MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Learning Gender with Support Faces

Baback Moghaddam   Ming-Hsuan Yang

TR-2002-12   January 2002

Abstract

Nonlinear Support Vector Machines (SVMs) are investigated for appearance-based gender classification with low resolution "thumbnail" faces processed from 1,755 images from the FERET face database. The performance of SVMs (3.4% error) is shown to be superior to traditional pattern classifiers (Linear, Quadratic, Fisher Linear Discriminant, Nearest-Neighbor) as well as more modern techniques such as Radial Basis Function (RBF) classifiers and large ensemble-RBF networks. Furthermore, the difference in classification performance with low resolution "thumbnails" (21-by-12 pixels) and the corresponding higher resolution images (84-by-48 pixels) was found to be only 1%, thus demonstrating robustness and stability with respect to scale and degree of facial detail.

IEEE Transactions on Pattern Analysis and Machine Intelligene (PAMI), Vol. 24, No. 5, May 2002

Learning Gender with Support Faces

Baback Moghaddam, *Member*, *IEEE*, and
Ming-Hsuan Yang

Abstract—Nonlinear Support Vector Machines (SVMs) are investigated for appearance-based gender classification with low-resolution "thumbnail" faces processed from 1,755 images from the FERET face database. The performance of SVMs (3.4 percent error) is shown to be superior to traditional pattern classifiers (linear, quadratic, Fisher linear discriminant, nearest-neighbor) as well as more modern techniques such as Radial Basis Function (RBF) classifiers and large ensemble-RBF networks. Furthermore, the difference in classification performance with low-resolution "thumbnails" (21-by-12 pixels) and the corresponding higher resolution images (84-by-48 pixels) was found to be only 1 percent, thus demonstrating robustness and stability with respect to scale and degree of facial detail.

Index Terms—Support vector machines, gender classification, linear, quadratic, Fisher linear discriminant, RBF classifiers, face recognition.

1 Introduction

As Human-Computer Interaction technology (HCI) evolves, computer vision systems for people monitoring will play an increasingly important role in our lives. Examples include human (face) detection, face/body tracking, action (gesture) recognition, person identification (face recognition) and estimation of age, ethnicity and perhaps most fundamentally gender. This information will not only enhance existing HCI systems but can also serve as a basis for passive surveillance and control in "smart buildings" (e.g., restricting access to certain areas based on gender) and collecting valuable demographics (e.g., the number of women entering a retail store on a given day). We have developed an appearance-based gender classifier for low-resolution images (extracted by an automatic face detection system) which uses Support Vector Machine (SVM) learning. This system exhibits performance far superior to existing classical classifiers.

In recent years, SVMs have been successfully applied to various tasks in computational face-processing, including face detection [22], face pose discrimination [19], and face recognition [24]. The good empirical results can be explained by the fact that SVM is an optimal discriminant method based on the Bayesian learning theory. For the cases where it is difficult to estimate the density model in high-dimensional space, e.g., images, the discriminant approach is preferable than the generative approach. Furthermore, SVMs provide an efficient discriminant method, not only to handle the patterns that are not linearly separable, but also achieve lower generalization error for the unseen test examples. In this paper, we develop an appearance-based method to classify gender from facial images using nonlinear SVMs and compare their performance with traditional classifiers (e.g., linear, quadratic, Fisher linear discriminant, and nearest-neighbor) as well as more modern techniques such as RBF networks and large ensemble-RBF classifiers. We have focused our study on very low-resolution "thumbnail" images in which only the main frontal facial regions (inside the "oval" of the face) are visible and almost completely excluded hair information (outside the "oval"). The motivation for using these particular images is two-fold. First, hair styles can change in appearance easily and frequently. Therefore, in a robust face recognition system face images are usually cropped to keep only the main facial regions. Second, we wished to investigate the minimal amount of face information (resolution) required to learn male and female faces by various classifiers. Previous studies on gender classification have used high-resolution images with hair information and relatively small data sets for their experiments. In our study, we demonstrate that SVM classifiers are able to learn and classify gender from a large set of hairless low-resolution images with very high accuracy. Furthermore, SVM classifiers showed negligible difference between their error rates with low and high-resolution facial images. Finally, in our experimental study, little or no hair information was used as input to the classifiers. This is in contrast to previous results reported in the literature where almost all methods include some hair information.

