Soheila Hesaraki -r086612

Face Recognition

In this assignment, we'll apply Facial Recognition techniques for verification and identification scenarios. The project consists of two parts, a mandatory and two optional tasks.

1. Mandatory tasks

1.1 Face Recognition with different feature extraction techniques

In this part, We'll use the CALTECH Faces dataset in combination with the Haar face detector to detect faces in the images. The detected faces are of size 47x47 pixels. To handle image variations lighting, and positioning, we employ four global feature extraction techniques to process images.

- **PCA:** PCA, or Principal Component Analysis, extracts the direction of maximum variance in a distribution by computing the eigenvectors of its covariance matrix. The first principal component, associated with the largest eigenvalue, is crucial. While commonly used for dimensionality reduction, PCA is also employed in Face Recognition to extract eigen faces, representing facial variation. These eigenfaces can be combined linearly to depict a face. The initial principal component captures global features, while subsequent eigenfaces capture local details. However, PCA fails to differentiate between shape and appearance and does not consider class information, such as whether two images belong to the same class.

- **LDA:** Linear Discriminant Analysis (LDA) is the most known discriminant analysis technique. Its aim is to find a subspace in the feature representation that aligns same-class sample vectors and maximizes the separation between different-class vectors, making it suitable for classification tasks. Fisherfaces refer to the basis vectors obtained when applying LDA to face image sets, which can be used for face recognition tasks.

- **LBP:** One alternative approach to analyzing images is by considering local features rather than global features. This can be achieved by dividing the face into regions and extracting features from each region. One method commonly used is Local Binary Patterns (LBP), which compares the value of each point to that of its neighboring point**s**

- **DL:** Deep Learning (DL) is capable of extracting features, making it advantageous due to its ability to optimize the user-defined objective function. Convolutional Neural Networks (CNNs) have shown remarkable performance in extracting features from images. In the context of Face Recognition, Siamese networks are frequently employed to construct embeddings that maximize the dissimilarity between embeddings of different classes and minimize the distance between embeddings of the same class.

Q1. Compute distance-based pairwise matching scores.

We created a pairwise distance matrix by defining a Python function to generate a pairwise distance matrix. The resulting matrix has 0 values on the diagonal, as it represents the self-distance of each image. To calculate a matching score, we normalized the matrix while preserving its symmetry by using the minimum and maximum values of the entire matrix. The similarity score is given by the equation below, indicating the similarity between pictures. A higher value signifies greater similarity.

$$similarity\_score = 1 - normalized\_matching\_scores$$

1.2 Validation as verification system

Q2. Compute F1 and accuracy scores for variable (and optimal) thresholds

By having the similarity score, we can get the predicted labels, in a way that if the similarity score between two images is above a certain threshold we can consider those images the same. Figure 1 shows the accuracy and F1 score in different thresholds. As the figure shows, although the F1 score differs a lot across different techniques, accuracy does not. The reason is that Accuracy represents the ratio of correct predictions to all predictions, but it disregards dataset imbalances. For example, as shown in Figure

1, in our verification scenario, we have a limited number of images for person_003 compared to the total dataset. Consequently, a system always predicting a different person would still achieve a high accuracy rate. On the other hand, the F1 score combines precision and recall, providing a more robust metric. In our analysis, F1-scores decrease to 0 when rejecting all input samples, indicating its reliability.

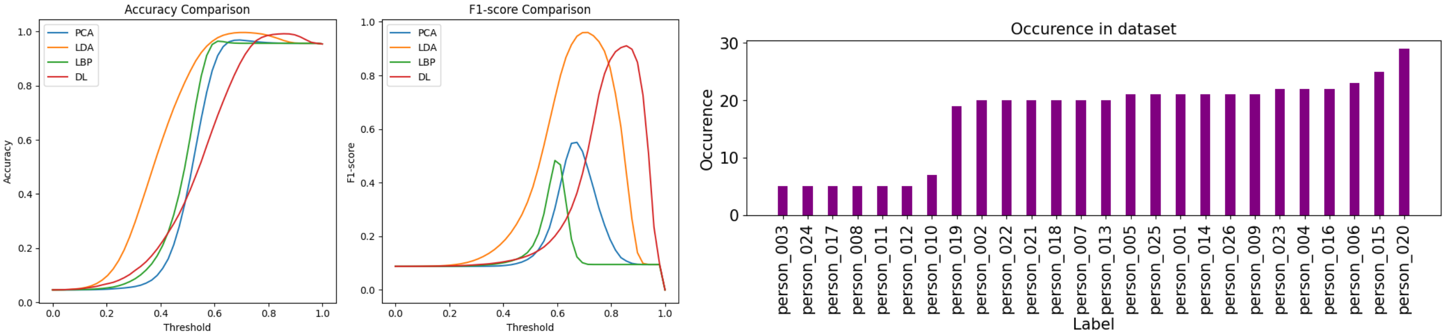

Figure1. (a) Comparison of Accuracy and F1 score, (b) data distribution

Comparing different methods, depending on the threshold value, one method can have a higher F1 score than the other. However, the maximum F1-score of LDA is higher than the others.

