Pattern Recognition Letters 23 (2002) 463–475

Retinal Vision applied to Facial Features Detection and Face Authentication

F. Smeraldi and J. Bigun
Halmstad University, Box 823, S-301 18 Halmstad, Sweden*

May 2001

Abstract

Retinotopic sampling and the Gabor decomposition have a well established role in computer vision in general as well as in face authentication. The concept of *Retinal Vision* we introduce aims at complementing these biologically inspired tools with models of higher-order visual process, specifically the Human Saccadic System. We discuss the *Saccadic Search* strategy, a general purpose attentional mechanism that identifies semantically meaningful structures in images by performing "jumps" (*saccades*) between relevant locations. Saccade planning relies on *a-priori* knowledge encoded by SVM classifiers. The raw visual input is analysed by means of a log-polar retinotopic sensor, whose receptive fields consist in a vector of modified Gabor filters designed in the log-polar frequency plane. Applicability to complex cognitive tasks is demonstrated by facial landmark detection and authentication experiments over the M2VTS and Extended M2VTS (XM2VTS) databases.

Keywords: Facial Feature Detection, Face Authentication, Human Saccadic System, Log-Polar Mapping, Support Vector Machine

1 Introduction

The spreading of Internet commerce, tele-banking and similar services which require privileged and possibly remote access to resources on the part of accredited users has put a strong emphasis on the development of reliable biometrics for person authentication. A wide choice of techniques has been proposed to meet this demand. Despite its intrinsic complexity face based authentication, admittedly a "natural" choice, still remains of particular interest because it is perceived as psychologically and/or physically "non-invasive" by users. Its failure modes (identical twins, scarce illumination, disguise, etc.) are furthermore essentially unrelated to the failure modes of other biometrics such as speech (acoustic noise, professional imitators, etc.) or fingerprints (missing or damaged fingerprints, etc). Therefore, face authentication can be advantageously included in multi-modal systems, which get an edge in robustness through the fusion of several authentication modalities (Bigun et al., 1998).

A variety of face authentication techniques has been developed in recent years. While certain approaches such as Eigenfaces (Sirovich and Kirby, 1990; Turk and Pentland, 1991) and Fisherfaces (Belhumeur et al., 1997) are entirely statistic in nature, some of the proposed methods support or are influenced by theories of human perception (Chellappa

[*] Part of this research was carried out in the years 1997-99 while the authors were employed at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

et al., 1995). Several algorithms rely on direct biological analogy by the use of *Gabor filters*, which are known to model the responses of the so-called simple cells in the visual cortex (Orban, 1984), and of *retinotopic sampling* (Takács and Wechsler, 1995; Tistarelli and Grosso, 1998). The use of the Gabor decomposition for face recognition purposes has been introduced by van der Malsburg in the context of elastic graph matching (Lades et al., 1993). This technique has then been substantially developed in the framework of face authentication (Bigun et al., 1998; Duc et al., 1999).

The intrinsic difficulty of the problem along with the demonstrated effectiveness of multi-scale tools and nonlinear sampling make face authentication the ideal test-bed for a new biologically inspired vision paradigm, which we call *Retinal Vision.* With it we pursue the integration of low-level biologically inspired signal conditioning with models of higher-order visual processes that constitute the interface to cognitive tasks in human beings. The aim is implementing attentional mechanisms that would allow a cognitive task to steer visual acquisition and processing. At an early stage of the chain that goes from visual stimuli to symbolic representations we find the *Human Saccadic System.* Humans and primates do not scan a visual scene in a raster-like fashion: they rather perform large jumps, known as *saccades* (Yarbus, 1967), between the different points of interest in the scene, on which the gaze is kept centred for a fraction of a second (*fixations*). Saccades are known to play a role in the underlying cognitive processes (Pelz, 1995); the saccadic pattern, as has been demonstrated by Yarbus already in the fifties, depends both on the visual scene and on the task to be performed. The main regions of interest for the face recognition/authentication task are, as is well known, the eyes and the mouth of a subject (Keating and Keating, 1993). We therefore propose to locate such facial landmarks using a *Saccadic Search* strategy built around a *log-polar retina*, which is used to sample the Gabor decomposition of the image (Smeraldi et al., 1999a).

Computational models of saccades based on the responses of multi-scale orientation selective filters such as derivatives of Gaussians (Bigun et al., 1991) have already been proposed (Rao et al., 1997), although no explicit description of a target was used to drive the saccades. Our Saccadic Search strategy makes use of *a-priori knowledge* in the form of appearance-based models of the eyes and the mouth. The models, which are implemented by means of Support Vector Machine (SVM) classifiers, describe the Gabor signature of the target features.

Face authentication is achieved using three independent machine experts to process the Gabor features extracted in the facial regions surrounding the eyes and the mouth. Each expert, implemented as an SVM, outputs an authentication score. These are then combined to achieve the final decision on the identity claim being processed.

We report experimental results on both the M2VTS and the Extended M2VTS (XM2-VTS) databases, featuring images from 37 and 295 subjects respectively. Authentication tests were performed according to the standard test protocols established by the European M2VTS consortium, thus allowing direct comparison of our results with those published by other research groups working on the same databases.