- *B. Moghaddam is with Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139. E-mail: baback@merl.com.*
- *M.-H. Yang is with Honda Fundamental Research Laboratories, 800 California St., Mountain View, CA 94041. E-mail: myang@hra.com.*

Manuscript received 27 Sept. 2000; revised 18 June 2001; accepted 21 Oct. 2001.
Recommended for acceptance by R. Sharma.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112913.

2 Background

Although gender classification has attracted much attention in the psychological literature [3], [6], [23], relatively few learning-based machine vision methods have been proposed. In this section we briefly review and summarize the prior art in visual gender classification. The studies referred to are also summarized in Fig. 1 where the final entry [21] reports some of the preliminary results reported in this paper.

Gollomb et al. [16] trained a fully connected two-layer neural network, SEXNET, to identify gender from 30-by-30 face images. Their experiments on a set of 90 photos (45 males and 45 females) gave an average error rate of 8.1 percent compared to an average error rate of 11.6 percent from a study of five human subjects. Cottrell [9] also applied neural networks for emotion and gender classification. The dimensionality of a set of 160 64-by-64 face images (10 males and 10 females) was reduced from 4,096 to 40 with an auto-encoder. These vectors were then presented as inputs to another one layer network for training. They reported perfect classification (albeit for only 20 individuals). Brunelli and Poggio [4] developed HyperBF networks for gender classification in which two competing RBF networks, one for male and the other for female, were trained using 16 geometric features as inputs (e.g., pupil to eyebrow separation, eyebrow thickness, and nose width). The results on a data set of 168 images (21 males and 21 females) show an average error rate of 21 percent. Using similar techniques as Golomb et al. [16] and Cottrell [9], Tamura et al. [29] used multilayer neural networks to classify gender from face images at multiple resolutions (from 32-by-32 to 8-by-8 pixels). Their experiments on 30 test images show that their network was able to determine gender from 8-by-8 images with an average error rate of 7 percent. Instead of using a vector of gray levels to represent faces, Wiskott et al. [31] used labeled graphs of two-dimensional views to describe faces. The nodes were represented by wavelet-based local "jets" and the edges were labeled with distance vectors. They used a small set of controlled model graphs of males and females to encode "general face knowledge," in order to generate graphs of new faces by elastic graph matching. For each new face, a composite reconstruction was generated using the nodes in the model graphs. The gender of the majority of nodes used in the composite graph was used for classification. The error rate of their experiments on a gallery of 112 face images was 9.8 percent. Recently, Gutta et al. [17] proposed a hybrid classifier based on neural networks (RBFs) and inductive decision trees with Quinlan's C4.5 algorithm with 3,000 FERET faces of size 64-by-72 pixels. The best average error rate was found to be 4 percent.

Study	Method	Size	Format	% Error
Gollomb 1991	NN	90	30-by-30	8.10
Wiskott 1995	EGM	112	Full	9.80
Tamura 1996	NN	30	8-by-6	7.00
Gutta 1998	RBF/DT	3000	64-by-72	4.00
Moghaddam 2000	SVM	1800	21-by-12	3.38

Fig. 1. Comparison of representative gender classification studies (see text).

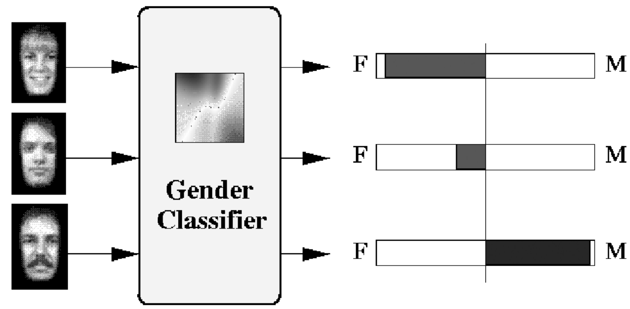

Fig. 2. Gender classifier.

