*LETTER* **Communicated by Peter Hancock**

# Classification of Faces in Man and Machine

**Arnulf B. A. Graf***
*arnulf.graf@nyu.edu*
**Felix A. Wichmann**
*felix.wichmann@tuebingen.mpg.de*
**Heinrich H. Bülthoff**
*heinrich.buelthoff@tuebingen.mpg.de*
**Bernhard Schölkopf**
*bernhard.schoelkopf@tuebingen.mpg.de*
*Max Planck Institute for Biological Cybernetics, D 72076 Tübingen, Germany*

**We attempt to shed light on the algorithms humans use to classify images of human faces according to their gender. For this, a novel methodology combining human psychophysics and machine learning is introduced. We proceed as follows. First, we apply principal component analysis (PCA) on the pixel information of the face stimuli. We then obtain a data set composed of these PCA eigenvectors combined with the subjects' gender estimates of the corresponding stimuli. Second, we model the gender classification process on this data set using a separating hyperplane (SH) between both classes. This SH is computed using algorithms from machine learning: the support vector machine (SVM), the relevance vector machine, the prototype classifier, and the K-means classifier. The classification behavior of humans and machines is then analyzed in three steps. First, the classification errors of humans and machines are compared for the various classifiers, and we also assess how well machines can recreate the subjects' internal decision boundary by studying the training errors of the machines. Second, we study the correlations between the rank-order of the subjects' responses to each stimulus—the gender estimate with its reaction time and confidence rating—and the rank-order of the distance of these stimuli to the SH. Finally, we attempt to compare the metric of the representations used by humans and machines for classification by relating the subjects' gender estimate of each stimulus and the distance of this stimulus to the SH. While we show that the classification error alone is not a sufficient selection criterion between the different algorithms humans might use to classify face stimuli, the distance of these stimuli to the SH is shown to capture essentials of the internal decision**

*Present address: Center for Neural Science, New York University, New York, NY, USA.

 

**space of humans. Furthermore, algorithms such as the prototype classifier using stimuli in the center of the classes are shown to be less adapted to model human classification behavior than algorithms such as the SVM based on stimuli close to the boundary between the classes.**

## 1 Introduction

Bringing together theoretical modeling and behavioral data is arguably one of the main challenges when studying the "computational brain" (Churchland & Sejnowski, 1992). The aim of this letter is to obtain a better understanding of the algorithms responsible for the classification of visual stimuli by humans. For this, we combine machine learning and psychophysical techniques to gain insights into the algorithms human subjects use during visual classification of images of human faces according to their gender. In this "machine-learning-psychophysics" approach, we substitute a complex system that is very hard to analyze—the human brain—with a reasonably complex system—a learning machine (Vapnik, 1998). The latter is complex enough to capture some essentials of the human behavior but is still amenable to close analysis (Poggio, Rifkin, Mukherjee, & Niyogi, 2004). The research presented in this article is focused on a novel methodology that bridges the gap between human psychophysics and machine learning by extracting quantitative information from a (high-level) human behavioral experiment.

The past decade has seen important technological advances in neuroscience from a microscopic scale (e.g., multiunit recordings) to a macroscopic scale (e.g., functional magnetic resonance imaging), yielding novel insights into visual processing. However, on an algorithmic level, the methods and understanding of brain processes involved in visual recognition are still limited, although numerous attempts have been made since this problem was pointed out by Marr (1982).

Recently various computational models for visual recognition have been proposed. For instance, a network of Gabor wavelet filters was used to describe the processing of visual information (Mel, 1997). Independent component analysis was combined with a nearest-neighbor classifier to model face recognition (Bartlett, Movellan, & Sejnowski, 2002). The computations done by the human visual system for facial expression recognition were described using Gabor wavelets, principal component analysis, and artificial neural networks (Dailey, Cottrell, Padgett, & Adolphs, 2002). Object recognition and classification was also modeled using a hierarchical model composed of a network of nonlinear units combined using a maximum operation (Riesenhuber & Poggio, 1999, 2002). While each of these methods is successful for its own task, they illustrate the divergence of the approaches used to understand human category learning as pointed out, for example, in the overview by Ashby and Ell (2001). In this letter, we propose a novel

method combining machine learning and human psychophysics to shed light on the algorithms humans use to classify visual stimuli. Our framework allows us to compare directly the classification behavior of different algorithms to that of humans.

While the results obtained in this letter have no claim to be biologically inspired or to explain a specific function of the visual system (see, e.g., Rolls & Deco, 2002, for an overview of such computational methods), we instead ask the following questions: Can we generate testable hypotheses about the algorithms humans use to classify visual inputs? Can we find a classifier whose behavior reflects human classification behavior significantly better than others? Current high-level vision research, with its intrinsically complex stimuli, is hampered by a lack of methods to answer such questions at the algorithmic level. The method presented here has the potential to contribute to overcoming this obstacle.

An initial attempt using machine learning to help understand the algorithms humans use to classify the gender of faces was presented by Graf and Wichmann (2004). This letter extends that work. In section 2 we present a psychophysical gender classification experiment of images of human faces and analyze the subjects' responses—the gender estimate with its reaction time and confidence rating. Section 3 introduces several algorithms from machine learning that will be used to model the classification behavior of humans. Our analysis of the classification behavior of humans proceeds in three steps. First, the classification performance of humans and machines is compared in section 4, and the findings are related to those described in the literature. Second, we correlate in section 5 the rank-order of the subjects' responses to each stimulus with the rank-order of the distance of this stimulus to the separating hyperplane (SH) of the machine. The success of these studies encourages us to perform the third step in section 6: a metric comparison of the representations used by humans and machines for classification, using the subjects' gender estimate of each stimulus and the corresponding distance to the SH of the machine. Section 7 summarizes our results and discusses their implications.