Q3. Plot genuine and impostor scores

The genuine distribution represents similarity scores between images of the same class, while the impostor distribution represents scores between images of different classes. By examining the genuine and impostor score distributions, we can determine if an input is genuine or an impostor based on a user-defined threshold η. In our case, the distributions overlap, resulting in a non-zero False Rejection Rate and False Acceptance Rate.

The overlapping areas for LDA and DL, as depicted in Figure 2, are much smaller compared to PCA and LBP. This indicates that LDA and DL offer lower FAR and FRR rates, consistent with the results in Figure 1.

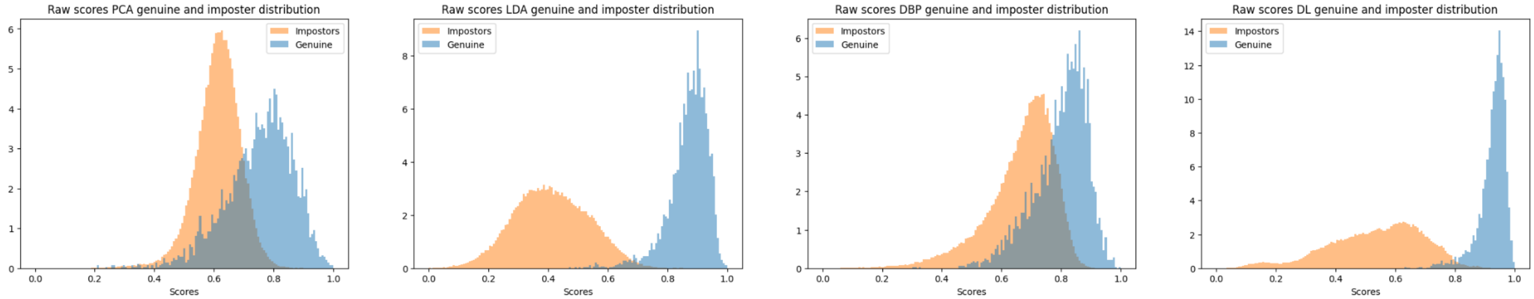

Figure 2. Comparison of Genuine and impostor distributions

Q4. Perform a full-on verification assessment based on the scores obtained. Interpret the results.

The ROC curve compares the True Acceptance Rate (TAR) to the False Acceptance Rate (FAR), enabling the selection of an operating point with high TAR and low FRR. As Figure 2 shows, all systems are better than a random binary classification (above line y = x). LDA and DL consistently outperform PCA and LBP, as shown in Figure 3 as they have lower FAR and higher TAR. The Precision-Recall curve assesses Precision against Recall, with higher values indicating better performance. LDA and DL also excel in this aspect, with LDA slightly surpassing DL (Figure 4). The Average Precision values are 0.6542 for PCA, 0.9803 for LDA, 0.5841 for LBP, and 0.984 for DL.

