Learning Support Vectors for Face Verification and Recognition

K. Jonsson[1,2], J. Matas[1,2], J. Kittler[1] and Y. P. Li[1]

[1] CVSSP, University of Surrey
Guildford, Surrey GU2 5XH, United Kingdom

[2] CMP, Czech Technical University
121 35 Prague, Czech Republic

Abstract

The paper studies Support Vector Machines (SVMs) in the context of face verification and recognition. Our study supports the hypothesis that the SVM approach is able to extract the relevant discriminatory information from the training data and we present results showing superior performance in comparison with benchmark methods. However, when the representation space already captures and emphasises the discriminatory information (e.g. Fisher's linear discriminant), SVMs loose their superiority. The results also indicate that the SVMs are robust against changes in illumination provided these are adequately represented in the training data. The proposed system is evaluated on a large database of 295 people obtaining highly competitive results: an equal error rate of 1% for verification and a rank-one error rate of 2% for recognition (or 98% correct rank-one recognition).

1. Introduction

High-security verification systems based on biometric modalities such as iris, retina and fingerprints have been commercially available for some time. However, one of the most attractive sources of biometric information is the human face since highly discriminative measurements can be acquired without user interaction. The recognition of faces is a well established field of research and a large number of algorithms have been proposed in the literature. Popular approaches include the ones based on Eigenfaces [11], dynamic link matching [5], and active appearance models [2]. These techniques vary in complexity and performance and the choice of algorithm is typically dependent on the specific application. The verification problem, on the other hand, is less explored. Recent examples include [4] in which a robust form of correlation is applied to face authentication.

The effectiveness of the SVM approach to face verification was evaluated in [3]. In this paper we extend this evaluation and also apply the approach to the recognition scenario. An earlier study of SVMs in face verification has been reported by Phillips [9]. An SVM verification system design was compared with a standard Principal Component Analysis (PCA) face authentication method and the former was found to be significantly better. In this approach the SVM was trained to distinguish between the populations of within-client and between-client difference images respectively, as originally proposed by Moghaddam [8]. This method gives client non-specific support vectors.

In our approach we adopt a client-specific solution which requires learning client-specific support vectors. However, this is not the main distinguishing feature of our work: it only reflects the choice of representation which is different from [9]. Our primary motivation for carrying out a similar study was to establish why the performance of the SVM approach is superior. We want to investigate the inherent potential of SVMs to extract the relevant discriminatory information from the training data irrespective of representation and pre-processing. In order to achieve this objective we have designed experiments in which faces are represented in both Principal Component (PC) and Linear Discriminant (LD) subspaces. The latter basis (Fisherfaces) is used as an example of a face representation with focus on discriminatory feature extraction while the former achieves simply data compression. We also study the effect of image photometric normalisation on the performance of the SVM method.

A number of criteria have been considered as a basis for the SVM approach evaluation, using other standard techniques as benchmarks. We have included as benchmark verification and recognition methods not only the classical PC variants with the L^2 norm and correlation coefficient respectively, but also the LD space with the same two decision schemes. As criteria for evaluating SVMs in relation to the benchmark methods we have concentrated on

the following: ability to extract discriminatory information, robustness (sensitivity to input data conditioning), and performance in different identity classification scenarios (verification and recognition). These three criteria are expressed quantitatively in terms of the verification and recognition error rates.

The findings of our study strongly support the hypothesis that the SVM approach is powerful in the sense of being able to extract the relevant discriminatory information from the training data. This is the main reason for the large difference between the observed performance of the classical Eigenface classification methods and SVMs (factor of almost 3). When the representation space already captures and emphasises the discriminatory information content as in the case of the LD bases, SVMs cease to be superior to the simple Euclidean distance or correlation decision rules.

SVMs also show a superior capability to cope with illumination changes, provided these are adequately represented in the training data. However, on data which has been sanitised by feature extraction (Fisherfaces) and/or normalisation, SVMs may suffer from over-training, resulting in a loss of the ability to generalise. SVMs involve many parameters and can employ different kernels. This makes the optimisation space rather extensive, without the guarantee that it has been fully explored to find the best solution.

