**MAIS 202 - Final Project Deliverable 1**

The basic idea of my final project is to create a software that can tell, given the faces of two individuals, if these two individuals belong to the same family tree or not (i.e. if they are related or not).

For the dataset, I choose one from Kaggle called Families in the Wild (FIW) Database. This dataset perfectly fits the requisites of the kind of dataset needed for my final project. First, it provides a training dataset where the data is divided into a category “Family”. Inside “Family”, there is another category, “Individuals”. Images from the same folder in “Individual” represent the same person and images from the same folder in “Family” belong to the same family. There is also a csv file containing the training labels, that is a file that lists the relationship between two members of the same family. It contains a test data containing face images of unknown individuals. Moreover, there is a sample file containing pairs of individuals associated to number 0 or 1 where 0 means unrelated and 1 means related.

I think the dataset I chose is very feasible since it has already been preprocessed for the purpose of my project by separating individuals into family members. Moreover, FIW contains 11 932 family photos that span 1000 different families, with the data distributed evenly. There is an average of about 12 images per family and the number of family members rangers from 3 to 38. In addition, there are families from different ethnicities i.e. Caucasian, Spanish/Latino, Asian, African/AA, Arabic and Mix. However, the distribution of these ethnicities is not even. Nevertheless, I don’t think it will be fatal to training an unbiased model. Moreover, the training set contains a total of eleven types of relationship which would make the model’s predictions more accurate: father-daughter (F-D), father-son (F-S), grandfather-granddaughter (GF-GD), grandfather-grandson (GF-GS), grandmother-granddaughter (GM-GD), grandmother-grandson (GM-GS), mother-daughter (M-D), mother-son (M-S), brother-sister/mixed gender siblings (SIBS), brother-brother (B-B) and sister-sister (S-S).

I want to predict if two individuals selected randomly are related to each other. I think using a convolutional neural network algorithm will be much easier to handle the large and complex data that I have since it uses relatively little pre-processing compared to the other image classification algorithms. Moreover, there is also the method using eigenfaces. However, one of the cons of implementing eigenfaces is that there is a lot of restrictions on the training set of images: the images has to be taken under the same lighting and normalized to have the eyes and mouths aligned across all images. The pros of having such restrictions might be that the algorithm would be easier to implement.

To integrate my model, I would like to build a webapp which showcases what my model does. It could take as input two pictures of different individuals and it could predict if the two individuals belong to the same family tree.

On the website of the database, it says that their top performing benchmark scored an average accuracy of about 71% (SMILE lab). I would really like to beat this score, but I would be more than happy to match with it.

Citations

SMILE lab, “Families In The Wild (FIW) Database*”, Families In the Wild: A Kinship*

*Recognition Benchmark*, <https://web.northeastern.edu/smilelab/fiw/database.html>