2 Image Representation

2.1 The retinotopic sampling grid

The Saccadic Search strategy and the face authentication algorithm are based on a sparse retinotopic grid obtained by *log-polar mapping* (Schwartz, 1980; Bigun, 1993). The grid is used to sample the Gabor decomposition of the image, each retinal point being associated with a *receptive field* represented by the support of a set of Gabor filters. Such a sensor can be viewed as a simple model of the ganglion cell lattice and the simple cells in V1 (Hubel, 1988; Takács and Wechsler, 1995). In our experiments we used a grid consisting of 50

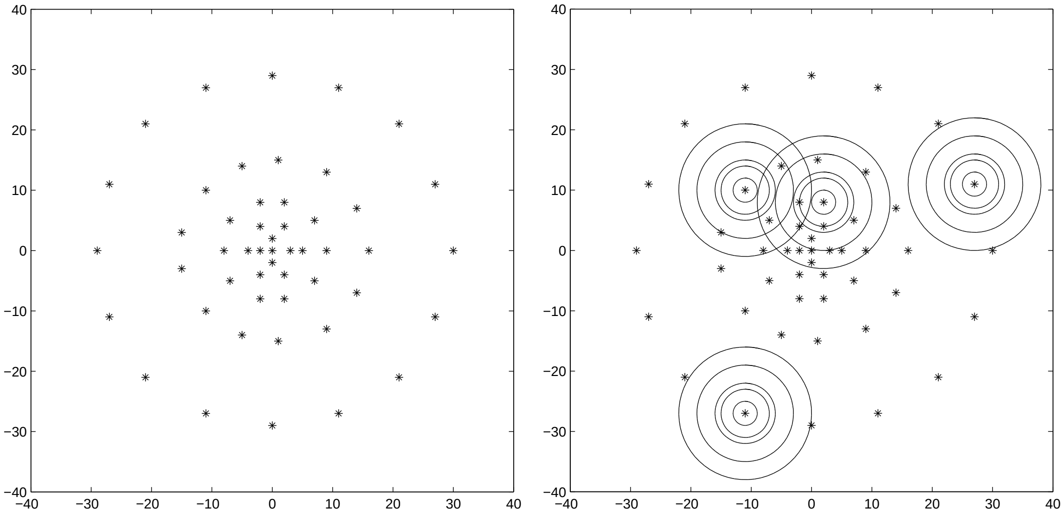

Figure 1: Left: the retinotopic sampling grid (axes graded in pixels). The slight asymmetry in the centre of the image is due to discretization (the radius of the inner circle is only 3 pixels). Right: a few receptive fields are represented as sets of concentric circles. For each Gabor frequency channel, the circles show the pass–band of the corresponding filters.

points arranged in 5 concentric circles (Figure 1). We set the radius of the innermost circle to 3 pixels and that of the outermost circle to 30 pixels, which is about the average inter-eye distance in the images we used for testing.

Receptive fields are modelled by computing a vector of 30 Gabor filter responses at each point of the retina. The filters are organised in 5 frequency channels and 6 equally spaced orientation channels. Filter wavelengths span the range from 4 to 16 pixels in half-octave intervals.

The same sensor is used both for facial landmark localisation and for person authentication. The sparse nature of the sensor and the relatively low number of fixations required for the saccadic search to converge to the facial landmarks make the computation of Gabor responses by direct filtering in the image domain feasible.

2.2 Log-polar mapping in the Fourier domain

The local power spectrum of the image is sampled at each retinal point by the vector of Gabor filters that constitute the receptive field associated with that point. In order for the filter responses to be descriptive of the image neighbourhood and to carry a maximal information content, i.e. to be uncorrelated, the local frequency plane must be covered as uniformly as possible.

Gabor filters are optimal in that they minimise the joint image and frequency plane spread. When only a small number of frequency channels is employed, however, the Gaussian spectrum of the filters results in an excessive overlap towards the (densely sampled) origin of the frequency plane, while the high frequency regions are poorly covered. This is due to the fact that each Gaussian weights high as well as low frequencies in its support in a symmetric manner, whereas the decomposition itself is coarser at high frequencies. In order to compensate for that, we employ a set of *modified Gabor filters* that are defined as Gaussians in the *log-polar* frequency plane (Knutsson, 1982; Bigun, 1994). That is, for

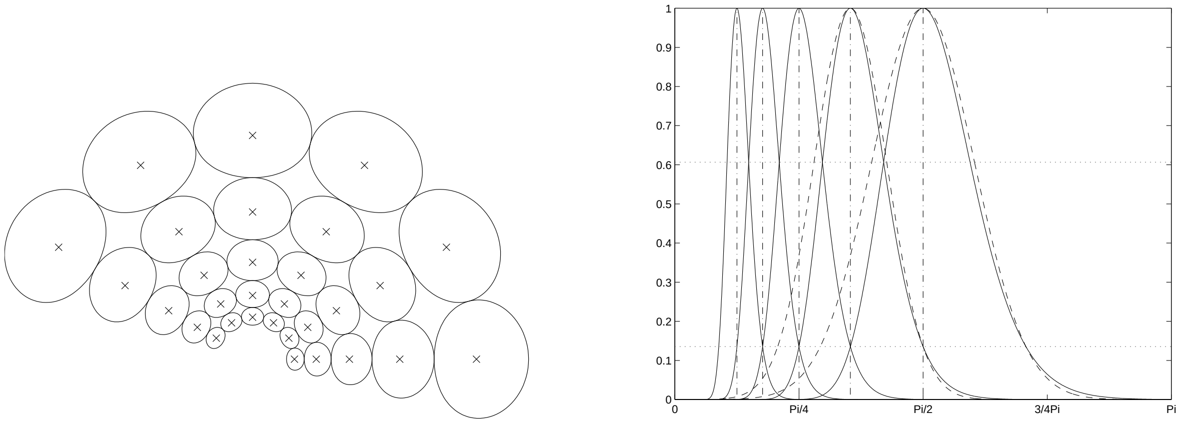

Figure 2: Modified Gabor filters in the Fourier domain. Left: level curves; right: radial cross-sectional plot. On the right, two standard Gabor filters are superimposed as dashed lines. Modified filters have a steeper cut-off on the low frequency side, which effectively reduces the overlap between adjacent channels.

a filter tuned to orientation φ_0 and angular frequency $\omega_0 = exp(\xi_0)$:

$$\hat{G}(\xi, \varphi) = Ae^{-\frac{(\xi-\xi_0)^2}{2\sigma_\xi^2}} e^{-\frac{(\varphi-\varphi_0)^2}{2\sigma_\varphi^2}} \tag{1}$$

where A is a normalisation constant and (ξ, φ) are the log-polar frequency coordinates:

$$(\xi, \varphi) = (\log(|\vec{\omega}|), \tan^{-1}(\omega_x, \omega_y)). \tag{2}$$

The new coordinate system has the property that translations along φ represent rotations in the image domain, while translations along ξ correspond to scaling. A complete bank of filters can then be designed by arranging a set of *identical* Gaussians in a *rectangular lattice* in the log-polar frequency plane (Smeraldi, 2000). When seen in the standard Fourier domain the filters reproduce the familiar “daisy” structure, but the overlap towards low frequencies is significantly reduced (Figure 2). These filters have been previously used in texture analysis, showing high discrimination power (Bigun, 1993).

