Individual discriminative face recognition models based on subsets of features

Line H. Clemmensen[1], David D. Gomez[2], and Bjarne K. Ersbøll[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark,
DK-2800 Lyngby, Denmark. lhc@imm.dtu.dk and be@imm.dtu.dk.
[2] Computational Imaging Lab, Pompeu Fabre University, Barcelona, Spain.
david.delgado@upf.edu.

Abstract. The accuracy of data classification methods depends considerably on the data representation and on the selected features. In this work, the elastic net model selection is used to identify meaningful and important features in face recognition. Modelling the characteristics which distinguish one person from another using only subsets of features will both decrease the computational cost and increase the generalization capacity of the face recognition algorithm. Moreover, identifying which are the features that better discriminate between persons will also provide a deeper understanding of the face recognition problem. The elastic net model is able to select a subset of features with low computational effort compared to other state-of-the-art feature selection methods. Furthermore, the fact that the number of features usually is larger than the number of images in the data base makes feature selection techniques such as forward selection or lasso regression become inadequate. In the experimental section, the performance of the elastic net model is compared with geometrical and color based algorithms widely used in face recognition such as Procrustes nearest neighbor, Eigenfaces, or Fisherfaces. Results show that the elastic net is capable of selecting a set of discriminative features and hereby obtain high classification rates.

1 Introduction

Historical facts (New York, Madrid, London) have put a great emphasis on the development of reliable and ethically acceptable security systems for person identification and verification. Traditional approaches such as identity cards, PIN codes, and passwords are vulnerable to falsifications and hacking, and such security breaks thus also appear frequently in the media.

Another traditional approach is biometrics. Biometrics base the recognition of individuals on the intrinsic aspects of a human being. Examples are fingerprint and iris recognition [1][2]. However, traditional biometric methods are intrusive,

[0] B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 6171, 2007, 2nd edition.

i.e. one has to interact with the individual who is to be identified or authenticated. In some cases, however, iris recognition is implemented as a standard security check in airports (e.g. New York JFK). Recognition of people from facial images on the other hand is non-intrusive. For this reason, face recognition has received increased interest from the scientific community in the recent years.

Face recognition consists of problems with a large number of features (of geometrical or color related information) in relation to the number of face images in the training sets. In order to reduce the dimensionality of the feature space we propose to use *least angle regression - elastic net* (LARS-EN) model selection to select discriminative features that increase the accuracy rates in facial identification. LARS-EN was introduced by Zou et. al in 2005 [3]. It regularizes the *ordinary least squares* (OLS) solution with both the Ridge regression and Lasso constraints. The method selects variables into the model where each iteration corresponds to loosening the regularization with the Lasso constraint. The ridge constraint ensures that the solution does not saturate if there are more variables in the model than the number of observations.

The rest of the paper is organized as follows: In section two, a review of the standard face recognition techniques is presented. Section three describes the LARS-EN algorithm. In section four, we describe and state the results for several experiments which we conducted to test the discriminative capacity of the obtained features. Finally, section 5 gives a conclusion of the conducted experiments and discusses some future aspects of the research.

2 Face recognition review

The first techniques developed for face recognition aimed at identifying people from facial images based on geometrical information. Relative distances between key points such as mouth or eye corners were used to characterize faces [4][5]. At this first stage of facial recognition, many of the developed techniques focused on automatic detection of individual facial features. The research was notably strengthened with the incursion of the theory of statistical shape analysis. Within this approach, faces were described by landmarks or points of correspondence on an object that matches between and within populations. In a 2D-image, a landmark $\mathbf{l}$ is a two dimensional vector $\mathbf{l} = (x, y)$ that, to obtain a more simple and tractable mathematical description, is expressed in complex notation by $\mathbf{l} = x + iy$, where $i = \sqrt{-1}$. In this framework, a face in an image is represented by a configuration or a set of n landmarks $[\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_n]$ placed on meaningful points. Geometrical face recognition based on landmarks is conducted by evaluating the similarity of the configuration of a test face with respect to the configurations in a facial database. In order to achieve this, different measures of similarity have been proposed, see e.g. [6]. Among all the proposed metrics, the Procrustes distance has been the most frequently used. Given two configurations w and z, the Procrustes distance between them is defined by

