Face Tracking and Recognition in a Conference context

Felix De Mûelenaere
Faculteit Ingenieurswetenschappen en Architectuur (UGent)
Master of Science in de industriële wetenschappen: elektronica-ICT (Major Beeldtechnologie)
Gent, België
Felix.DeMuelenaere@UGent.be

***Abstract*—This paper discusses the design of a Facial Recogniton System to classify a new dataset recorded at a conference by a camera module of the company Televic. The embedded hardware to process the audio & video and the constrained conference setting leads to the use of fast algorithms that work well in a constrained environment on frontal head poses instead of State of the art Deep Learning (DL) neural networks. The value of a video sequence is examined and a novel method based on a weighted average of probabilty scores is presented with success.**

***Index Terms*—Face recognition, PLDA, LDA, QDA, FisherFace, Computer vision**

I. Introduction

The domain of facial recognition has a long-standing history of research behind it, and a vast amount of methods have been designed to tackle the problem. These will be discussed briefly in section II.

Televic is a Belgian based international company who commissioned this master's thesis. The aim of this work is to roll out a facial recognition system (FRS) to identify attendees to conferences. Televic has numerous communication modules in use by for example the ”United Nations: Room XVII” Geneva, Switzerland.

Their products are a great tool to moderate and facilitate conferences, along with their CoCon software suite. In Geneva Televic has deployed their ”uniCOS” units, which have a camera, a microphone and a touchscreen as is depicted in Fig. 1.

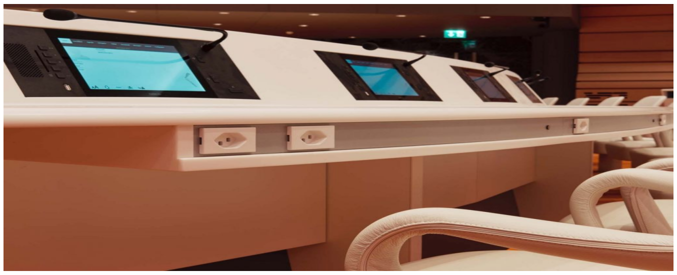


Fig. 1: The *United Nations: Room XVII* - Televic's uniCOS units with Cocon Voting & PLIXUS® system

The modules are connected to a central processing node which handles the audio & video. The central unit runs on embedded hardware and this in turn limits the available algorithms that can be deployed, so that the FRS can operate in real-time.

Additionally research was conducted to determine the most suited classifier to classify a new dataset being made during the first conference at a new location or where a dataset of the participants isn't available yet. The research experiments are discussed in section V.

II. Face recognition system

The general approach to facial recognition can be seen in Fig. 2, where an input image (potentially from a video sequence) is first scanned to detect the presence of a face. When a face is detected, a set of features is extracted and compared to a stored set of features describing the faces the algorithm was trained on. If there is a match with enough confidence between the features, the subject in the image is labeled as a person from the set if the purpose is identification. Verification is a slightly different process where the output is a binary yes or no, in this case the system is asked if the person contained in the image is a specific person.

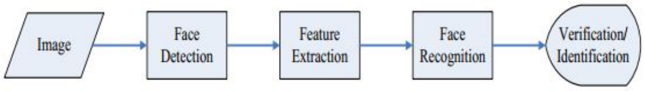


Fig. 2: Flowchart of the general approach to facial recognition systems [2]

There are a number of assumptions being made about the input images in the setting of conferences namely:

1) Conference rooms are well lit
2) There is little to no variation in the scale of faces in the camera-feed, because of near fixed distance between the camera-modules and the chairs
3) There will be enough frontal head-poses in the camera-feed

These assumptions specify that we're dealing with a constrained environment and the system doesn't have to be illumination- or pose-invariant.

The face recognition problem for constrained environments has been solved for a long time. The aforementioned problems of head pose and illumination have been longstanding issues for many real-world applications like CCTV camera security and have only been solved fairly recently with human-like performance [3].

The following sections will focus on each step of the FRS, where most of the discussion will be spent on the feature extraction step.

III. Face Detection

Face Detection is a well-studied subject and can be divided into four categories according to Chao [2].