## 2 Human Classification

In a human psychophysical classification experiment, 55 human subjects were asked to classify a random gender-balanced subset of 152 out of 200 realistic human faces according to their gender. The stimuli were presented sequentially once to each subject. The temporal envelope of stimulus presentation was a modified Hanning window (a raised cosine function with a raising time of 500 ms and a plateau time of 1000 ms, for a total presentation time of 2000 ms per face). After the presentation of each stimulus, a blank screen with mean luminance was shown to the subjects for 1000 ms before the presentation of the following stimulus. We recorded the subjects' estimated gender (female or male) together with the reaction time (RT) and

a confidence rating (CR) on a scale from 1 (unsure) to 3 (sure). No feedback on the correctness of the subjects' answers was provided. Subjects were asked to classify the faces as fast as possible to obtain perceptual, rather than cognitive, judgments. Most of the time they responded well before the presentation of the stimulus had ended (mean reaction time over all stimuli and subjects was approximately 900 ms). A training phase of 8 faces (4 male and 4 female faces) preceded the actual classification experiment in order to acquaint the subjects with the stimuli and the experimental procedure. Subjects viewed the screen binocularly with their head stabilized by a headrest. All subjects had normal or corrected-to-normal vision and were paid for their participation. Most of them were students from the University of Tübingen, and all of them were naive to the purpose of the experiment.

Each stimulus was an 8-bit grayscale frontal view of a Caucasian face with a nominal size of $256 \times 256$ pixels. All faces were centered on the display, had the same pixel-surface area and the same mean intensity, and they came from a processed version of the MPI face database[1] (Blanz & Vetter, 1999). The details of the image processing are described in Graf and Wichmann (2002). The stimuli were presented against the mean luminance (50 cd/m$^2$) of a linearized Clinton Monoray CRT driven by a Cambridge Research Systems VSG 2/5 display controller. Neither the presentation of a male nor of a female face changed the mean luminance of the screen.

The subjects' gender estimates were analyzed using signal detection theory (Wickens, 2002). We assume that on the decision axis, the internal class representations are corrupted by gaussian distributed noise with same unit variance but different means. We define correct response probabilities for male (+) and female (−) stimuli as $P_+ = P(\hat{y} = 1|y = 1)$ and $P_- = P(\hat{y} = -1|y = -1)$, where $\hat{y}$ is the estimated class and $y$ the true class of the stimulus. The discriminability of both classes can then be computed as $d' = Z(P_+) + Z(P_-)$, where $Z = \Phi^{-1}$, and $\Phi$ is the cumulative normal distribution with zero mean and unit variance. Averaged across all subjects, we obtain a high discriminability, $d' = 2.63 \pm 0.57$, suggesting that the classification task is comparatively easy for the subjects, albeit not trivial (no ceiling effect). Furthermore, the subjects exhibit a pronounced male bias in the responses defined as $\log(\beta) = \frac{1}{2}(Z^2(P_+) - Z^2(P_-)) = 1.49 \pm 1.15$, indicating that more females are classified as males than males as females.

In Figure 1 we show the relation between the average across all subjects of the subjects' responses for each stimulus, each point in these plots representing one stimulus. We can first see that for $P(\hat{y} = +1|x) \approx 1$, all the stimuli are male and that for $P(\hat{y} = +1|x) \approx 0$, all the stimuli are female. Second, we can observe the male bias already mentioned: a higher density of responses near $P(\hat{y} = +1|x) \approx 1$. Furthermore, there are female stimuli for which $P(\hat{y} = +1|x) > \frac{1}{2}$, but no male stimuli for which $P(\hat{y} = +1|x) < \frac{1}{2}$.

[1] To be found online at http://faces.kyb.tuebingen.mpg.de.

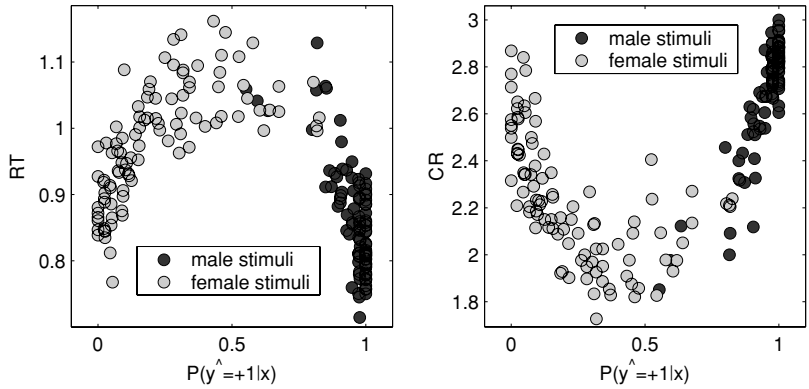


Figure 1: Relation between the subjects' responses—the probability $P(\hat{y} = +1|x)$ to answer male, the reaction time RT, and the confidence rating CR—on a stimulus-by-stimulus basis (responses averaged across subjects).

Clearly the threshold for male-female discrimination depends on the male bias and is located in $[\frac{1}{2}, 1]$. Third, we notice that for stimuli with a high probability to belong to either class ($P(\hat{y} = +1|x) = 0$ or 1), the corresponding RTs are short and the CRs are high. In other words, when the subjects make a correct gender estimate, they answer fast, and they are confident of their response. For the stimuli where the subjects have difficulty choosing a class ($P(\hat{y} = +1|x) \approx 0.5$), they take longer to respond (long RT) and are unsure of their response (low CR). Subjects thus have a rather good knowledge of the correctness of their gender estimate.

## 3 Machine Classification

To model the subjects' classification behavior using machine learning, we first need to preprocess the stimuli to reduce their "apparent" dimensionality. We use principal component analysis PCA (Duda, Hart, & Stork, 2001), a widely used linear preprocessor from unsupervised machine learning, to preprocess the data. PCA is an eigenvalue decomposition of the covariance matrix associated with the data matrix $D = BE$ along the directions of largest variance where the columns of the basis matrix $B$ are constrained to be orthonormal and the rows of the encoding matrix $E$ are orthogonal. The rows of $B$ are termed *eigenfaces* according to one of the first studies to apply PCA to human faces (Sirovich & Kirby, 1987). PCA has also been successfully applied to model face perception and classification in a large number of studies, from psychophysics (O'Toole, Abdi, Deffenbacher, & Valentin, 1993: Valentin, Abdi, Edelman, & O'Toole, 1997; O'Toole, Deffenbacher, Valentin, McKee, Huff, & Abdi, 1998; O'Toole, Vetter, & Blanz, 1999; Furl,

