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

Gender Classification with Support Vector Machines

Baback Moghaddam
Ming-Hsuan Yang

TR-2000-01 January 2000

Abstract

Support Vector Machines (SVMs) are investigated for visual gender classification with low resolution "thumbnail" faces (21-by-12 pixels) processed from 1,755 images from the FERET face database. The performance of SVMs (3.4traditional 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. SVMs also out-performed human test subjects at the same task: in a perception study with 30 human test subjects, ranging in age from mid-20s to mid-40s, the average error rate was found to be 32images. The difference in performance between low and high resolution tests with SVMs was only 1%, demonstrating robustness and relative scale invariance for visual classification.

Publication History:–

1. First printing, TR-2000-01, January 2000

Gender Classification with Support Vector Machines

Baback Moghaddam
Mitsubishi Electric Research Laboratory
201 Broadway
Cambridge, MA 02139 USA
baback@merl.com

Ming-Hsuan Yang
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801 USA
mhyang@vision.ai.uiuc.edu

Abstract

Support Vector Machines (SVMs) are investigated for visual gender classification with low resolution "thumbnail" faces (21-by-12 pixels) 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. SVMs also out-performed human test subjects at the same task: in a perception study with 30 human test subjects, ranging in age from mid-20s to mid-40s, the average error rate was found to be 32% for the "thumbnails" and 6.7% with higher resolution images. The difference in performance between low and high resolution tests with SVMs was only 1%, demonstrating robustness and relative scale invariance for visual classification.

1 Introduction

This paper addresses the problem of classifying gender from thumbnail faces in which only the main facial regions appear (without hair information). The motivation for using such 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. It has been shown that better *recognition* rates can be achieved using hairless images [10]. 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 datasets 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.

In recent years, SVMs have been successfully applied to various tasks in computational face-processing. These include face detection [16], face pose discrimination [14] and face recognition [18]. In this paper, we use SVMs for gender classification of thumbnail facial images and compare their performance with traditional classifiers (*e.g.,* Linear, Quadratic, Fisher Linear Discriminant, and Nearest Neighbor) and more modern techniques such as RBF networks and large ensemble-RBF classifiers.

We also compare the performance of SVM classifiers to that of human test subjects with both high and low resolution images. Although humans are quite good at determining gender from generic photographs, our tests showed that they had difficulty with hairless high resolution images. Nevertheless, the human performance at high resolution was deemed adequate (6.5% error), but degraded with low resolution images (31% error). SVM classifiers showed negligible changes in their average error rate. In our study, little or no hair information was used in both human and machine experiments. This is in contrast to previous results reported in the literature where almost all methods used include some hair information in gender classification.

2 Background

Gender perception and discrimination has been investigated from both psychological and computational perspectives. Although gender classification has attracted much attention in psychological literature [2, 5, 9, 17], relatively few learning based vision methods have been proposed.

Gollomb *et al.* [12] 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% compared to an average error rate of 11.6% from a study of five human subjects. Cottrell and Metcalfe [7] also applied neural networks for face emotion and gender classification. The dimensionality of a set of 160 64-by-64 face images (10 males and 10 females) was reduced from 4096 to 40 with an auto-encoder. These vectors were then presented as inputs to another one layer network for training. They reported perfect classification.[1] Brunelli and Poggio [3] developed

[1] However one should note that a dataset of only 20 unique individuals may be insufficient to yield statistically significant results.

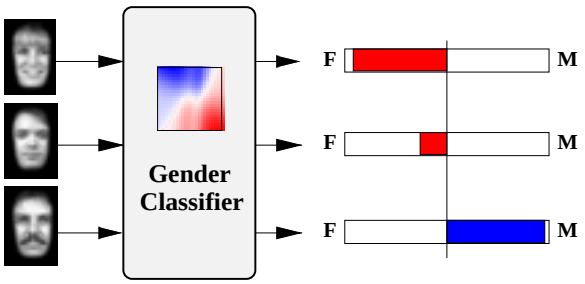

Figure 1. Gender classifier

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%. Using similar techniques as Golomb *et al.* [12] and Cottrell and Metcalfe [7], Tamura *et al.* [20] used multi-layer 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%. Instead of using a vector of gray levels to represent faces, Wiskott *et al.* [22] 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 similar to the geometric features in [4]. 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%. Recently, Gutta *et al.* [13] proposed a hybrid classifier based on neural networks (RBFs) and inductive decision trees with Quinlan's C4.5 algorithm. Experiments were conducted on 3000 FERET faces of size 64-by-72 pixels. The best average error rate was found to be 4%.