1) Knowledge-based
2) Feature invariant
3) Template matching
4) Appearance-based

The *Haar* features by Viola et al [4] belong to the last category and is the method that is used for the detection step of the FRS. This method has proven itself to be very effective and is employed very often for frontal face poses.

IV. Feature extraction

A large number of taxonomies have been proposed to explain the variety of methods for facial recognition. A recent multi-level taxonomy proposed by Moghaddam et al [5] has been adopted.

After carefully considering each feature extraction method's advantages and disadvantages, the Holistic/Appearance based linear methods were the most attractive in the situation of a controlled environment with frontal face pose.

The reasons to implement these kind of algorithms are the following:

1) Easy to implement, no key-points to determine or need for special datasets
2) Works well for frontal faces according to Zhao et al [6]
3) Allows the dimensionality reduction of features

To explain the benefit of dimensionilty reduction it is important to observe that an m by n pixel image is treated as a m*n feature row or column matrix, where every pixel is a feature or dimension. An image of 100 by 100 pixels thus has 10,000 dimensions and therefore a lot of methods suffer from the "curse of high dimensionality".

Not all of these dimensions contain useful information, that's where dimensionality reduction comes into play: we want to represent the data in a compact way without discarding potentially important discriminating information.

Three methods have been implemented and will be discussed individually in the next sub-sections.

A. *EigenFace and FisherFace*

The EigenFace method is based on the statistical tool Principal Component Analysis (PCA) that finds the directions that account for the most variance in the data. The method transforms a set of possibly correlated variables into a smaller set of correlated variables.

The statistical method at the heart of the FisherFace algorithm is Fisher's Discriminant Analysis (FDA), presented by Fisher in his orginal paper [7]. Linear Discriminant Analysis (LDA) is the generalization of FDA to multiple classes.

Where PCA looks for the directions that account for the most variance in the data, it doesn't take

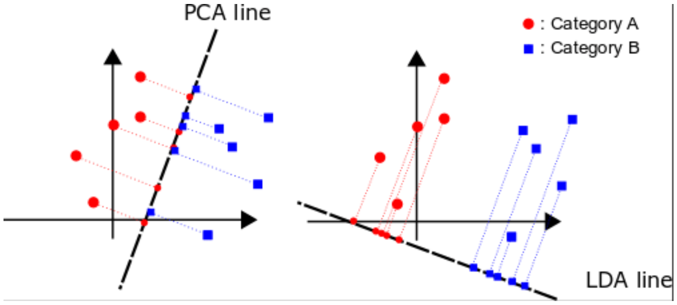


Fig. 3: 2D to 1D dimensionality reduction by PCA and LDA

class information into account.

LDA is a supervised dimensionality reduction tool that maximizes a class separability criterion. The ratio of the within class scatter to the between class scatter is maximized.

Samples are projected to a subspace or ”Face-Space” where samples belonging to the same class are clustered together and different classes are spread out as much as possible.

The distinction between PCA and LDA is evident when looking at Fig. 3.

B. Probalistic Linear Discriminant Analysis (PLDA) and Quadratic DA (QDA)

PLDA is an extension of LDA that can be used for dimensionality reduction and for classification using a Bayesian framework [8].

The famous rule of Bayes states the folowing:

$$\begin{aligned} P(y=k|X) &= \frac{p(X|y=k)P(y=k)}{P(X)} \\ &= \frac{p(X|y=k)P(y=k)}{\sum_l p(X|y=l)\cdot P(y=l)} \end{aligned} \tag{1}$$

Where a small p denotes a distribution and a capital P is used for a probability. To use PLDA as a classifier we assume that the images per class follow a Gaussian distribution with mean μ and covariance matrix Σ:

$$\begin{aligned} p(X|y=k) &= \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \\ \times \exp &\left(-\frac{1}{2}(X-\mu_k)^t\Sigma_k^{-1}(X-\mu_k)\right) \end{aligned} \tag{2}$$

The distinction between PLDA and QDA is that PLDA assumes that the classes share one covariance matrix $\Sigma_k = \Sigma \ \forall k$, this leads to Linear Decision Surfaces (LDS).