Phillips, & O'Toole, 2002), to artificial recognition systems (Turk & Pentland, 1991; Golomb, Lawrence, & Sejnowski, 1991; Gray, Lawrence, Golomb, & Sejnowski, 1995; O'Toole, Phillips, Cheng, Ross, & Wild, 2000; Bartlett et al., 2002) and facial expression modeling (Calder, Burton, Miller, Young, & Akamatsu, 2001). Like all previous studies, we apply PCA to the vectors obtained when reshaping the intensity matrix of the pixels of each face into a single $256^2 \times 1$ vector. We keep the full space of the data, that is, the 200 nonzero components of the PCA decomposition of the data, and obtain a PCA-encoding data matrix $E$ of size $200 \times 200$, where each row is the encoding corresponding to a face stimulus. By construction, these encodings are already centered. Subsequently these encodings are also normalized since this has been shown to be quite effective in real-world applications for some classifiers (Graf, Smola, & Borer, 2003). Since we consider the full encoding space of dimension 200, the choice of PCA as a preprocessor is of little consequence, and the face stimuli can be reconstructed perfectly from these encodings.

In this letter, we consider two types of stimulus data sets for each subject: the *true* and the *subject* data sets. The patterns in both data sets are represented by their (centered and normalized) PCA encodings. The true data set contains the $p = 152$ encodings $\vec{x}_i \in \mathbb{R}^{200}, i = 1, \ldots, p$ of the stimuli seen by the subject, combined with the true labels $y_i = \pm 1$ of these stimuli—their true gender as given by the MPI face database. The subject data set is composed of the same encodings $\vec{x}_i$, combined this time with the labels $\hat{y}_i$ of the stimuli as estimated by the subject in the psychophysical classification experiment. This data set represents what we assume to be the subject's internal representation of the face space. Altogether we thus have 55 true and subject data sets.

We use methods from supervised machine learning to model classification. The classifiers are applied to the true and the subject data sets and thus classify in the PCA space of dimension 200. We consider classifiers that are linear: they classify using a separating hyperplane (SH) defined by its normal vector $\vec{w}$ and offset $b$. Furthermore, these classifiers can all be expressed in dual form: the normal vector is a linear combination of the patterns of the data set $\vec{w} = \sum_i \alpha_i \vec{x}_i$. Since we cannot investigate all such classifiers in an exhaustive manner, we consider the most representative member of each one of four families of classification principles: the support vector machine, the relevance vector machine, the prototype classifier and the K-means classifier. Figure 2 shows these classifiers applied on a two-dimensional toy data set. These classifiers are presented and discussed in further detail below.

The support vector machine SVM (Vapnik, 2000; Schölkopf & Smola, 2002) is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory. SVMs classify by maximizing the margin separating both classes while minimizing the classification errors. This trade-off between maximum margin and misclassifications is controlled by a

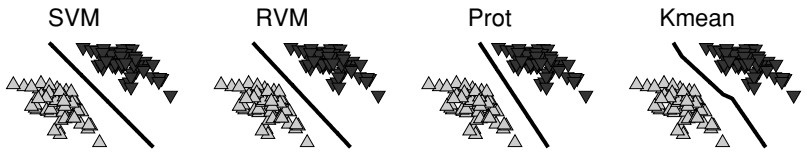


Figure 2: Classification of a two-dimensional toy data set using the classifiers considered in this study. The dark lines indicate the SHs.

parameter $C$ set by cross-validation.[2] The optimal dual space parameter $\vec{\alpha}$ maximizes the following expression,

$$\sum_i \alpha_i - \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j \langle \vec{x}_i | \vec{x}_j \rangle,$$

subject to

$$\sum_i \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C,$$

where $\langle \cdot | \cdot \rangle$ stands for the inner (or scalar) product between two vectors. The offset is computed as $b = \langle y_i - \langle \vec{w} | \vec{x}_i \rangle \rangle_{i|0<\alpha_i<C}$. Patterns of the data set satisfying $\alpha_i \neq 0$ are called *support vectors*, and they lie on the boundary or inside of the margin stripe between the classes. Both perceptrons (Rosenblatt, 1958) and adaboost (Freund & Schapire, 1995) can be interpreted as maximum margin classifiers—a maximum margin is a property of these algorithms but no notion of margin appears in their definition (see Graepel, Herbrich, & Williamson, 2001, and Schapire, Freund, Bartlett, & Lee, 1998, respectively) and thus belong to the same family of algorithms as SVMs. Also SVMs can be thought of as a more principled version of two-layered feedforward artificial neural networks (LeCun, Bottou, Orr, & Müller, 1998; Haykin, 1999).

Probabilistic Bayesian classification is represented by the relevance vector machine RVM (Tipping, 2001), which belongs to the family of gaussian processes (Williams & Barber, 1998). The RVM classifies patterns by maximizing a conditional probability of class membership $P(\vec{y}|X, \vec{\beta})$ given the data $X = \{\vec{x}_i\}_{i=1}^p$ and some hyperparameter $\vec{\beta}$. The class membership $P(\vec{y}|X, \vec{\alpha})$ is modeled using a Bernoulli distribution. The sparseness of $\vec{\alpha}$ is

[2] Cross-validation is used to assess in an unbiased manner the classification error of an algorithm on a given data set. An $N$-fold cross-validation scheme separates the data set into $N$ subsets where $N-1$ are used for training and the remaining one is used for testing. The average over all $N$ possibilities is then an estimate of the classification error of the classifier on the considered data set. When estimating optimal parameters, the parameter yielding the minimal classification error is chosen.

introduced using a gaussian distribution for $P(\vec{\alpha}|\vec{\beta})$. Learning then amounts to maximizing with respect to $\vec{\beta}$ the following conditional probability:

$$P(\vec{y}|X,\vec{\beta}) = \int P(\vec{y}|X,\vec{\alpha})P(\vec{\alpha}|\vec{\beta})d\vec{\alpha}.$$

The value of $\vec{\beta}$ maximizing the above probability is then used to compute $\vec{\alpha}$ using $P(\vec{\alpha}|\vec{\beta})$, and thus also $\vec{w}$ and $b$. Since this integral cannot be solved analytically, the Laplace approximation (local approximation of the integrant by a gaussian) is used for solution, yielding an iterative update scheme for $\vec{\beta}$.

