Deep Learning Approach For Classifying Spontaneous and Deliberate Smiles

Liang Wang
University of San Francisco
San Francisco, CA 94105, USA
lwang89@usfca.edu

David Guy Brizan
University of San Francisco
San Francisco, CA 94105, USA
dgbrizan@usfca.edu

Beste Filiz Yuksel
University of San Francisco
San Francisco, CA 94105, USA
byuksel@usfca.edu
author4@hchi.anotherco.com

Every submission will be assigned their own unique DOI string to be included here.

Abstract

We conducted experiments to classify spontaneous and deliberate smiles. In the first experiment, we extract features related to human smile dynamics. We evaluated and compared results from Linear Support Vector Machine (Linear-SVM), Random Forest, Stochastic Gradient Descent(SGD), k-Nearest Neighbors(kNN), AdaBoosting and Gradient-Boosting. An accuracy of 83% was obtained using a kNN classifier. Next, we designed and implemented a LSTM model to do classification. We got an accuracy of 99.92%.

Author Keywords

Expressions classification; Natural versus acted data; Natural dataset; Deep learning

1. INTRODUCTION

Recognizing human facial expressions correctly is a hard human computer interaction and machine learning problem. Most of the previous exploratory studies have attempted to classify so-called âĂIJbasic emotionsâĂİ (anger, disgust, fear,happiness, sadness, and surprise) from images and videos([**?**], [**?**] as reported in [**?**]). 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 [**?**]. Automatic analysis and classification of emotional facial expressions have been an active research topic since the Facial Action Coding system (FACS) was proposed by [**?**].

In recent studies, analysis of spontaneous facial expressions have gained more interest. For social interaction analysis, it is necessary to distinguish genuine (spontaneous/felt) expressions from the posed (deliberate) ones since they convey different meanings. Spontaneous expressions can reveal states of attention, agreement and interest, as well as deceit. The foremost facial expression for spontaneity analysis is the smile as it is the most frequently performed expression [**?**]. A smile can signal enjoyment, embarrassment, politeness, etc. [**?**]. It is also used to mask other emotional expressions, since it is the easiest emotional facial expression to pose voluntarily [**?**], [**?**].

In this study, we want to figure out several questions:

- Is there a difference when people smile under frustration as opposed to being genuinely delighted [**?**]?
- How do humans perform in correctly labeling smiles elicited under frustrated and delighted stimuli [**?**]?
- How do the statistic machine learning classifiers perform on recognizing mental states such as frustration and delight when acted, as well as when naturally elicited [**?**]?
- Can we use deep learning model to improve the accuracy?
- Which features are really important in making this classification?
- What is the least number of instances we need when a excellent performance is maintained?

Our contributions are: 1) we re-implement algorithms and statistic machine models to classify spontaneous smiles and deliberate smiles, using the largest spontaneous/deliberate smile database in the literature; 2) we report an accurate deep learning smile classification method, which outperforms the state-of-the-art methods.

The remaining part of the paper is structured as follows: In Section 2, related work in smile and spontaneity analysis is given. Section 3 describes methods for smile classification. Section 4 presents results from human, statistic machine learning models and a deep learning model. In Section 5, the findings of this study are discussed. Section 6 concludes the paper.

2. RELATED WORK

We will firstly explain smile physiognomy, and then reported work on automatic smile analysis.

Smile Physiognomy

The smile is the easiest emotional facial expression to pose voluntarily [**?**]. Broadly, a smile can be identified as the upward movement of the lip corners, which corresponds to Action Unit 12 (AU12) in the facial action coding system (FACS) [**?**].In terms of anatomy, the zygomatic major muscle contracts and raises the corners of the lips during a smile [**?**]. In terms of dynamics, smiles are composed of three non-overlapping phases; the onset (neutral to expressive), apex, and offset (expressive to neutral), respectively. Ekman individually identified 18 different smiles (such as enjoyment, fear, miserable, embarrassment, listener response smiles) by describing the specific visual differences on the face and indicating the accompanying action units, however temporal dynamics for each smile type were not described [**?**].

