Sahith Nallapareddy
12/7/18
Machine Learning

Classifying Facial Expression in Sign Language

The problem attempted to solve is classifying facial expressions in the context of sign language. Facial expressions are important to sign language as they help decide the grammar (can help determine context) and allow for disambiguation between phrases. These are called the Grammatical Facial Expressions. Facial expressions in sign language can indicate questions or different intonations like how we speak with differing tones for a question and statement. These are critical in adding modifiers to adjectives like "very important" and the same sign can have different meaning just based on facial expression. Facial grammar is an important part of sign language which extends the language and clarifies meaning. The data set used is from UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions . The data was captured via a Microsoft Kinect sensor with people performing Brazilian Sign Language. Then using a single frame, points are gathered for the face. Each image is labeled as well providing the response value in the classification problem. Two people were used in this experiment each with their own dataset. There are 27,000 data points and 100 features. The features are the placement of the features of the face in terms of coordinates. The x and y are in pixels of the Kinect camera for the frame. The z value is in millimeters which gives depth. There are 33 points describing the eyes, eyebrows, nose, mouth, and face contour. The response value is just whether facial expression was present meaning was the expression used in a grammatical context.

Dataset Exploration

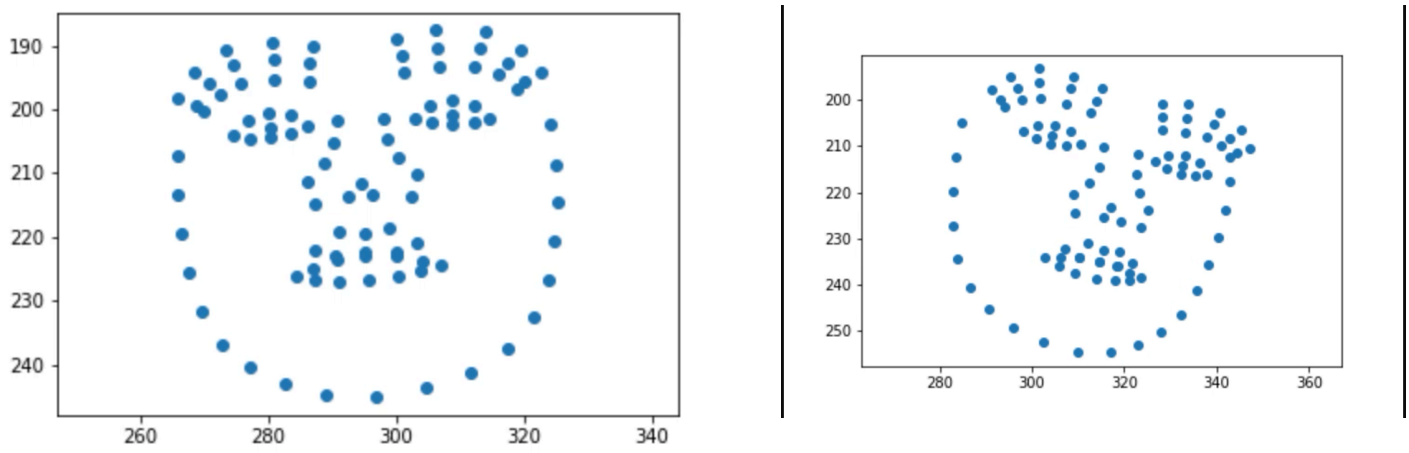
The first step in understanding the data was understanding the data points available. There were two people recorded so there is data for both. Then, the sentences said by either were signed same and trying to convey the same meaning (implying that the expressions should be similar). The types of expressions said were a wh question (who, what, where, when, and how question), yn question (yes and no question), doubt question (emphasize that information is required), topic (constitutes beginning of sentence), negation, emphasis, conditional clause, relative, and affirmative. This means there is a dataset for each of these questions and for each

person saying them meaning there are 18 datasets that could be individually used or combined depending on the model trying to be built. Below is a table of the breakdown for each dataset. In each experiment, the data was split 75-25 for training and testing set.

Dataset	Positive Frames	Negative Frames	Total Frames
a_affirmative	414	648	1062
a_conditional	548	1359	1907
a_doubt_question	491	821	1312
a_emphasis	330	1073	1403
a_negative	528	596	1124
a_relative	644	1686	2330
a_topics	360	1436	1796
a_wh_question	609	677	1286
a_yn_question	532	858	1390
b_affirmative	528	546	1074
b_conditional	589	1445	2034
b_doubt_question	780	717	1497
b_emphasis	531	813	1344
b_negative	712	870	1582
b_relative	550	1354	1904
b_topics	467	1358	1825
b_wh_question	549	779	1328
b_yn_question	715	1023	1738

Table 1: Positive, meaning the presence of EFG, and Negative frame breakdown for each dataset

An easy way to visualize the data was to graph the points to see how the face was moving. Since each data point is just a still frame, each frame can be plotted to get an image of the face. Using the x and y coordinates, it is possible to create a crude representation of the data. This is not an exact representation of the face because the z value gives depth so certain features such as rises and contours do not get represented well in this case. However, visually it is valuable to try to understand the expression through each frame. Gifs were created of the frames put together to see the progression of expressions.