3 Eyes and Mouth Detection

3.1 Modelling the Facial Landmarks

The *Saccadic Search* strategy requires the use of two levels of modelling for each of the targets (Smeraldi et al., 1999b). The coarser, or *local*, model is obtained for each facial landmark from the vector of Gabor responses extracted at the location of that landmark in the images of the training set. The finer, or *extended* model is obtained by placing the whole retina at the location of the facial landmark on the training images (Figure 3) and collecting the set of Gabor filter responses from all of the retinal points. These Gabor features are then arranged into a single vector.

In both cases, the example vectors for each facial landmark are normalised for contrast invariance, complemented with negative examples and used to train an SVM classifier (Cortes and Vapnik, 1995; Vapnik, 1995). This choice is motivated by the non–parametric nature of SVMs, which frees us from the need of modelling the highly non–trivial distributions of feature vectors that describe the facial landmarks. A training set

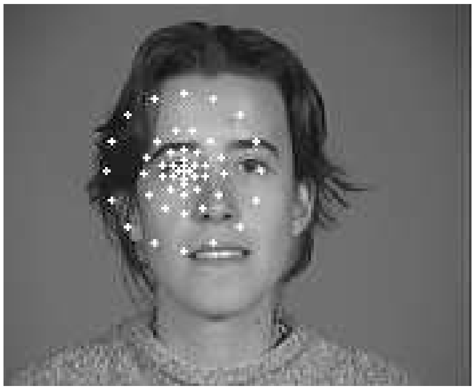

Figure 3: The extended model of a facial landmark is obtained by centring the retina on that landmark in the images of the training set.

of 202 frontal images of 13 subjects has been extracted from the M2VTS database (see section 5). This provides 202 positive examples for each landmark. Negative examples are collected by extracting Gabor vectors at 10 random locations in each image. Also, right eyes and mouths have been included as negative examples in the models of the left eye, and so on. The SVM engine employed was developed following the ideas in (Joachims, 1998) and (Osuna et al., 1997).

3.2 Selecting the Kernel Function

SVM classifiers are characterised by a decision surface that can be written as a hyper-plane in a high dimensional space $\mathcal{H}$. The mapping between the input space and $\mathcal{H}$ is handled implicitly by means of a kernel function K, which is a (generally nonlinear) symmetric scalar function of two input vectors. The decision function can then be written as

$$f(\vec{v}) = \sum_j \alpha_j y_j K(\vec{s}_j, \vec{v}) + b \gtrless 0, \tag{3}$$

where $\vec{v}$ is the input to be classified. The support vectors $\vec{s}_j$ constitute a subset of the training data which is determined through an optimisation process. The optimisation also defines the weight α_j; the y_j are constant and have a value $+1$ for the support vectors of the "accept" class, -1 for those of the "reject" class. Finally, b is fixed so that the hyper-plane in $\mathcal{H}$ cuts exactly halfway between the closest training examples of the two antagonistic classes. This choice is optimal provided that the training set is equally descriptive of both populations. In the case that one class is poorly represented in the training set, we might want to compensate for that by shifting the hyper-plane further away from its support vectors.

In order to determine the most appropriate kernel function for the local and extended models, we used the 349 images from the remaining 24 subjects in the M2VTS database as a probe set. Probe images were used to feed the classifiers with the appropriate set of Gabor responses extracted at the location of the facial landmark they were supposed to model, together with a set of negative examples they were supposed to reject. The negative examples included, for each SVM, the remaining two facial landmarks plus ten random points per image.

Since the optimisation process depends on K, training had to be performed anew for each choice of the kernel. In order to speed up the procedure, the Gabor response set

EER (τ)	**Kernel type**			
Local Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	11% (-5.0)	11% (-2.0)	11% (-2.0)	**10% (-1.9)**
Right eye	13% (-7.0)	11% (-2.3)	11% (-2.0)	**11% (-1.9)**
Mouth	6% (-3.5)	8% (-1.8)	8% (-1.7)	**7% (-1.6)**
Extended Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	0.5% (-0.30)	0.3% (-0.15)	0.3% (-0.05)	**0.3% (-0.03)**
Right eye	0.8% (-0.43)	0.5% (-0.30)	0.4% (-0.25)	**0.3% (-0.22)**
Mouth	2.0% (-0.90)	1.8% (-0.82)	1.7% (-0.80)	**0.3% (-0.25)**

Table 1: EER and corresponding value of τ for various kernel types. Notice how especially the extended mouth model benefits from a higher order kernel. Differences between the two eye models are due to asymmetric reflections on eyeglasses.

for the extended models was "pruned" by eliminating the high frequency channels at the periphery of the retina and the low frequency channels in the foveal region.