3 Gender Classifiers

A generic appearance-based gender classifier is shown in Fig. 2. An input facial image x generates a scalar output $f(\mathbf{x})$ whose polarity—sign of $f(\mathbf{x})$—determines class membership. The magnitude $\|f(\mathbf{x})\|$ can usually be interpreted as a measure of belief or certainty in the decision made. Nearly all binary classifiers can be viewed in these terms; for density-based classifiers (linear, quadratic and Fisher) the output function $f(\mathbf{x})$ is a log-likelihood ratio, whereas for kernel-based classifiers (nearest-neighbor, RBFs, and SVMs) the output is a "potential field" related to the distance from the separating boundary.

3.1 Support Vector Machines

A Support Vector Machine is a learning algorithm for pattern classification, regression and density estimation [30], [8], [11]. The basic training principle behind SVMs is finding the optimal linear hyperplane such that the expected classification error for unseen test samples is minimized—i.e., good generalization performance. According to the structural risk minimization inductive principle [30], a function that classifies the training data accurately and which belongs to a set of functions with the lowest VC dimension [8] will generalize best, regardless of the dimensionality of the input space. Based on this principle, a linear SVM uses a systematic approach to find a linear function with the lowest VC dimension. For linearly nonseparable data, SVMs can (nonlinearly) map the input to a high-dimensional feature space where a linear hyperplane can be found. Although there is no guarantee that a linear solution will always exist in the high-dimensional space, in practice it is feasible to find a working solution.

Given a labeled set of M training samples $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in R^N$ and y_i is the associated label $(y_i \in \{-1, 1\})$, a SVM classifier finds the optimal hyperplane that correctly separates (classifies) the largest fraction of data points while maximizing the distance of either class from the hyperplane (the margin). Vapnik [30] shows that maximizing the margin distance is equivalent to minimizing the VC dimension in constructing the optimal hyperplane. Computing the optimal hyperplane is posed as a constrained optimization problem and solved using quadratic programming techniques. The discriminant hyperplane is defined by:

$$f(\mathbf{x}) = \sum_{i=1}^{M} y_i \alpha_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b,$$

where $k(\cdot,\cdot)$ is a kernel function, b is a bias term and the sign of $f(\mathbf{x})$ determines the class membership of $\mathbf{x}$. Constructing an optimal hyperplane is equivalent to finding all the nonzero α_i and is formulated as a quadratic programing (QP) problem with linear constraints [5]. Any vector $\mathbf{x}_i$ that corresponds to a nonzero α_i is a *support vector* (SV) of the optimal hyperplane. A desirable feature of SVMs is that the number of training points which are retained as support vectors is usually quite small, thus providing a compact classifier. Solving the constraint optimization problem for a SVM with a large data set is a nontrivial task, many methods have been proposed to tackle such problems. In our study, we used a public-domain SVM package which uses conjugate gradients for solving the QP optimization in training [26]. For more recent advances in fast optimization methods for SVMs, see [27].

For a linear SVM, the kernel function is just a simple dot product in the input space while the kernel function in a nonlinear SVM effectively projects the samples to a feature space of higher (possibly infinite) dimension via a nonlinear mapping function:

$$\Phi : R^N \rightarrow F^M, M \gg N$$

and then constructs a hyperplane in F. The motivation behind this mapping is that it is more likely to find a linear hyperplane in the high-dimensional feature space. Using Mercer's theorem [10], the expensive calculations required in projecting samples into the high-dimensional feature space can be replaced by a simpler kernel function satisfying the condition

$$k(\mathbf{x}, \mathbf{x}_i) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i),$$

where Φ is the nonlinear projection function. Several kernel functions, such as, polynomials and radial basis functions, have been shown to satisfy Mercer's theorem and have been used successfully in nonlinear SVMs:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}_i) &= ((\mathbf{x} \cdot \mathbf{x}_i) + 1)^d \\ k(\mathbf{x}, \mathbf{x}_i) &= \exp(-\gamma |\mathbf{x} - \mathbf{x}_i|^2), \end{aligned}$$