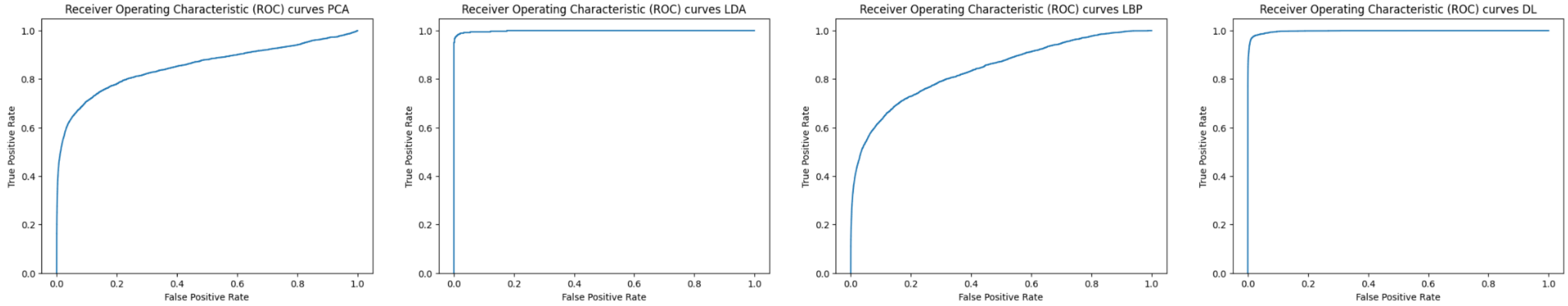

Figure 3. Comparison of ROC curves

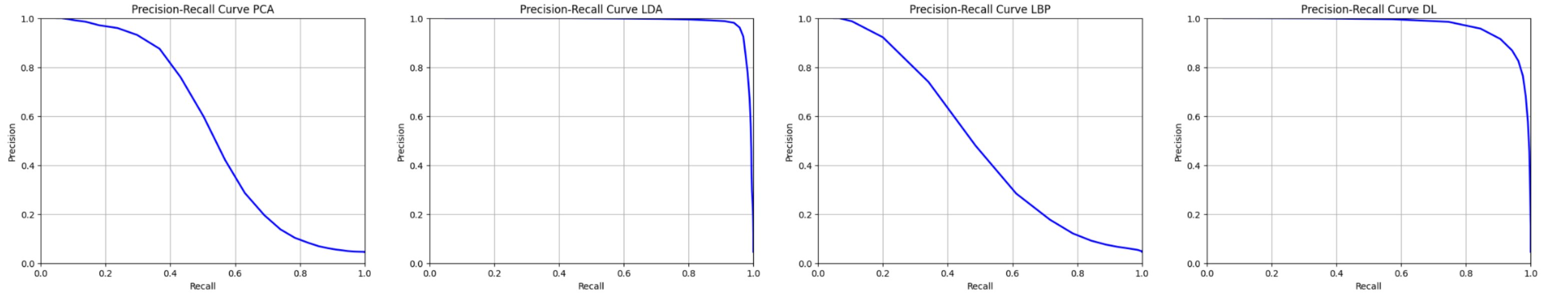

Figure 4. Comparison of Precision-Recall Curves

Furthermore, the Equal Error Rate (EER) represents the point where FAR equals FRR. Lower EER implies better performance at that specific point. However, choosing an operating point depends on the application's security requirements and convenience. LDA and DL exhibit lower EER, making them superior choices (Table 1.), with DL achieving the lowest EER overall.

Table1.

	PCA	LDA	LBP	DL
EER	0.6702	0.6702	0.7459	0.7901
AUC	0.5271	0.9615	0.4752	0.9453
Average Precision	0.6545	0.9803	0.5841	0.9745

In terms of the AUC (Area Under the Curve) values, which measure overall performance, LDA achieved the highest value of 0.9615, indicating superior performance. DL also performed well with an AUC of 0.9453. However, PCA and LBP had lower AUC values of 0.5271 and 0.4752, respectively, indicating comparatively weaker performance. Figure 11 visualizes the comparison of EER, AUC, and average precision across different feature extraction methods and datasets.

1.3. Validation as an identification system

The CMC curve illustrates the probability of correctly identifying a person within the top "t" ranked matching scores. For instance, an 85% probability at rank 5 implies that in 85% of cases, the correct identity is found among the top 5 predictions. Figure 5 showcases the resulting CMC curves for each method. Achieving the highest rank-1 recognition rate is desirable. It is observed that the DL and LDA methods yield the highest rank-1 recognition rates.

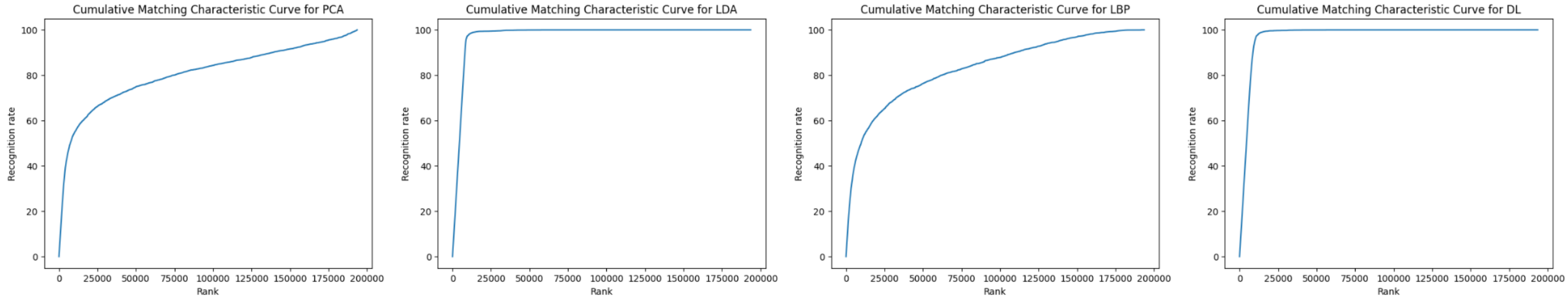

Figure 5. Comparison of CMC Curves

2. Optional Tasks

2.1 Evaluate your system on the other two datasets (AT&T, LFW). Feel free to subsample datasets if they are too memory-consuming on your system. (1pt.)