Due to the wider availability of negative examples during the training process, all SVMs showed a higher tendency towards making errors when classifying a positive probe (False Rejection, FR) rather than a negative probe (False Acceptance, FA). In order to compensate for that, we modified the decision function (3) with the introduction of a threshold τ: $f(\vec{v}) \gtrless \tau$. We adjusted the value of τ in order to obtain an Equal Error Rate (EER) on the probe set; we also recorded the corresponding value of the threshold and the value of τ at which the FR goes to zero. The dependence of these quantities on the kernel type is displayed in tables 1 and 2 for low-order polynomial kernels of the form $K(\vec{v}, \vec{w}) = (1 + \vec{v} \cdot \vec{w})^d$ (higher-order kernels were not considered because of the somewhat longer time required for computation). Since our algorithm employs a competition scheme rather than a fixed decision surface (see section 3.3), the behaviour of the classifiers as the decision surface is shifted by varying τ is of interest to us. A small value of τ both at the EER and at the point where FR $= 0$ indicates that the outputs of the SVM classifier are "appropriately grouped". Also, it is convenient that the classifiers trained on the three facial landmarks have approximately the same behaviour. As evidenced by tables 1 and 2, these requirements are best met by SVMs based on a fourth order polynomial kernel. We therefore decided to employ this kernel function in both the local and the extended models. We can here observe that the local models are much cruder than the extended ones, that turn out to be approximately 30 times more accurate. This is in line with our expectations, since local models describe a facial landmark based only on a neighbourhood of a single point (a receptive field), whereas each extended model employs the entire set of receptive fields that make up the retina.

3.3 The Saccadic Search strategy

The facial landmark detection procedure consists in a saccadic exploration, starting with the retinotopic sensor placed at random on the image. Search is initially aimed a randomly chosen facial landmark. Gabor response vectors are computed at all retinal points; vector $\vec{v}_{c\gamma}$ extracted at point γ on circle c is rated by the local model of the selected landmark according to the output of the corresponding SVM, $f_{loc}(\vec{v}_{c\gamma})$. The retina is subsequently centred at the position of its sampling point $s_{c\gamma}$ that maximises f_{loc}. This procedure is iterated until the sensor is centred on a local maximum.

Note that one advantage of this search strategy is that the search automatically becomes finer as a local maximum is approached, since the artificial retina is denser at the

τ at FR $= 0$	**Kernel type**			
Local Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	−14.0	−5.5	−5.0	**−5.0**
Right eye	−19.0	−8.0	−7.0	**−6.5**
Mouth	−8.0	−9.0	−8.0	**−6.0**
Extended Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	−0.5	−0.2	−0.1	**−0.2**
Right eye	−0.7	−0.6	−0.5	**−0.5**
Mouth	−1.4	−1.3	−1.2	**−0.4**

Table 2: Maximum value of τ for which FR $= 0$ for various kernel types. Although this measure is sensitive to the presence of outliers, its marked tendency to increase with kernel order indicates that fluctuations in the SVM outputs become smaller with a higher order kernel.

centre (*fovea*) than at the periphery. As a consequence of this and of our experimental results we think that the acuity gradient existing between the peripheral and the foveal vision in the topology of the human retina plays a role in achieving fast convergence (*homing*) of the saccades.

After saccades have converged to a maximum, the retinotopic grid is displaced in a pixel-by-pixel fashion to maximise the output f_{ext} of the more accurate extended model for the detected facial landmark. Matches that score less than the support vectors of the reject class ($f_{ext} < -1$) are discarded at this stage; the others are pushed onto a stack along with their score.

Once a match for a facial landmark has been found, a saccade to the assumed location of one of the others is performed based on a simple probabilistic model. An attempt at detection is made directly with the corresponding extended model. If this fails, the search is restarted at random to look for the feature for which the fewest candidates have been detected.

The probabilistic model describes the relative position of each facial landmark with respect to the next (in a cyclic order) by means of a Gaussian distribution, the parameters of which are estimated from the training data. The correlation between the resulting three vector distances has been neglected for simplicity; a more detailed model which exploits such correlation to compute better estimates of the position of the missing landmarks can be found in (Leung et al., 1995).

A global matching score is computed based exclusively on the quality of the best matches detected; this procedure has a linear cost with respect to the number of facial landmarks employed (always three in our experiments). Saccadic search is continued until a complete set of facial landmarks which has a very high score is found or until the retina has been centred on the maximum allowed number of different locations. If at this point no set of facial landmarks with a satisfactory score has been found, the search is restarted from scratch for a second attempt (we have allowed a maximum of four independent trials). Sets of fewer than three facial landmarks are considered if no candidates are detected for one or more of the search targets.

4 Face Authentication

Face authentication is performed using a multiple expert approach. Three classifiers are employed to independently authenticate each client based on the three sets of Gabor responses obtained with the retinotopic sensor centred on the eyes and the mouth of the subject, as detected by the Saccadic Search algorithm.

We have experimented with different implementations of the experts, namely Nearest Neighbour (NN), K Nearest Neighbours (KNN) and SVM classifiers with various choices of the kernel function. All classifiers are used in the two-class context of face authentication, with examples from the client representing the "accept" class and examples from the training impostors representing the "reject" class, irrespective of their differing identities (note how this differs from face recognition, which is intrinsically a multi-class problem). NN experts output a discrete score of $+1$ (acceptance) or -1 (rejection). KNN experts also yield a discrete but graded score equal to the signed difference between the number of positive and negative examples that appear among the K nearest neighbours; this in a way gives a measure of "how sure" the expert is feeling about its output.

Expert fusion is achieved by majority voting in the case of NN classifiers. KNN experts are combined by considering the sign of the total score. When fusing together the outputs of SVM experts, we have found it convenient to use a nonlinearity to limit the influence of any single expert $\mathcal{E}_j$ on the final outcome by gradually saturating the scores outside the $(-1, 1)$ interval. The decision on identity claim I is therefore given by the inequality

$$1/3 \sum_{j=1}^{3} \tanh\left(\kappa \mathcal{E}_j(I)\right) \gtrless \tau \tag{4}$$

where κ is a constant and τ can be varied to obtain FA/FR curves and determine the EER.