Some classifiers used in neuroscience, cognitive science, and psychology are variants of the mean-of-class prototype classifier Prot (Reed, 1972; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Its popularity may be due in part to its intuitiveness—representing the class by its mean tendency—as well as due to its simplicity: it classifies according to the nearest mean-of-class prototype. In its simplest form, all dimensions are weighted equally, but variants exist where the weight of each dimension is inversely proportional to the class variance along that dimension. As we cannot estimate class variance along all 200 dimensions from only 200 stimuli, we chose to implement the simplest prototype classifier with equal weights along all dimensions where the prototypes are defined as

$$\vec{p}_{\pm} = \frac{\sum_i \vec{x}_i(y_i \pm 1)}{\sum_i (y_i \pm 1)}.$$

The weight vector and the offset are then computed respectively as

$$\vec{w} = \vec{p}_+ - \vec{p}_- \quad \text{and} \quad b = \frac{\|\vec{p}_-\|^2 - \|\vec{p}_+\|^2}{2}.$$

The expression for $\vec{w}$ can be rewritten as a linear combination of the patterns $\vec{x}_i$, and thus Prot can be viewed as a classifier in dual form. Note that due to the homogeneity of the faces in the MPI face database (Graf & Wichmann, 2002) this classifier is likely to be close to the "best" possible prototype classifier. The popularity of prototype classification has led to several variants. For instance, the general context model (Palmeri, 2001; Nosofsky, 1991) is a classifier where instead of computing $\|\vec{x} - \vec{p}_{\pm}\|$ as for the prototype classifier, the quantity $\sum_{i|y_i=\pm 1} \|\vec{x} - \vec{x}_i\|$ is used for classification. Moreover, the Fisher linear discriminant classifier FLD (Fisher, 1936) is a whitened variant of the prototype classifier. Indeed, the FLD weight vector can be written as $\vec{w} = S_w^{-1}(\vec{p}_+ - \vec{p}_-)$, where $S_w = S_+ + S_-$ and $S_{\pm} = \sum_{\vec{x}_i|y_i=\pm 1} |\vec{x}_i - \vec{p}_{\pm}\rangle\langle\vec{x}_i - \vec{p}_{\pm}|$ is the within-class covariance matrix of the positive and negative data, respectively (Duda et al., 2001), the notation $|\cdot\rangle\langle\cdot|$ standing for the outer product of two vectors. Consequently, if we disregard the constant offset $b$, we can write the decision

function as $\langle \vec{w}|\vec{x}\rangle = \langle S_w^{-1}(\vec{p}_+ - \vec{p}_-)|\vec{x}\rangle = \langle S_w^{-1/2}(\vec{p}_+ - \vec{p}_-)|S_w^{-1/2}\vec{x}\rangle$, which is a prototype classifier using the prototypes $\vec{p}_\pm$ after whitening the space with $S_w^{-1/2}$. Finally, we may mention that FLD is prone to overfitting when considering fewer patterns $p$ than dimensions $n$, which is the case for us: $p = 152 \leq n = 200$. This makes FLD not suited as a classifier for our studies.

An extension of prototype classification is to consider for each class multiple "prototypes" computed, for instance, using the K-means clustering algorithm (Duda et al., 2001). By combining these prototypes with a nearest-neighbor classifier, we obtain the K-means classifier Kmean. The number of means $K$ is assumed to be the same for both classes, and its value is determined using cross-validation. The SH obtained here is piecewise linear, and Kmean represents the family of piecewise linear SH algorithms. Every portion of the SH of Kmean is computed using the Prot algorithm, which makes Kmean a classifier in dual form "by parts." The extension of the prototype algorithm to a multiprototype one has been suggested by Edelman (1995) in the context of his "chorus of prototype" approach, which cannot be directly applied to our study. Our Kmean classifier is close in spirit, however.

## 4 Classification Errors of Man and Machine

First we assess the classification errors of humans and machines using cross-validation, a method involving multiple training and testing sets, which allows us to estimate the generalization ability of the classifiers. Second, we show that for the particular task we chose, training on the entire data set using a single training and testing set does not lead to overfitting since the classification errors obtained with and without cross-validation are not significantly different. Finally, we study the training error of the classifiers, which is a measure of how well the classifiers can recreate the subjects' internal decision boundary for faces.

For humans, the classification error on the true data set is simply obtained by considering the mean and standard deviation over all 55 human subjects of the individual mean classification error computed by comparing the true gender of a stimulus with its estimate. The classification error on the subject data set cannot be computed directly since the subject's labels are not known beforehand. To obtain this error, we use a method derived from cross-validation where for each stimulus shown to a particular subject, we compute the mean error the other subjects made on this stimulus by defining as an error when the other subjects responded differently than the considered subject did. The classification error on the subject data set is thus computed by treating each subject's responses in turn as being "correct" and calculating the classification error of all the other subjects by this standard. In other words, we compare the subjects' gender responses on common stimuli and determine the mean consistency between subjects. We then

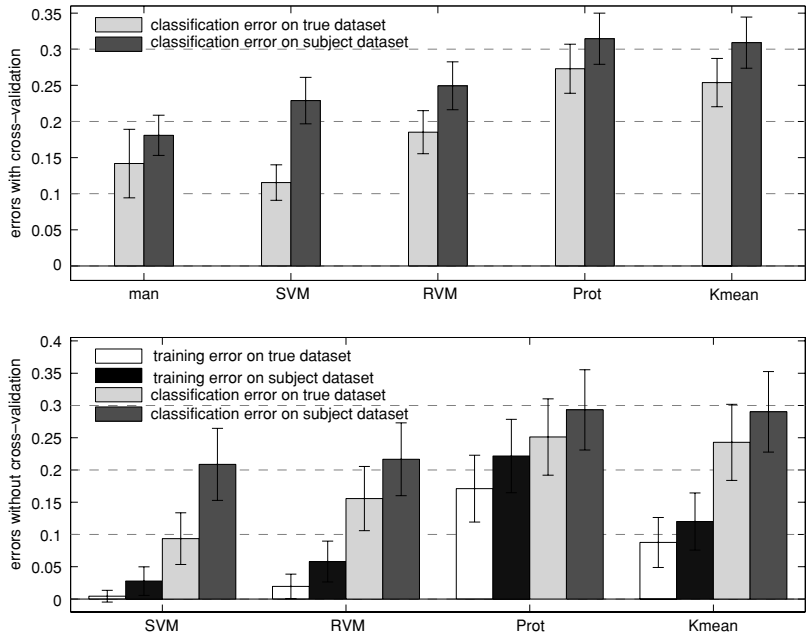


Figure 3: (Top) Classification error of humans and machines on both data sets assessed using cross-validation (multiple training and testing sets). (Bottom) Training and classification errors of the machines on both data sets computed without cross-validation (single training and testing set).