The paper is organised as follows. In the next section we introduce the two face representation spaces used in our study, namely Eigenfaces and Fisherfaces. In Section 2 we overview the SVM approach to face identity verification and recognition, and summarise the benchmark classification methods. Section 3 introduces the face database used in experimentation and describes the experiments carried out, their objectives and the results obtained. Finally, in Section 4 conclusions are drawn.

2. Face verification and recognition

Any verification or recognition process involves two basic computational stages. In the first stage a suitable representation is derived with the multiple objective of making the subsequent, decision-making stage, computationally feasible, immune to environmental changes during the biometric data acquisition, and effective by providing it only with information which is pertinent to the decision-making task. In the case of verification, the purpose of the second stage is to accept or reject the identity claim corresponding to a probe biometric measurement. This is basically a two-class pattern recognition problem. In the case of recognition, there is no claim and the objective is to establish the identity corresponding to the measurement. This is an n-class problem where n is the number of subjects in the database. In the following subsections we introduce the methods adopted for the design of each of these two stages

in the context of the face verification and recognition study pursued in this paper.

2.1. Representation of faces

The first step in the face representation process involves image pre-processing in order to establish correspondence between face images to be compared. Once an image is registered, it can further be normalised photometrically. In our study we set out to investigate the resilience of different decision-making methods to varying illumination and thus this step was applied only in a subset of experiments. In the final step of processing, the image is projected into a coordinate system which facilitates the decision making process computationally and possibly emphasises the important attributes for face verification.

Geometric normalisation. As the focus of the paper is on the decision-making aspects of face verification and recognition we have tried to eliminate the dependency of our experiments on processes which may lack robustness. For this reason we have performed face registration semi-automatically. This is the only part of the method in which manual intervention is required and it allows us to separate the issues of localisation and verification/recognition (cf the FERET face recognition test [10]). The procedure is based on manually localised eye positions. Four parameters computed from the eye coordinates (rotation, scaling and translation in the horisontal and vertical directions) are used to crop the face part from the original image and scale it to any desired resolution.

Photometric normalisation. Two different approaches to photometric normalisation were evaluated. In the fist approach, the pixel values of the geometrically normalised image were shifted using the mean and scaled by their standard deviation, estimated over the whole cropped image. In the second approach, the distribution of image intensity values was flattened using histogram equalisation.

Image projection. Suppose that we have c clients and M training face images $x_i, i = 1, \ldots, M$, $x_i \in R^D$ each belonging to one of the client classes $\{C_1, C_2, \ldots, C_c\}$. Then we can define the following second-order statistics:

- Between-class scatter matrix:

$$S_B = \frac{1}{c} \sum_{k=1}^{c} (\mu_k - \mu)(\mu_k - \mu)^T \qquad (1)$$

- Within-class scatter matrix:

$$S_W = \frac{1}{M} \sum_{k=1}^{c} \sum_{i \mid x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \quad (2)$$

where μ is the grand mean and μ_k is the mean of class C_k. The sum of these scatter matrices is typically denoted the total scatter, S_T.

The aim of the Principal Component Analysis is to identify the subspace of the image space spanned by the training face image data and to decorrelate the pixel values. This can be achieved by finding the eigenvectors W_{pca} of matrix S_T associated with nonzero eigenvalues Λ by solving

$$S_T W_{pca} - W_{pca} \Lambda = 0 \quad (3)$$

These eigenvectors are referred to as Eigenfaces. The classical representation of a face image is obtained by projecting it to the coordinate system defined by the Eigenfaces.

The projection of face images into the Principal Component (Eigenface) subspace achieves information compression, decorrelation and dimensionality reduction to facilitate decision making. If one is also interested in identifying important attributes (features) for face verification and recognition, one can adopt a feature extraction mapping. A popular technique is to find the Fisher linear discriminants (Fisherfaces) by solving

$$S_B W_{lda} - S_W W_{lda} \Lambda = 0 \quad (4)$$

The projection of a face image into the system of Fisherfaces associated with nonzero eigenvalues will yield a representation which will emphasise the discriminatory content of the image. The solution of the generalised eigenvalue problem in Equation 4 is known, but due to the high dimensionality many standard methods fail and the choice of a stable numerical algorithm is non-trivial [6].