where d is the degree of freedom in a polynomial kernel and γ is the spread of a Gaussian cluster. In fact, by using different kernel functions, SVMs can implement a variety of learning machines, some of which coincide with classical classifiers, e.g., Bayesian classifier, radial basis function networks, maximum entropy approaches. Nevertheless, automatic selection of the "right" kernel function and its associated parameters remains problematic and in practice one must resort to trial and error with validation set for model selection. However, see [7] for a recently proposed method for multiple parameter selection for SVMs.

3.2 Radial Basis Function Networks

A radial basis function (RBF) network is also a kernel-based technique for improved generalization, but it is based instead on regularization theory [25], [18]. A typical RBF network with K Gaussian basis functions is given by

$$f(\mathbf{x}) = \sum_{i}^{K} w_i \mathcal{G}(\mathbf{x}; \mathbf{c}_i, \sigma_i^2) + b,$$

where the $\mathcal{G}$ is the ith Gaussian basis function with center $\mathbf{c}_i$ and variance σ_i^2 and b is a bias term. The weight coefficients w_i combine

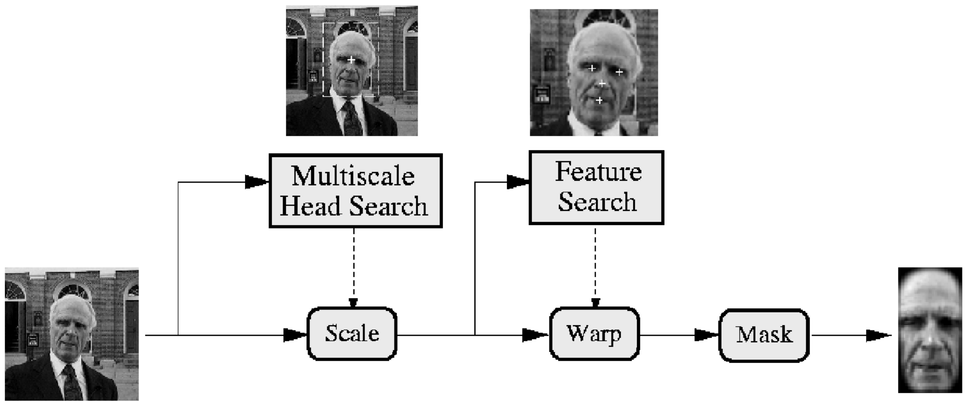

Fig. 3a. Face alignment system.

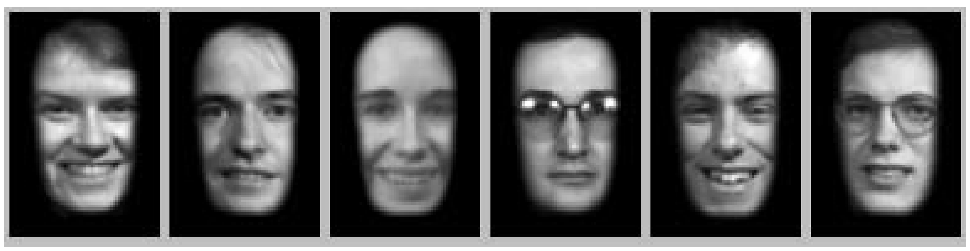

Fig. 3b. Some processed faces.

the basis functions into a single scalar output value, with b as a bias term. Training a Gaussian RBF network for a given learning task involves determining the total number of Gaussian basis functions, locating their centers, computing their corresponding variances, and solving for the weight coefficients and bias (which can be computed by using the regularization theory [25]. Judicious choice of K, $\mathbf{c}_i$, and σ_i^2, can yield RBF networks which are quite powerful in classification and regression tasks. The number of radial bases in a conventional RBF network is predetermined before training, whereas the number for a large ensemble-RBF network is iteratively increased until the error falls below a set threshold. The RBF centers in both cases are usually determined by k-means clustering. In contrast, a SVM with the same RBF kernel will automatically determine the number and location of the centers, as well as the weights and threshold that minimize an upper bound on the expected risk. Recently, Evgeniou et al. [14] have shown that both SVMs and RBF networks can be formulated under a unified framework in the context of Vapnik's theory of statistical learning [30]. As such, SVMs provide a more systematic approach to classification than classical RBF and various other neural networks.

