Acted vs. natural frustration and delight: Many people smile in natural frustration

Mohammed (Ehsan) Hoque
MIT Media Lab
Cambridge, MA, USA
mehoque@media.mit.edu

Rosalind W. Picard
MIT Media Lab
Cambridge, MA, USA
picard@media.mit.edu

***Abstract*—This work is part of research to build a system to combine facial and prosodic information to recognize commonly occurring user states such as delight and frustration. We create two experimental situations to elicit two emotional states: the first involves recalling situations while expressing either delight or frustration; the second experiment tries to elicit these states directly through a frustrating experience and through a delightful video. We find two significant differences in the nature of the acted vs. natural occurrences of expressions. First, the acted ones are much easier for the computer to recognize. Second, in 90% of the acted cases, participants did not smile when frustrated, whereas in 90% of the natural cases, participants smiled during the frustrating interaction, despite self-reporting significant frustration with the experience. This paper begins to explore the differences in the patterns of smiling that are seen under natural frustration and delight conditions, to see if there might be something measurably different about the smiles in these two cases, which could ultimately improve the performance of classifiers applied to natural expressions.**

Keywords-natural vs. acted data; smile while frustrated; machine learning;

I. INTRODUCTION

Recognizing emotion using facial expressions or prosody (the patterns of intonation and stress in a language) in speech or fusion of multiple modalities remains an active area of exploration. This area of research not only holds promise to reshape the ways we interact with machines today, but also helps us to think of innovative ways to help people with communication difficulties (e.g., people diagnosed with autism, and people with nonverbal learning disabilities). However, as we realize, expressions come in many varieties; some intense and continual, while others are subtle and momentary [1]. Therefore, developing a computational model that can capture all the intrinsic details of human emotion would require natural training data containing all the inherent details so that the model can learn from it.

Given the difficulty of collecting natural data, a majority of past research has focused on data collected through acting or posing an expression. An alternate approach is to have participants watch emotionally stimulating video clips while videotaping their facial expressions. The obvious limitation of this approach is that there would not be any speech data. Even for face data, some may argue that such a dataset does not provide a task dependent environment where context becomes an inevitable part of elicited emotional states. To simplify the classification and to establish a common benchmark, there has been a trend to use and analyze basic emotional states (neutral, happiness, sadness, surprise, fear, anger, and disgust; ([7], [8] as reported in [2]) and to correlate certain Facial Action Units (FACS) [3] with emotional states.

Our hypothesis in this study is that 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 express their difficulty in recognizing emotions in appropriate context. 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 1 as "delight". But in reality, half of the images in Figure 1 were from participants who were in frustrating situations and self-reported to be strongly frustrated. To further stimulate the rest of the content of this paper, the readers are requested to look at Figure 1 and guess the images where the participants were frustrated and delighted. Answers are provided at the "Acknowledgement" section of this paper.

How big are the differences between natural and acted expressions of frustration and delight? The work in this paper finds huge differences, especially with large numbers of smiles appearing in natural frustration but not in acted. This work presents two new ways of getting these data, and begins to look at how the smiles in both frustration and delight conditions occur and unfold over time.