2.1.1 AT&T dataset:

The AT&T dataset, also known as the Olivetti Research Lab Database of Faces, consists of ten images per each of the 40 subjects. So, the dataset is balanced. The images exhibit variations in lighting, facial expressions, and details such as glasses.

2.1.2 LFW dataset:

The dataset consists of over 13,000 JPEG pictures of famous people obtained from the internet. Each image is labeled with the corresponding person's name. Among the individuals included, 1,680 have multiple distinct photos in the dataset. The images are diverse in terms of pose, illumination, expression, and occlusion, as they were detected using the Viola-Jones face detector. Additionally, each picture is focused on a single face.

To evaluate the system on verification and identification scenarios on these two datasets, I used the same metric as I used in section 1, using the CALTECH Faces dataset. As Figure 6 shows, for an unbalanced dataset(LFW) F1 score is a better metric than accuracy as it shows the difference in performance between different methods. With both datasets, LDA outperforms the other feature extraction methods as it has a higher F1 score, in almost all ranges of thresholds.

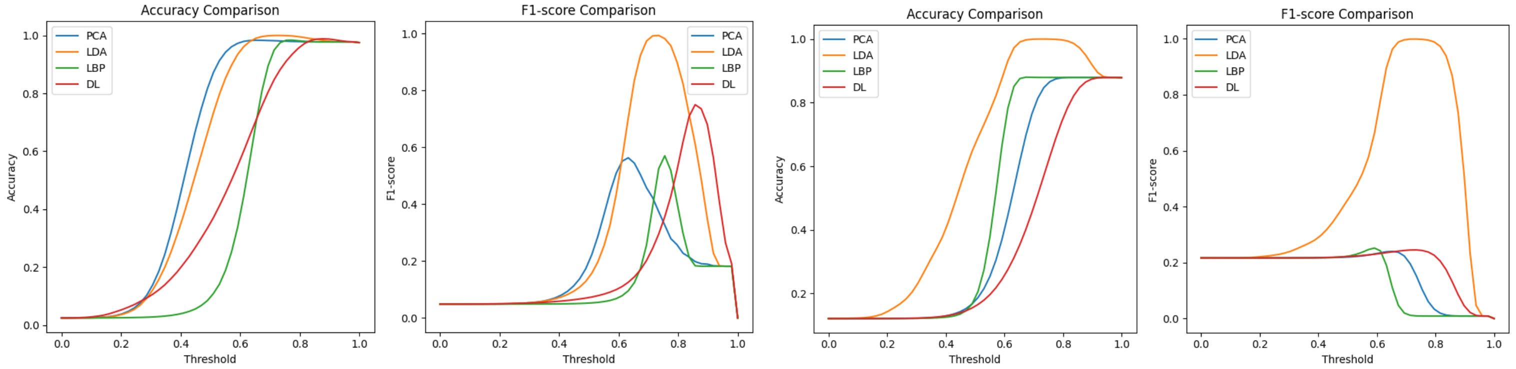

Figure 6. Comparison of Accuracy and F1 score (a) Using AT&T dataset, (b) Using LFW dataset