3 Gender Classifiers

A generic gender classifier is shown in Figure 1. An input facial image $\mathbf{x}$ 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. We will now briefly review the details of the various classifiers used in our study.

3.1 Support Vector Machines

A Support Vector Machine is a learning algorithm for pattern classification and regression [21, 6]. 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 [21], a function that classifies the training data accurately and which belongs to a set of functions with the lowest VC dimension [6] 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 non-separable 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 quite feasible to construct 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 [21] shows that maximizing the margin distance is equivalent to minimizing the VC dimension in constructing an optimal hyperplane. Computing the best hyperplane is posed as a constrained optimization problem and solved using quadratic programming techniques. The discriminant hyperplane is defined by the level set of

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

where $k(\cdot, \cdot)$ is a kernel function and the sign of $f(\mathbf{x})$ determines the membership of $\mathbf{x}$. Constructing an optimal hyperplane is equivalent to finding all the nonzero α_i. Any vector $\mathbf{x}_i$ that corresponds to a nonzero α_i is a *supported 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.

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:

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

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 [8], the expensive calculations required in projecting samples into the high dimensional feature space can be replaced by a much 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. In fact, by using different kernel functions, SVMs can implement a variety of learning machines, some of which coincide with classical architectures. Nevertheless, automatic selection of the "right" kernel function and its associated parameters remains problematic and in practice one must resort to trial and error for model selection.

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 [19]. 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. The weight coefficients w_i combine 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. 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.* [11] 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 [21]. As such, SVMs provide a more systematic approach to classification than classical RBF and various other neural networks.

3.3 Fisher Linear Discriminant

Fisher Linear Discriminant (FLD) is an example of a class specific subspace method that finds the optimal linear projection for classification. Rather than finding a projection that maximizes the projected variance as in principal component analysis, FLD determines a projection, $y = \mathbf{W}_{\mathcal{F}}^T \mathbf{x}$, that maximizes the ratio between the between-class scatter and the within-class scatter. Consequently, classification is simplified in the projected space.

Consider a c-class problem, with the between-class scatter matrix given by

$$S_B = \sum_{i=1}^{c} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

and the within-class scatter matrix by

$$S_W = \sum_{i=1}^{c} \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T$$

where $\boldsymbol{\mu}$ is the mean of all samples, $\boldsymbol{\mu}_i$ is the mean of class i, and N_i is the number of samples in class i. The optimal projection $\mathbf{W}_{\mathcal{F}}$ is the projection matrix which maximizes the ratio of the determinant of the between-class scatter to the determinant of the within-class scatter of the projections

$$\mathbf{W}_{\mathcal{F}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T S_B \mathbf{W}|}{|\mathbf{W}^T S_W \mathbf{W}|} = [w_1 \; w_2 \; \ldots w_m]$$

where $\{w_i | i = 1, 2, \ldots, m\}$ is the set of generalized eigenvectors of S_B and S_W, corresponding to the m largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \ldots, m\}$. However, the rank of S_B is $c - 1$ or less since it is the sum of c matrices of rank one or less. Thus, the upper bound on m is $c - 1$. To avoid the singularity, one can apply PCA first to reduce the dimension of the feature space to $N - c$, and then use FLD to reduce the dimension to $c - 1$. This two-step procedure is used in computing "FisherFaces" [1], for example. In our experiments, we used a single Gaussian to model the distributions of male and female classes 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.

3.4 Linear and Quadratic Classifiers

The decision boundary of a quadratic classifier is defined by a quadratic form in $\mathbf{x}$, derived through Bayesian error minimization. Assuming that the distribution of each class is Gaussian, the classifier output is given by

$$\begin{aligned} f(\mathbf{x}) \quad = \quad & \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \\ & \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) + \tfrac{1}{2} \ln \tfrac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \end{aligned}$$

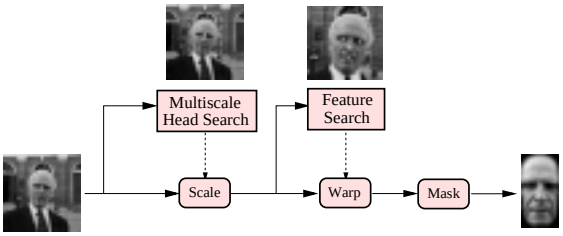

Figure 2. Face alignment system

Figure 3. Some processed FERET faces.

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ ($i = 1, 2$) are the mean and covariance matrix of the respective Gaussian distributions.