$$D_P(w,z) = \inf_{\beta,\theta,a,b} \| \frac{z}{\|z\|} - \frac{w}{\|w\|}\beta e^{i\theta} - a - ib\| \quad , \tag{1}$$

where $\|\cdot\|$ represents the l_2 norm, and the parameters β, θ, a, and b, which denotes a scaling, a rotation, and a translation of configuration w, are chosen to minimize the distance between w and z. Several extensions of this measure have been proposed. For instance, Shi et. al [7] has recently proposed a refined Procrustes distance based on principal component analysis. The configurations (the landmark representations of the faces) are first centered at the origin and transformed to have unit size. Then a complex principal component analysis is conducted to reduce the dimensionality. The similarity measure is defined in this lower m-dimensional space by

$$D_{RP}(w,z) = \sum_{k=1}^{m} \| \frac{\hat{z}_k}{\sqrt{\lambda_k^{(z)}}} - \frac{\hat{w}_k}{\sqrt{\lambda_k^{(w)}}} \| \quad , \tag{2}$$

where $\hat{z}_k$ is the k^{th} eigenvector of configuration y, $\hat{w}_k$ is the k^{th} eigenvector of configuration w, and $\lambda_k^{(z)}$ and $\lambda_k^{(w)}$ the corresponding eigenvalues.

The publication of Eigenfaces by Turk and Pentland [8] showed that it was possible to obtain better classification rates by using the color intensities. Since then, geometrical face recognition was gradually declining until the extent that, nowadays, it principally remains to support color face recognition. The appearance of Eigenfaces provided an excellent way of summarizing the color information of the face. The facial images in a training database were first registered to obtain a correspondence of the pixels between the images. Then, a principal component analysis was conducted to reduce the high data dimensionality, to eliminate noise, and to obtain a more compact representation of the face images. When a new test image was desired classified, the same data reduction was applied to obtain a comparable compact test image representation. The similarity of the compact test image representation was measured with each of the compact training image representations based on the Euclidean distance. The test image was associated with the training image with the smallest Euclidean distance. Based on Eigenfaces, Fisherfaces obtained higher classification rates by applying a Fisher Linear discriminant on the obtained principal components. As a result of the publication of Fisherfaces a considerable percentage of the current research in the field is devoted to find more discriminative projections [9][10].

In this paper, an approach to increase the discrimination among individuals is proposed. However, instead of looking for more discriminative projections as the previous methods, it aims at finding more discriminative features. This is in line with the face detector of Viola and Jones [11] that selects Haar features which are important for the face detection task. Basing the identification on only a subset of the features will make the system work faster for future identifications. The approach is described in next section.

3 Elastic net model selection

We consider the linear model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad , \tag{3}$$

where each $\epsilon_i \sim N(0, \sigma^2)$. We assume $\boldsymbol{y}$ centered (i.e. $\sum_{i=1}^{n} y_i = 0$) and the columns of $\mathbf{X}$ normalized to zero mean and unit length.

The LARS-EN method is used to make multiple individual discriminative models by the use of dependent variables with ones and zeros discriminating one individual from the remaining people in the data set. In the case of one image per individual the k^{th} individual model is:

$$\text{center}\left(\begin{bmatrix} \mathbf{0}_{k-1} \\ 1 \\ \mathbf{0}_{n-k} \end{bmatrix}\right) = \text{normalize}\left(\begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}\right) \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad , \tag{4}$$

where n is the number of individuals (there are $n-1$ individuals distinct from individual k), and p is the number of features. $\mathbf{0}_{k-1}$ denotes a vector of $k-1$ zeros. The geometrical features used in this work were the x and the y coordinates of the landmarks. The color based features were the gray scale intensities of the facial images after warping.

3.1 The elastic net

Least angle regression - elastic net (LARS-EN) model selection was proposed by Zou et. al [3] to handle $p \gg n$ problems. The method regularizes the *ordinary least squares* (OLS) solution using two constraints, the 1-norm and the 2-norm of the coefficients. These constraints are the ones used in the *least absolute shrinkage and selection operator* (Lasso) [12] and Ridge regression [13], respectively. The naive elastic net estimator is defined as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta}\{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\} \quad , \tag{5}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\beta_i|$, $|\cdot|$ denoting the absolute value, and $\|\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{p} \beta_i^2$. Choosing $\lambda_1 = 0$ yields Ridge solutions, and likewise choosing $\lambda_2 = 0$ yields Lasso solutions. For the Lasso method it is likely that one or more of the coefficients is zero at the solution, while for the Ridge regression it is not very likely that one of the coefficients is zero. Hence, we obtain a sparsity in the solution by using the Lasso constraint. The Ridge constraint ensures that we can enter more than n variables into the solution before it saturates.