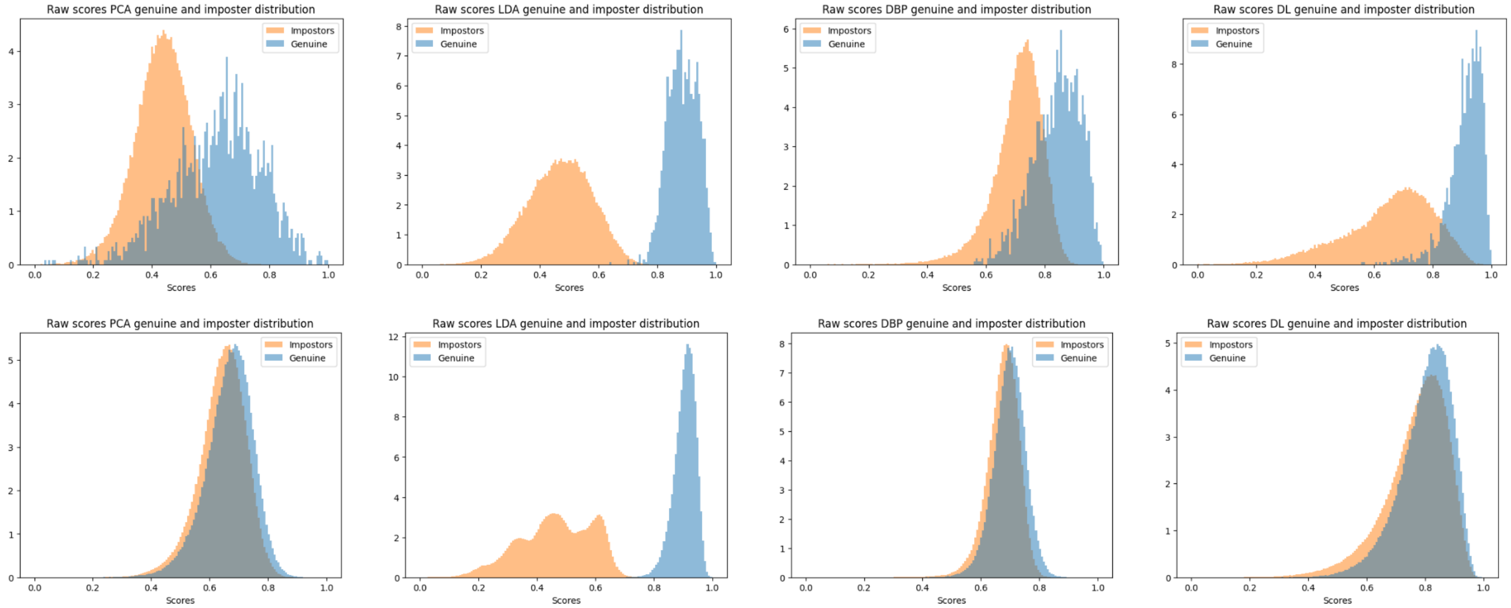

Figure 7. Comparison of Genuine and impostor distributions (top) AT&T dataset (down) LFW dataset

Based on Figure 7, LDA is the best feature extractor for both datasets, specially LFW. Because it has the least overlapping areas between genuine and impostor distributions.

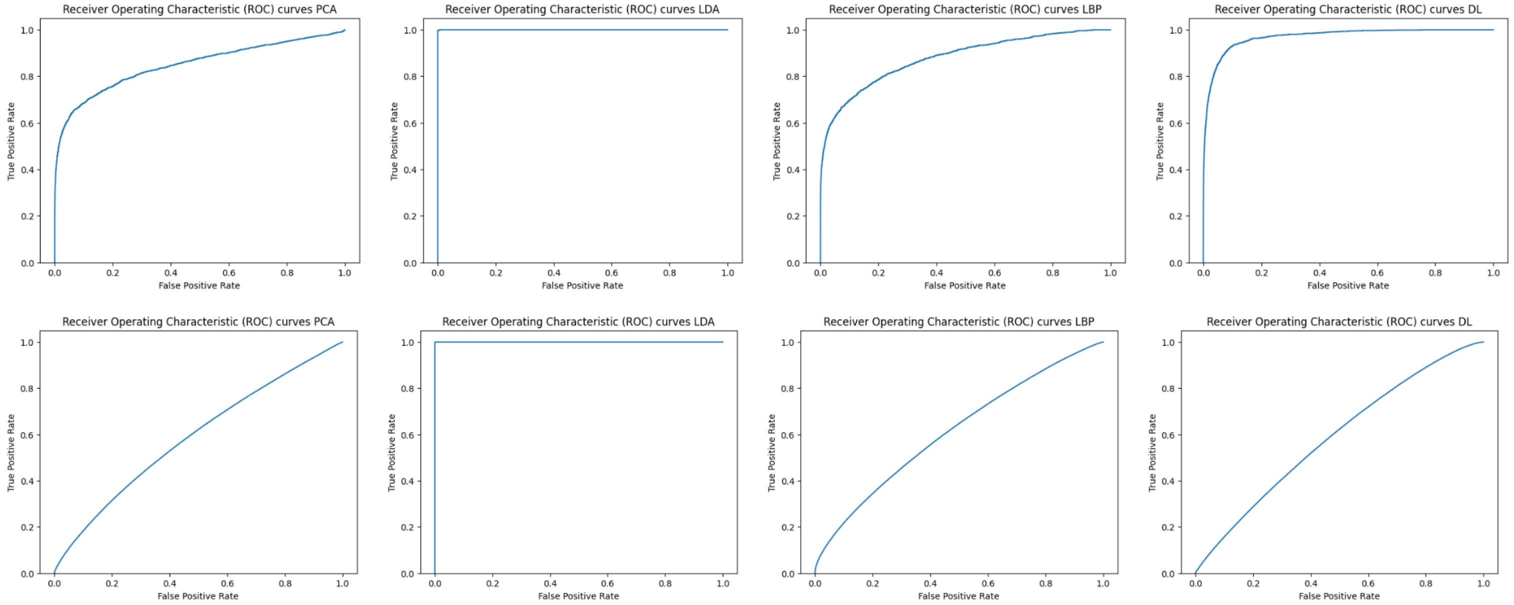

Figure 8. Comparison of ROC curves (top) AT&T dataset (down) LFW dataset

LDA and DL consistently outperform PCA and LBP when using AT&T dataset, as shown in Figure 8 as they have lower FAR and higher TAR. For the LFW dataset, LDA outperforms the other methods. Based on Table 2, LDA and DL exhibit lower EER, making them superior choices, with DL achieving the lowest EER overall.