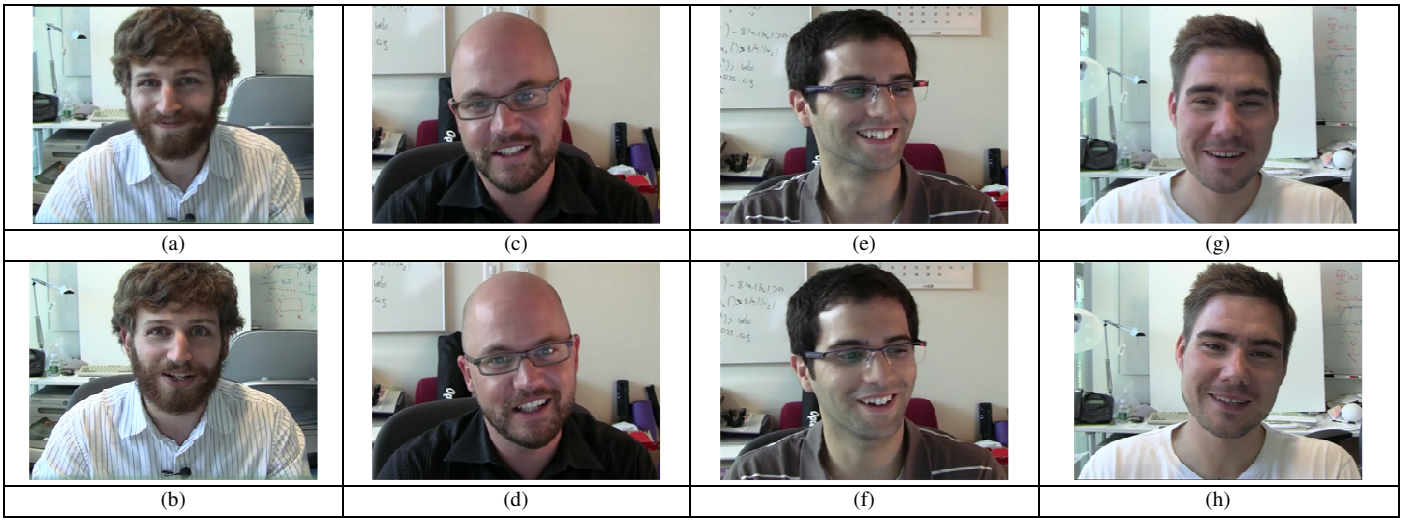


Figure 1. Four participants, each smiling while being in either a (i) frustrated or (ii) delight state. Can you tell which smile is which state? Answers are provided in the Acknowledgement section.

II. DATA COLLECTION

A. *STUDY 1: Acted data Experiment:*

We brought participants into an empty room to interact with a computer program. The computer program consisted of a 2d image of an avatar (Figure 2), with which participants were expected to interact. The avatar asked a sequence of questions, and the participants were expected to respond to those questions by speaking directly at the avatar. Participants wore a Headset to speak with the avatar and there was a video camera to capture the face of the participant. The sequence of the 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)*

Figure 2. 2d image of the computer program used in the "Acted experiment"

Participants: There were 15 participants in "Acted Data Experiment", 10 male and 5 female. Their age ranged from 25-40 and all were office employees at a major corporation. From 15 participants, we gathered 45 clips of frustrations, delight and neutral expressions. The average duration per clip for frustration and delight was a little over 20 seconds. The average duration for neutral was around 10 seconds per participant. We used a Logitech ClearChat Comfort USB Headset for the participant to speak to the avatar. We used a Logitech 2 MP Portable Webcam C905 to record the face data of the participants.

B. *STUDY 2: Natural data Experiment:*

This study involved 27 new participants who were not part of the "Acted data experiment". For this study, we recruited subjects to fill out a tedious web form on a computer.

After the participant entered the experiment room, they were told that they would be asked to fill out a form, and based on how the task progresses the participant may or may

not be asked to speak to the camera. The form contained a bunch of biographical questions (details in Table 1) along with fields where participants were asked to enter current date and time without providing any hint on the format. They were also told not to exit the room until they reach the confirmation screen of the form (screen 16 of Table 1). The exact sequence of interaction 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.

Screens	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 at 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 messages in Table 1 were embedded as .wav files into the form. In other words, the text messages from each screen were read out loud by the form as the user navigated from one screen to another. We used ATT's publicly available text to speech engine to produce those utterances using a female voice on American accent. Initially, the users were prompted with two questions (screens 3 and 4 of Table 1), one after another, by the form. The purpose of those questions was to elicit statements from the participant which were more likely to be neutral. The reason we opted for two questions, rather than one is because during our pilot study, we noticed that the very first time people felt awkward to provide feedback to the camera and in most cases, either they laughed out of embarrassment or provided only a very brief statement, when asked "why are you interested in this study?" Adding a follow up question in the next screen helped most of them to ease off and provide a more neutral expression for the second answer. We have seen this "first expression" effect dominate expressed emotions regardless of which emotion the stimuli were designed to elicit, and it is important that scientists consider this when designing emotion elicitation experiments.