We can transform the naive elastic net problem into an equivalent Lasso problem on the augmented data (c.f. [3, Lemma 1])

$$\mathbf{X}^* = (1 + \lambda_2)^{-1/2} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p \end{bmatrix} \quad , \quad \boldsymbol{y}^* = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0}_p \end{bmatrix} \quad . \tag{6}$$

The normal equations, yielding the OLS solution, to this augmented problem are

$$\begin{gathered}\left(\frac{1}{\sqrt{1+\lambda_2}}\right)^2 \begin{bmatrix}\mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p\end{bmatrix}^T \begin{bmatrix}\mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p\end{bmatrix} \hat{\boldsymbol{\beta}}^* = \frac{1}{\sqrt{1+\lambda_2}} \begin{bmatrix}\mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p\end{bmatrix}^T \begin{bmatrix}\boldsymbol{y} \\ \mathbf{0}_p\end{bmatrix} \Leftrightarrow \\ \frac{1}{\sqrt{1+\lambda_2}} \left(\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I}_p^T \mathbf{I}_p\right) \hat{\boldsymbol{\beta}}^* = \mathbf{X}^T \boldsymbol{y} \quad . \end{gathered} \tag{7}$$

We see that $\frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$ is the Ridge regression estimate with parameter λ_2. Hence, performing Lasso on this augmented problem yields an elastic net solution. The *least angle regression* (LARS) model selection method proposed by [14] can be used with advantage to compute the Lasso solution on the augmented problem. The LARS algorithm obtains the Lasso solution with a computational speed comparable to computing the OLS solution of the full set of covariates.

The algorithm uses the LARS implementation with the Lasso modification as described in the following section. Hence, we have the parameter λ_2 to adjust, but also the number of iterations for the LARS algorithm can be used. The larger λ_2, the more weight is put on the Ridge constraint. The Lasso constraint is weighted by the number of iterations. Few iterations corresponds to a high value of λ_1, and vice versa. The number of iterations can also be used to ensure a low number of active variables like the forward selection procedure.

3.2 Least angle regression

The least angle regression selection (LARS) algorithm method proposed by Efron et. al [14] finds the predictor most correlated with the response, takes a step in this direction until the correlation is equal to another predictor, then it takes the equiangular direction between the predictors of equal correlation (*the least angle direction*) and so forth.

By ensuring that the sign of any non-zero coordinate β_j has the same sign as the current correlation $\hat{c}_j = \boldsymbol{x}_j^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, the LARS method yields all Lasso solutions[3]. This result is obtained by differentiating the Lagrange version of the Lasso problem. For further details see [14].

3.3 Distance measure

By introducing a distance measure we obtain a measure of how close a new image is to the different individuals in the database. We used the absolute difference between the predicted value $\hat{y}_k$ for model k and the true value y_k for an image belonging to individual k as a measure of the distance between the new image and individual k.

[3] $\boldsymbol{y}$ is centered and normalized to unit length, $\mathbf{X}$ is normalized so each variable has unit length, and $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

4 Results and comparison

In order to test the performance of LARS-EN with respect to the previously commented geometrical and color face recognition technique, two identification experiments were conducted. The difference of the experiments is in the used features. In the first experiment, only the landmarks were used. The second experiment considered only the color. In order to conduct the experiments, the XM2VTS database was used [15]. Eight images for each of the first 50 persons were selected. For all experiments a 4-2-2 strategy was chosen: 4 images of each person to train the model, 2 images of each person to adjust the parameters in the model, and 2 images of each person to verify the model.

To evaluate the performance of the algorithms we used rank plots of the cumulative match scores as proposed in [16]. The horizontal axis of the rank plots is the rank itself (referring to the sorted distance measure) and the vertical axis is the cumulated probability of identification. Hence, we obtain an answer to the question: "Is the correct match in the top n matches?".

