Express Recognition

Exploring Methods of Emotion Detection

Yasemin Ersoy
yasemin.ersoy@ti.com
Stanford CS229, Autumn 2013

***Abstract*— Computer, robotic and mobile interfaces are beginning to use expression recognition to give a more human experience. To make an interface more dynamic and seamless to human interaction, understanding of emotions is key. A robust application to recognize certain facial expressions in real time has many obstacles starting from correctly identifying a face and extracting necessary features of the face to then mapping these features to the right expression. There has been a thorough study of how to extract key points on faces and OpenCV even provides this code freely [1]. A topic that has been over looked is how to best manipulate a minimal number of facial key points efficiently or even if facial key points are the fastest route to facial expression identification. The goal will be to compare different methods of extracting expressions and gain more intuition about this system to further improve how expressions are defined and modeled.**

I. Introduction

Facial expression recognition has been studied since the 1980's and has been spreading to a variety of fields such as psychology, art, robotics, and computer vision. Though initially studies targeted smaller subsets such as those with impaired understanding of emotions and lie detection, in our current technology these valuable expressions can be utilized to create a more fluid user interface. Cameras surround our world, and with the development of image processing even the lower resolution cameras on laptops or phones are enough to robustly extract the necessary features for expression recognition.

The majority of expression recognition research involves using a grand assembly of facial key points to mimic and understand the entire face: i.e. representing the eyebrow using at least five points to digitally visualize the curvature. My goal is to utilize fewer points and still capture the expression. The 12 points that will be examined are: top center of face, middle of eyebrows, sides of the eyes, tip of the nose, corners of the mouth, center bottom of the bottom lip, and the tip of the chin (Figure 1). Since the focus will be on the expression side of events, the algorithm will be inputted features directly. Unfortunately this means manual feature extraction which is time consuming.

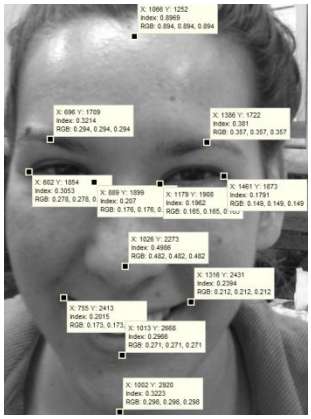

Figure 1: Key Points

II. Understanding Emotions

Artists refer to triangles of shadowing and warping these triangular areas surrounding the eyes nose and mouth to reflect emotions in drawings, as Pamela Davis at Stanford University explains. These triangles can also be viewed as angles between different corners of the face such as between the chin and the corners of the mouth. The happy, sad, angry, and surprised emotions have been studied in the greatest detail so these four emotions will be tested and detected along with the neutral face. The lower face angles surrounding the mouth are more relevant indicators for happiness and sadness and the upper face angles regarding the eyebrows are more prominent in detecting anger and surprise [2].

Using angles rather than distances (Figure 3, 4) allows different sized faces and rotated faces to be treated the same and is a novel and speed optimized method based on past expression detection methods which look at the direction of face flow, distance of features, and relative facial key point location training [2, 3, 4]. If angles are not used then normalizing the features using top and bottom of face is crucial.

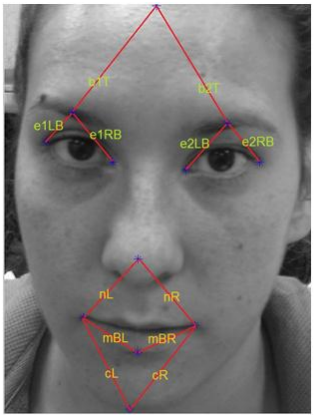

Figure 2: Neutral Face Angles

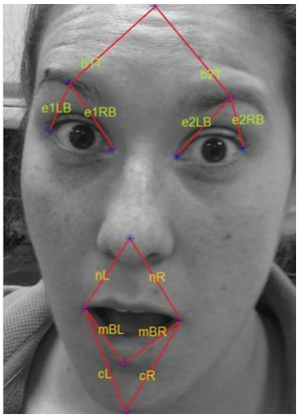

Figure 3: Surprised Face Angles

The neutral face is an important foundation when examining expressions. The variation based on a neutral face on each subject rather than a global neutral face is much smaller, so it is best to calibrate each user with his/her own neutral face then look at the distance of the expression face from the relative neutral point to detect the expression from this difference. Plus each person's neutral face is slightly different depending on the shape of their facial features so this method makes up for facial variations in users. Below is an

example of expression angles of a single subject (Figure 4) and variation of 5 different subjects (Figure 5).