compute the mean and standard deviation of this error over all the stimuli presented to that subject. For machines, the mean and standard deviation of the classification error is obtained, for both the true and the subject data sets, using a single five-fold cross-validation on the classification error for the RVM and Prot and a double five-fold cross-validation to determine also the optimal values of $C$ for the SVM and $K$ for Kmean. The mean and standard error over all 55 subjects of the mean and standard deviation of the above "individual" classification errors are computed for both data sets and are shown in the top row of Figure 3.

When considering the classification error of humans, we notice that the standard error is smaller for the subject data set than for the true one. This is due to our method of assessing the classification error on the subject data set: it is computed using the consistency between each subject's responses and the other subjects' responses on the same set of stimuli. As the subjects' responses tend to agree—a stimulus whose gender is difficult to assess by one subject is likely to be difficult to classify also by the other subjects—the average gender response over all subjects will also be similar. Hence the

standard error of the subjects' responses will be small. However, on the true data set, we do not have this "average of average" consistency effect, which is why the standard error of the classification error is larger on the true than on the subject data set.

On the true data set, while the classification errors are not significantly different for humans versus the SVM and for humans versus the RVM (the error bars overlap), humans significantly outperform Prot and Kmean. On the subject data set, however, all the machines perform on average worse than humans, at least given our method of assessing the human classification error on the subject data set. This suggests that at least on the subject data set, humans and machines may be using different image features for classification. In all considered cases, the classification error on the subject data set is higher than on the true data set, which suggests that the subjects' labels make classification more difficult. This may be due to the inherent variability (noise or jitter) in the subjects' labeling. Prot and Kmean perform much worse than humans on both data sets, suggesting that either humans do not use Prot and Kmean for classification, or they do not use the PCA representation, or none of these.

The above results can be compared to those obtained by Graf and Wichmann (2004) where instead of applying PCA directly on the pixel information, PCA was applied to a representation of the faces that uses correspondences between the images such as texture and shape maps (e.g., a nose is mapped to a nose). Although the conclusions were similar, the classification errors of machines are higher in this study, which suggests that a representation using correspondences, that is, an additional amount of information, makes classification an easier task for the machines. There have also been numerous attempts to compare the classification performance of humans and machines in the context of gender classification. Most of them used artificial neural network (ANN) classifiers applied on a PCA representation of the image intensity information. The so-called holons, computed from the PCA representation, were used by Cottrell and Metcalfe (1991) as inputs to an ANN in the EMPATH recognition system to predict the identity, the emotion, and the gender of the face stimuli. This system was shown to perfectly classify gender and to outperform humans for assessing emotion. In Golomb et al. (1991), ANNs were shown to classify gender better than humans, although not much, using the so-called SEXNET architecture. Contrary to the above findings, in our case the SVMs, although a principled version of ANNs, do not perform significantly better than humans, which may be due to the fact that we use linear SVMs. Other studies (Gray et al., 1995) using face stimuli at different resolutions but without the PCA stage indicate that the gender classification problem seems to be linearly separable since a simple perceptron yielded results similar to a multilayer ANN. We have obtained a similar result: some linear classification algorithms are good models for gender classification as testified by their relatively low classification errors.

A cross-validation scheme involving multiple training and testing sets is useful to assess the generalization ability of a classifier by giving an estimate of its classification error on a given data set. However, training on the entire data set may not always yield to overfitting. In particular, if we can show that the classification error assessed using cross-validation is not significantly different from the one obtained by training on the entire data set and testing on a separate testing set, we can then assert that the classifiers are not overfitting, even if trained on the entire data set. Moreover, we then also have a gain of interpretability of the classification process: training on the entire data set yields a single SH, while cross-validation amounts to using multiple SHs in a piecewise linear manner. In the case of the SVM and Kmean, in order to determine the optimal value of the parameters $C$ and $K$, respectively, we still need to proceed to a single 10-fold cross-validation on the classification error. However, the classifiers are still trained on the whole data set using these optimal values, and therefore each classifier has a single SH. We then compute for each subject the mean and standard deviation of the following errors for the various classification algorithms:

- The training error on the true and on the subject data set
- The classification error on the true data set computed using the unseen stimuli with their true labels
- The classification error on the subject data set determined using the unseen stimuli with, as labels, the sign of the mean of the other subjects' responses for each of these unseen stimuli

The unseen stimuli are the remaining 48 stimuli out of the 200 that have not been seen by the considered subject. These training and testing errors are then averaged, and the standard error is computed over all subjects. The resulting values are shown in the bottom row of Figure 3.

We compare the generalization ability of the classifiers when trained once on the entire data set (no cross-validation) or multiple times on parts of it as done for cross-validation by comparing, respectively, the classification errors of the bottom and top rows of Figure 3. Although the classification errors are slightly lower without cross-validation, which may be due to overfitting, these errors do not differ significantly. Moreover, although the errors themselves are slightly changed, their relation among each other is unchanged. Therefore, for the considered task, we do not need to use cross-validation.

Most important, the training errors on both data sets are a measure of how well the classifiers can recreate the subjects' internal decision boundary for face representation. While the SVM and the RVM perform quite well at this task, Prot and Kmean are rather poor candidates. As for the classification errors, the machines have on average more difficulty learning the subject data set than the true one.