A linear classifier is a special case of the quadratic form, based on the assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, which simplifies the discriminant to

$$f(\mathbf{x}) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)$$

For both classifiers, the sign of $f(\mathbf{x})$ 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 pre-processed using an automatic face-processing system which compensates for translation, scale as well as slight rotations. Shown in Figure 2, this system is described in detail in [15] 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 Figure 2 were standardized to 80-by-40 (full) resolution. These "face-prints" were further sub-sampled to 21-by-12 pixel "thumbnails" for our low resolution experiments. Figure 3 shows a few examples of processed face-prints (note that these faces contain little or no hair information). A total of 1755 thumbnails (1044 males and 711 females) were used in our experiments. For each classifier, the average error rate was estimated with 5-fold cross validation (CV) — *i.e.,* a 5-way dataset split, with 4/5th used for training and 1/5th used for testing, with 4 subsequent rotations. The average size of the training set was 1496 (793 males and 713 females) and the average size of the test set was 259 (133 males and 126 females).

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%
Quadratic classifier	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%

4.1 Machine Classification

The SVM classifier was first tested with various kernels in order to explore the space of possibilities and performance. 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 1289 which corresponds to 86% of the training set. The number of radial bases for classical RBF networks was heuristically set to 20 prior to actual training and testing. Quadratic, Linear and Fisher classifiers were implemented using Gaussian distributions and 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 Figure 4.

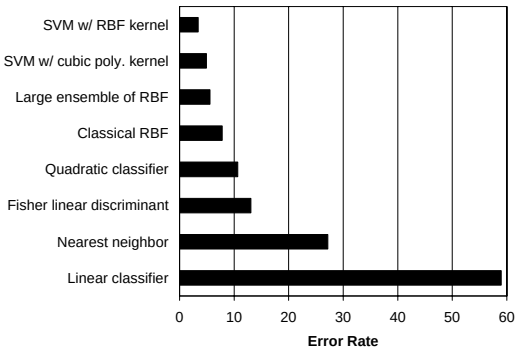

Figure 4. Error rates of various classifiers

The SVMs out-performed all other classifiers, although the performance of large ensemble-RBF networks was close to SVMs. However, nearly 90% 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% of the training set. We also applied SVMs to classification based on high resolution

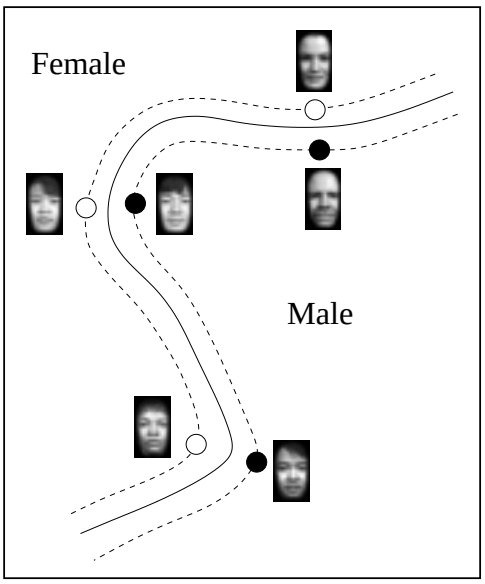

Figure 5. Support faces at the boundary

images. The Gaussian and cubic kernel SVMs performed equally well at both low and high resolutions with only a slight 1% error rate difference. Figure 5 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 (hyper-surface) 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. As seen in Table 1, all the classifiers tested had higher error rates in classifying females, most likely due to the less prominent and distinct facial features present in female faces.

4.2 Human Classification

In order to calibrate the performance of SVM classifiers, human subjects were also asked to classify gender using both low and high resolution images. A total of 30 subjects (22 males and 8 females) ranging in age from mid-20s to mid-40s participated in an experiment with high resolution images and 10 subjects (6 males and 4 females) with low resolution images. All subjects were asked to classify the gender of 254 faces (presented in random order) as best as they could without time constraints. Although these tests were not as comprehensive as the machine experiments, the test set used with humans was identical to one of the 5-fold CV partitions used in Section 4.1.

Table 2. Human error rates

Gender of human subject	Error Rate	
	High resolution	Low resolution
Male	7.02%	30.87%
Female	5.22%	30.31%
Combined	**6.54%**	**30.65%**

Figure 6. SVM vs. Human performance

The human error rates obtained in our study are tabulated in Table 2. Comparing Tables 1 and 2, it is clear that SVMs also perform better than humans with both low and high resolution faces — this is more easily seen in Figure 6. These results suggest that the concept of gender is more accurately modeled by SVMs than any other classifier. It is not surprising that human subjects perform better with high resolution images than with low resolution images. SVM performance, however, was mostly unaffected by the change in resolution.