4.1 Geometrical face recognition

In order to conduct this first experiment, a set of 64 landmarks were placed along the face, eyes, nose and mouth of each of the 400 selected images. Figure 1 displays the landmarks used in the experiment.

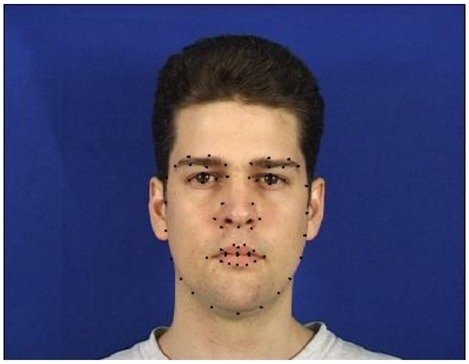

Fig. 1. Illustration of the landmarks used in the experiment.

Table 1 summarizes the classification rates obtained using only the landmarks. The LARS-EN method has higher classification rates than Procrustes, Refined Procrustes, and PCA, but not the Fisher method.

The LARS-EN models included on average 52 of the 128 shape features (x and y coordinates of the landmarks). It should be noted that the mean square error of both the training and the test set in LARS-EN were of the same size, i.e. no severe overfitting was observed. Furthermore, LARS-EN seems to be more

Method/Classification rate	Training	Validation	Test
Procrustes	1.00	-	0.67
Refined Procrustes	1.00	0.76	0.52
PCA	1.00	0.73	0.63
PCA+Fisher	1.00	0.88	0.81
LARS-EN	0.96	0.76	0.71

Table 1. Summary of the classification rates for the models based solely on the landmarks.

honest in the training error and in that sense overfit less than the other methods compared. Figure 2 illustrates a rank plot of the performances of the landmark models. We see a good performance for LARS-EN better than PCA, Refined Procrustes, and Procrustes, and also based on fewer features.

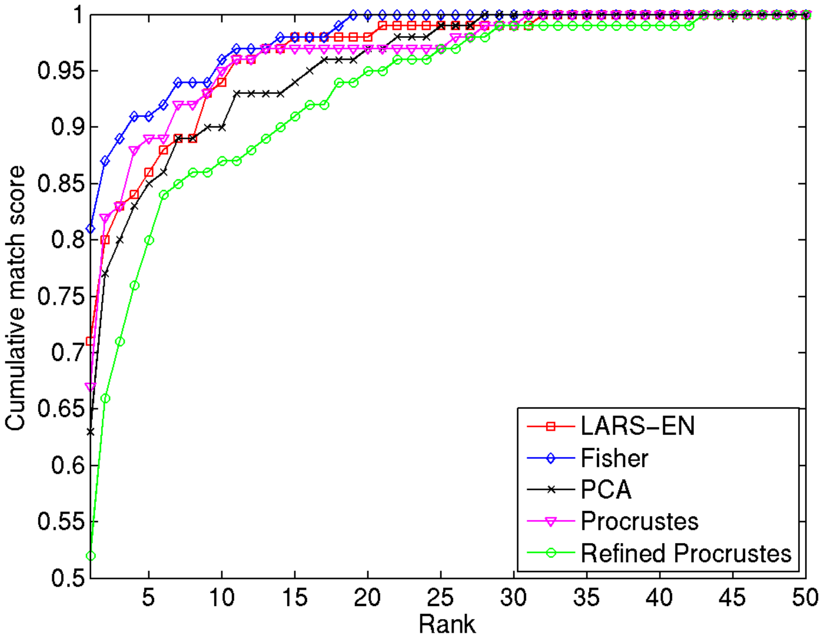

Fig. 2. Identification performance of the models based solely on landmarks.

Figure 3 illustrates which landmarks are selected for four of the individual models. Observe how the selected landmarks depend on the facial characteristics of each person.

