Predicting Active Facial Expressivity in People with Parkinson's Disease

Ajjen Joshi
Department of Computer Science
Boston University
ajjendj@bu.edu

Margrit Betke
Department of Computer Science
Boston University
betke@bu.edu

ABSTRACT

Our capacity to engage in meaningful conversations depends on a multitude of communication signals, including verbal delivery of speech, tone and modulation of voice, execution of body gestures, and exhibition of a range of facial expressions. Among these cues, the expressivity of the face strongly indicates the level of one's engagement during a social interaction. It also significantly influences how others perceive one's personality and mood. Individuals with Parkinson's disease whose facial muscles have become rigid have difficulty exhibiting facial expressions. In this work, we investigate how to computationally predict an accurate and objective score for facial expressivity of a person. We present a method that computes geometric shape features of the face and predicts a score for facial expressivity. Our method trains a random forest regressor based on features extracted from a set of training videos of interviews of people suffering from Parkinson's disease. We evaluated our formulation on a dataset of 727 20-second video clips using 9-fold cross validation. We also provide insight on the geometric features that are important in this prediction task by computing variable importance scores for our features.

CCS Concepts

•Computing methodologies → Computer vision tasks; Machine learning algorithms;

Keywords

Facial expressivity prediction; Geometric facial features; Random Forest regression; Feature importance

1. INTRODUCTION

The ability to effectively express ourselves is an important part of daily living as it affects every aspect of our social lives, such as forming and maintaining relationships and creating impressions. The human face is "one of the most powerful channels of nonverbal communication" [5]. Computational analysis of facial expressivity has a wide range of applications. It can assist behavioral scientists in automatically annotating data, an otherwise expensive and laborious process, and includes automated lie detection. Automated analysis of facial expressivity can also help assess symptoms in people suffering from behavioral or motor disabilities such as Parkinson's disease.

Parkinson's disease is a progressive neurodegenerative disease affecting over 1,000,000 people in the United States [7]. Patients with Parkinson's disease suffer from rigidity of facial musculature which worsens with time. Because Parkinson's affects both verbal and non-verbal channels of communication, the ability to accurately measure communication abilities becomes paramount for both patients and caregivers. A computational tool capable of doing so would help therapists diagnose and evaluate patients and help provide individualized therapy.

Automatic analysis of facial expressions and affect has been an active research topic in the fields of computer vision and machine learning [5, 9]. The task is either recognizing facial expressions from static images [8] or from a sequence of images [4]. Many approaches focus on the detection and recognition of facial action units [12]. An action unit (AU) describes the movement of one or more facial muscles. It is used in describing facial activity through the Facial Action Coding System (FACS) [6], which measures perceptible facial movement through an accurate anatomical taxonomy. Facial events, such as the expression of an emotion, can thus be described by a combination of AUs.

A common pipeline in automated affect analysis consists of first detecting the face. Facial landmark detectors are used to detect head-pose as well as points describing the eyes, nose and mouth. After normalization to account for rigid head motion and variance in distance to camera, shape and appearance features can be extracted. Some examples of geometric shape features include distance between the inner brow and eye, distance between the outer brow and eye, distance that measures the height and width of the mouth, and the angle between mouth corners. Appearance features represent textural components, such as wrinkles, of various facial parts. Feature descriptors are used to train classification or regression models in a supervised setting and the trained models are evaluated in their performance of tasks such as expression or action unit classification.

Advances in automated facial analysis and machine learning has enabled new applications such as evaluation of neuromuscular impairment or assessment of psychopathology.

PETRA '16 Corfu, Greece

DOI: 10.1145/1235

Cohn et al. [3] explored the feasiblity of detecting depression using facial actions and vocal prosody. Wang et al. [13] used automated analysis of video-based expressions to analyze neuropsychiatric disorders such as Asperger's Syndrome and Schizophrenia. Wu et al. [14] attempted to quantify facial expressivity of patients with Parkinson's by comparing the occurrence and intensity of 11 different Action Units between a group of control participants and Parkinson's patients.

In this paper, we propose a framework for automatic analysis of expressivity. Expressivity is an overall measure of the capacity to express emotions. Facial expressivity is used as one of 20 indicators of the Interpersonal Communication Rating Protocol (ICRP), which is a manual for objectively measuring the expressive behavior of individuals [10] suffering from Parkinson's disease. According to the ICRP, active expressivity in the face is measured along a 5-point Likert scale. A group of raters provide a "gestalt" rating of each ICRP indicator based on intensity, duration, and frequency of the variables of expressive behavior.

We here propose a framework for computing geometric shape feature descriptors based on facial landmarks from a video sequence. We trained a random forest regression model from features to expressivity scores. The ground-truth annotations were given by domain experts. We tested the expressivity prediction capabilities of our system on a dataset of 727 videos using 9-fold cross validation. Finally, we provide insight on the geometric features that are important in this prediction task by computing the local variable importance from our feature set.

