UPPSALA UNIVERSITET

Report for Intelligent Interactive Systems

Project
Emotion attribution from facial landmarks
Group 7

Nikhil Karthik Punnam Davide Rendina Vishnu Sharma Srivijay Manjunath
Kughan Krishnaswami

May 27, 2019

Abstract

The goal of this project is to predict one of the six basic emotions to a set of 3D facial landmark positions. The datasets used were from the Bosphorus Database that contains 3D face data under various poses, expressions, and occlusions. The various facial features or landmarks used were agreed upon with the two other groups namely Computer vision and Emotion synthesis. The outcome of the project is a model which predicts an emotion with an accuracy around 65 %.

1 Introduction

The objective of our project is, given a set of facial landmarks in a 3D environment, to assign to the given input one of the six basic emotions which include: happiness, disgust, fear, surprise, anger and sadness.
After training we expect our system to be able to recognize the six type of emotions and output a label with an associated probability score.
During the first phase of the project we worked together to plan our work and pre-process the given data in order to extract the most relevant features.
Once we selected the features to train our models on, we chose five classifiers and each of us worked on one of them, we also kept updating each other on the performances
The chosen classifiers include : *K-Nearest Neighbors* (Kughan Krishnaswami), *Support Vector Machines* (Davide Rendina), *Random Forest* (Srivijay Manjunath), *XGBoost* (Nikhil Karthik Punnam) and *Multilayer Perceptron* (Vishnu Sharma) .
Each one of us worked on training one of the chosen classifier, testing them with different parameters and gathering the results.
We first experimented on the data to fit and were unsuccessful in getting good results by using just the co-ordinates of the landmarks and then used the distances between the landmarks using the coordinates as features and it has resulted in a better performance. We used a voting method to combine good performing models to predict the final emotion.

We also worked with the Computer Vision (group 1) and Emotion Synthesis (group 8) groups in order to co-operate and develop a pipeline, where the CV group would provide the input features on which our system would then output a labelled emotion to send to the ES group.

2 Methodology

2.1 Feature Engineering

In the dataset provided to us, 22 facial feature co-ordinates were marked. We made use of only the most important co-ordinates which were most important for each emotional expression. The following are the set of co-ordinates which contributes the most to predict an expression:

- Happiness- 'Left mouth corner', 'Right mouth corner','Upper lip outer middle', 'Lower lip outer middle'.

- Surprise- 'Left mouth corner', 'Right mouth corner','Middle left eyebrow','Middle right eyebrow','Chin middle'.
- Fear-'Left mouth corner', 'Right mouth corner','Inner left eyebrow','Inner left eye corner'. 'Inner right eyebrow', 'Inner right eye corner"Outer left eyebrow', 'Middle left eyebrow'.
- Sadness-'Left mouth corner', 'Right mouth corner','middle right eyebrows','middle left eyebrows'.
- Anger-'Left mouth corner', 'Right mouth corner',left eyebrow', 'Middle left eyebrow','middle right eyebrow','inner right eyebrow','middle right eyebrow'.
- Disgust-'Left mouth corner', 'Right mouth corner','middle right eyebrow','middle left eyebrow'.

TODO

2.2 Landmarks distances

We chose 15 most important landmarks from these for training our model on. We decided to use distances between the landmarks using their 3-D coordinates instead of directly using the coordinates [1]. Following is the list of pairs of landmarks we have decided to compute distance between,
('Outer left eyebrow', 'Middle left eyebrow'), ('Outer left eyebrow', 'Inner left eyebrow'), ('Inner left eyebrow', 'Middle left eyebrow'), ('Inner right eyebrow', 'Middle right eyebrow'), ('Outer right eyebrow', 'Inner right eyebrow'), ('Outer right eyebrow', 'Middle right eyebrow'), ('Middle left eyebrow', 'Middle right eyebrow'), ('Outer left eyebrow', 'Outer left eye corner'), ('Outer right eyebrow', 'Outer right eye corner'), ('Inner left eyebrow','Inner left eye corner'), ('Inner right eyebrow', 'Inner right eye corner'), ('Left mouth corner', 'Right mouth corner'), ('Left mouth corner', 'Lower lip outer middle'), ('Left mouth corner', 'Outer left eye corner'), ('Left mouth corner', 'Chin middle'), ('Right mouth corner', 'Lower lip outer middle'), ('Right mouth corner', 'Outer right eye corner'), ('Right mouth corner', 'Chin middle'), ('Upper lip outer middle', 'Lower lip outer middle'), ('Chin middle', 'Middle left eyebrow') and ('Chin middle', 'Middle right eyebrow').