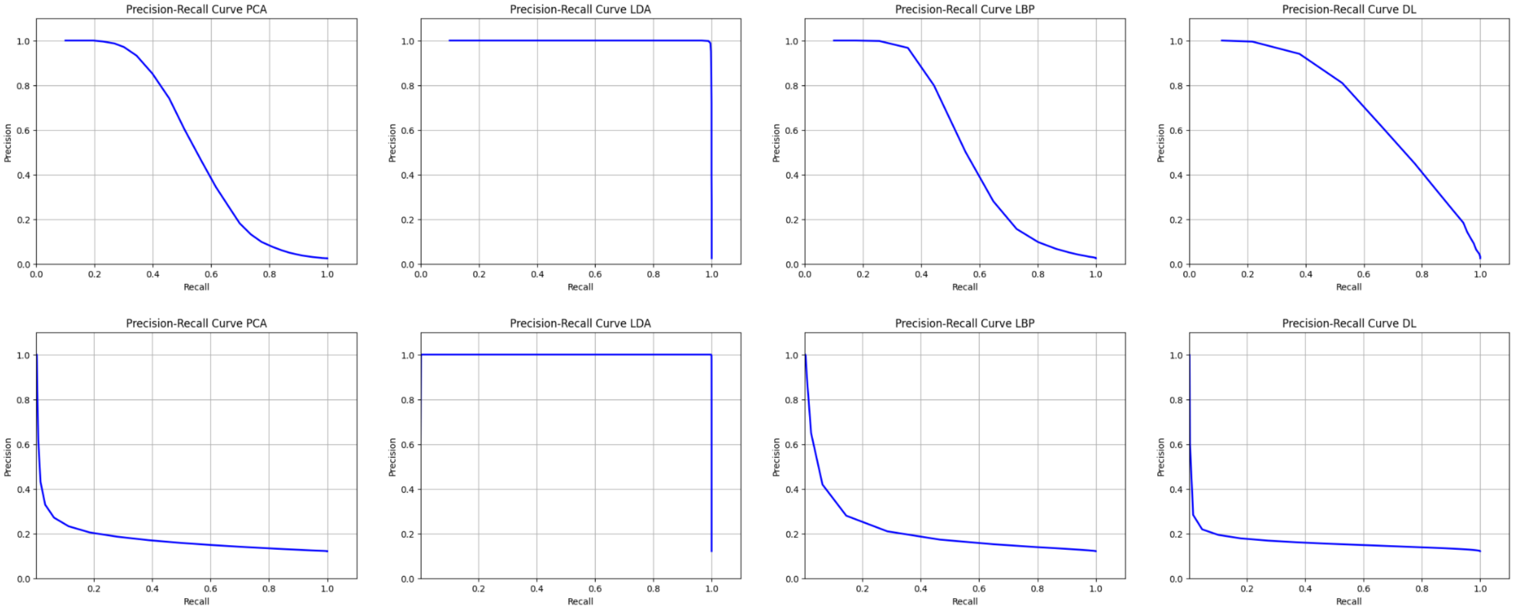

Figure 9. Comparison of Precision-Recall Curves (top) AT&T dataset (down) LFW dataset

The Precision-Recall curve assesses Precision against Recall, with higher values indicating better performance. LDA excels in this aspect in both datasets (Figure 9). Also as you see Class imbalance can have an impact on precision and recall in the verification task. In imbalanced datasets (LFW), where one class is significantly more prevalent than the other (Like saying image 003 is not verified), the classifier can be biased towards the majority class. As you see, as the recall increases, the precision drops more rapidly. This indicates that the classifier is more prone to including false positives as it tries to capture the rare instances of the minority class.