Guillaume Duchenne experimented on muscle activities during smiles, and proposed that smiles resulting from felt joy not only utilize the zygomaticus major muscle, but also the orbicularis oculi (a circular muscle around the eyes). Duchenne claimed that the orbicularis oculi could not be controlled voluntarily during posed smiles [**?**]. This kind of joy smiles are called Duchenne smiles (D-smiles) in his honor. In [**?**], a strong correlation between D-smiles and felt enjoyment smiles were found. However, the definition of D-smiles was updated as the combined contraction of zygomaticus major and the outer strands (pars lateralis) of orbicularis oculi, since fewer people can voluntarily contract the outer strands of orbicularis oculi, as compared to its inner strands [**?**].

Contraction of the orbicularis oculi, pars lateralis raises the cheek, narrows the eye aperture, and forms wrinkles (crowsfeet) on the external side of the eyes [**?**]. This activation corresponds to Action Unit 6 (AU6) and is named as the Duchenne marker (D-marker) in the literature [**?**]. [**?**] indicates that most people cannot voluntarily contract orbicularis oculi, pars lateralis and the ones who can do it usually cannot activate this muscle on both sides of their face simultaneously. However, new empirical findings question the reliability of the D-marker [**?**] [**?**] [**?**]. Recently, it has been shown that orbicularis oculi, pars lateralis can be active or inactive under both spontaneous and posed conditions with similar frequencies [**?**]. On the other hand, untrained people consistently use the D-marker to recognize genuine and posed enjoyment smiles [**?**].

Automatic Smile Analysis
In [**?**], Valstar propose a method to automatically discriminate between spontaneous and deliberate brow actions using intensity, duration, trajectory, symmetry, and occurrence order of the actions. In [**?**], a multimodal system is presented to classify posed and genuine smiles. GentleSVM-Sigmoid classifier is used with the fusion of shoulder, head and inner facial movements.

In [31], a spatio-temporal method is proposed using both natural and infrared face videos to discriminate between spontaneous and posed facial expressions.

Recently, DibeklioÄ§lu proposed a system which use a generic descriptor set which can be applied to different facial regions to enhance the indicated facial cues with detailed dynamic features. Additionally, they focus on the dynamical characteristics of eyelid movements (such as duration, amplitude, speed, and acceleration), instead of simple displacement analysis, motivated by the findings of [9] and [25].

In [**?**], they extracted local and global features related to human smile dynamics and then developed an automated system to distinguish between naturally occurring spontaneous smiles under frustrating and delightful stimuli by exploring their temporal patterns given video of both.

In this paper, the aim is to use a much smarter deep learning model to classify spontaneous smiles and deliberate smiles with higher accuracy in a much simpler way.

3. METHOD

This paper proposes some classic machine learning smile classification models and an deep learning neural network smile classification model. In this section, details of the proposed spontaneous/posed enjoyment smile classification system are summarized. The flow of the system is as follows. Facial landmark points are located in the first frame, and tracked during the rest of the smile video.

In classic machine learning smile classification models part, we select special landmark couples and calculate mean distance and standard deviations of them, which are used to train classic machine learning models (linear-SVM, Random Forest, kNN, AdaBoost, GradientBoost, etc.).

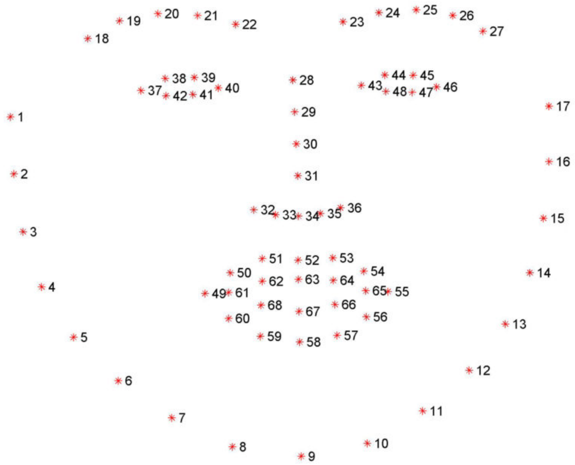

Figure 1: Visualizing the 68 facial landmark coordinates

In deep learning model part, we use all facial landmarks because we want to input all details to classifier first and then find really important landmarks.