If this assumption is not made, it leads to Quadratic Decision Surfaces (QDS) with omission of the mathematical background. When you compare log-probabilities of PLDA you obtain the following equation:

$$\begin{aligned} \log\left(\frac{P(y=k|X)}{P(y=l|X)}\right) &= \log\left(\frac{P(X|y=k)P(y=k)}{P(X|y=l)P(y=l)}\right) \\ &= 0 \Leftrightarrow (\mu_k-\mu_l)^t\Sigma^{-1}X = \\ \frac{1}{2}(\mu_k^t\Sigma^{-1}\mu_k &- \mu_l^t\Sigma^{-1}\mu_l) - \log\frac{P(y=k)}{P(y=l)} \end{aligned} \tag{3}$$

Where the equation has the shape $y = a * x + b$ in 2D, in multiple dimension the equation will describe a hyperplane.

Dimensionality reduction is inherent to LDA and test-samples are first reduced then classified according to the highest probability.

V. Experiments

Three datasets were used in the following experiments, the AT&T datset, the FaceScrub dataset and for the third experiment a video dataset was recorded on campus which mimics a conference setting.

Three experiments have been carried out, the first one researches whether it is useful to apply a robustly trained FDA projection to classify a new dataset.

The second experiment examines how the PLDA & QDA method scale with an increasing number of persons to classify. And which of the two methods performs the best.

The third and final experiment examines whether the availability of a video sequence can enhance the classification accuracy. A novel method is proposed, using a weighted average of probabilities in a window of detected faces spaced in time.

The f1 accuracy metric is used to evaluate the classification performance in all experiments along with classification reports. Experiment II & III also makes use of cross-validation and in III confusion matrices were also generated.

A. Experiment I

For this experiment, the selected dataset is subdivided into two subsets A & B. In a first approach, only 3 persons are used (the 3 in A are different than those in B) so that 2D plots can easily be generated to analyse the data. A FDA projection is trained on Subset A which is composed of all the available images of the three persons, the learned transformation matrix and can project samples to a subspace or FaceSpace that we denote by F1 from now on.

Subset B consists of three other persons and only x with x $\in [2, 4, 5, 7]$ images per person are used to train a PLDA classifier (C1).

Subset B is also projected into F1 where a second classifier, a nearest neighbour classifier (C2) with mahalanobis as distance metric is trained to the class centroids of subset B in F1.

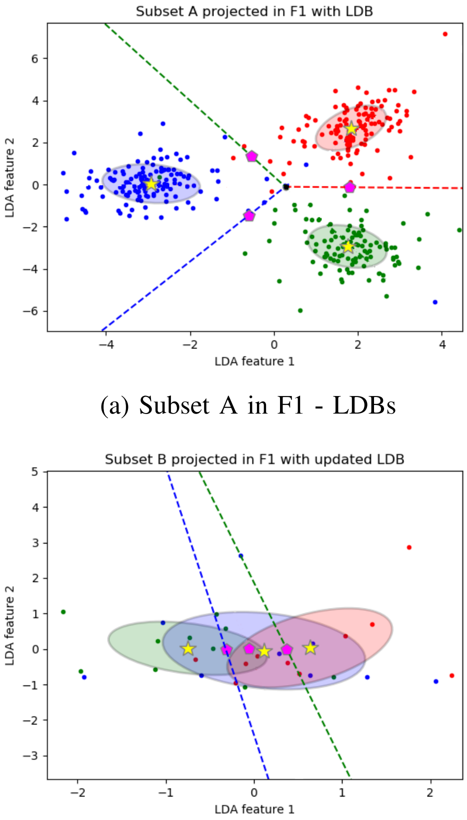
The Linear Decision Boundaries (LDB) are plotted in the F1 space as dashed lines, once for subset A and recomputed for subset B as is depicted in Fig. 4. The yellow stars are the class means μ and the magenta pentagons are the points halfway between each pair of class means. The ellipses represent the covariance of the classes.

Notice that the classes overlap when Subset B is projected in F1 and classification in this space will be inaccurate for new samples. Thus the PLDA classifier C1 outperform the nearest neighbour classifier C2 as seen in 5, surprisingly the difference is not as big as was expected.