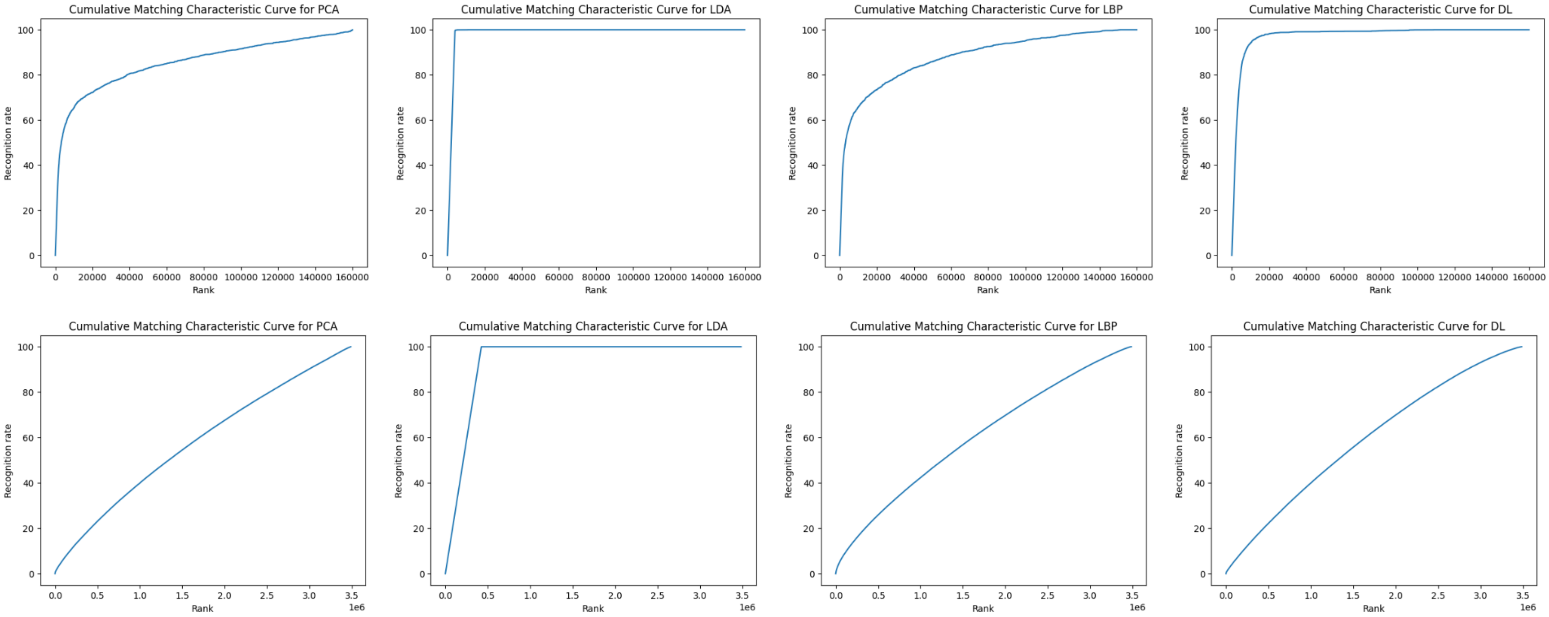

Figure 10. Comparison of CMC Curves (top) AT&T dataset (down) LFW dataset

Based on Figure 10, It is observed that LDA methods yield the highest rank-1 recognition rates for both datasets.

Table 2, compares different Average precision and AUC of different face recognition systems (based on feature extraction methods) on the three datasets used so far and Figure 11 visualizes it. As it shows, the LDA method has the highest AUC and average precision regardless of the chosen dataset. Considering this and the other metrics discussed so far, LDA is the best feature extraction method among all other methods investigated in this report.

Table 2.

	PCA	LDA	LBP	DL
EER Threshold (LFW dataset)	0.3595	0.2945	0.420463	0.2648
AUC(LFW dataset)	0.179	0.9978	0.2149	0.1644
Average precision(LFW dataset)	0.2151	0.9997	0.2411	0.1912
EER Threshold (AT&T dataset)	0.5224	0.3156	0.32973	0.2099
AUC(AT&T dataset)	0.515	0.9491	0.5271	0.6618
Average precision(LFW dataset)	0.6507	0.9999	0.6774	0.7602

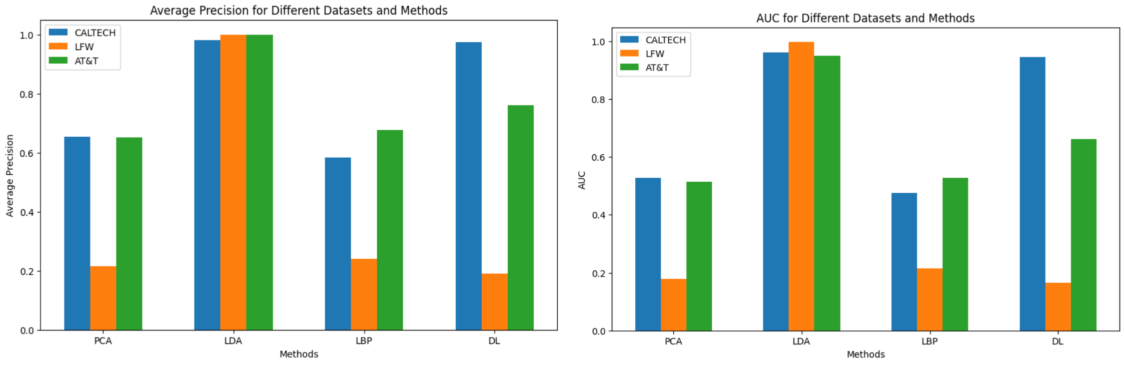

Figure 11. Comparison of Average precision and AUC for different systems and datasets

2.2 Implement a classification-based scoring method, using an advanced classifier of your choice. Evaluate this system in an identification and verification scenario. (Hint: Follow the steps introduced in section IV. Distance-based and classification-based scoring) (2pt.)