In Section 3, we perform experiments with different number of basis vectors taken from either the Eigenface or Fisherface systems. In the following, for the sake of notational simplicity, we shall not distinguish between the two different basis systems, nor shall we explicitly denote the dimensionality of the representation space. The actual representation used will be clear from the experiment description. Thus in general, in each experiment we shall work with some transformation matrix W. A sample face image $\mathbf{y}$ will then be represented by a projection $\mathbf{x}$ obtained as $\mathbf{x} = W^T \mathbf{y}$. Similarly, the client model μ_k will be projected into a vector ω_k in the appropriate representation space.

2.2. Classification

Support vector machines. The main decision making tool investigated in this paper is the Support Vector Machine. Below we give a brief presentation of the basic theory. The reader is referred to [1] for a more comprehensive

introduction. SVMs are based on the principle of structural risk minimisation. The aim is to minimise the upper bound on the expected (or actual) risk defined as[1]

$$R(\alpha) = \int \frac{1}{2} |z - f(\mathbf{x}, \alpha)| dP(\mathbf{x}, z) \quad (5)$$

where α is a set of parameters defining the trained machine, z a class label associated with a training sample $\mathbf{x}$, $f(\mathbf{x}, \alpha)$ a function providing a mapping from training samples to class labels, and $P(\mathbf{x}, z)$ the unknown probability distribution associating a class label with each training sample. Let l denote the number of training samples and choose some η such that $0 \leq \eta \leq 1$. Then, with probability $1 - \eta$, the following bound on the expected risk holds:

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}} \quad (6)$$

where $R_{emp}(\alpha)$ is the empirical risk as measured on the training set and h is the so called Vapnik Chervonenkis (VC) dimension. The second term on the right hand side is called the VC confidence. There are two strategies for minimising the upper bound. The first one is to keep the VC confidence fixed and to minimise the empirical risk and the second one to fix the empirical risk (to a small value) and minimise the VC confidence. The latter approach is the basis for SVMs and below we will briefly outline this procedure.

First consider the linear separable case. We are looking for the optimal hyperplane in the set of hyperplanes separating the given training samples. This hyperplane minimises the VC confidence and provides the best generalisation capabilities. Giving a geometric interpretation, the optimal hyperplane maximises the sum of the distances to the closest positive and negative training samples. This sum is called the *margin* of the separating hyperplane. It can be shown that the optimal hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ (where $\mathbf{w}$ is normal to the hyperplane) is obtained by minimising $\|\mathbf{w}\|^2$ subject to a set of constraints. This is a quadratic optimisation problem.

These concepts can be extended to the non-separable and non-linear case. The separability problem is solved by adding a term to the expression subject to minimisation. This term is the sum of the deviations of the non-separable training samples from the boundary of the margin. This sum is weighted using a parameter controlling the cost of misclassification. The second problem is how to handle non-linear decision boundaries. This is solved by mapping the training samples to a high-dimensional feature space using kernel functions. In this space the decision boundary is linear and the techniques outlined above can be directly applied. The kernel functions used in the experiments reported

---

[1]The notation is similar to the one in [1].

in Section 3 are radial basis functions defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) \quad = \quad e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2} \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ denote two samples. The γ-value is a user-controlled parameter. The reader is referred to [3] for an evaluation of the relative performances of different kernels.

In addition to the SVMs we have also implemented the following standard classification rules as baselines for experimental comparison.

Euclidean distance. The most commonly used decision rule is based on the Euclidean distance between the sample projection $\mathbf{x}$ and the projection of the k-th client mean ω_k, i.e.

$$d_E(\mathbf{x}, \omega_k) = \sqrt{(\mathbf{x} - \omega_k)^T (\mathbf{x} - \omega_k)} \tag{8}$$

In verification, the claimed client identity is accepted if $d_E(\mathbf{x}, \omega_k)$ is below a threshold τ_{Ek}. Otherwise it is rejected.