In biographical forms (screens 5, 7, 9 in Table 1), there was a timer that started counting the time in a bigger font in the middle of the form to indicate elapsed time. Right mouse click, as well the CTRL keys of the keyboard were disabled so that participants could not copy content from one screen to the next one. Also, the claim that "94.5% of the previous participants were able to fill out the form in less than 2 minutes" was made-up to add additional stress to the participants. In screen 10 of the interface, after trying to submit the form three times, the form prompts the user to solve a CAPTCHA to move forward. The CAPTCHAs were taken from Google images of ones that were nearly impossible to solve. Therefore, regardless of whatever the participant typed, the interface kept on prompting to solve a new CAPTCHA. It went on for 3 trials. After that, in screen 13 of the interface, participants were asked to provide feedback about what they have done wrong with the interface for which they could not finish the form as quickly as most of the participants. In this phase, we expected the participants to be somewhat frustrated and to reveal signs of frustration either through their face, or speech or both.

In the second phase of the interaction, with screen 14, users were given a bit of time to relax and think about a funny experience to share to the camera momentarily. To help them ease off, the interface showed a YouTube video of a baby laughing uncontrollably. The video has more than 150 million views in YouTube since 2006 and can be watched at http://tiny.cc/xovur. We specifically picked this video because we felt that laughing specially from a baby, is contagious and it could potentially distract the participants from their bitter experience of filling out the web form. After the experiment, majority of the participants mentioned that they had already seen the video, and they still found it funny and exhilarating. After the end of the interaction with the web form, we set up a post briefing with the participant to document their self reported measure of how frustrated and delighted they were when they had provided their feedback through the camera.

Participants: There were a total of 27 graduate students who participated in this study. Five of them were female and 22 male. In post-experimental briefing, three participants informed us that they were able to figure out the forms were intentionally designed to be buggy to provoke frustration from them. Since they were able to figure out the objective of the study, we eliminated their data, resulting in 24 clips of frustration. 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 post-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 neutral expressions, we only considered expressions from screen 4, as indicated in Table 1, and ignored the expressions elicited in screen 3 of the interface. Therefore, we had 27 instances of neutral expressions from 27 participants. The average length of each clip for frustration and delight was a little over 30 seconds. The average length for neutral expression from our participants was around 15 seconds. For this experiment, we used Canon VIXIA HF M300 Camcorder and Azden WMS-PRO Wireless Microphone.

III. METHODS

We had 45 clips from study 1 and 72 clips from study 2 containing both audio and video that we needed to be analyzed. After extracting features from audio and video channels, we concatenated the speech and facial features per clip in a vector such that in each clip's feature vector, V_{clip} ={ $A_1,\ldots A_n$, $F_1,\ldots F_m$}, $A_1,\ldots A_n$ are n speech features, and $F_1,\ldots F_m$ are m facial features. In this study, n was equal to 15 and m was equal to 25; features are described below.

C. Face analysis

We used Google's facial feature tracker, formerly known as Nevenvision, to track 22 feature points (9 points surrounding the mouth region, 3 points for each eye, two points for each eye-brow, and three points for two nostrils and nose tip) of the face. The local distances among those points as well as their standard deviations were measured in every frame and used as features [10]. Additionally, we used Sophisticated Highspeed Object Recognition Engine (SHORE) [5] API by Fraunhofer to detect features such as eye blinks and mouth open. Shore API also provides a probability score (0-100%) of smile by analyzing mouth widening, and Zygomaticus muscles of face in every frame. In this paper, this score is referred to as the strength of smile or in other words, probability of smile. All the features were tracked in every frame and then were averaged to form a 1d vector per clip. In the first study 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 study with natural data, we also averaged all the features across each clip. We have investigated temporal patterns of the features per clip and will report on that in a separate publication.

D. 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, we extracted features related to pitch (mean, standard deviation, maximum, minimum), perceptual loudness, pauses, rhythm and intensity, per clip.

IV. RESULTS

We used five classifiers (BayesNet, SVM, RandomForest, AdaBoost, and Multilayer Perceptron,) from the weka toolbox [6], to compare the classification accuracy between natural data and acted data. Figure 3 shows all the classifiers performed significantly better with acted data compared to natural data (using 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 natural 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 27 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 natural data.