So far, we have utilized pairwise matching scores, typically computed using distance metrics like L2-distance. In verification mode, the matching score is compared to a decision threshold in a 1-to-1 setting. In identification mode, the matching score is used to rank templates in the database in a 1-to-N setting, possibly with thresholding. Now, we will explore an advanced method (SVM) involving classification algorithms that provide classification scores or probabilities, indicating the likelihood of an image belonging to each subject in the dataset.

Support Vector Machine (SVM) scores can be utilized as a reliable matching score in biometric applications. SVMs are well-suited for classification tasks and provide decision scores that indicate the distance of an image from the decision boundary. These scores can be interpreted as the confidence or probability of the image belonging to a particular subject.

One reason why SVMs are a good choice for biometric recognition is their ability to handle high-dimensional feature spaces effectively. Biometric data, such as facial or fingerprint images, often consists of numerous features. SVMs employ a kernel function to map the input features into a higher-dimensional space, where they can be more easily separated by a hyperplane. This allows

SVMs to capture complex relationships and patterns in the data, enhancing their performance in distinguishing between different subjects.

The dataset used for this part is the CALTECH dataset used in section 1. To implement a classification-based system using SVM, the dataset was divided into train and test sets. For each person in the dataset, one image was left out as a test sample. This approach ensured that the test set size matched the number of individuals in the dataset, enabling comprehensive evaluation.

The SVM was then trained on the remaining images in the train set. Subsequently, classification probabilities were generated for each image in the test set using the trained SVM model. By comparing the predicted labels with the actual labels, the performance of the system was assessed.

The next step involved constructing a similarity matrix of size 26 by 26, where 26 represents the number of test images (which is equal to the number of classes in the CALTECH dataset). Each row in the matrix corresponded to a test image, while each column represented an individual in the dataset. This similarity matrix facilitated further analysis and evaluation of the system's effectiveness.

Figure 12 shows the performance of the SVM system in the verification scenario.

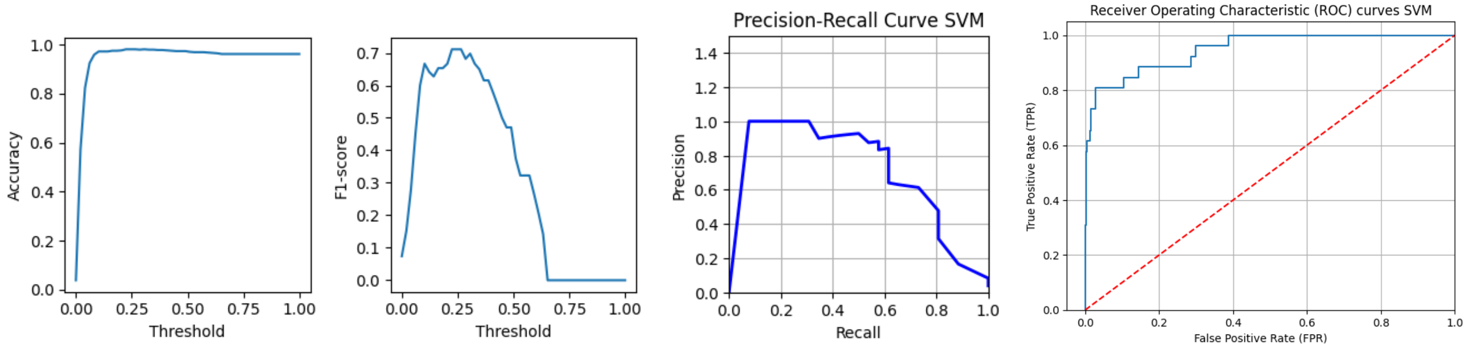

Figure 12. Performance of the SVM system in verification scenario

Figure 13 shows its performance in an identification scenario.

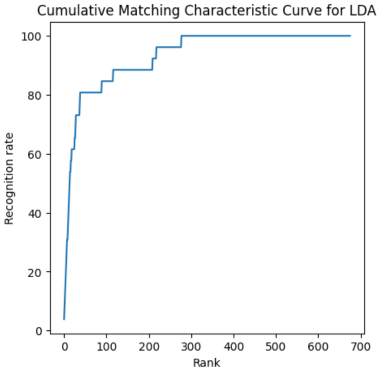

Table 3.

	EER Threshold	AUC	Average precision
SVM system	0.0453	0.6917	0.8149