Optimal performance is already achieved using linear SVM classifiers, which separate the client and impostor classes with a hyper-plane. This confirms the remarkably good behaviour of SVM classifiers when a very low number of training examples is available. Although one could argue that the lack of statistically significant improvements using higher order kernels is due to the scarcity of training examples, the fact that a remarkably low EER can be achieved by means of *linear* decision surfaces confirms the discriminating power of the (orientation) features employed.

5 Experimental Results

We present experimental results on the subsets of frontal images from both the M2VTS (37 persons) and Extended M2VTS (295 persons) databases.

These multi-modal databases, featuring audio and video sequences of volunteers, have been collected by the M2VTS consortium specifically for identity verification purposes. Each subject has been acquired on four different occasions separated by a significant time interval (four months for XM2VTS). This constitutes the major advantage of these datasets over other databases including FERET, as it allows performing more client tests, which are needed to draw significant False Rejection curves.

The set of frontal images from the M2VTS database we employed contains a total of 551 images of the 37 subjects, that is 3–4 portraits of each person in each of the four sessions. The frames were converted to grey level and down-sampled to a resolution of 174×143, corresponding to an average inter-eye distance of 29 pixels.

A set of frontal images from the XM2VTS database is released by the M2VTS consortium as a CD-ROM distribution. It includes 2 images per session for each of its 295 subjects, yielding a total of 2360 frames. Again we converted these images to grey level and we down-sampled them to 180×144 resolution, which resulted in an average inter-eye distance of 27 pixels.

In both data-sets, appearance of subjects varies widely across the different sessions. For the same person differences include changes in tan, expression, and hair style. Some subjects changed their glasses or did not wear them in all of the sessions in favour of contact lenses.

5.1 Facial Landmark Detection

Facial landmark detection tests were run on a total of 349 images from the M2VTS database and on 2388 images from XM2VTS (we also included in the test set the images of a few subjects who are not part of the official XM2VTS distribution). Training was performed in both cases on the training set from the M2VTS database (see section 3.1); all parameters were left unchanged for tests on the M2VTS and XM2VTS data-sets.

Detection performance was evaluated by visual inspection. Experimental results show reliable eyes and mouth detection performance, with comparable results on both databases. The error rate for the M2VTS test set is 0.7% for each eye and 0.0% for the mouth, which is always detected correctly. Therefore, 5 images out of 349 present a misdetection of one eye; the three facial landmarks are correctly detected in all the others (98.6%). Error rates on the XM2VTS database are 1.0% for each eye and 1.3% for the mouth. This last figure reflects the higher variability in the appearance of the mouth due to changes in expression. Completely erroneous detection has occurred in 0.3% of the images (i.e. 6 images out of 2388); in 99.5% of the cases at least two features have been detected correctly. Correct detection of both the eyes and the mouth has been achieved on 97.4% of the test set. The system appears to be robust to the presence of eyeglasses, partial occlusion and even significant pose changes (Figure 4). It also generalises quite well to the presence of beards and of non-Europeans, that are not represented in the training set (the M2VTS database only includes Europeans, and only two subjects have a beard). This generalisation ability is likely a consequence of the contrast invariance obtained by normalising Gabor feature vectors prior to classification.

Results on facial landmark detection on the M2VTS database by another research group have been published in (Kotropoulos and Pitas, 1997), where an 86.5% success rate for simultaneous detection of eyebrows/eyes, nostrils/nose and mouth over 37 frontal images from the database is reported.

5.2 Face Authentication on the M2VTS Database

Face authentication experiments on the M2VTS database were conducted according to a "leave-one-out" rotation scheme (jack–knife bootstrapping) (Efron and Tibshirani, 1993) also known as "Brussels Protocol" (Bigun et al., 1998). At each step one person is removed from the database to act as impostor, while one entire session is set aside for client tests. The 36 persons remaining act as registered clients, and their three image series are available for training. Client tests are performed taking, for each client, a non-training image. Likewise, a single image of the impostor is used for impostor tests. This amounts to 36 client tests and 36 impostor tests. By choosing the impostor and the test series in all the possible ways, $36 \times 37 \times 4 = 5328$ client tests and the same number of impostor tests can be generated.

In our training, each expert (independently of its implementation) uses the three training series of images from the client which it should learn as positive examples, and all of the images from the remaining 35 clients known to the system as negative examples (training impostors).

Since part of the M2VTS database has been used as the training set for the facial landmark detection algorithm, we only present authentication results obtained by manually identifying the position of the eyes and the mouth of the subjects in both the training and the test images, so that the performance of the authentication algorithm itself can be gauged.