2. SYSTEM OVERVIEW

We here explain the landmark detection, feature extraction, and regression components of our system.

Facial Landmark Detection

The input to our system consists of 20-second videoclips of interviews of subjects facing the camera. Most frames in the sequences of images contain full frontal faces of the subject along with the torso. In some videos the frontal face of the subject cannot be detected in a significant number of contiguous frames due to occlusion by the hand or severe out-of-plane rotation of the head. These videos were discarded from the training and testing procedure in our framework. For facial landmark detection, we used a robust facial landmark tracker [1]. For every frame, the tracker outputs the x and y coordinates of 59 facial landmarks, as well as pitch, roll and yaw angles to describe head pose. To produce a summary score for active expressivity of the face, our method extracts geometric features from the temporal signals associated with corresponding facial landmarks in a contiguous sequence of frames.

Geometric Feature Extraction

Our method extracts geometric shape features based on facial landmark coordinates. Aside from being informative about discriminative facial events, each geometric attribute has the advantage of being easily interpretable. Geometric features that measure the distance between the brows and the eyes, the height of the eye, the height of the mouth and the angle between the mouth corners have been commonly used in facial expression analysis [5]. Moreover, the facial dynamics associated with these features are studied by ICRP raters while determining the rating for facial expressivity. We aimed to not only maximize expressivity prediction accuracy but also provide insight on the geometric features that are most discriminative.

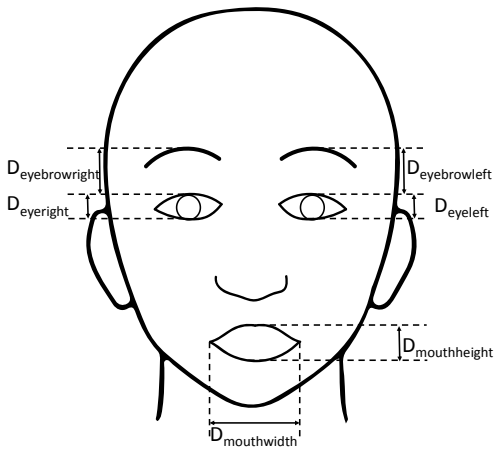

Figure 1: Geometric features, which capture facial dynamics, computed by our system

First, our method sets the landmark between the two eyes as the origin of the reference frame. To account for the variation in the distances between the subject and the camera and the dimensions of the faces of the different subjects, our method normalizes the coordinates of the landmarks by taking the inter-ocular distance of the subject as a reference. From the normalized coordinates of the facial landmarks, our method extracts distances between certain facial landmarks to describe the dynamics of the eyes ($\mathbf{D}_{\text{eyeright}}$, $\mathbf{D}_{\text{eyeleft}}$), eyebrows ($\mathbf{D}_{\text{eyebrowright}}$, $\mathbf{D}_{\text{eyebrowleft}}$) and mouth ($\mathbf{D}_{\text{mouthheight}}$, $\mathbf{D}_{\text{mouthwidth}}$) at each frame (Figure 1).

For each of the distance signals, first derivatives are approximated. Finally, for each signal channel, our method computes the three quartile values (Q1, Q2, Q3), the max (Q4), the standard deviation (std) and peak frequency (pf). We define peak frequency as the number of local maxima of a given signal channel per unit length. Finally, we concatenate these attributes to form a single representative feature vector that captures the intensity, duration and frequency of eye, eyebrow and mouth dynamics.

Random Forest Training

The training set is defined as $\mathcal{D} = \{(\mathbf{X}_1, y_1), ..., (\mathbf{X}_n, y_n)\}$. Here, $(\mathbf{X}_1, ..., \mathbf{X}_n)$ is the set of feature vectors representing the video samples in the training set, and $(y_1, ..., y_n)$ represents their corresponding expressivity scores.

A random forest regression model consists of several regression trees $\{t(\mathbf{X}, \phi_k), k = 1, \dots\}$ [2]. Here $\mathbf{X}$ is an input vector and ϕ_k is a random vector used to generate a bootstrap sample of objects from the training set $\mathcal{D}$.

At each internal node of the tree, m features are randomly selected from the available d, where d is the dimensionality of the feature vector of the inputs, such that $m < d$. From the m chosen features, the feature that most reduces the sum of squared errors is chosen to split the tree. We chose $m = \frac{d}{3}$. The sum of square errors (S) can be defined as:

$$S = \sum_{c \in leaves(T)} \sum_{i \in C} (y_{\text{truth,i}} - m_c)^2, \qquad (1)$$

where $y_{\text{truth,i}}$ is the true expressivity value and $m_c = \frac{1}{n_c}\sum_{i \in C} y_i$ is the prediction made at leaf c.