The accuracy of the PLDA method is low because it does not work well on the FaceScrub dataset. This is because the dataset contains too much variation in head pose, illumination, background, age of the persons and so on.

The experiment is also carried out on the well constrained AT&T dataset for 3, 10 and 20 persons with good results for the C1 classifiers as shown in Fig. 6.

It should be noted that it's possible to obtain even better results for the C1 classifier by using another training method for PLDA. The fastest method is Singular Value Decomposition (SVD), while the Least Squares Regression (LSQR) method is computationally expensive but more accurate.



(a) Subset A in F1 - LDBs

(b) Subset B in F1 - new LDBs

Fig. 4: The LDBs for three people of the FaceScrub dataset

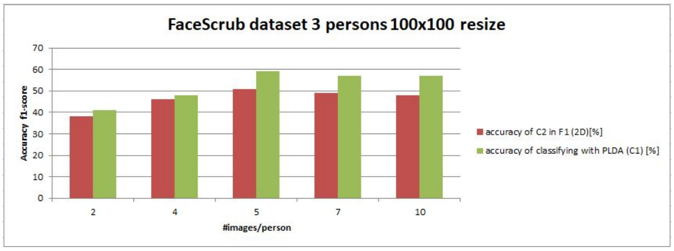


Fig. 5: f1-scores for C1 (green) and C2 (red) for three persons of the FS dataset

B. Experiment II

As mentioned in V, we will examine the scalability of PLDA and QDA, this time the accuracies are only evaluated on the AT&T dataset.

To apply QDA we first use a PLDA classifier to reduce the amount of dimensions, as QDA isn't able to perform dimensionality reduction. The optimal number of dimensions has been determined by comparing the accuracy of PLDA for an increasing

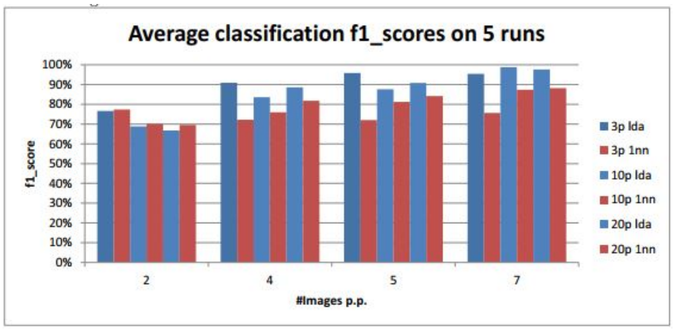


Fig. 6: f1-scores averaged on 5 runs with a different subset of persons (3,10 and 20 persons), C1 in red and C2 in blue

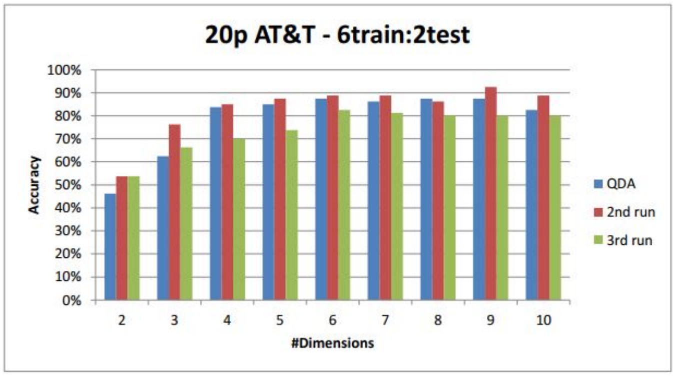


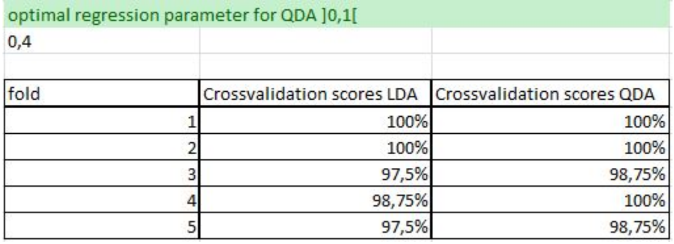
Fig. 7: Accuracies of 3 runs on the AT&T dataset with each time a different subset of 20 out of 40 persons, using 6 images to train and 4 to test

number of dimensions. Six dimensions produce the best results and this dimensionality will be used for QDA from now on.