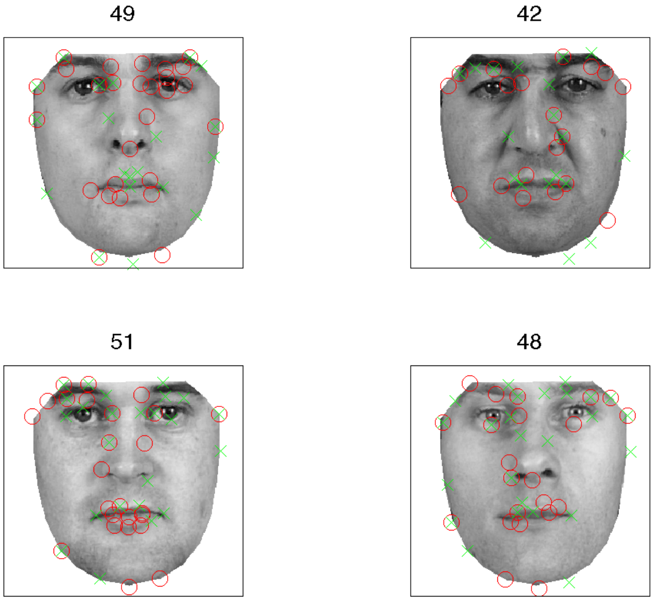

Fig. 3. Illustration of four persons and the selected landmarks in the individual LARS-EN models. x-coordinates are marked with crosses, and y-coordinates are marked with circles. From left to right the person are: No. 1, no. 13, no. 36, and no. 44.

4.2 Color face recognition

In order to obtain a one to one correspondence of pixels between the images the faces were aligned with warping. The same 4-2-2 validation strategy as before was applied and the Eigenfaces, Fisherfaces, and LARS-EN methods were compared. Table 2 summarizes the results.

Method/Classification rate	Training	Validation	Test
Eigenfaces	1	0.87	0.85
Fisherfaces	1	0.96	0.94
LARS-EN	1	0.97	0.92

Table 2. Summary of the classification rates for the models based solely on the color information.

Based on color information we observed higher classification rates than those for LARS-EN based on geometrical information. LARS-EN and Fisherfaces were comparable while both were better than Eigenfaces. The LARS-EN models included around 2000 features (pixels) out of approximately 47000. Figure 4 illustrates the performance of the color based methods. The performance of Fisherfaces was slightly better than for the other two methods which were comparable in performance.

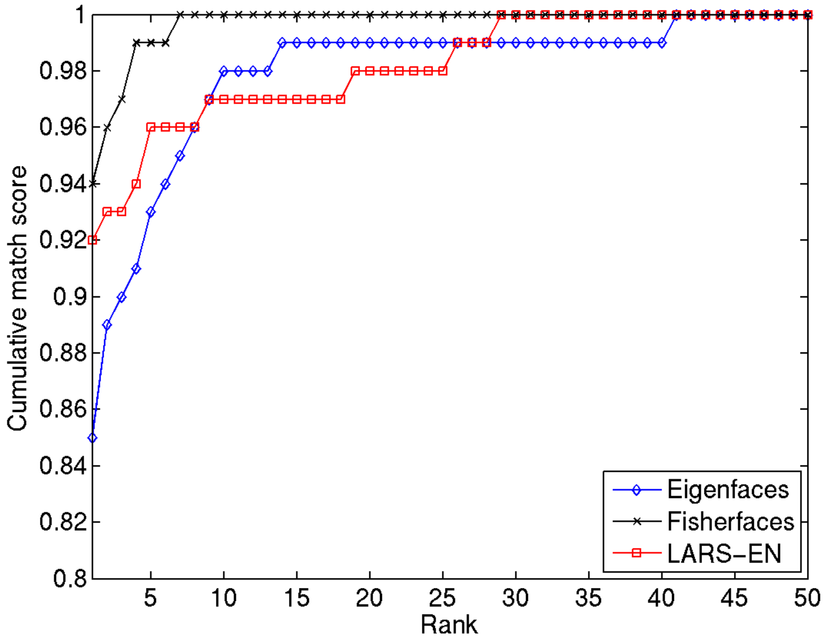

Fig. 4. Identification performance of the models based solely on color information.

Similar to what was done for the geometric features we now examine which features were selected in experiment two. Figure 5 shows the selected color pixels on four different persons. The selected pixels are to a high degree situated around the eyebrows, the eyes, the nose, and the mouth, but also on e.g. the cheeks and the chin. Furthermore, the features are individual from person to person. Observe e.g. the different selection of pixel features on and around the noses of the individuals.

5 Discussion and conclusion

The LARS-EN method performed better than the reference methods Procrustes, refined Procrustes, and PCA, but nor better than PCA+Fisher when based solely on information from landmarks.