Each decision tree $t(\mathbf{x}, \phi_k)$ in the forest is constructed until the leaves contain no more than 5 items. Finally, we tested our method using 9-fold cross validation.

Variable Importance

It is possible to estimate the importance of a feature variable using the random forest algorithm. After the construction of each tree, the values for a given feature variable in the out-of-bag-examples (examples not in the bootstrapped set used to construct the tree) are randomly permuted. The constructed tree is then used to predict the target variable after this random permutation. The difference in error before and after the process of noising up each of the feature variables is used as an estimate of the importance of the variable in the regression task [2].

3. DATASET

The dataset consists of 805 video samples. This dataset was collected by Tickle-Degnen et al. [11] to determine the effects of self-management rehabilitation on Health-Related Quality-Of-Living in Parkinson's disease. Participants (N = 117) in this study were divided randomly into three groups based on the type of rehabilitation in a 6-week intervention program. All participants in this study had previously been diagnosed with Parkinson's disease by a movement disorder specialist and had the ability to understand and communicate with personnel. Patients were videotaped participating in standardized social interactions, where cameras were placed to show a mostly frontal face and torso view. From the videotapes, a 20-second representative segment consisting of patients speaking about a positive or negative experience was chosen for analysis. For each video segment, four trained research assistants provided a measure of active expressivity of the face on a 5-point Likert scale. A composite score for each variable was computed by taking the average of the scores provided by each rater.

Videos, where facial landmarks of the subject could not be detected in a sufficient number (30) of contiguous frames, were discarded from the set used to build our expressivity prediction model. This reduced the size of the dataset from 805 to 727 video samples.

4. EXPERIMENTS

Here, we provide a description of the experiments performed on the dataset to evaluate our expressivity prediction framework. We trained and tested our framework using 9-fold cross-validation with a number of feature representations.

(a) Mouth shape statistics feature set (MS): For each sample video, we computed, for every frame, $\mathbf{D}_{\text{mouthheight}}$ and $\mathbf{D}_{\text{mouthwidth}}$ and their first derivatives $\mathbf{D}'_{\text{mouthheight}}$ and $\mathbf{D}'_{\text{mouthwidth}}$. For each signal channel, we computed the three quartile values, the max, the standard deviation, and peak frequency, and concatenated them to form a single representative 24-dimensional feature vector.

(b) Eye shape statistics feature set (ES): In this feature representation, we computed the vector of distances $\mathbf{D}_{\text{eyeleft}}$, $\mathbf{D}_{\text{eyeright}}$, $\mathbf{D}_{\text{eyebrowleft}}$ and $\mathbf{D}_{\text{eyebrowright}}$. For each vector of distances and their first derivatives, we computed the three quartile values, the max, the standard deviation, and peak frequency, and concatenated them to form a single representative 48-dimensional feature vector.

(c) Combined Geometric shape statistics feature set (MS+ES): In this feature representation, we concatenated the aforementioned feature vectors to produce a combined 72-dimensional feature vector.

Table 1: Average Mean Absolute Error (MAE) and its corresponding Standard Deviation (SD) on Expressivity Prediction for feature sets Mouth Shape (MS), Eye Shape (ES) and a combination (MS+ES)

Feature set	Average MAE	SD
MS	0.604	0.096
ES	0.560	0.084
MS+ES	0.566	0.074

Table 2: Average R^2 scores and its corresponding Standard Deviation (SD) on Expressivity Prediction for feature sets Mouth Shape (MS), Eye Shape (ES) and a combination (MS+ES)

Feature set	Average R^2	SD
MS	21.27	18.21
ES	42.33	12.94
MS+ES	40.68	12.77

For each feature set described above, we trained and tested our framework of random forests with 150 trees using 9-fold cross-validation. We determined the ideal number of trees in the forest by observing the average Out-Of-Bag (OOB) error rate while training our model with each feature set.

We computed the mean absolute errors (Table 1) and R^2 scores (Table 2) averaged over all folds along with their respective standard deviations for all feature sets.

The Mean Absolute Error (MAE) is given by:

$$\text{MAE} = \frac{\sum_{i=1}^{N} | y_{\text{truth,i}} - y_{\text{pred,i}} |}{N}, \tag{2}$$

where $y_{\text{truth,i}}$ and $y_{\text{pred,i}}$ correspond to ground truth and predicted scores and N is the number of test samples. The MAE score accounts for the average absolute error of the predicted scores.