2.3 Model Training and Parameter Tuning

Once we computed the distance between the landmarks, we split the data into train set and test set with test set being 20 % of the whole dataset. We then scaled the dataset [2]. We experimented on K-Nearest Neighbors and Support Vector Machine as an extension of Lab 2 and then tried our hand at RandomForest [3], XGBoost [4] and Multi-layer Perceptron [5].

To decide which parameters perform best given the training data, we used Scikit-Learn's GridSearchCV to search through the parameter grid to find the parameter that fit the training data best. We then initialized our models using those parameters and predicted on the test dataset.

After training our data on 5 different models, we used one of Scikit-Learn's Ensemble methods to provide better inference when predicting unknown data. VotingClassifier of Scikit-Learn helps to combine multiple machine learning models by averaging or by a majority vote of target probabilties to better predict the output by eliminating the individual models' weaknesses.

3 Results

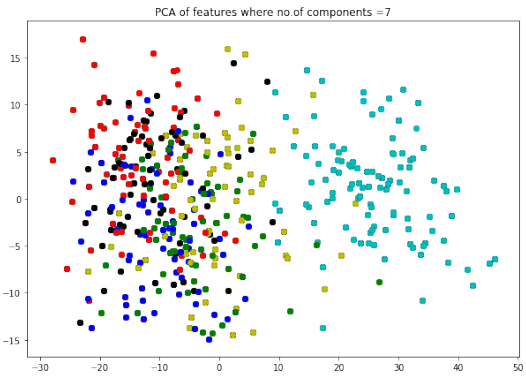

Figure 1: PCA Plot of features where no .of Components were 7

Legend for the plot is Happiness, Sadness, Disgust, Fear, Surprise, Anger .
As we can see from Figure 1, datapoints from happiness are easily separable from other classes while datapoints from disgust are partially separable from others.

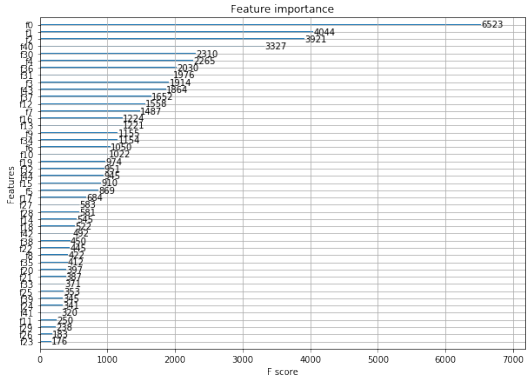

Figure 2: Plot showing feature importance for the model XGB

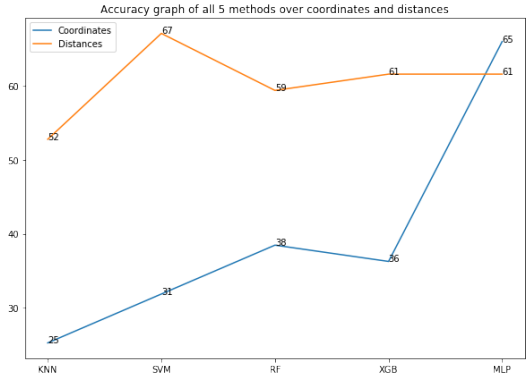

Figure 3: Graph showing accuracy of different models

Figure 2 shows us the importance of each feature for the model XGBoost while figure 3 tells the varied accuracy of different models over different data. We tried fitting the coordinates directly into the model and then computed distances into the model. We achieved the highest accuracy with SVM with sigmoidal kernel. By using Scikit-Learn's voting method, we used XGBoost, Multi-layer perceptron and SVM for final voting ensemble method with weights of 1,1 and 3 respectively and acheived an accuracy of 64.83 % .

	precision	recall	f1-score	support
SURPRISE	0.73	0.73	0.73	15
ANGER	0.50	0.64	0.56	14
HAPPINESS	0.91	0.95	0.93	21
SADNESS	0.50	0.69	0.58	13
DISGUST	0.57	0.39	0.38	14
FEAR	0.55	0.43	0.48	14
ACCURACY			0.65	91
MACRO AVG	0.63	0.62	0.61	91
WEIGHTED AVG	0.65	0.65	0.64	91

Table 1: Classification Report for weighted voting method

Table 2 gives us a confusion matrix for our weighted voting method.