Comparing the classification errors of humans and machines mainly describes the input-output mapping of the human brain and of the machine. This shows only what is available in a black-box approach, and, as we may guess, this is not enough to make strong claims about the algorithms that humans actually use to classify visual stimuli. To infer these algorithms from our machine-learning-psychophysics approach, we have to take a closer look at the inner workings of the classification behavior of humans and machines.

## 5 Rank-Order Relation Between Man and Machine

In this section we investigate the classification behavior of humans using machine learning. For this we study, on a stimulus-by-stimulus basis, the correlations between the average of the subjects' responses—the subjects' classification error, the corresponding reaction time (RT), and confidence rating (CR)—for a stimulus $\vec{x}$ and the average response of the machine represented by the distance,

$$\delta(\vec{x}) = \frac{\langle \vec{w} | \vec{x} \rangle + b}{\|\vec{w}\|},$$

of this stimulus to the SH of the machine in the PCA space, the averages being computed across all 55 subjects. The metric used to compute the above distance is the common and simple Euclidean 2-norm. The distance $\delta$ reflects how the learning machine structured the face space of the subjects. To link machine learning and human classification, we make the following conjecture: the closer a stimulus is to the SH (the smaller $|\delta|$), the harder the classification should be (more errors by the subjects, longer RTs, and lower CRs). The rank-order of both the responses of humans (classification error, RT, and CR) and machines ($|\delta|$) is considered so as to avoid having to specify the precise metric of how to relate humans and machines. If this approach is successful, we can then consider the full metric information given by the responses of humans and machines (see section 6).

Since the training errors on the true and the subject data sets do not differ significantly (see the bottom row of Figure 3), we may consider only the SHs obtained using the subject data set. Moreover, only these SHs reflect what we hypothesize to be the internal face representation of the subjects. Hence, for each subject, a "personal" SH is computed using the labels $\hat{y}_i$ estimated by this subject. The distance $\delta$ between the SH and each stimulus presented to this subject is then computed for each classification algorithm. In the case of Kmean, this distance is computed using the piece of hyperplane constructed using the "prototype" of each class nearest to the considered stimulus. We then assess the correlation between the average classification behavior of humans and machines on a stimulus-by-stimulus basis. For this, we compute, for each stimulus and classifier, the relation

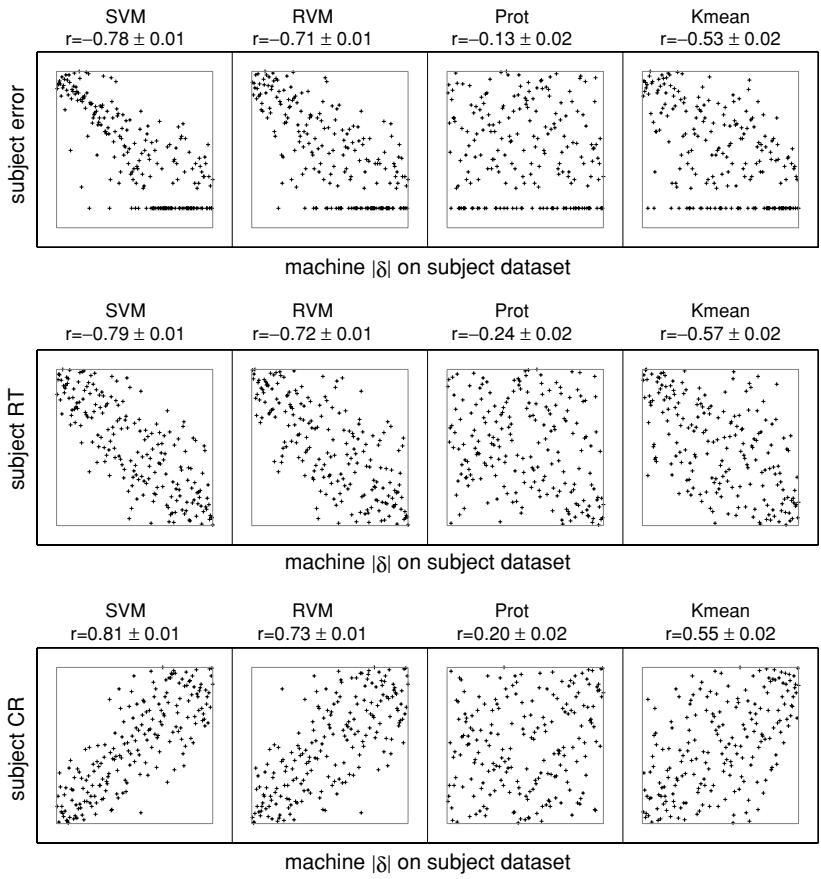


Figure 4: Rank-order analysis on a stimulus-by-stimulus basis between the responses of humans (the classification error, the corresponding RT and CR) and machines ($|\delta|$ computed on the subject data set). Both axes range from 1 to 200. In the plots of the top row, the horizontal aggregations are stimuli that have been perfectly classified by the subjects, which translates in a tied-rank analysis into a horizontal line with an offset.

between the absolute value $|\delta|$ of the average across all subjects of the distance of that stimulus to the SH and the mean response of the subjects for that stimulus. To assess this correlation, we perform a nonparametric rank-order correlation analysis using the tied-rank of the subject's response and of $|\delta|$ across the set of stimuli by computing Spearman's rank correlation coefficient $r$. The mean value of $r$ and its standard deviation are obtained using a bootstrap method by averaging over 1000 random poolings of 90% of the 200 stimuli. Figure 4 shows these rank-order correlation plots relating

humans and machines, each of the 200 scatter points representing one face stimulus. Considering the relation between the rank-order of the subjects' responses and $|\delta|$ of machine, we notice that stimuli far from the SH (high $|\delta|$) are classified more accurately (low subject error), faster (short subject RT), and with higher confidence (high subject CR) than stimuli close to the SH. These rather intuitive trends are present for all classifiers, albeit to different degrees, and illustrate that $|\delta|$ may indeed be a good measure to bridge the gap between human psychophysics and machine learning. Given a classifier, if the man-machine correlations are high for one of the subjects' responses, they can be expected to be high also for the other responses since the subjects' responses are related, as already pointed out in section 2.