The R^2 score is given by:

$$R^2 = (1 - \frac{\sum_{i=1}^{N}(y_{\text{truth,i}} - y_{\text{pred,i}})^2}{\sum_{i=1}^{N}(y_{\text{truth,i}} - \bar{y}_{\text{truth}})^2}) \times 100, \tag{3}$$

where $\bar{y}_{\text{truth}}$ corresponds to the mean of the ground truth. The R^2 score is based on the ratio of the error made by the model to the error made by a baseline predictor that always predicts the mean score of the training data. The R^2 score gives a measure of the relative improvement in the Mean Square Error (MSE) of our regression model with respect to the baseline mean expressivity predictor.

Our analysis shows that the feature set containing eye shape statistics (ES) has the lowest mean absolute error of 0.560, and the highest R^2 score of 42.336, averaged over 9

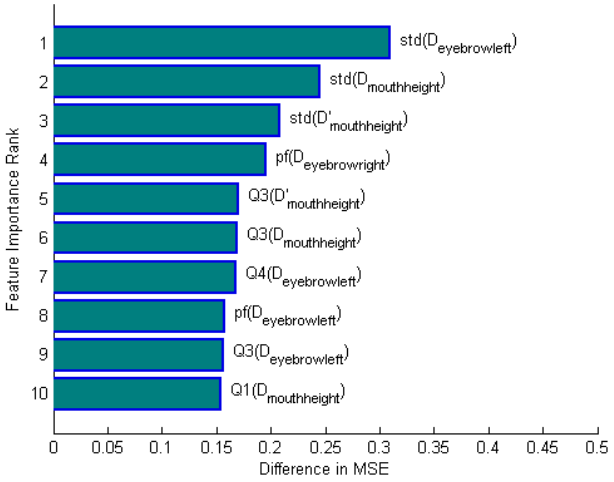

Figure 2: Bar graph displaying features with 10 highest feature importance scores. (std: standard deviation, pf: peak frequency, Q: quartile)

folds. The feature set based only on mouth shape dynamics perform worse on both measures.

For the MS+ES feature set, we computed the variable importance estimates, averaged over all folds, and sorted them. The ten features with the highest importance scores are shown in Figure 2. The average difference in MSE before and after randomly permuting this set of features is the highest among all features. We can observe that different attributes of $\mathbf{D}_{\text{eyebrowleft}}$, $\mathbf{D}_{\text{eyebrowright}}$ and $\mathbf{D}_{\text{mouthheight}}$ distance vectors populate this list, indicating their importance to the regression task.

5. CONCLUSION

We have presented a random forest regression framework for the problem of predicting active expressivity of the face. The method consists of first detecting facial landmarks for a sequence of continuous frames from an input video and extracting geometric shape features. Each sample is represented by a feature vector computed from statistics of the geometric shape signals. We evaluated our framework on a dataset of 727 videos using 9-fold cross validation. From our analysis, the dynamics of the eyes and eyebrows are better predictors of facial expressivity than dynamics of the mouth. Additionally, we computed importance scores for each feature to provide insight into what geometric shape features are most important in this challenging prediction task.

One possible extension to this work is to build gender and culture-specific models in order to explore the extent of rater bias in measuring expressivity. Another interesting research topic is automating the evaluation of an individual's general expressivity. In addition to facial expressivity, other attributes, such as body posture and movement as well as voice and prosody features, could be included in measuring the expressivity of a person during communication.

6. ACKNOWLEDGMENTS

The work was supported in part by NSF (1337866). We acknowledge Drs. Sarah Gunnery, Linda Tickle-Degnen, Theresa Ellis, and Stan Sclaroff for their helpful input.

7. REFERENCES

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[4] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.

[5] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual analysis of humans*, pages 377–409. Springer, 2011.

[6] P. Ekman and W. V. Friesen. *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[7] K. D. Lyons and L. Tickle-Degnen. Reliability and validity of a videotape method to describe expressive behavior in persons with Parkinson's disease. *American Journal of Occupational Therapy*, 59(1):41–49, 2005.

[8] M. Pantic and L. J. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[9] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[10] L. Tickle-Degnen. The interpersonal communication rating protocol: A manual for measuring individual expressive behavior. Technical report, Tufts University, 2010.

[11] L. Tickle-Degnen, T. Ellis, M. H. Saint-Hilaire, C. A. Thomas, and R. C. Wagenaar. Self-management rehabilitation and health-related quality of life in parkinson's disease: A randomized controlled trial. *Movement Disorders*, 25(2):194–204, 2010.

[12] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Computer Vision and Pattern Recognition Workshop. CVPRW'06*, pages 149–149, 2006.

[13] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of neuroscience methods*, 168(1):224–238, 2008.

[14] P. Wu, I. Gonzalez, G. Patsis, D. Jiang, H. Sahli, E. Kerckhofs, and M. Vandekerckhove. Objectifying facial expressivity assessment of Parkinson's patients: Preliminary study. *Computational and Mathematical Methods in Medicine*, ID 427826, 2014. 12 pages.