When implementing QDA, there is only one parameter to tune and that is the regression parameter with a value in [0-1], which regularizes the per-class covariance estimates. Similar to the optimal number of dimensions, a range of values were compared and the optimal one was selected.

The results of running the experiment three times with a different subset of 20 out of 40 persons, are shown in Fig. 7 with 6 train and 4 test images per person.

Conventional cross-valdation with 5 folds was carried out on the whole dataset, the results in Fig. 8 confirm our hypothesis that QDA scales better than PLDA, but at a greater computational cost, which leads us to select PLDA for the final implementation of the algorithm.



optimal regression parameter for QDA]0,1[

0,4

fold	Crossvalidation scores LDA	Crossvalidation scores QDA
1	100%	100%
2	100%	100%
3	97,5%	98,75%
4	98,75%	100%
5	97,5%	98,75%

Fig. 8: 5 fold cross-validation on the entire AT&T dataset using 6 discriminating dimensions for QDA and using the optimal regression parameter

C. *Experiment III*

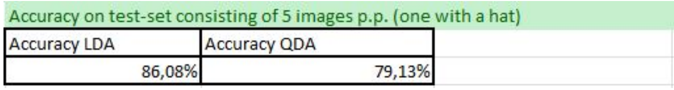
To evaluate the potential additional value of using a video dataset, one was recorded on campus where a conference setting was simulated. The dataset consists of 23 subjects with an average video length of 30 seconds.

The participants are talking and looking around to simulate a realistic conference setting. To carry out a rigorous test of the FRS, all participants wear a hat near the end of the recording. Participants with glasses were asked to remove them for the same purpose of obtaining variation in the appearance of the subjects.

Additionally, a train- and test-set were generated from the video sequence. The train set consists of 20 images spaced in time by 20 face detections. The test-set consists of 5 images, manually selected from the video. In one out of those five test images, the participant wears a hat, to examine whether the classifier can handle these types of situations and can generalize the appearance of a person.

Cross-validation was carried out on the train images with staggering results (100% for every fold), but this is to be expected as the images were taken form one video sequence and this the result of overfitting.

The accuracy of the classifiers on the test images are shown in Fig. 9, where the 80% accuracy signifies that the classifier usually failed on the image containing a hat, as is to be expected since important facial features are occluded, which is very unlikely to happen during a conference.



Accuracy on test-set consisting of 5 images p.p. (one with a hat)

Accuracy LDA	Accuracy QDA
86,08%	79,13%

Fig. 9: Accuracy of the PLDA and QDA classifier on the test-set

A novel method for classification will be discussed in the next subsection.

1) Averaged weighted probabilities: Until now classification by PLDA or QDA was carried out by assigning the label corresponding with the highest class probability to a test sample. With the abundance of images in a video sequence it is now possible to use a weighted average over a window of frames.

Instead of using one test sample, a window of 10 detected faces separated in time by 10 detections is taken, the probabilities per sample and per class are summed up and averaged by dividing by the size of the window.

There are many combinations possible for the size of the window and how to spread the samples in time. The aforementioned size of the window and time spacing of 10 and 10 were obtained by trial and error.