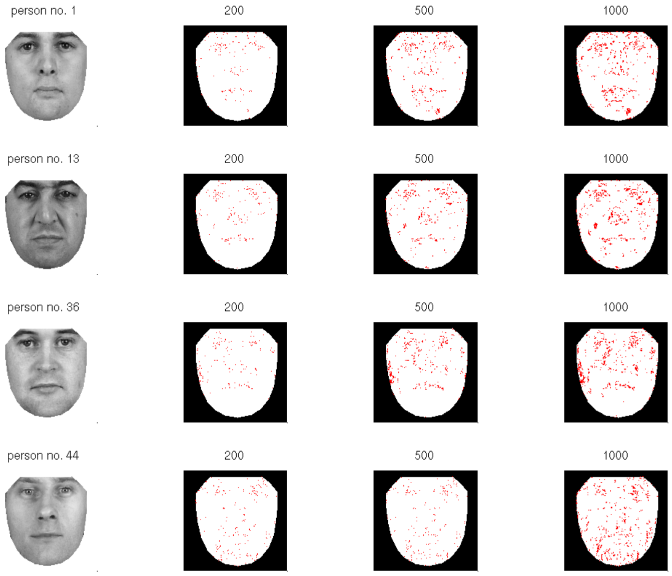

Fig. 5. Illustration of four persons with the first 200, 500, and 1000 selected pixels marked.

Based on color information the LARS-EN models obtained better classification rates than the Eigenfaces and classification rates comparable to Fisherfaces.

Additionally, we identified important features via the feature selection. For the landmarks, only 52 features were needed on average for the individual models. The color models were based on around 2000 features which were situated around the eyes, the nose, the mouth, and the eyebrows, but also on the cheeks and the chin. The selected features differ from individual to individual. Furthermore, the reduction of the feature space decreases the computational efforts for predictions.

Consequently, our results show that a limited number of geometrical or color features can suffice for face recognition, and emphasize that geometrical information should not be disregarded. There are several other possibilities of feature extraction from geometrical information of faces, such as ratios and angles between landmarks, which would be interesting to explore. The LARS-EN algorithm is a good tool for exploring new feature spaces and finding the more interesting ones.

In future work, it is furthermore of interest to examine the methods for a larger database.

6 Acknowledgements

The authors would like to thank Karl Sjöstrand who has implemented the LARS and LARS-EN methods in Matlab. The implementations are available at his homepage[4].

References

1. Daugman, J.: How iris recognition works. Proceedings of 2002 International Conf. on Image Processing **1** (2002)
2. Daugman, J.: High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(11) (1993) 1148–1161
3. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc. B **67**(Part 2) (2005) 301–320
4. Goldstein, A.J., Harmon, L.D., B., A.: Lesk and identification of human faces. Proc. IEEE **59**(5) (1971) 748–760
5. Craw, I., anf T. Kato, N.C., Akamatsu, S.: How should we represent faces for automatic recognition. IEEE Trans. Pattern Anal. Mach. Intell. **21**(8) (1999) 725–736
6. Dryden, I., Mardia, K.: Statistical Shape Analysis. Wiley series in probability and statistics (1998)
7. Shi, J., Samal, A., Marx, D.: How effective are landmarks and their geometry for face recognition? Computer Vision and Image Understanding (**102**(2006)) 117–133
8. Turk, M., Pentland, A.: Face recognition using eigenfaces. IEEE Conf. Computer Vision and Pattern Recognition (1991)
9. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence **19**(7) (1997) 711–720
10. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(1) (2005) 4–13
11. Viola, P., Jones, M.: Robust real-time object detection. In Proc. of IEEE Workshop on Statistical and Computational Theories of Vision (2001)
12. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B **58**(No. 1) (1996) 267–288
13. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12** (1970) 55–67
14. Efron, B., Hastie, T., Johnstore, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32** (2004) 407–499
15. Messer, K., Kittler, J.M.J., Luettin, J., Maitre, G.: Xm2vtsbd: The extended m2vts database. Proceedings 2nd Conference on Audio and Video-base Biometric Personal Verification (AVBPA99) (1999)
16. Philip, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence **22**(10) (2000) 1090–1104

[4] www.imm.dtu.dk/∼kas