Figure 7 shows the top 5 mistakes made by human test subjects (the true gender is F-M-M-F-M from left to right). Our results also indicated that there was a degree of correlation between the mistakes made by SVMs and those made by humans. Faces misclassified by SVMs were almost always misclassified by humans as well (for all the SVM mistakes, the average human error rate was more than 30%). On the other hand, the converse was not generally found to be true (humans made different mistakes than SVMs). Finally, we note that SVM classifiers performed better than any single human test subject, at either resolution.

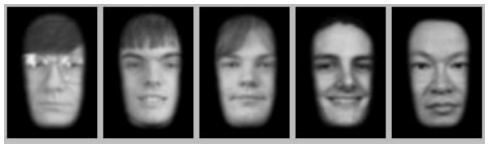

Figure 7. Top five human misclassifications

5 Discussion

In this paper we have presented a comprehensive evaluation of various classification methods for determination of gender from facial images. The non-triviality 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% (*i.e.*, worse than a random coin flip). Furthermore, an acceptable error rate ($< 5\%$) for the large ensemble-RBF network required storage of 86% of the training set (SVMs required about 20%). Storage of the entire dataset in the form of the nearest-neighbor classifier yielded too high an error rate (30%). 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.

The comparison of machine *vs.* human performance, shown in Figure 6, indicates that SVMs with low resolution images actually do better (3.4%) than human subjects with high resolution images (6.5%). This can be partly explained by the fact that hair cues (mostly missing in our dataset) are important for human gender discrimination. The fact that human performance degrades with lower resolution is not too surprising: as humans, our lifetime of "training" in gender classification has been carried out with moderate-to-high resolution stimuli. The various machine classifiers, on the other hand, were *re-trained* for each resolution. The relative invariance of SVMs to input resolution is due to the fact that their complexity (hence performance) depends primarily on the number of training samples and *not* their dimension [21].

Given the relative success of previous studies with low resolution faces it is re-assuring that 21-by-12 faces (or even 8-by-6 faces [20]) can in fact be used for reliable gender classification. Unfortunately, most of the previous studies used datasets of relatively few faces (and even fewer human subjects to test them on). The most directly comparable study to ours is that of Gutta *et al.* [13], which also used FERET faces. With a dataset of 3000 faces at a resolution of 64-by-72, their hybrid RBF/Decision-Tree classifier achieved a 4% error rate. In our study, with 1800 faces at a resolution of 21-by-12, a Gaussian kernel SVM was able to achieve a 3.4% 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 Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(7):711–720, July 1997.

[2] 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*, 22:131–152, 1993.

[3] R. Brunelli and T. Poggio. HyperBF networks for gender classification. In *Proceedings of the DARPA Image Understanding Workshop*, pages 311–314, 1992.

[4] R. Brunelli and T. Poggio. Face recognition : Features vs. templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10), October 1993.

[5] A. M. Burton, V. Bruce, and N. Dench. What's the difference between men and women? evidence from facial measurement. *Perception*, 22:153–176, 1993.

[6] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20, 1995.

[7] G. W. Cottrell and J. Metcalfe. EMPATH: Face, emotion, and gender recognition using holons. In *Advances in Neural Information Processing Systems*, pages 564–571, 1991.

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

[9] B. Edelman, D. Valentin, and H. Abdi. Sex classification of face areas: how well can a linear neural network predict human performance. *Journal of Biological System*, 6(3):241–264, 1998.

[10] H.-Y. L. *et al.* Face recognition using a face-only database: A new approach. In *Proceedings of Asian Conference on Computer Vision*, volume 1352 of *Lecture Notes in Computer Science*, pages 742–749. Springer, 1998.

[11] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. Technical Report AI Memo No. 1654, MIT, 1999.

[12] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. SEXNET: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems*, pages 572–577, 1991.

[13] S. Gutta, H. Wechsler, and P. J. Phillips. Gender and ethnic classification. In *Proceedings of the IEEE International Automatic Face and Gesture Recognition*, pages 194–199, 1998.

[14] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines. In *Proc. of 14th Int'l Conf. on Pattern Recognition (ICPR'98)*, pages 154–156, August 1998.

[15] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(7):696–710, July 1997.

[16] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[17] 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*, 26:75–84, 1997.

[18] P. J. Phillips. Support vector machines applied to face recognition. In M. S. Kearns, S. Solla, and D. Cohen, editors, *Advances in Neural Information Processing Systems 11*, volume 11, pages 803–809. MIT Press, 1998.

[19] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[20] S. Tamura, H. Kawai, and H. Mitsumoto. Male/Female identification from 8×6 very low resolution face images by neural network. *Pattern Recognition*, 29(2):331–335, 1996.

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

[22] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition and gender determination. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, pages 92–97, 1995.