3.3 Classical Discriminant Methods

Fisher linear discriminant (FLD) is an example of a class specific subspace method that finds the optimal linear projection for classification [15], [12], [2]. Rather than finding a projection that maximizes the projected variance as in principal component analysis, FLD determines a projection that maximizes the ratio between the between-class scatter and the within-class scatter. Consequently, classification is simplified in the projected space. In our experiments, we used a single Gaussian to model the distribution of male or female class in the resulting one dimensional space. The class membership of a sample was then determined using the maximum a posteriori probability, or equivalently by a likelihood ratio test. See [1], [13], [28] for face recognition methods using FLD.

The decision boundary of a quadratic classifier is defined by a quadratic form in $\mathbf{x}$, derived through Bayesian error minimization [15], [2], [12]. Assuming that the distribution of each class is Gaussian, the classifier output is given by finding the minimum Mahalanobis distance to a cluster center. A linear classifier is a special case of the quadratic form, based on the assumption all the clusters have the same covariance matrix. For both classifiers, the sign of the discriminant function determines class membership and is also equivalent to a likelihood ratio test.

4 Experiments

In our study, 256-by-384 pixel FERET "mug-shots" were preprocessed using an automatic face-processing system which compensates for translation, scale as well as slight rotations. Shown in Fig. 3a, this system is described in detail in [20] and uses maximum-likelihood estimation for face detection, affine warping for geometric shape alignment and contrast normalization for ambient lighting variations. The resulting output "face-prints" in Fig. 3a were standardized to 80-by-40 (full) resolution. These "face-prints" were further subsampled to 21-by-12 pixel "thumbnails" for our low-resolution experiments. Fig. 3b shows a few examples of processed face-prints (note that these faces contain little or no hair information). A total of 1,755 thumbnails (1,044 males and 711 females) were used in our experiments. For each classifier, the average error rate was estimated with five-fold cross validation (CV)—i.e., a five-way data set split, with 4/5th used for training and 1/5th used for testing, with four subsequent nonoverlapping rotations. The average size of the training set was 1,496 (793 males and 713 females) and the average size of the test set was 259 (133 males and 126 females).

The SVM classifier was first tested with various kernels in order to explore the space of possibilities and performance. In all the experiments, we set the soft margin C value to infinity so that no training error is allowed [30]. Meanwhile, each training and testing vector was scaled to be between -1 and 1, and each optimization problem was solved by the conjugate gradient method with a decomposition method similar to [22]. Fig. 4 shows the empirical results with various kernels and parameters (based on one training set). The mediocre results achieved by the first order polynomial kernel indicated that the linear decision surface is not able to

