E. Hoque, R. W. Picard, "I See You (ICU): Towards Robust Recognition of Facial Expressions and Speech Prosody in Real Time", *International Conference on Computer Vision and Pattern Recognition (CVPR)*, DEMO, San Francisco, CA, 2010.

[10] C. Kueblbeck and A. Ernst, "Face detection and tracking in video sequences using the modified census transformation", *Journal on Image and Vision Computing*, Vol. 24, Issue 6, pp. 564-572, 2006, ISSN 0262-8856

[11] C. Chang and C. Lin, LIBSVM : a library for support vector machines., *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 27:1--27:27, 2011.

[12] K. Murphy, "The Bayes Net Toolbox for Matlab" *Computing Science and Statistics*, Vol. 33, 2001.

[13] L-P. Morency, Hidden-state Conditional Random Field Library.

[14] J. Lafferty, A. McCallum, F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *18th International Conf. on Machine Learning*, pp. 282–289, Morgan Kaufmann, San Francisco, CA, 2001.

[15] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behavior", *Philosophical Transactions of Royal Society B*, Vol. 364, pp. 3505-3513, December 12, 2009.

[16] K. Schneider & I. Josephs, "The expressive and communicative functions of preschool children's smiles in an achievement situation", *Journal of Nonverbal Behavior*, Vol. 15, 1991.

[17] C. Küblbeck, T. Ruf, and A. Ernst, "A Modular Framework to Detect and Analyze Faces for Audience Measurement Systems", in Proc. GI Jahrestagung, 2009, pp.3941-3953.

[18] D. Messinger, A. Fogel, K. L. Dickson, "What's in a smile?" *Developmental Psychology*, Vol. 35(3), 1999, pp. 701-708.

[19] B. Reeves and C. Nass, "The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places", Cambridge University Press: 1996.

[20] E. L. Rosenberg, and P. Ekman, "Coherence between expressive and experimental systems in emotion", *Cognition and Emotion*, Vol. 8, 1994, pp. 201-229.

[21] J. M. Carroll and J. A. Russell, "Facial expressions in Hollywood's portrayal of emotion", *Journal of Personality and Social Psychology*, Vol. 72, No. 1, 1007, pp. 164-176.

[22] A. Ortony, adn T. J. Turner, "What's basic about basic emotions?", *Psychology Review*, Vol. 74, 1990, pp. 315-341.

[23] A. J. Fridlund, "Human facial expression: An Evolutionary view", San Diego, CA: Academic Press, 1994.

[24] J. Fernández-Dols, F. Sánchez, P. Carrera and M. Ruiz-Belda, "Are Spontaneous Expressions and Emotions Linked? an Experiment Test of Coherence", *Journal of Nonvebal Behavior*, Vol. 21, No. 3, 1997, pp. 163-177.

[25] P. Ekman, W. V. Friesen, and P. Ellsworth, Emotion in the human face. New York: Pergamon Press, 1972.

[26] A. J. Fridlund, "Sociality of social smiling: Potentiation by an implicit audience", *Journal of Personality and Social Psychology*, Vol. 60, 1991, pp. 229-240.

[27] U. Hess, R. Banse and A. Kappas, "The intensity of facial expression is determined by underlying affective state and social situation", *Journal of Personality and Social Psychology*, Vol. 69, 1995, pp. 280-288.

[28] R. E. Kraut and R. E. Johnston, "Social and emotional messages of smiling: an ethological approach", *Journal of Personality and Social Psychology*, Vol. 37, 1979, pp. 1539-1553.

[29] R. El Kaliouby, P. Robinson, P. and S. Keates, "Temporal context and the recognition of emotion from facial expression", *In 10th International Conference on Human-Computer Interaction (HCI' 2003)*, 22-27 Jun 2003, Crete, Greece.

[30] M. E. Hoque, R. W. Picard, Acted vs. natural frustration and delight: Many people smile in natural frustration, 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), Santa Barbara, CA, USA, March 2011.

[31] S. Baron-Cohen, H. A. Ring, E. T. Bullmore, S. Wheelright, C. Ashwin, and S. C. R. Williams, "The Amygdala Theory of Autism", *Neuroscience and Behavioual Reviews,* Vol. 24, 2000, pp. 355–364.

[32] M. A. Howard, P. E. Cowell, J. Boucher, P. Brooks, A. Mayes, A. Farrant, N. Roberts, "Convergent Neuroanatomical and Behavioural Evidence of an Amygdala Hypothesis of Autism", *Neuroreport*, Vol. 11, pp. 2931–2935.

[33] FaceTracker. *Facial Feature Tracking SDK*. Neven Vision, 2002.

Mohammed Ehsan Hoque is PhD candidate at the Affective Computing Group of MIT Media Lab. Hoque has earned his bachelor degree in Computer Engineering from Pennsylvania State University and a master's degree in Electrical Engineering from the University of Memphis. He has previously held positions at Goldman Sachs, Walt Disney Imagineering Research & Development, and IBM T. J. Watson Research Center. Hoque's research efforts are about helping people with their communication difficulties in face-to-face interaction by developing and drawing techniques from computer vision, speech processing, machine learning and multimodal data fusion. Hoque was the recipient of 2009 IEEE Gold Humanitarian Fellowship award. Contact him at mehoque@media.mit.edu.

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

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