These rank-order correlations allow us also to get a first hint at the algorithms humans may use to classify visual stimuli. The SVM shows the highest man-machine correlations for all responses. Moreover, it has the lowest training error on the subject data set (see the bottom row of Figure 3). In other words, the SVM can almost perfectly recreate the subjects' internal decision space, and it also gives the best man-machine correlations. The SVM is thus a good candidate to model algorithmically visual gender classification in humans. Although the RVM has a slightly higher training error on the subject data set, its good man-machine correlations make it also a good candidate for this enterprise. The prototype classifier shows the lowest man-machine correlation for all responses. Under the assumptions of this study (in particular, no nonlinear preprocessing), a mechanism akin to prototype learning seems to be a poor model of human classification behavior. A piecewise extension of Prot such as Kmean also shows low man-machine correlations and is not nearly as good as the SVM or the RVM. It is thus unlikely that humans use this type of piecewise linear decision function. However, we cannot draw any definite conclusions for Prot and Kmean since both classifiers have rather high training errors on the subject data set.

Comparing these results to those reported in Graf and Wichmann (2004), we notice that the man-machine correlations are higher in the present study. We may then conclude that using the correspondence information between the face images, although reducing the classification errors as mentioned in section 4, decreases the man-machine correlations. The latter may hint at the fact that a texture-shape correspondence representation may not be used by humans to encode visual information. It further emphasizes that classification performance per se and man-machine correlations are not equivalent measures. At first sight, the result that the correspondence information reduces the man-machine correlations may contradict the one obtained by Hancock, Bruce and Burton (1998), where it is shown that applying PCA on the texture and shape information separately increases the man-machine correlations for face recognition. The setting of the two studies is, however, different: while we focus here on gender classification,

the study by Hancock et al. (1998) mainly considers face recognition. The task performed by humans and machines is thus quite different between the two studies. While correspondence information should improve face recognition by better relating the face stimuli and removing artifacts, it may at the same time also degrade some gender-specific cues that are necessary for gender classification. Furthermore, it is difficult to compare both studies directly because of the difference in the implementation of the preprocessing stage: in the study by Graf and Wichmann (2004), PCA is applied to the concatenation of the texture and shape vectors, while in the study by Hancock et al. (1998), PCA is applied to the texture and shape vectors separately.

Our results can also be related to those of Ashby, Boynton, and Lee (1994) where the reaction time RT for the classification of low-level stimuli is shown to decrease with the distance of the stimuli to the "categorization decision bound"—the SH in this study. The RT is also shown to be independent of the distance of the stimuli to the prototypes of each class. In this study, we corroborate those findings and also extend them in two ways. First, we consider the gender estimate and the confidence rating corresponding to the reaction time and find that these responses are related. Second, we investigate different algorithms rooted in machine learning to compute this SH. Third, in the next section, we gain insights into the actual metric humans use for visual gender classification.

From the above rank-order correlation studies, we conclude that our data are orderly and that there is structure in the data that the machines can uncover. Even when removing the metric information presented in the responses of humans and machines by computing their tied-rank, some distinct trends can be seen in the data, these trends allowing us to compare humans and machines. In the next section, we proceed to a more quantitative analysis of the classification algorithms humans use for gender classification by removing the absolute value and the rank-order operations. We thus assess directly the metric of the internal decision space in humans using machine learning.

## 6 Metric Relation Between Man and Machine

The success of the above rank-order analysis suggests to us that the distance $\delta$ of the stimuli to the SH of a classifier is a meaningful measure to compare the classification behavior of humans and machines. Moreover, $\delta$ seems to capture more information about man-machine comparisons than the classification error. In this section, we proceed with a metric analysis by relating on a stimulus-by-stimulus basis the probability that a stimulus is classified as male to the distance of this stimulus to the SH for each classifier, this time, however, without taking the rank-order of both quantities and by considering $\delta$ instead of $|\delta|$.

The subjects' gender responses $\hat{y}$ are used to define the mean probability $P(\hat{y} = +1|x)$ that a stimulus $\vec{x}$ is classified as male across all 55 subjects. This probability has the characteristic of a smooth psychometric function: it is near 0 for stimuli classified predominantly as females, increases to 1/2 for stimuli where the classification is more difficult, and approaches 1 for stimuli classified mainly as males. This situation is typical for virtually all psychophysical tasks where human performance is a smooth, monotonic function of task difficulty. If any of the machines has captured more than just the input-output (classification error) mapping of the human subjects but instead captured some aspects of the human internal representation for gender classification, then the distance of a face to the SH should reflect the human classification difficulty. Thus, a regression of a monotonic function against the responses of machines $\delta$ on the $x$-axis and the responses of humans $P(\hat{y} = +1|x)$ on the $y$-axis should yield a good fit: an averaged psychometric function. We fit that subject-averaged psychometric function to the responses of humans and machines using a constrained maximum-likelihood method (Wichmann & Hill, 2001). The goodness of fit is assessed using the variance explained $\sigma_{\text{exp}}$, which compares the amount of information captured from the data by the fitted function to the amount captured by a horizontal fit through the data. A high value of $\sigma_{\text{exp}}$ indicates a good fit, whereas a low value indicates a poor fit; $\sigma_{\text{exp}}$ ranges from 1.0 (perfect fit) to 0.0 (no explanatory gain over a horizontal line, that is, no relation between the variables). We fit either a clipped linear, a Weibull, or a logistic function to the data, selecting the one for each classifier that maximizes $\sigma_{\text{exp}}$. The plots of Figure 5 relate on a stimulus-by-stimulus basis $P(\hat{y} = +1|x)$ to $\delta$, which is scaled to $[0, 1]$ and computed, as in section 5, using the subject data set.