A Face Analysis
We used OpenCV and dlib [**?**] to track 68 feature points (Figure **??**).

B Analysis via Classic ML Models
We calculated raw distances (in pixels) as well as their standard deviations across facial feature points [**?**].For example, distances and standard deviations tween 37 and 43, 43 and 46, 39 and 21, 44 and 24, 22 and 40, 23 and 43, 49 and 55, 52 and 58 etc. were calculated.

The local distances among those points as well as their standard deviations were measured in every frame and used as features [**?**].

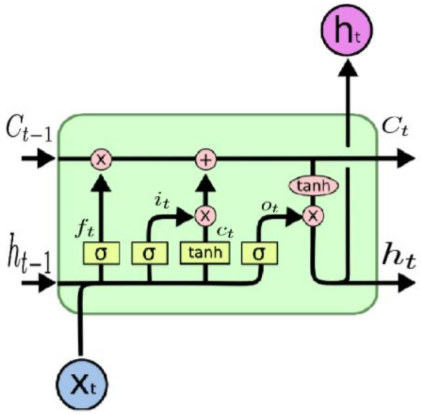

Figure 2: Long Short-Term Memory Neural Network

B Analysis via A Deep learning Model
We compared a bunch of deep learning models, such R-CNN, RNN, GRU and LSTM. We choose to use LSTM at last for it is a mature model which has good performance on dealing with sequential data. We reshape all 68 landmark points coordinates and pad all video instances to make sure them having the same length(Figure **??**).

4. DATABASE

We applied and used UvA-NEMO Smile Database [**?**] from University of Amsterdam to analyze the dynamics of spontaneous/posed smiles.This database is composed of videos (in RGB color) recorded with a Panasonic HDC-HS700 3MOS camcorder, placed on a monitor, at approximately 1.5 meters away from the recorded subjects. Videos were recorded with a resolution of $1920 * 1080$ pixels at a rate of 50 frames per second under artificial D65 daylight illumination. Additionally, a color chart is present on the background

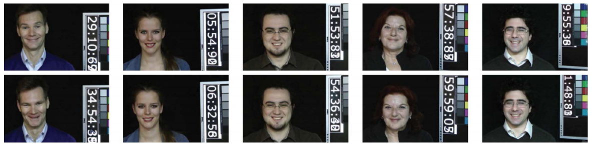

Figure 3: Sample frames from the UvA-NEMO Smile Database showing posed enjoyment smile (top), and spontaneous enjoyment smile (bottom) [**?**].

of the videos for color normalization. Fig. **??** shows sample frames from the UvA-NEMO Smile Database.

The database has 1240 smile videos (597 spontaneous, 643 posed) from 400 subjects (185 female, 215 male). The ages of subjects vary from 8 to 76 years, and there are 149 young people (235 spontaneous, 240 posed) and 251 adults (362 spontaneous, 403 posed). 43 subjects do not have spontaneous smiles and 32 subjects have no posed smile samples [**?**].

5. RESULTS

In this section, we firstly present the performance of five statistic models (Linear-SVM, Random Forest, kNN, AdaBoost and GradientBoost). We use a 10-fold cross-validation scheme: each time a test fold is separated, a 9-fold cross-validation is used to train the system. The results are in Fig. **??**. The best result comes from kNN model, we get 83.53% accuracy for recognizing deliberated smiles and 82.05% accuracy for recognizing spontaneous smiles.