Normalised correlation. Alternatively, the decision can be based on the correlation score

$$d_C(\mathbf{x}, \omega_k) = \frac{|\mathbf{x}^T \omega_k|}{|\mathbf{x}||\omega_k|} \tag{9}$$

In the case of the correlation measure the claimed identity is accepted if $d_C(\mathbf{x}, \omega_k)$ exceeds a pre-specified threshold τ_{Ck}.

Thresholding. The client-specific threshold τ_k is determined using an independent evaluation set. Given the mean μ_k and the standard deviation σ_k of the impostor distances in the evaluation set, the threshold is computed as

$$\tau_k = \mu_k + \tau \cdot \sigma_k \tag{10}$$

where τ is a global threshold. In the case of recognition, there are no thresholds involved but the distances are normalised in a similar client-specific way:

$$d_n = \frac{d - \mu_k}{\sigma_k} \tag{11}$$

where d and d_n are the original and normalised distances, respectively.

3. Experimental results

The experiments summarised below were all performed on frontal-face images from the extended M2VTS multimodal database [7]. This publicly available database contains face images and speech recordings of 295 persons.

The subjects were recorded in four separate sessions uniformly distributed over a period of 5 months, and within each session a number of shots were taken including both frontal-view and rotation sequences. In the frontal-view sequences the subjects read a specific text (providing synchronised image and speech data), and in the rotation sequences the head was moved vertically and horizontally (providing information useful for 3D surface modeling of the head).

The verification experiments were conducted according to the Laussane evaluation protocol [7]. This protocol provides a unified framework within which the performances of vision- and speech-based person authentication systems running on the extended M2VTS database can be measured. The protocol specifies a partitioning of the database into three disjoint sets: a training set (200 clients), an evaluation set (200 clients and 25 impostors) and a test set (200 clients and 70 impostors). The training set is used to build client models, the evaluation set to establish client-specific verification thresholds, and the test set to obtain estimates of the verification rate on independent data. This partitioning of the database was also used in the recognition experiments (excluding the impostors).

3.1. Verification and recognition results

Verification and recognition experiments were performed for the two different representations (PC and LD basis) and for three different photometric normalisation techniques (no normalisation, zero mean and unit variance, and histogram equalisation). The results are listed in Tables 1 and 2 for verification and recognition, respectively. In the case of verification, the performance is measured in terms of the false acceptance and the false rejection. The operating point where these two error rates equal each other is typically referred to as the equal error rate. Starting with the evaluation set, we analyse the receiver operating characteristics (see Figure 1) to find the equal error rate. The corresponding threshold is then applied to the test set yielding the false acceptance and false rejection values listed in Table 1. Also shown is the mean error rate which is simply the average of the false acceptance and the false rejection. For recognition, we measure the percentage of misclassifications corresponding to a given rank. This error measure was chosen to be consistent with the verification experiments. Also, we keep the same partitioning of the data set but there are no thresholds involved in the performance evaluation. Note that the error measure is directly related to the cumulative rank score which is typically used in the object recognition literature.

Let us first of all look at the relative performances of the different classifiers. One can see that, in the PCA subspace, the error rates decrease monotonically with the classifier (Euclidean distance, normalised correlation and SVMs) and

PCA SUBSPACE

Cls	Nrm	Evaluation set		Test set		
		Thr	EER	FA	FR	ME
EDs	×	0.63	9.92	9.82	6.08	7.95
NCr		3.91	7.68	8.32	6.50	7.41
SVM		4.34	3.00	3.65	1.29	2.47
EDs	ZMn	0.40	8.45	9.39	4.50	6.95
NCr		4.13	6.60	7.39	4.25	5.82
SVM		4.40	3.00	3.64	2.00	2.82
EDs	HEq	0.09	5.36	6.46	4.00	5.23
NCr		4.45	4.32	5.01	3.75	4.38
SVM		4.79	1.50	2.19	1.50	1.85

LDA SUBSPACE

Cls	Nrm	Evaluation set		Test set		
		Thr	EER	FA	FR	ME
EDs	×	0.66	9.47	10.29	6.22	8.26
NCr		5.44	1.69	2.23	1.25	1.74
SVM		5.07	1.18	1.54	1.25	1.40
EDs	ZMn	-0.92	1.72	1.96	1.00	1.48
NCr		5.55	1.50	1.95	1.25	1.60
SVM		5.03	1.25	1.58	1.62	1.60
EDs	HEq	0.34	5.30	7.58	4.25	5.92
NCr		5.66	1.25	1.56	0.75	1.15
SVM		5.07	1.00	1.37	0.75	1.06