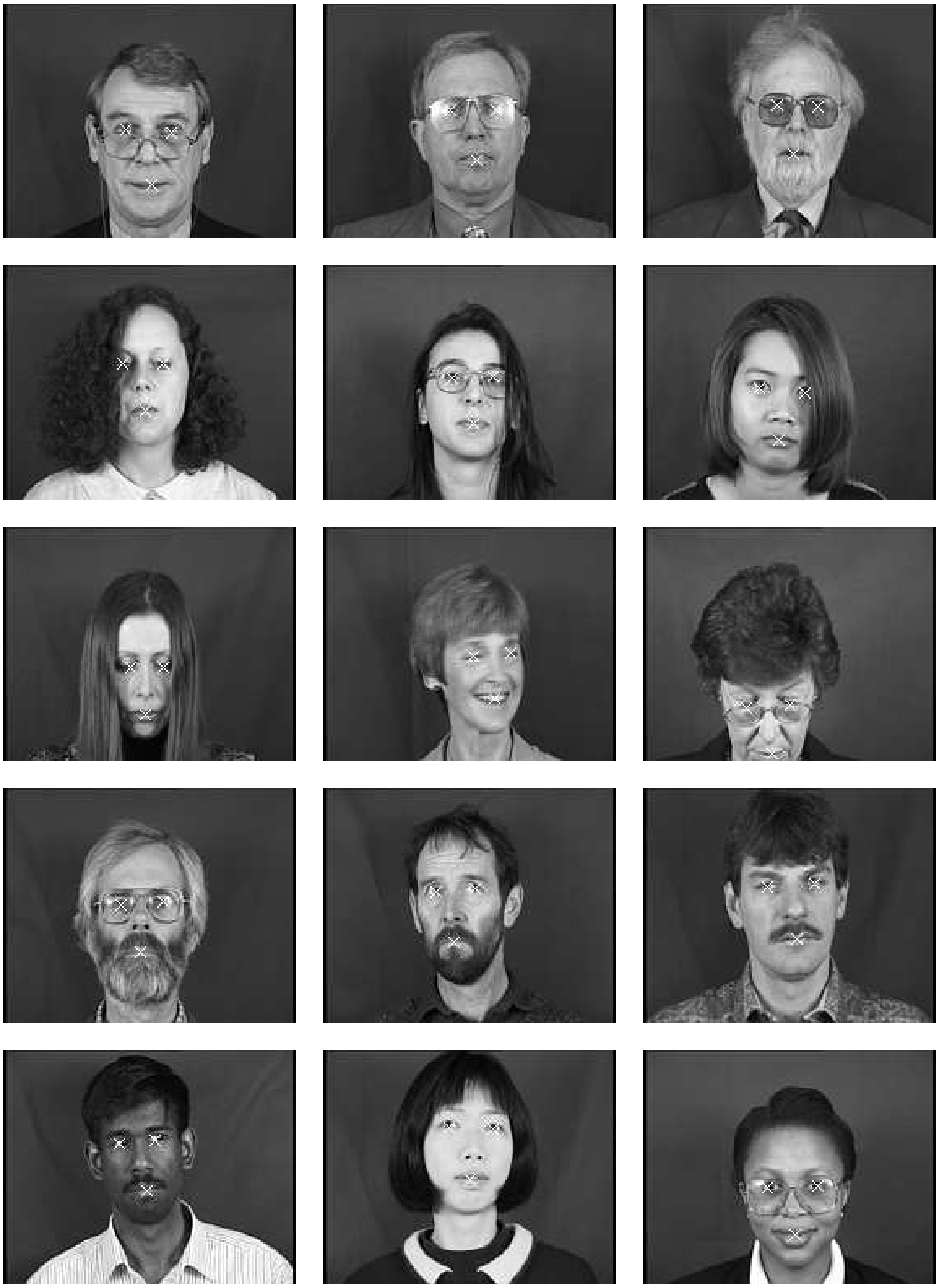

Figure 4: Reliable eyes and mouth detection can be achieved even in the presence of eyeglasses (1st row), partial occlusion (2nd row), pose changes (3rd row) and facial hair (4th row). The system generalises well to non-Europeans (5th row). These images belong to the XM2VTS database.

(K)NN	FA	FR	$\sqrt{\text{FA} \cdot \text{FR}}$	SVM	FA	FR	EER
NN	1.1%	2.0%	1.5%	Linear	0.0%	10.0%	0.17%
K=3	0.5%	4.0%	1.3%	Poly - deg 2	0.0%	10.8%	0.15%
K=5	0.2%	8.0%	1.4%	Poly - deg 3	0.0%	11.5%	0.15%
				Poly - deg 4	0.0%	11.5%	0.13%

Table 3: FA, FR and EER with KNN and SVM experts (M2VTS database).

Experiments indicate that a very low error rate can be achieved with both SVM based and (K)NN based experts (table 3). As can be observed, all classifiers have a marked tendency towards false rejection, due to the low number of training images available for each client as compared to impostors. In the case of SVM experts, this can be compensated by tuning the threshold τ in equation (4), which allows determining the EER. To provide some form of comparison, for (K)NN experts we listed the geometric average error. Variations of the EER due to the choice of different kernel functions for the SVM classifiers are not statistically significant. The FA/FR curve for linear SVM experts is reported in Figure 5. For increased readability, the curve has been plotted on a Normal Deviate scale (Martin et al., 1997) using the software provided by NIST.

Summing up, (linear) SVM experts practically yield the best performance with an EER around 0.15%. For comparison, we refer the reader to published results on the M2VTS database obtained using Elastic Graph Matching (EER = 6.1%, Brussels Protocol) (Duc et al., 1999) and the Morphological Dynamic Link Architecture (EER = 3.7%, Brussels Protocol) (Tefas et al., 1998). The latter technique is based on pyramids of morphological differences that have been extensively used in texture analysis, e.g. (Veenland et al., 1998). Finally, EERs of 5.4% for frontal image authentication and 3.1% for image sequence based authentication are reported in (Matas et al., 1997), where a warping and correlation algorithm is employed. Although none of these approaches require the facial landmarks to be detected (which we did manually in our test since part of the M2VTS data was used to train the detector), the comparable results (EER = 0.25%, section 5.3) achieved by our procedure in entirely automated tests over the larger XM2VTS database suggest that the above comparison has at least an indicative value.

5.3 Face Authentication on the XM2VTS Database

Face authentication experiments on the XM2VTS database were carried out using the training and test sets stated by the Lausanne Protocol, Configuration II, as established by the M2VTS Consortium (Messer et al., 1999). The Protocol specifies 200 identities to be used as clients and 70 impostor identities. For each client, two sessions (4 images) are available for training; one is used for testing (2 images) and two images are set aside as "evaluation set" (unused). All the 8 images of each impostor are used to generate impostor tests. This gives a total of 400 client tests and $70 \times 8 \times 200 = 112'000$ impostor tests.

As was the case with the M2VTS database, we trained each expert using the available images from the client it is supposed to authenticate as positive examples and the images of all the other registered clients as negative examples (training impostors).

Experiments are performed in an *entirely automatic* fashion, that is we let the facial landmark detection algorithm (which has not been trained on any image of the XM2VTS database) find the coordinates of the eyes and the mouth both in the training images and in the images of the test set. An EER of 0.50% has been achieved using linear SVM experts (dash-dot line, Figure 5). This corresponds to two client tests failed out of 400.