		PREDICTED					
		SURPRISE	ANGER	HAPPINESS	SADNESS	DISGUST	FEAR
ACTUAL	SURPRISE	11	0	0	0	1	3
	ANGER	1	9	0	1	1	2
	HAPPINESS	0	0	20	0	1	0
	SADNESS	0	4	0	9	0	0
	DISGUST	0	3	2	5	4	0
	FEAR	3	2	0	3	0	6

Table 2: Confusion Matrix for weighted voting method

4 Discussion

4.1 Challenges faced

We were unable to get good performance from models when trained directly on the coordinates of the landmarks itself. Also, the data did not show any clear pattern or clusters in it. We then decided to find the most frequent or highly used landmarks for each emotion and compute the distance between them to be used as features to the model [6] . Using the distances between the landmarks rather than the landmarks themselves, we were able to visualize the clusters present in the dataset for different emotions with clear distinctions. Since, the dataset was exhibiting clear patterns, it was ready to be fed to machine learning models. Next challenge was to fine tune the parameters for these models so as to get the best performance out of them.
Also, the time taken to fine tune the parameters using GridSearchCV was too long hence it made difficult to test and train the models on the dataset.

4.2 Integration in the final pipeline

There has not been any major issues with the final integration in the pipeline. The Computer Vision group provided us with all the features agreed upon and we obtained satisfactory results. The only problem we have encountered is in the format the features were delivered. In fact, while we processed our data in a 3D environment, the Computer Vision processed the features in 2D. To solve the issue, in the pipeline implementation the CV group set the value for the z-axis to 0 so that the features would be sent in 3D. Once, our model is trained, we dump it long with the scaler to serialized object files. Computer Vision group provides us inputs by printing it on standard output. We read these inputs from the standard output, parse them into the desired format. Then, we load our model and scaler from the object files saved earlier. Using scaler, we scale the inputs and feed them to the model, which in turn predicts the emotion. Finally, we print the emotion over standard output, which is then read by the emotion synthesis group. In a sense, computer vision group's code and our code run as a sub-process for the emotion synthesis group.

4.3 Ethical issues

The bosporous database was provided by the Boğaziçi University. This dataset consists of various emotions portrayed by 105 people, in-order for us to train our model. The dataset was provided to us on an accordance to not to share it with the outside world as a whole or even a part of it. It is very important as a student and as a team to respect the privacy and use the dataset wisely.

5 Conclusion

The final model predicts the emotion with an accuracy of 64.83 %. While this is not the result we were hoping for, accuracy can be improved by better understanding the data and enhance the level of feature engineering. With more amount of data samples for training, one can implement a neural network model to fit the data better in hopes of getting an accurate prediction of emotions.

References

[1] Deepak Ghimire, SungHwan Jeong, Joonwhoan Lee, and Sang Hyun Park. Facial expression recognition based on local region specific features and support vector machines. *CoRR*, abs/1604.04337, 2016.

[2] Lennart Eriksson, Nouna Kettaneh-Wold, Johan Trygg, Conny Wikström, and Svante Wold. *Multi- and Megavariate Data Analysis : Part I: Basic Principles and Applications*. Umetrics Inc, 2006.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[5] H. Ramchoun, M. A. Janati Idrissi, Y. Ghanou, and M. Ettaouil. Multilayer perceptron: Architecture optimization and training with mixed activation functions. pages 71:1–71:6, 2017.

[6] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *CoRR*, abs/1604.03225, 2016.

6 Appendix

6.1 Objectives

The aim of our project is to build a system, which is able to recognize emotions. The input would be an image of one individual showing some type of emotion, given the image the system should be able to identify the emotion (given the set of emotions on which the system was trained). The output of the system would be an emotion and its associated probability scores.

6.2 Methods

Step 1 Data Preprocessing: Gather the relevant images along with 22 facial features and 25 AU Sets. Once we have gathered them, we try and clean/refine the data into useful features and extract the important features from them.

Step 2 Model Training: Given the data from the previous step, explore different models which can fit the data well. The models that will be trained will be KNN and SVM (which will be taken from the lab 1). In addition, we will implement a RandomForest classifier and a Feed Forward Neural Network based on ResNet or AlexNet. Once we find a model which fits the data with a high enough accuracy, we aim to tweak the model to produce better results.

Step 3 Evaluate the model and discuss and explain the obtained results and ethical implications of the system. We plan to user accuracy, F1-score and precision values of each model to evaluate the system.

6.3 Timeline

For the whole duration of the project we are planning to meet 2-3 times per week and do most of the work together. In doing so, since we are coming from different background, we aim to get the most of it.

Week 18

After the feedback session, we will start working on step 1. Preprocessing the data in order extract important features and feed it to the models.

Week 19

(Second Feedback session) By week 19, we are planning to start training the models. We will start from the two models used in Lab 1 (KNN and SVM) in order to have them ready for the feedback session. Finally, move to the two additional model chosen for this project.

Week 20

The first half of the week will be spent on optimizing the chosen model. After that, we will start evaluating the models. Finally, we will provide an interface to the Emotion Synthesis group.

Week 21

Week 21 will be mainly used to refine the code and prepare the presentation. We are also leaving some time in case we have some delays during previous weeks.

Week 21

(Presentation) Submitting project and report (28/5) and presentation (29/5).