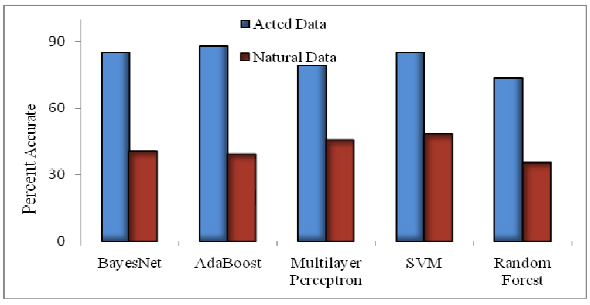


Figure 3. Classification accuracy of recognizing frustration, delight and neutral states using various classifiers with natural and acted data. The accuracy is reported using the leave-one-out method.

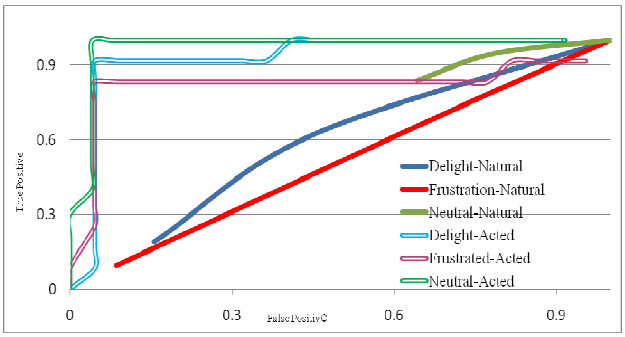


Figure 4. The ROC curves for recognition of delight, frustration and neutral expressions in natural data and acted data. These show superior performance for the acted data set over the natural one.

Additional analysis on the feature vectors for participants from study 1 and study 2 revealed that in acted data, close to 90% of the participants did not smile when they were frustrated. On the contrary, in natural dataset of study 2, close to 90% of the participants did smile when they were frustrated.