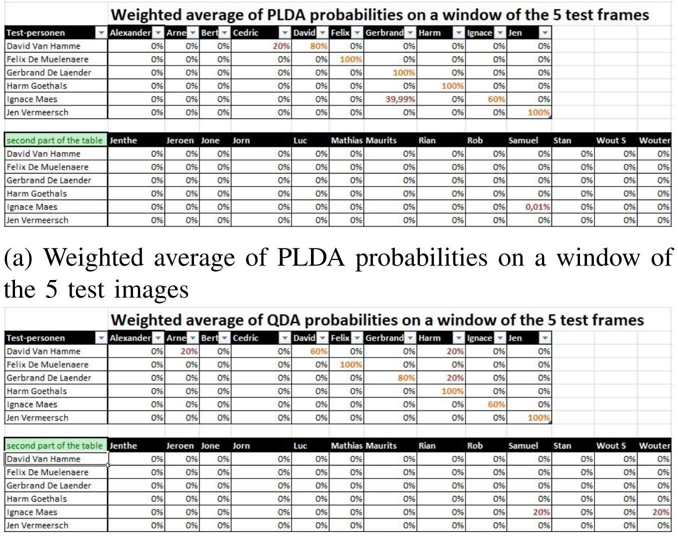
The results of using our method of weighted probabilities on a few persons can be seen in Fig. 10a and 10b where PLDA and QDA obtain 100% accuracy on all subjects. The window that was used are the 5 test-images, to show that the proposed method is robust even when there is significant change in appearance when the images are spaced by enough detections.

In practice the system will be retrained on subsequent conferences up to 10 and after that on the last 10 sessions, this will be representative of the results we obtained for the AT&T dataset in the preceding experiments.

VI. Conclusion

The best way to classify a new dataset was researched, where we can conclude that using a PLDA classifier is the best option in an embedded hardware setting and a constrained environment with frontal head poses.

The value of a video sequence was demonstrated as well as the proposed novel method of using an average of weighted probabilities to classify images.



Weighted average of PLDA probabilities on a window of the 5 test frames

Test-personen	Alexander	Arne	Bert	Cedric	David	Felix	Gerbrand	Harm	Ignace	Jen
David Van Hamme	0%	0%	0%	20%	80%	0%	0%	0%	0%	0%
Felix De Muelenaere	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Gerbrand De Laender	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%
Harm Goethals	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%
Ignace Maes	0%	0%	0%	0%	0%	0%	39,99%	0%	60%	0%
Jen Vermeersch	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%

second part of the table	Jenthe	Jeroen	Jone	Jorn	Luc	Mathias	Maurits	Rian	Rob	Samuel	Stan	Wout S	Wouter
David Van Hamme	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Felix De Muelenaere	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Gerbrand De Laender	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Harm Goethals	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ignace Maes	0%	0%	0%	0%	0%	0%	0%	0%	0%	0,01%	0%	0%	0%
Jen Vermeersch	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

(a) Weighted average of PLDA probabilities on a window of the 5 test images

Weighted average of QDA probabilities on a window of the 5 test frames

Test-personen	Alexander	Arne	Bert	Cedric	David	Felix	Gerbrand	Harm	Ignace	Jen
David Van Hamme	0%	20%	0%	0%	60%	0%	0%	20%	0%	0%
Felix De Muelenaere	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Gerbrand De Laender	0%	0%	0%	0%	0%	0%	80%	20%	0%	0%
Harm Goethals	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%
Ignace Maes	0%	0%	0%	0%	0%	0%	0%	0%	60%	0%
Jen Vermeersch	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%

second part of the table	Jenthe	Jeroen	Jone	Jorn	Luc	Mathias	Maurits	Rian	Rob	Samuel	Stan	Wout S	Wouter
David Van Hamme	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Felix De Muelenaere	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Gerbrand De Laender	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Harm Goethals	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Ignace Maes	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	20%
Jen Vermeersch	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

(b) Weighted average of QDA probabilities on a window of the 5 test images

Future work should explore other feature extraction methods to compare their accuracy to this method and a more elaborate video dataset recorded on a conference should be used to execute the experiments.

References

[1] Televic, “Congreso de mexicali,” https://www.televic-conference.com/en/references/congreso-de-mexicali, accessed: 2020.

[2] W.-L. Chao, “Face recognition,” *GICE, National Taiwan University*, 2007.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[4] P. Viola and M. J. Jones, “Robust real-time face detection,” *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[5] A. Sepas-Moghaddam, F. M. Pereira, and P. L. Correia, “Face recognition: A novel multi-level taxonomy based survey,” *IET Biometrics*, 2019.

[6] W. Zhao, R. Chellappa, and P. J. Phillips, *Subspace linear discriminant analysis for face recognition*. Citeseer, 1999.

[7] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936. [Online]. Available: http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

[8] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.