This result can be further improved by adding the remaining two images of each client

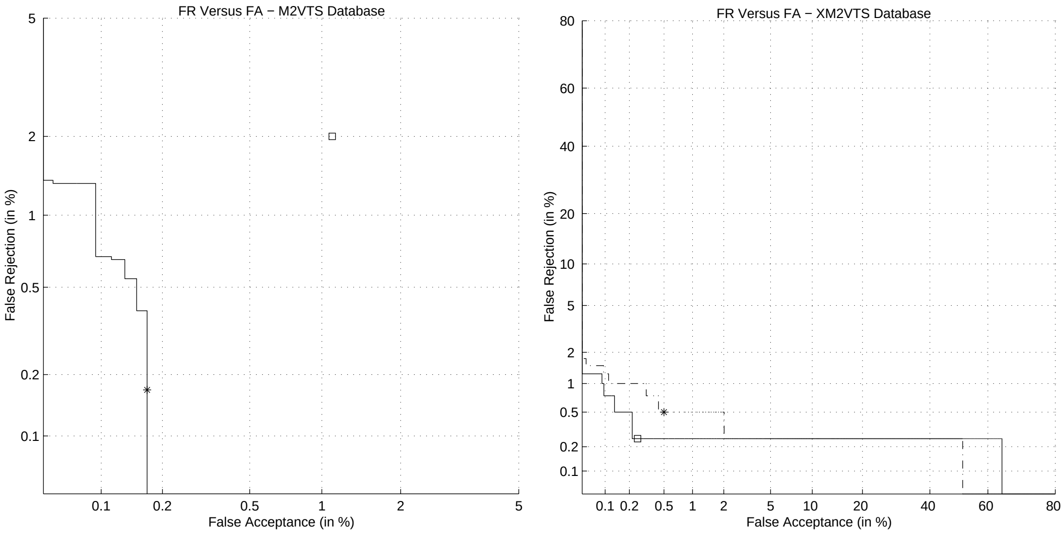

Figure 5: Left: FR vs FA, M2VTS database (linear SVMs). The EER (∼ 0.15%) is indicated by a star. The square marks the fixed operation point of the system when NN experts are used. Right: FR vs FA, XM2VTS database (linear SVMs). The dash-dot line refers to Lausanne Protocol, Config. II tests (4 training images). The EER (0.50%) is marked with a star. The solid line is obtained by using 6 training images for each client. The square indicates the EER (0.25%). The FR plateau to the right of the square is due to the false rejection of a single image.

that are labelled “evaluation set” in the Lausanne Protocol to the training set, which then contains six images of each registered subject. In that case system operation is described by the solid line in Figure 5. The EER (0.25%) corresponds to only one client test failed out of 400. Comparison of these results with those obtained on the M2VTS database by manually identifying the position of the eyes and the mouth (EER = 0.15%) shows that relying on automatic facial landmark detection for both training and tests does not significantly degrade authentication performance. This is due both to the reliability of the facial landmark detection procedure and to the intrinsic robustness of the authentication algorithm, which is achieved through the use of multiple training images and of a mixture of experts approach.

A detailed comparison of our results with the performance of several other algorithms developed by independent research groups was published following the face verification contest organised in conjunction with the International Conference on Pattern Recognition 2000 (Matas et al., 2000). The seven algorithms reviewed all achieved EERs in excess of 1% over the same training and test sets.

6 Discussion and Conclusions

In this paper we introduced a new vision paradigm, namely a concept of *Retinal Vision*. It consists of an artificial retina driven by saccades that is able to locate sought-for landmarks and to perform authentication. Our experiments indicate that a good integration between cognitive tasks and low-level visual processing can be achieved if the biological analogy that underlies retinotopic sampling and the Gabor decomposition is coherently pursued at higher stages of the chain that goes from the visual input to symbolic representations. The idea is that of providing a simple view–based mechanism for modelling objects to be located and possibly authenticated. This intermediate layer would allow cognitive processes to be more perspicuous and effective in their use of visual information.

The model of saccadic search we have presented attempts to implement such a concept by providing a flexible *general purpose* attentional mechanism. The algorithm is in no way specific to eye or mouth detection, and more primitive versions have already been successfully applied in such different contexts as real time head detection and tracking (Smeraldi et al., 2000) and robot navigation (Arleo et al., 2000).

We have chosen to present the *Retinal Vision* paradigm with an application to face authentication because of the intrinsic complexity of this problem. Also, the low-level visual processing tools involved, retinotopic sampling and the Gabor decomposition, have already been widely applied in face authentication, which makes the specific contribution of the more organic *Retinal Vision* approach plausible.

The very low error rates achieved on the two largest verification oriented databases available (EER $\lesssim 0.25\%$) as well as on privately owned industrial databases support the viability of our approach. Although the databases used included scale changes up to 10% and severe pose disturbances in the supposedly frontal images and the results were good, we believe that scale and pose invariance should yet be included into the Retinal Vision approach. In future work, we would like to address those issues by introducing generalisation ability with respect to scale and pose at the level of the saccadic planning, thus providing the cognitive level with a more powerful interface to the world.

7 Acknowledgements

While at EPFL, this research has been carried out within the European Community IT-LTR Project VIRSBS. Research at Halmstad University has been funded by the Chalmers Foundation, Gothenburg.

References

Arleo, A., Smeraldi, F., Hug, S., and Gerstner, W. (2000). Place cells and spatial navigation based on vision, path integration, and reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 13, pages 89–95. MIT-Press.

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *"IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence"*, 19(7):711–720.

Bigun, J. (1993). Gabor phase in boundary tracking and region segregation. In *proceedings of DSP & CAES conf. Nicosia, Cyprus*, pages pp. 229–234. Univ. of Nicosia.