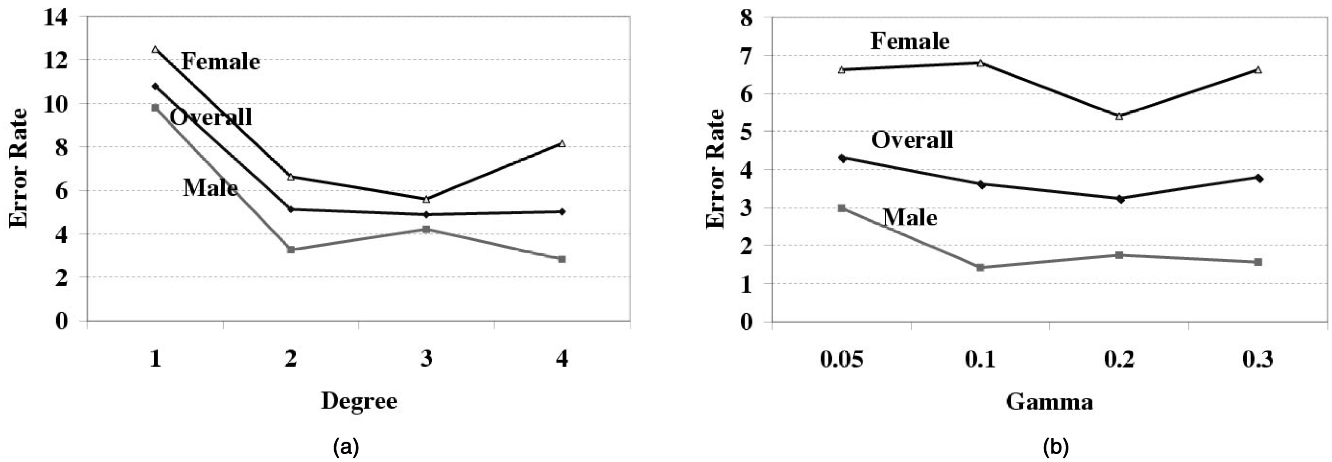

Fig. 4. Empirical results with thumbnails using various kernels. (a) Polynomial kernel. (b) RBF kernel.

effectively classify all the data points, which also indicated the "hardness" of this data set (also, see the results of linear and Fisher linear discriminant classifiers in Table 1). On the other hand, nonlinear decision surfaces constructed by second, third and fourth order polynomial kernel achieved good results. Meanwhile, the variance of the overall error rate among these nonlinear SVMs was not significant. For RBF kernels, we found that the variance of the overall error rate was not significant when we chose reasonable γ values.

A Gaussian RBF kernel was found to perform the best (in terms of error rate), followed by a cubic polynomial kernel as second best. In the large ensemble-RBF experiment, the number of radial bases was incremented until the error fell below a set threshold. The average number of radial bases in the large ensemble-RBF was found to be 1,289 which corresponds to 86 percent of the training set. The number of radial bases for classical RBF networks was heuristically set to 20 prior to actual training and testing. The quadratic, linear, and Fisher classifiers were all implemented using single Gaussian distributions. In each case, a likelihood ratio test was used for classification. The average error rates of all the classifiers tested with 21-by-12 pixel thumbnails are reported in Table 1 and summarized in Fig. 5.

The SVMs out-performed all other classifiers, although the performance of large ensemble-RBF networks was close to SVMs. However, nearly 90 percent of the training set was retained as radial bases by the large ensemble-RBF. In contrast, the number of support vectors found by both SVMs was only about 20 percent of the training set. We also applied SVMs to classification based on high-resolution images (84-by-48 pixels). The Gaussian and cubic kernel SVMs performed equally well at both low- and high-resolutions with only a slight 1 percent error rate difference. Fig. 6 shows three pairs of opposite (male and female) support faces from an actual SVM classifier. This figure is, of course, a crude low-dimensional depiction of the optimal separating hyperplane (hypersurface) and its associated margins (shown as dashed lines). However, the support faces shown are positioned in accordance with their basic geometry. Each pair of support faces across the boundary was the closest pair of images in the projected high-dimensional space. It is interesting to note not only the visual similarity of a given pair but also their androgynous appearance. Naturally, this is to be expected from a face located near the boundary of the male and female domains. We also note that as seen in Table 1, all the classifiers had higher error rates in classifying females. This phenomenon is most likely due to the general lack of prominent and distinct facial features in female faces.