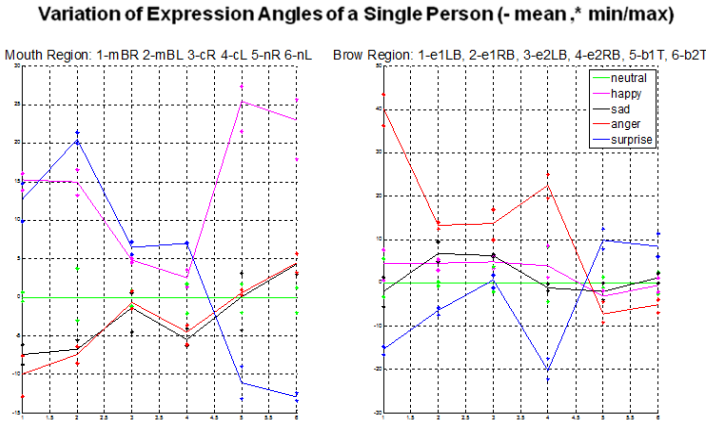

Figure 4: Single Subject Variation

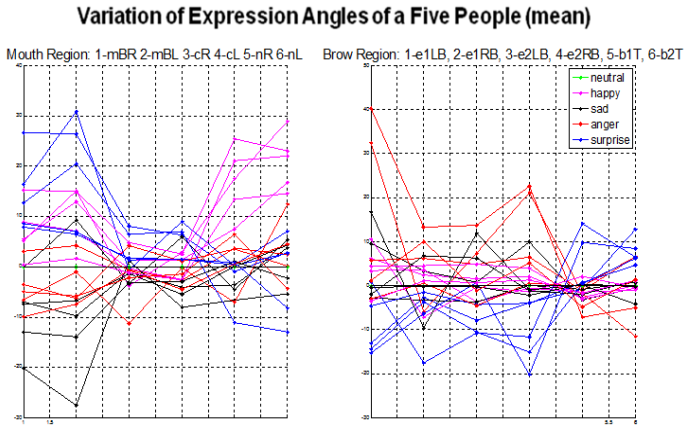

Figure 5: Five Subject Variation

Using the correlation of the current face to the neutral face, a reliable model can be created for expression recognition.

III. Modeling and Machine Learning

A. *Brute-force Modeling*

Since angles between features appear to be more stable reference points than exact locations, it is best to get a deeper understanding of the expression model using these angles to create an initial algorithm. Based on the troubles or observations of this initial "brute-force" algorithm the best method of continuing with machine learning can be chosen.

The thresholds for each expression were estimated by experimenting using a collected database of faces (examples in Figure 6). The most prominent angles for each expression are given the heaviest weighting for that expression, where prominent is defined as the least overlapping points with another expression and greatest change from neutral. The collected data showed that combinations of angles are also useful identifiers such as the first 4 angles in the lower face of surprise being positively different with respect to neutral while the last two are negatively placed (Figure 4).

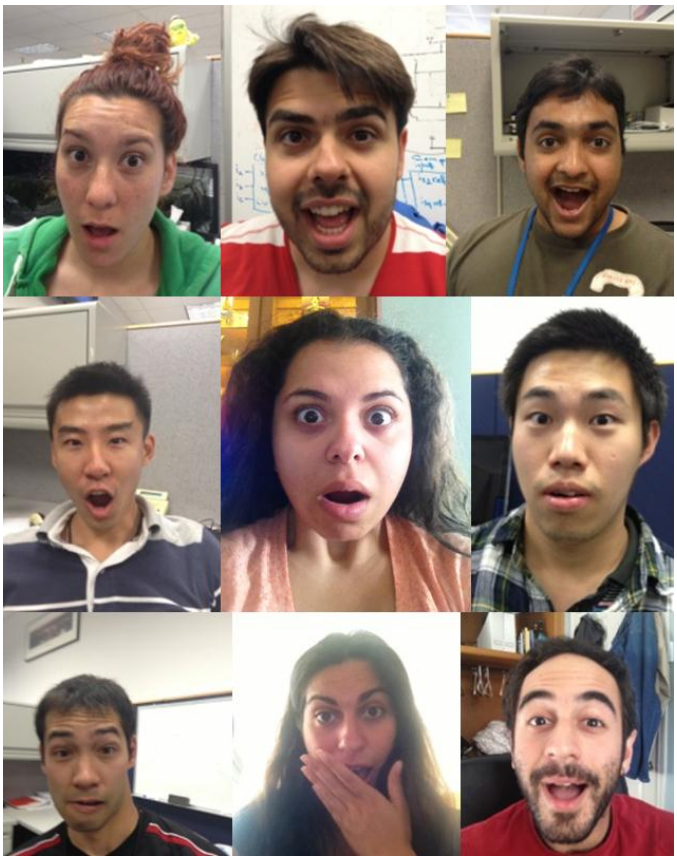

Figure 6: Sample Surprised Subjects

Given the face angles and using the calculated thresholds for each expression, facial expression at each instance is calculated as follows:

1) Add chin and center top of forehead to key points using face bounding box, eyebrow center and lip corner keypoints.