Results – Binary

Using one person and the data points for just the affirmative dataset, a couple of classifiers were built to identify the presence of an expression. The classifiers that were trained are Decision Tree, SVM, and Random Forest. The results are shown below.

Model	Accuracy	Precision	Recall	Error
Random Forest	**92.90%**	**93.20%**	**86.40%**	**7.10%**
SVM	85.30%	85.70%	77.00%	14.70%
Decision Tree	86.50%	86.10%	79.80%	13.50%

Table 2: Metrics on models build using person A's affirmative dataset

The SVM was the worst with 81% accuracy and Decision Tree had 83%. To confirm the accuracy of the Random Forest model, cross validation was used to obtain an accuracy of 92% with standard deviation of 0.06. Building more Random Forest models on the other sets of data, namely affirmative, conditional, doubt, emphasis, negative, relative, topics, wh_question, yn_question, we get accuracies ranging from 90% to 97% and using cross validation we get accuracies averaging at 90%. Looking at the important features of the Random Forest model, we can rank the top 10 as such:

Rank	Feature	Importance Value
1	62y	0.049817

2	50y	0.048881
3	86y	0.04576
4	12y	0.045387
5	88y	0.043885
6	14y	0.042077
7	39y	0.039565
8	56y	0.039245
9	8y	0.039164
10	89y	0.039035

Table 3: Random Forest model important features ranking

All the features are y coordinates, meaning the up and down movement seemed the most important, and were concentrated around mouth and eyes. A feed forward neural network with 3 layers, a 300-node layer with sigmoid activation function, 200-node layer with relu activation function, and finally a 1 node soft max layer for predicting. With this architecture, this model was used to obtain an accuracy of 60%. Other similar architectures were tried with lower results.

Results – Multiclass

This dataset can also lend itself to a multiclass problem by joining all the datasets together. By taking all the positive classifications for a single person, and classifying them based on the phrase, the dataset would now have 9 classes, one for each phrase, for each data point. This was then tested on Random Forest and Decision Trees with the results shown below:

Model	Accuracy	CV Accuracy	Precision	Recall
Random Forest	**97.30%**	90.0%	**97.30%**	**97.30%**
Decision Tree	93.50%	**92.0%**	93.50%	93.80%

Table 4: Models made for multiclass classification problems on person A's data

Random Forest still performed very well on this dataset, but the cross-validation score was lower than Decision Trees. This could imply that the Random Forest over fit on the dataset and would not perform as well as the testing set indicates. Looking at the features this time the Random Forest model used we obtain this ranking:

Rank	Feature	Importance Value
1	45y	0.138088
2	76y	0.118896
3	68x	0.108831
4	44z	0.090481
5	77z	0.078817
6	38x	0.074893
7	41y	0.04553
8	80z	0.045471
9	99y	0.025123
10	92x	0.022713

Table 5: Ranking of important features from Random Forest Model

These features are more diverse than the last set. Most are still y coordinates and the features are concentrated around the nose, eyebrow, and mouth. A feed forward neural network was also tested, but the results were poor as accuracy was below 30%. Many different architectures were tried, but it seems that more data would be necessary for the model to perform better in this classification problem.

Results – Generalizing Model

Because there were two people recorded, there are two datasets per phrase. This gives rise to an interesting problem of predicting the presence of person B's facial expression using only data from person A. However, this is difficult as person A and person B have different facial features, meaning the placement of their eyes, nose, and eyebrows are in different places. This is understandable as if all humans had the same exact face it would be frightening. To solve this, the displacement was taken between two consecutive frames. This was done to try to capture a crude representation of the motion of the face rather than the features. Using this, a model was trained for binary classification on person A and tested against a testing set from A and the dataset from B.

Person	Accuracy	Precision	Recall
A (testing set)	81.60%	78.70%	71.80%
B	68.30%	79%	48.40%

Table 6: Random Forest Model metrics on person A, test data, and person B

The accuracy on person A dropped predictably from the Random Forest as this data is less defined as the points on the facial features were representing displacement. This means that there was less definite bounds that separated the two classes implying that the movement of facial features is harder to predict than still frames. A multiclass classification was also attempted but the results were even worse. Using a Random Forest model, the accuracy against A's testing set was 53% and then against B's data was 17%.

Conclusion

There was success in both the binary and multiclass classifications. In both cases, Random Forest and Decision Trees performed the best. They gave good indicators on what features were the most important as in the binary case the up and down motion of the mouth and eyes were valuable to the model. While in the multi class case, any motion related to nose and eyebrows were more favored. They had good metrics with cross validation, accuracy, and precision for both cases. However, neural networks did not bode well in this experiment. Testing different architectures, it was difficult to bring the accuracy up and the loss function usually did not reduce. The most likely cause was that there was not enough data in this case as for binary classification between 1000-2000 frames did not seem enough to train the model. In the multiclass situation, there were only 400-800 frames per class which was insufficient for the neural net to train on. Overall, the conclusion was that Random Forest was the best model in either case as it distinguished the best in accuracy, precision, and recall over all sets of data.