Table 1. Verification performance: equal error rate (EER), false acceptance (FA), false rejection (FR) and mean error rate (ME) as functions of classification method (Cls) and normalisation (Nrm). Classifiers: Euclidean distance (EDs), normalised correlation (NCr) and support vector machines (SVM). Normalisations: no normalisation (×), zero mean and unit variance (ZMn) and histogram equalisation (HEq).

PCA SUBSPACE

| Cls	Nrm	Evaluation set			Test set		
		R1	R2	R10	R1	R2	R10
EDs	×	23.75	17.75	9.00	24.75	16.75	6.00
NCr		24.75	17.75	8.75	21.75	15.25	7.00
SVM		8.75	5.25	2.00	6.75	3.25	1.25
EDs	ZMn	19.75	14.00	6.75	18.50	14.75	5.00
NCr		18.25	13.25	6.50	16.00	12.00	5.75
SVM		9.00	5.00	2.25	7.00	4.50	1.50
EDs	HEq	13.00	9.50	4.25	13.00	10.00	3.50
NCr		10.75	7.75	4.25	12.75	9.00	3.00
SVM		3.50	2.25	0.50	3.75	2.00	0.75

LDA SUBSPACE

Cls	Nrm	Evaluation set			Test set		
		R1	R2	R10	R1	R2	R10
EDs	×	3.25	1.75	0.75	2.50	1.75	0.50
NCr		3.50	2.75	0.75	2.25	1.75	0.75
SVM		3.75	2.25	0.75	2.50	1.75	0.50
EDs	ZMn	3.00	2.00	0.75	2.50	1.50	0.50
NCr		3.25	2.75	1.00	3.25	2.25	0.75
SVM		3.75	2.25	1.00	3.25	1.75	0.50
EDs	HEq	2.00	1.25	0.75	2.25	0.75	0.25
NCr		1.75	1.50	1.00	2.00	1.50	0.25
SVM		2.00	1.25	1.00	2.25	1.25	0.25

Table 2. Recognition performance: rank 1 (R1), rank 2 (R2) and rank 10 (R10) error rates as functions of classification method and normalisation (see caption of Table 1).

the SVMs clearly outperform the other two decision rules. This is true for both verification and recognition. However, in the LDA subspace, the relationships between the different techniques are less clear. The verification experiments show that, independent of photometric normalisation, the normalised correlation perform at more or less the same level as the SVMs. Indeed, for the case of zero-mean normalisation, all classifiers produce very similar error rates. This trend is even more pronounced in the recognition scenario and here the SVMs do no longer outperform the benchmark methods.

Comparing the Euclidean distance and the normalised correlation, we know that the former decision rule is particularly sensitive to deviations from the implicit model underlying the approach, i.e. client clusters being very compact and roughly spherical. The correlation coefficient can cope better with deviations from the sphericity. However, once the data is of that form as in the case of the LD bases with normalised data, the inherent flexibility of this classifica-

tion method does not result in any performance advantages in relation to the Euclidean distance.

Looking at photometric normalisation, one can see a similar relationship between the PCA and LDA subspaces. In the case of the PC basis, the error rates drop monotonically with normalisation (no normalisation, zero mean and unit variance, and histogram equalisation). The only exception is the SVMs which exhibit the same performance for unnormalised data and zero-mean normalisation. Again this is true for both verification and recognition. In the case of the LD basis, the verification rates do improve with the normalisation but there are two exceptions. First, the Euclidean distance show high error rates for histogram equalisation and, second, the SVM performance does not change for zero-mean normalisation. For recognition, the error rates are very similar across all photometric normalisations. However, the histogram equalisation is still performing slightly better. A somewhat surprising result is that the Euclidean distance produces very low error rates even for unnormalised data. This is probably due to the ranking as opposed to the absolute thresholding used in the verification experiments.