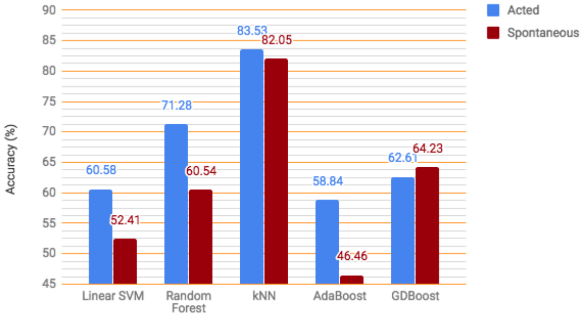

Figure 4: Classic machine learning model results

Then we used the same 10-fold cross-validation protocol for LSTM model. We conducted several sets of experiments on this model. The first one was that we used a 9-fold cross-validation to train the system and a fold to test. There is no subject overlap between folds. The results are shown in Table. **??**. We got an unbelievable high average accuracy of 99.92%. Then we made a flip: we used a separated fold to train the system and other 9 folds to test. We still got amazing results with a high accuracy of 99.80%. Detailed results are shown in Table. **??**.

Then we check if number of epochs affect the accuracy of prediction. Results (Table. **??**) shows increasing number of epochs cannot affect accuracy significantly.

	Average	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Accuracy	0.9992	1	1	1	1	1	1	1	1	1	0.9920
Errors	0.1	0	0	0	0	0	0	0	0	0	1

Table 1: train:test = 9:1, epochs = 300

6. DISCUSSION

In our statistic experiment part, we use the method M.E.Hoque proposed [**?**]. However, we got totally different results compared with his paper. None of our statistic model can reach

	Average	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Accuracy	0.9980	0.9982	0.9964	0.9991	0.9973	1	0.9955	0.9991	0.9973	0.9973	1
Errors	2.2	2	4	1	3	0	5	1	3	3	0

Table 2: train:test = 1:9, epochs = 300

	Average	*300*	*500*	*800*	*1500*	*3000*
Accuracy	0.9978	0.9982	0.9973	0.0.9991	0.9964	0.9982
Errors	2.4	2	3	1	4	2

Table 3: train:test = 1:9, fold = 0

a 90% accuracy for recognizing deliberate smiles. Highest accuracy (83.53%) comes from kNN. Meanwhile, all results of recognizing spontaneous smiles from his paper are below 45%. Most of our results are over 50% which means better than guessing. Our kNN model gives us an accuracy of 82.05% for recognizing spontaneous smiles.

There are two reasons lead to different results. Firstly, M.E Hoque mentioned he extracted 25 facial features while he alse used 15 audio features. But we can only find 16 facial features from his paper. Less features we used may be a reason leads to a lower accuracy. Secondly, the video database he used has only 116 videos with 27 participants while UvA-NEMO has 1240 videos with 400 participants. Too less videos and records may be an important reason leads to Over-fitting. Table. **??** comparison of two video databases.

We usually do computation and aggregation to fit our data to statistic machine learning models which cause lossing of a lot of details of raw data. Deep learning model avoid this weakness and help us to find important features. In this experiment, we input 246k records to LSTM model while there are only 6971 records for statistic models. We have to split every video instance to small segments (30 frames) for statistic models while we input whole video instance data to LSTM model.

Database	*Participants*	*Resolution & Frame Rate*	*Total Videos*	*Spontaneous*
MIT	27	Unknown @30HZ	116	72
UvA-NEMO	400	1920 * 1080 @50HZ	1240	597

Table 4: MIT Database & UvA-NEMO Database

Besides, LSTM model not only shows incomparable accuracy but also impressively powerful performance. With less than 100 videos, we can classify over 980 videos with an average accuracy of 99.80%.

7. CONCLUSION

In this study, we re-implement algorithms using statistic machine learning models to classify spontaneous smiles and deliberate smiles from UvA-NEMO Smile Database. Later we design and implement a deep learning neural network to do classification.

Our results show that LSTM model has a far better accuracy and performance than anyone of statistic machine learning models.