5 Discussion

We have presented a comprehensive evaluation of various classification methods for determination of gender from facial images. The nontriviality of this task (made even harder by our "hairless" low-resolution faces) is demonstrated by the fact that a linear classifier had an error rate of 60 percent (i.e., worse than a random coin flip). Furthermore, an acceptable error rate (< 5 percent) for the large ensemble-RBF network required

TABLE 1
Experimental Results with Thumbnails

Classifier	Error Rate		
	Overall	Male	Female
SVM with Gaussian RBF kernel	**3.38%**	**2.05%**	**4.79%**
SVM with cubic polynomial kernel	**4.88%**	**4.21%**	**5.59%**
Large ensemble-RBF	5.54%	4.59%	6.55%
Classical RBF	7.79%	6.89%	8.75%
Bayesian (Quadratic)	10.63%	9.44%	11.88%
Fisher linear discriminant	13.03%	12.31%	13.78%
Nearest neighbor	27.16%	26.53%	28.04%
Linear classifier	58.95%	58.47%	59.45%

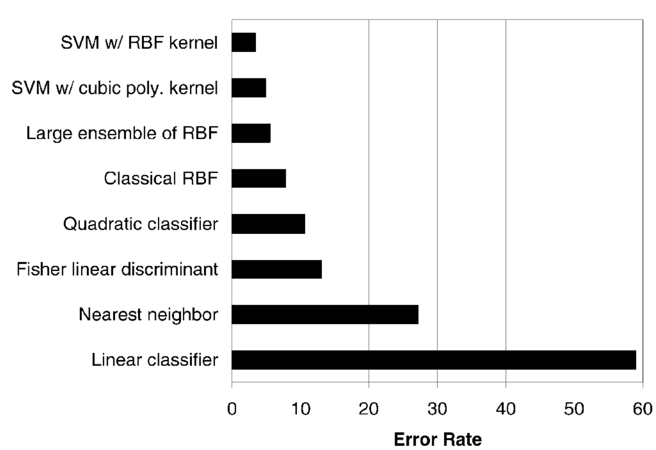

Fig. 5. Error rates of various classifiers.

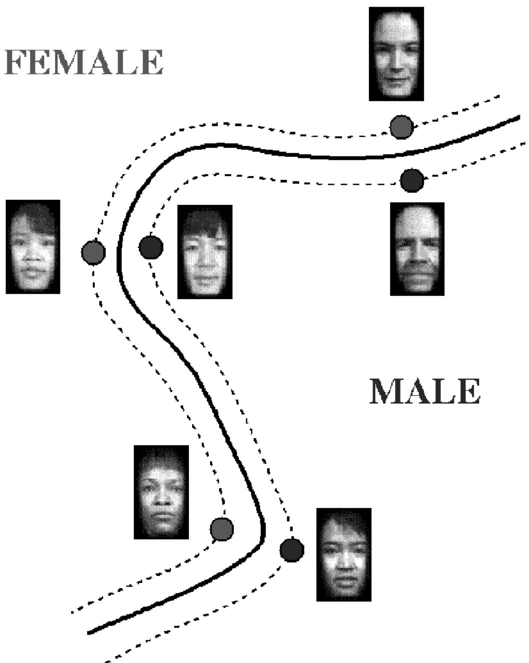

Fig. 6. Support faces at the boundary.

storage of 86 percent of the training set (SVMs required about 20 percent). Storage of the entire data set in the form of the nearest-neighbor classifier yielded too high an error rate (30 percent). Clearly, SVMs succeeded in the difficult task of finding a near-optimal gender partition in face space with the added economy of a small number of support faces.

Given the relative success of previous studies with low-resolution faces it is re-assuring that 21-by-12 faces can in fact be used for reliable gender classification. Unfortunately, most of the previous studies used data sets of relatively few faces. The most directly comparable study to ours is that of Gutta et al. [17], which also used FERET faces. With a data set of 3,000 faces at a resolution of 64-by-72, their hybrid RBF/Decision-Tree classifier achieved a 4 percent error rate. In our study, with 1,800 faces at a resolution of 21-by-12, a Gaussian kernel SVM was able to achieve a 3.4 percent error rate. Both studies use extensive cross validation to estimate the error rates. Given our results with SVMs, it is clear that better performance at even lower resolutions is made possible with this learning technique.

REFERENCES

[1] V. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[2] C.M. Bishop, *Neural Networks for Pattern Recognition.* Oxford Univ. Press, 1995.

[3] V. Bruce, A.M. Burton, N. Dench, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney, "Sex Discrimination: How do We Tell the Difference between Male and Female Faces?" *Perception,* vol. 22, pp. 131-152, 1993.