2) Calculate facial expresion angle using all key points

3) Check if past expression was inputed

- Yes: If past angles are within a threshold to equal new input, expression has not changed - output past expression
- No: Continue to face detection

4) Check if calibrating

- Yes: Record current expression angles as neutral (Figure 6) and check if previous average neutral angles were imported: if so average the two neutral angle sets and output as new neutral expression angle set
- No: Continue to face detection

5) Detect expression

a) Upper Face

a. Compare upper face thresholds to calculated emotion thresholds of eyebrows

b. If anger and surprise are detected in upper face give a higher weight

c. Check if movement is in all negative or all positive direction

b) Lower Face
   a. Weigh mouth corners to nose heavily for happiness
   b. Weigh mouth corners to bottom lip middle heavily for sadness

c) Check for angle thresholds for surprise and If the maximum weighted expression is not strong enough label face neutral

d) If contrasting emotions are detected (ie angry eyebrows with smiling face) label as confused

e) If two emotions are equally rated, which happens in anger and sadness quite often, further differentiate the two expressions
   a. Shape of eyebrow movement relative to eyes and top of forehead to differentiate sadness and anger
   b. General direction of mouth relative to nose and chin to differentiate happiness from surprise and sadness from anger

6) Output detected expression along with the probability of detecting each expression.

Using the estimated algorithm above a small sample set of nine separate collections were tested to give the results shown in Table I, indicating that sadness is the most difficult emotion to detect.

TABLE I. INITIAL EXPRESSION RECOGNITION ACCURACY

Detected Emotion	**Actual Emotion**			
	Happy	***Sad***	***Anger***	***Surprise***
Happy	**100%**	0	0	0
Sad	0	**65%**	12.5%	0
Anger	0	20%	**75%**	0
Surprise	0	0	0	**100%**
Confused or Neutral	0	15%	12.5%	0

B. Machine Learning

The sample algorithm that was used to test the dynamics of the expression model is not a convex model to optimize so it is best to start with simplifying the recognition model. The input vector can be considered as the distance from neutral normalized to face size (24 points for x and y coordinates) or the difference of the 12 chosen angles between features from the neutral angles; the output will be a classification of neutral, happy, sad, angry, or surprise.

$$\theta_d x_d = y$$

$$\theta_a x_a = y$$

Where m is the number of inputs, x_d is the matrix of normalized distance of the features from the neutral of size m by 24, and θ_d are to be estimated to capture the outputs y, a size m vector indicating the expression; x_a is the matrix of the distance of the angles from the neutral angles of size m by 12 and θ_a are to be estimated to capture output classification y.

The two input methods above can be tested using different classification methods to see the most robust system. Since the data is sparse, Naïve Bayes should give intuitive results, Discriminant Analysis could capture the characteristics of the limited data input more accurately for better results, and K-nearest neighbor can give a localized estimation of the data since there are overlapping emotions that we do not accommodate for in the simplified algorithm. The features are not independent so the assumption is that Naïve Bayes performs the worst; 1, 2, and 3 nearest neighbors are checked for K-nearest neighbor and give surprisingly similar results most likely due to the need for further depth in the training set. In general, the prediction before testing was that angles would give better results than distances, and K-nearest neighbors would give superior fitting in comparison to Discriminant Analysis which in turn would be more accurate than Naïve Bayes.

IV. RESULTS

The data used to test the best fit and input method was derived from 11 separate subjects creating the four proposed expressions plus neutral. The features were recorded manually and multiple collections per subject were made to test variation as shown in Figure 4. The mean variance of angles was 1.5° and the average variance for normalized face distances was 2%. Using this information, a larger pool of data was created by varying each subject output angles by ±1.5° which created 3 possible inputs from a single collection (ϕ, $\phi+\Delta$, $\phi-\Delta$) and each x and y coordinate for distance by 2% creating 7 possible inputs from a single collection ([x,y], [x+Δ,y], [x,y+ Δ], [x-Δ,y], [x,y+-Δ], [x+Δ,y+Δ], [x- Δ,y- Δ],). Therefore the total data set for distance training was 11 subjects with 7 variations and 5 separate faces: 385 with input vectors of size 24 and for angle training only 3 variations: 165 with input vectors of size 12.