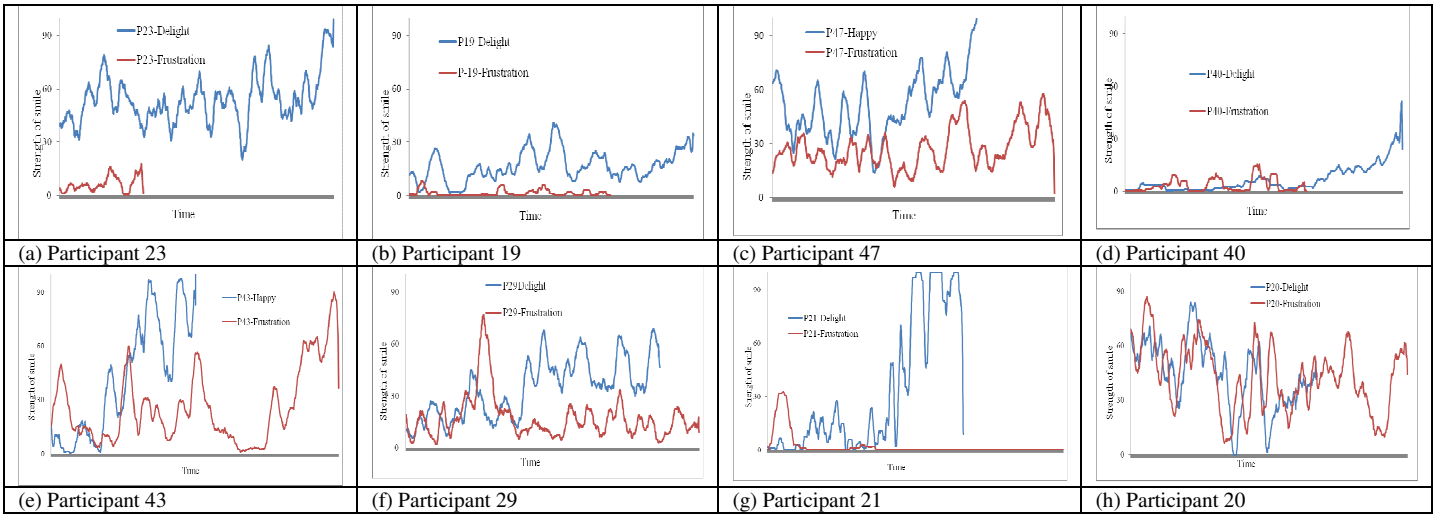
V. DISCUSSIONS

The results shown in Figures 3 and 4 demonstrate significant differences in correctly classifying instances when the expressions are acted as opposed to being natural. One possible explanation is that acted expressions seem to contain the prototypical unique facial features, whereas natural 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 natural data. In this study, we were primarily interested to explore, in the context of our dataset, what properties of natural expressions make it more difficult to recognize them using machine learning algorithms. Thus, we felt that along with analyzing peoples' expressions and reporting the average, it might be worthwhile to zoom more into subtle individual differences in terms of expressions. As a result, 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. Given the page limit of this paper, we here zoom into a narrow set of features, namely smiles, to analyze the intrinsic dynamics of expressions conveyed under natural conditions. Analyzing other face and speech features in the context of individual differences will be aimed at a future publication.

Analyzing each individual clip, from study 2, for all the participants revealed interesting findings. We noticed that almost all the participants, despite self-report to be extremely frustrated, did not illustrate 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 natural frustration may help explain why some people diagnosed with Autism Spectrum Disorder (ASD) have difficulties in reading facial expressions. If one is taught that smiles mean happiness then it would be easy to mistake smiling expressions of a frustrated person as "things are going great – they look delighted!" This misunderstanding could cause real problems in a high pressure workplace.

As mentioned previously, almost all of our participants from study 2, whether frustrated or delighted, demonstrated signatures of smile during their interaction. This is problematic data for those who promote the belief that smile is a strong disambiguating feature between delight and other basic emotions. 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 patterns are plotted in Figure 5. A small subset of the participants, as shown in Figure 5(a, b, c), have clear separation of their smiles in terms of magnitude or strength 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 intra level differences rather than reporting the average. Figures 5(d, e, f, g) are symbolic in a way in all cases participants, in context of delight, gradually progressed into their peaks in terms of smile. This finding is very insightful because now it stimulates the need to analyze the smiling patterns that progress through time. The prevalence of smile 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". Looking at Figures 5(e, f, g), the social smiles usually appear as spikes, which is very consistent with what exists in the literature [9]. Another interesting occurrence to observe, especially in Figure 5 (g) and 5(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.



(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 5: (a-h) graphs of 8 participants whose patterns are representative of rest of the participants. In this graph, x axis is the time and y axis is the strength of smile or probability of smiling. Figures 5(a, b, and c) are examples of participants who have distinct patterns of strength in smile when they are frustrated and delighted. Figures 5(d, e, f, and g) provide examples of how the state of delight builds up in terms of smile through time. Figures 5(f, g) are examples of participants who initiated their frustration with a social smile. Figure 5(h) is an example of a few people who exhibit similar smile patterns whether they were delighted or frustrated.

A smile is such a universal and multifaceted expression in our daily life that one may err by equating its occurrence to a particular positive emotion. People smile to express rapport, polite disagreement, delight, sarcasm, and more. Detecting the lip-corner puller (AU12) and cheek raiser (AU 6) thus do not reliably recognize a happy state.

We demonstrate in this work that it is useful to explore how the patterns of smile evolve through time, and that while a smile may occur in positive and in negative situations, its dynamics may help to disambiguate the underlying state. Our immediate extension of this work would be to explore other facial and speech features for 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.

ACKNOWLEDGEMENTS

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 1. 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 for this work.

REFERENCES

[1] Z. Ambadar and J. Schooler, & J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpretating subtle facial expressions", *Psychological Science*, Vol. 16, No. 5, pp. 403-410., 2005.

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

[3] P. Ekman and W. Friesen. "Facial Action Coding System: A Technique for the Measurement of Facial Movement", Consulting Psychologists Press, Palo Alto, 1978.

[4] P. Boersma and D. Weenink. "Praat: Doing Phonetics by Computer." Internet: www.praat.org, [January, 2011].

[5] 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

[6] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition ed. San Francisco: Morgan Kaufmann, 2005.

[7] 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.

[8] 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.

[9] M.F. Valstar, H. Gunes and M. Pantic, "How to Distinguish Posed from Spontaneous Smiles using Geometric Features", in *Proceedings of ACM International Conference on Multimodal Interfaces (ICMI'07)*, pp. 38-45, Nagoya, Japan, November 2007.

[10] 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.