The error rates listed in Table 1 correspond to single points in the receiver operating characteristics. By varying
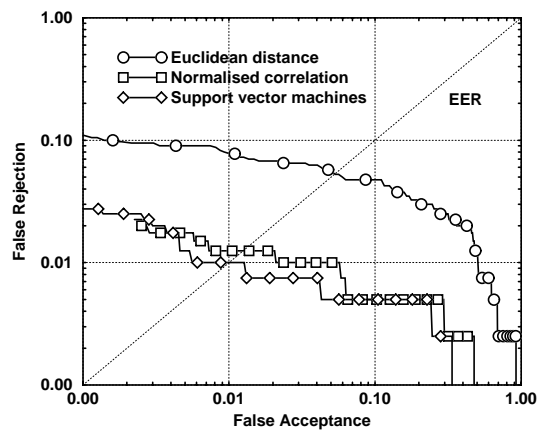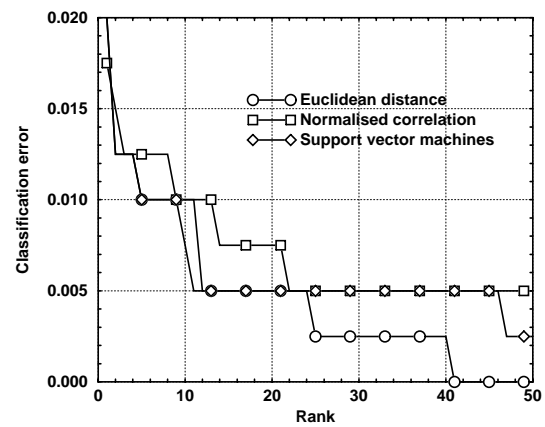
Figure 1. Verification performance: false rejection versus false acceptance for three different classifiers.

Figure 2. Recognition performance: classification error versus rank for three different classifiers.

the verification threshold, we obtain a set of points showing the trade-off between the false rejection and the false acceptance. This is shown in Figure 1 for the LDA subspace using histogram equalisation. The corresponding analysis for the recognition scenario is shown in Figure 2. The points were obtained by thresholding on the rank and computing the classification error.

4. Conclusions

The paper studied SVMs in the context of face verification and recognition. Our study proved the hypothesis that the SVM approach is able to extract the relevant discriminatory information from the training data. This is the main reason for the large difference between the observed performance of the standard Eigenface classification method used as a benchmark and the SVMs (factor of almost 3). When the representation space already captures and emphasises the discriminatory information content as in the case of the LD bases, the SVMs cease to be superior to the simple Euclidean distance or correlation decision rules. The SVMs also show a superior capability to cope with illumination changes, provided these are adequately represented in the training data. We presented an evaluation on a large face database showing competitive error rates for both the verification and recognition scenarios.

The research reported in this paper was carried out within the framework of the European Union ACTS project M2VTS and ESPRIT RETINA.

References

[1] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[2] T. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H. Burkhardt and B. Neumann, editors, *ECCV'98*, pages 484–498, 1998.

[3] K. Jonsson, J. Kittler, Y. P. Li, and J. Matas. Support vector machines for face authentication. In T. Pridmore and D. Elliman, editors, *BMVC'99*, pages 543–553, 1999.

[4] K. Jonsson, J. Matas, and J. Kittler. Learning salient features for real-time face verification. In S. Akunuri and C. Kullman, editors, *AVBPA'99*, pages 60–65, 1999.

[5] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, Mar 1993.

[6] Y. P. Li, J. Kittler, and J. Matas. Effective implementation of linear discriminant analysis for face recognition and verification. In A. Leonardis and F. Solina, editors, *CAIP'99*, pages 234–242, 1999.

[7] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In S. Akunuri and C. Kullman, editors, *AVBPA'99*, pages 72–77, 1999.

[8] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. In *FG'98*, pages 30–35, 1998.

[9] P. J. Phillips. Support vector machines applied to face recognition. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *NIPS'98*, 1998.

[10] S. A. Rizvi, P. J. Phillips, and H. Moon. The FERET verification testing protocol for face recognition algorithms. In *FG'98*, pages 48–53, 1998.

[11] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.