TABLE II. INITIAL CLASSIFICATION TESTING

Method	**Error Rate**	
	Distance	***Angle***
Brute	N/A	29.7%
Naive	0%	6.7%
Discriminant	0%	6.7%
K-nearest	0%	0%

Initial results were just to see how well the three classification methods fit the entire data set and also to see how the brute-force method reacted to the variations, displayed above in Table II. Next, the training set was decreased by 1/3 of the data (used for testing) with 40 separate combinations of testing and training data. The training set and the testing set were then made of equal size again with 40 separate random combinations. Finally the training set was dropped to 1/3 of the data and 2/3 of the data was used for testing, again with 40 combinations of training and predicting.

TABLE III. CLASSIFICATION TESTING TRAIN:TEST RATIO

Method	Error Rate					
	Distance 2:1	***Angle 2:1***	***Distance 1:1***	***Angle 1:1***	***Distance 1:2***	***Angle 1:2***
Naïve Min	0%	5.45%	0%	6.1%	0.39%	6.36%
Discr. Min	0%	5.45%	0%	6.1%	0.39%	6.36%
K-near Min	0%	0%	0%	0%	1.82%	1.21%
Naïve Max	9.38%	23.64%	10.42%	24.39%	14.84%	24.55%
Discr. Max	9.38%	23.64%	10.42%	24.39%	14.84%	24.55%
K-near Max	1.82%	3.64%	3.64%	4.85%	11.17%	9.7%
Naïve Mean	*3.46%*	*12.77%*	*4.31%*	*12.93%*	*6.29%*	*16.4%*
Discr. Mean	*3.46%*	*12.77%*	*4.31%*	*12.93%*	*6.29%*	*16.4%*
K-near Mean	*0.47%*	*0.38%*	*2.05%*	*1.77%*	*6.28%*	*4.71%*

The simplified clustering results are shown in Table III. As expected, K-nearest neighbor performed the best. However, Naïve Bayes and Discriminant Analysis led to the same solution, which was not predicted. Plus surprisingly, for these two weaker solutions interpreting the feature points with normalized distances created a better fit for the data. Adding further rotational variations to the key points would cause the need for a larger training set for distance measurements and create errors that the angle method can bypass. K-nearest neighbor not only proved superior results but also was efficiently able to use the angle training set. Even by using only a third of the data to train the maximum error was still below 10% and the mean below 5%. The only problem with this method is that it is memory based: the more data that is stored the more accurate the system will (be up to a plateau).

Overall, if memory is limited and a quick solution is desired then Naïve Bayes is optimal, but if memory is available and robust predictions are preferred then K-nearest neighbor is definitely the right choice.

V. CONCLUSION AND FUTURE WORK

Expression detection is certainly achievable by means of a few number of key points and can handle slight variations in the key point locations. Without using a large database, a K-nearest neighbor system can be quickly trained to calculate predictions using angles of these facial features which allows for a distance and rotation robust algorithm.

The next step for expression recognition is in two directions. One problem is calibrating to a neutral face and another is facial feature detection accuracy, which drops dramatically in poor lighting. A generalized neutral face can be determined using an extended database but this would cause an increase in false detection because of the variation of natural facial features. Removing facial features for emotion detection all together can be a solution to both of these problems.

The face can be simplified even further than the minimal facial key points by local averaging. All humans have eyes, nose and mouth located in relatively similar locations. For expression recognition using feature extraction precise locations of these features are vital- but if gradients of the face were taken then the general trend of the face can be captured: a crude example shown in Figure 7. The averaged shades of the face can then be used to train a new system that can calculate thresholds for emotion detection.

Expression recognition has plenty of room to grow in our ever increasingly technology intertwined world. By continuing research to model facial expressions effectively and discover efficient solutions with the help of machine learning, soon the line between cold electronics and warm emotional beings will begin to blend.

ACKNOWLEDGMENT

Thanks to Pamela Davis for artistic expression modeling support.

REFERENCES

[1] M. Uricar, V. Franc and V. Hlavac, Detector of Facial Landmarks Learned by the Structured Output SVM, *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, 2012

[2] Neeta Sarode, Shalini Bhatia. Facial Expression Recognition. *International Journal on Computer Science and Engineering 2*, 2010, p. 1552-1557.

[3] Wu, Yuwen, Hong Liu, and Hongbin Zha. "Modeling facial expression space for recognition." *Intelligent Robots and Systems*, 2005.(IROS 2005).

[4] Yang, Y., S. Ge, et al. (2008). "Facial expression recognition and tracking for intelligent human-robot interaction." *Intelligent Service Robotics 1*(2): 143-157.

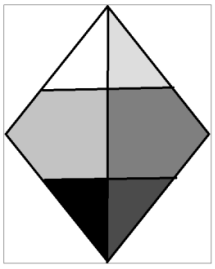

Figure 7: Facial Frame