Face Identification Project

By:
Veeraraju Elluru (B22CS080), Souvik Maji (B22CS089), Pranjal Tiwari(B22CS095), Parth Mina(B22EE074), Ram Suthar(B22EE081)

1. Introduciton

Labeled Faces in the Wild is a public benchmark for face verification, also known as pair matching. The goal of our project is to utilize extracted features from the dataset and bring about the best possible results for the classification task.

2. Dataset Description

2.1 General Overview

- The data set contains more than 13,000 images of the faces of 5749 people collected from the web (13233 to be precise). Each face has been labeled with the name of the person pictured. Out of the 5749 people, 1680 of them have two or more distinct photos in the data set. The only constraint on these faces is that they were detected by the Viola-Jones face detector, using *openCV* library. In the original dataset itself, there are six incorrectly labeled matched pairs.
- Further, there are some other errata too, like shuffling of first and last names. Examples: Wang_Yingfan and Yingfan_Wang are the same person, Wang_Nan and Nan_Wang are the same person. Talisa_Bratt and Talisa_Soto are also the same person, even though the names are completely different, and so on. hence, the dataset is not perfect and users have to be conscious.

2.2 Image Information

- Image file format: Each image is available as *lfw/name/name_xxxx.jpg*. Here *xxxx* represents the sample number of the person with name as *name*.
- Every image is a $(250, 250)$ sized jpg file, with color channels = 1 (greyscale images).

2.3 Folder structure and contents

- pairs.csv: Contains randomly generated splits for 10-fold cross-validation specifically for pairs, to be used for image-restricted configuration when forming training sets.
- people.csv: Contains randomly generated splits for 10-fold cross-validation specifically for individual faces, to be used for image-unrestricted configuration when forming training sets.

- matchpairsDevTest.csv: Contains 500 matched pairs of faces for the testing set. For *pairs* configuration.
- matchpairsDevTrain.csv: Contains 1100 matched pairs of faces for the training set. For *pairs* configuration.
- mismatchpairsDevTest.csv: Contains 500 mismatched pairs of faces for the testing set. For *pairs* configuration.
- mismatchpairsDevTrain.csv: Contains 1100 mismatched pairs of faces for the training set. For *pairs* configuration.
- peopleDevTest.csv: Contains 1711 people and 3708 images. For *people* configurations
- peopleDevTrain.csv: Contains 4038 people and 9525 images. For *people* configuration.

3. Early Results

The features extracted from the boiler-plate code were concatenated and fit into all the classifiers used. The feature extraction procedure was done using 3 methods - CNN, HoG, LBP. The classifiers are - kNN, Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Multi-layer Perceptron. This improves the learning of algorithm and reduces the extent of over-fitting. The early results obtained for every classifier are illustrated below. However, it must be noted that no hyperparameter-tuning has been performed yet. This is our major task for the remainder of the project.

KNN Classifier
Accuracy: 0.3643410852713178
Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	0.17	0.10	0.12	21
Colin_Powell	0.23	0.37	0.28	41
Donald_Rumsfeld	0.12	0.10	0.11	20
George_W_Bush	0.51	0.60	0.55	114
Gerhard_Schroeder	0.18	0.13	0.15	23
Hugo_Chavez	0.00	0.00	0.00	10
Tony_Blair	0.40	0.14	0.21	29
accuracy			0.36	258
macro avg	0.23	0.20	0.20	258
weighted avg	0.34	0.36	0.34	258

Figure 1: Classification Report for kNN classifier

Logistic Regression Classifier
Accuracy: 0.42248062015503873
Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	0.50	0.19	0.28	21
Colin_Powell	0.23	0.32	0.27	41
Donald_Rumsfeld	0.10	0.05	0.07	20
George_W_Bush	0.54	0.77	0.63	114
Gerhard_Schroeder	0.00	0.00	0.00	23
Hugo_Chavez	0.00	0.00	0.00	10
Tony_Blair	0.20	0.10	0.14	29
accuracy			0.42	258
macro avg	0.22	0.20	0.20	258
weighted avg	0.34	0.42	0.36	258

Figure 2: Classification Report for Logistic Regression classifier

Decision Tree Classifier
Accuracy: 0.5736434108527132
Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	0.71	0.48	0.57	21
Colin_Powell	0.53	0.51	0.52	41
Donald_Rumsfeld	0.36	0.45	0.40	20
George_W_Bush	0.74	0.70	0.72	114
Gerhard_Schroeder	0.50	0.48	0.49	23
Hugo_Chavez	0.07	0.10	0.08	10
Tony_Blair	0.47	0.55	0.51	29
accuracy			0.57	258
macro avg	0.48	0.47	0.47	258
weighted avg	0.60	0.57	0.58	258

Figure 3: Classification Report for Decision Tree classifier

4. Future Goals

We plan to do the following:

- Understand in greater depth which set of features are to be used and how. Consider pair-wise features (all possible pairs from CNN, HoG, LBP) and also individual features separately.
- Implement few-shot learning for feature extraction.
- Perform rigorous hyperparameter tuning to get the best results for every classifier.

Random Forest Classifier
Accuracy: 0.6976744186046512
Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	0.00	0.00	0.00	21
Colin_Powell	0.90	0.93	0.92	41
Donald_Rumsfeld	1.00	0.45	0.62	20
George_W_Bush	0.61	1.00	0.75	114
Gerhard_Schroeder	1.00	0.26	0.41	23
Hugo_Chavez	0.00	0.00	0.00	10
Tony_Blair	1.00	0.45	0.62	29
accuracy			0.70	258
macro avg	0.64	0.44	0.47	258
weighted avg	0.69	0.70	0.63	258

Figure 4: Classification Report for Random Forest Classifier

Naive Bayes Classifier
Accuracy: 0.46124031007751937
Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	0.72	0.62	0.67	21
Colin_Powell	0.45	0.68	0.54	41
Donald_Rumsfeld	0.21	1.00	0.34	20
George_W_Bush	1.00	0.27	0.43	114
Gerhard_Schroeder	0.80	0.35	0.48	23
Hugo_Chavez	1.00	0.30	0.46	10
Tony_Blair	0.42	0.55	0.48	29
accuracy			0.46	258
macro avg	0.66	0.54	0.49	258
weighted avg	0.75	0.46	0.47	258

Figure 5: Classification Report for Naive Bayes classifier

KNN Classifier
Accuracy: 0.3643410852713178
Classification Report:

	precision	recall	f1-score	support
Ariel_Sharon	0.17	0.10	0.12	21
Colin_Powell	0.23	0.37	0.28	41
Donald_Rumsfeld	0.12	0.10	0.11	20
George_W_Bush	0.51	0.60	0.55	114
Gerhard_Schroeder	0.18	0.13	0.15	23
Hugo_Chavez	0.00	0.00	0.00	10
Tony_Blair	0.40	0.14	0.21	29
accuracy			0.36	258
macro avg	0.23	0.20	0.20	258
weighted avg	0.34	0.36	0.34	258

Figure 6: Classification Report for Multi-layer Perceptron classifier