[4] R. Brunelli and T. Poggio, "Hyperbf Networks for Gender Classification," *Proc. DARPA Image Understanding Workshop,* pp. 311-314, 1992.

[5] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, no. 2, 1998.

[6] A.M. Burton, V. Bruce, and N. Dench, "What's the Difference Between Men and Women? Evidence from Facial Measurement," *Perception,* vol. 22, pp. 153-176, 1993.

[7] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing Multiple Parameters for Support Vector Machines," *Machine Learning,* 2001, To appear.

[8] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning,* vol. 20, 1995.

[9] G.W. Cottrell, "Empath: Face, Emotion, and Gender Recognition Using Holons," *Advances in Neural Information Processing Systems 3,* R.P. Lippmann, J.E. Moody, and D.S. Touretzky, eds. pp. 564-571, 1991.

[10] R. Courant and D. Hilbert, *Methods of Mathematical Physics.* vol. 1, New-York: Interscience, 1953.

[11] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines.* Cambridge Univ. Press, 2000.

[12] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* Wiley Interscience, 2001.

[13] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *J. Optical Soc. of Am.,* vol. 14, no. 8, pp. 1724-1733, 1997.

[14] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization Networks and Support Vector Machines," *Advances in Computational Math.,* vol. 13, no. 1, pp. 1-50, 2000.

[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[16] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski, "Sexnet: A Neural Network Identifies Sex from Human Faces," *Advances in Neural Information Processing Systems 3,* R.P. Lippmann, J.E. Moody, and D.S. Touretzky, eds. pp. 572-577, 1991.

[17] S. Gutta, H. Wechsler, and P.J. Phillips, "Gender and Ethnic Classification," *Proc. Third IEEE Int'l Automatic Face and Gesture Recognition,* pp. 194-199, 1998.

[18] S. Haykin, *Neural Networks : A Comprehensive Foundation.* Prentice Hall, 1998.

[19] J. Huang, X. Shao, and H. Wechsler, "Face Pose Discrimination Using Support Vector Machines," *Proc. 14th Int'l Conf. Pattern Recognition,* pp. 154-156, 1998.

[20] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[21] B. Moghaddam and M.-H. Yang, "Gender Classification with Support Vector Machines," *Proc. Fourth Int'l Conf. Automatic Face and Gesture Recognition,* pp. 306-311, 2000.

[22] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition,* pp. 130-136, 1997.

[23] A.J. O'Toole, T. Vetter, N.F. Troje, and H.H. Bulthoff, "Sex Classification is Better with Three-Dimensional Structure than with Image Intensity Information," *Perception,* vol. 26, pp. 75-84, 1997.

[24] P.J. Phillips, "Support Vector Machines Applied to Face Recognition," *Advances in Neural Information Processing Systems 11,* M.S. Kearns, S. Solla, and D. Cohen, eds., MIT Press, vol. 11, pp. 803-809, 1998.

[25] T. Poggio and F. Girosi, "Networks for Approximation and Learning," *Proc. IEEE,* vol. 78, no. 9, vol. 1481-1497, 1990.

[26] C. Saunders, M.O. Stitson, J. Weston, L. Bottou, B. Schoelkopf, and A. Smola, "Support Vector Machine Refernce Manual," Technical Report CSD-TR-98-03, Dept. Computer Science, Royal Holloway, Univ. of London, 1998.

[27] *Advances in Kernel Methods.* B. Schlkopf, C.J.C. Burges, and A.J. Smola, eds. MIT Press, 1998.

[28] D. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 8, pp. 831-836, Aug. 1996.

[29] S. Tamura, H. Kawai, and H. Mitsumoto, "Male/Female Identification from 8×6 Very Low Resolution Face Images by Neural Network," *Pattern Recognition,* vol. 29, no. 2, pp. 331-335, 1996.

[30] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995.

[31] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face Recognition and Gender Determination," *Proc. First Int'l Workshop Face and Gesture Recognition,* pp. 92-97, 1995.