Bigun, J. (1994). Speed, frequency, and orientation tuned 3-D Gabor filter banks and their design. In *Proceedings of International Conference on Pattern Recognition, ICPR, Jerusalem*, pages C–184–187. IEEE Computer Society.

Bigun, J., Duc, B., Fischer, S., Makarov, A., and Smeraldi, F. (1998). Multi modal person authentication. In Wechsler et al., H., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 26–50. Springer.

Bigun, J., Granlund, G. H., and Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE-PAMI*, 13(8):775–790.

Chellappa, R., Wilson, C. L., and Sirohey, S. (1995). Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, 83(5):705–740.

Cortes, C. and Vapnik, V. (1995). Support–Vector Networks. *Machine Learning*, 20:273–297.

Duc, B., Fischer, S., and Bigun, J. (1999). Face authentication with Gabor information on deformable graphs. *IEEE Transactions on Image Processing*, 8(4):504–516.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the Bootstrap.* Chapman & Hall, New York.

Hubel, D. (1988). *Eye, brain and vision.* Scientific American Library.

Joachims, T. (1998). *Making Large-scale SVM Learning Practical*, chapter 11 of Advances in Kernel Methods - Support Vector Learning. MIT Press. Eds. B. Schölkopf, C. J. C. Burges, A. J. Smola.

Keating, C. F. and Keating, E. G. (1993). Monkeys and mug shots: cues used by Rhesus monkeys (Macaca mulatta) to recognize a human face. *Journal of Comparative Psychology*, 107(2):131–139.

Knutsson, H. (1982). *Filtering and reconstruction in image processing.* Number 88 in Linköpings Studies in Science and Technology: Dissertations. Linköping University. ISBN 91-7372-595-1.

Kotropoulos, C. and Pitas, I. (1997). Rule–based face detection in frontal views. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume IV, pages 2537–2540.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., Malsburg, C. v. d., Hurtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architectures. *IEEE Trans. on Computers*, 42(3):300–311.

Leung, T. K., Burl, M. C., and Perona, P. (1995). Finding faces in cluttered scenes using random labeled graph matching. In *Proceedings of ICCV95, Cambridge MA*, pages 637–644.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings of EuroSpeech '97*, volume IV, pages 1895–1898.

Matas, J., Hamouz, M., Jonsson, K., Kittler, J., Li, Y., Kotropoulos, C., Tefas, A., Pitas, I., Tan, T., Yan, H., Smeraldi, F., Bigun, J., Capdevielle, N., Gerstner, W., Ben-Yacoub, S., Abdeljaoued, Y., and Mayoraz, E. (2000). Comparison of face verification results on the XM2VTS database. In *Proceedings of the 15th International Conference on Pattern Recognition, Barcelona (Spain), September 2000*, volume 4, pages 858–863. IEEE Comp. Soc. Order No. PR00750.

Matas, J., Jonsson, K., and Kittler, J. (1997). Fast face localisation and verification. In Clark, A. F., editor, *Proceedings of the British Machine Vision Conference (BMVC97).*

Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G. (1999). XM2VTSDB: the Extended M2VTS database. In *Proceedings of the 2nd international conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99), Washington DC, U.S.A.*, pages 72–77.

Orban, G. A. (1984). *Neuronal operations in the visual cortex.* studies of brain functions. Springer.

Osuna, E., Freund, R., and Girosi, F. (1997). Improved training algorithm for Support Vector Machines. In *Proceedings of IEEE NNSP'97*, pages 276–285.

Pelz, J. B. (1995). *Visual representations in a natural visuo-motor task.* PhD thesis, Carlson Center for Imaging Science, Rochester Institute of Technology.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., and Ballard, D. H. (1997). Eye movements in visual cognition: a computational study. Technical Report 97.1, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Visual Research*, 20:645–669.

Sirovich, L. and Kirby, M. (1990). Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108.

Smeraldi, F. (No. 2153 (2000)). *Attention–driven pattern recognition.* PhD thesis, Swiss Federal Institute of Technology – Lausanne, Switzerland.

Smeraldi, F., Capdevielle, N., and Bigun, J. (1999a). Face authentication by retinotopic sampling of the Gabor decomposition and Support Vector Machines. In *Audio and Video based Biometric Person Authentication – AVBPA99*, pages 125–129.

Smeraldi, F., Capdevielle, N., and Bigun, J. (1999b). Facial features detection by saccadic exploration of the Gabor decomposition and support vector machines. In *Proceedings of the 11th Scandinavian Conference on Image Analysis – SCIA 99, Kangerlussuaq, Greenland*, volume I, pages 39–44.

Smeraldi, F., Carmona, O., and Bigun, J. (2000). Saccadic search with Gabor features applied to eye detection and real-time head tracking. *Image and Vision Computing*, 18(4):323–329.

Takács, B. and Wechsler, H. (1995). Face localization using a dynamic model of retinal feature extraction. In Bichsel, M., editor, *Proceedings of the International workshop on automatic face and gesture recognition, Zurich*, pages 243–247. Multimedia Laboratory; Univ. of Zurich; Winterthurers. 190 CH-8057 Zurich.

Tefas, A., Kotropoulos, C., and Pitas, I. (1998). Variants of dynamic link architecture based on mathematical morphology for frontal face authentication. In *Proceedings of the IEEE Computer Society conference on Computer Vision and Pattern Recognition (CVPR 98), Santa Barbara (CA), U.S.A.*, pages 814–819.

Tistarelli, M. and Grosso, E. (1998). Active vision-based face recognition: issues, applications and techniques. In Wechsler et al., H., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 262–286. Springer.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *J. Cognitive Neuroscience (Winter)*, 3(1):71–86.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Springer–Verlag.

Veenland, J. F., Grashuis, J. L., and Gelsema, E. S. (1998). Texture analysis in radiographs: the influence of mtf and noise on the discriminative ability of texture features. *Medical Physics*, 25(6):922–936.

Yarbus, A. L. (1967). *Eye movements.* Plenum, New York.