Since we use a linear preprocessor (PCA), a linear classifier, and the Euclidean norm for the computation of $\delta$, we may expect that the linear fit would be the best type of fit. For the SVM, we indeed find that the best-fitting function is a clipped linear regression, whereas for the other classifiers, we require nonlinear sigmoidal functions. The SVM, which has a low training error on the subject data set and also exhibits the highest man-machine correlations in the rank-order analysis, provides here the best fit (highest values of $\sigma_{\text{exp}}$) and is also the only one of the studied classifiers that allows a linear interpolation between the responses of humans and machines. The SVM thus again creates the gender classification space for faces closest to that of humans. The RVM has a lower quality of fit (lower value of $\sigma_{\text{exp}}$), and the interpolation function follows a Weibull function. It seems thus less appropriate for our purpose, although it is still a possible candidate. Consistent with the previous rank-order results, the prototype classifier exhibits the least structure in the data, and consequently also the poorest goodness of fit $\sigma_{\text{exp}}$. Its piecewise linear extension Kmean shows slightly more structure, but is still far worse than the SVM or the RVM. Similarly to the previous findings of this study, it thus seems unlikely that

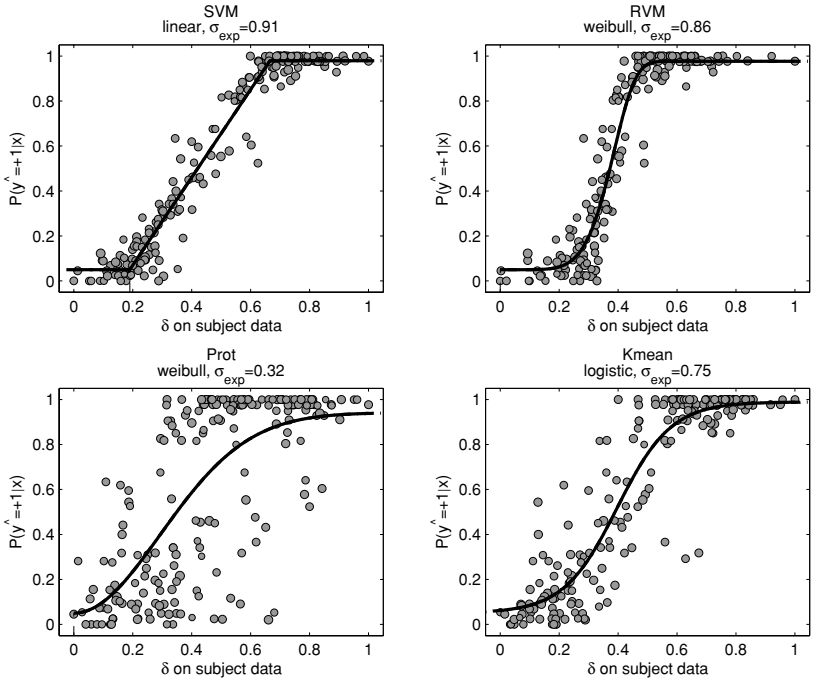


Figure 5: Metric analysis on a stimulus-by-stimulus basis between the response of humans $P(\hat{y} = +1|x)$ and machines $\delta$ (computed on the subject data set and rescaled to $[0, 1]$).

humans use algorithms based on the concept of prototype to classify the gender of faces.

## 7 Conclusions

Estimating the human internal metric representation of objects and categories is one of the central problems in cognitive psychology, and many previous investigations exist using, for instance, the geometrical relations between objects in feature spaces (Edelman, 1999) or the aftereffects induced in humans by face stimuli (Leopold, O'Toole, Vetter, & Blanz, 2001). There have also been previous attempts to study human and machine classification behavior by comparing, for instance, the generalization ability of humans and machines using the so-called other-race effect (Furl et al., 2002). In this article, we introduced a unified algorithmic approach based on machine learning techniques to gain insights into both the internal decision space of humans and the classification behavior of

humans in the context of a gender classification task of images of human faces.

Our research introduces a novel methodology and tests it by applying it to visual gender classification. Understanding the algorithms humans use in classification tasks shows what computations a more biologically realistic model should perform. We hope that our results provide further guidance for the construction of neural population models that could explain human classification behavior on a more microscopic scale (Dayan & Abbott, 2001; Gerstner & Kistler, 2002). However, before dealing with these microscopic aspects, a better knowledge of the macroscopic classification behavior is necessary. In this letter, we hope to have given a framework to study human classification of visual stimuli quantitatively.

The main aspect of our letter is that we study the subjects' internal decision space for face stimuli. First, the input-output characteristics of human and machine gender classification were compared using the classification error as a measure. Second, the classification behavior of humans and machines was related by comparing the rank-orders of the responses of humans (the subjects' classification error with the corresponding reaction and confidence rating) and machines (the distance of the stimuli to the SH of the machines). First trends in the data were obtained from these rank-order studies: stimuli far from the SH are classified more accurately, faster, and with higher confidence than stimuli closer to the SH. In other words, the distance of stimuli to a hyperplane separating both classes is demonstrated to be a useful measure to compare humans and machines. Third, we considered the full metric information contained in the responses of humans and machines and studied the subjects' internal decision space for gender classification of images of faces. From this, we concluded that combining a linear preprocessor (PCA) with a linear classifier in a Euclidean metric space gave exceedingly good fits for the SVM: the distance of a face to the SH was an almost perfect predictor of the human classification performance averaged across all our subjects. In contrast, the prototype classifier behaved in the least human-like manner. This finding supports the arguments against the concept of prototype outlined by Földiák (1998). Here we show that more sophisticated algorithms such as the SVM better capture the human internal face space, at least given our gender classification task.

Both the rank-order and the metric studies on the subjects' internal decision space for faces gave similar results: the SVM, and to some extent the RVM, are the best candidates to model the classification algorithms in humans, while the prototype classifier as well as its piecewise linear extension Kmean seem to be least adapted for this task. A classification algorithm using the center of the classes such as for the prototype classifier seems thus less adapted to model human classification behavior than a classifier maximizing the margin between the classes such as the SVM. In other words, when making decisions about the gender of faces, humans may rely more on androgynous faces that are difficult to classify (such as

the support vectors, that is, stimuli lying on or in the margin stripe) rather than on the prototypical faces that are easy to classify.

We have focused here on the classification algorithms using a single preprocessor, PCA. This allowed us to study in depth the algorithmic models for gender classification that humans use. However, the preprocessing stage cannot be ignored in a complete model of visual gender classification. While such a model would be beyond the scope of this letter, our future studies derived from Graf (2004) will include the use of other preprocessors, such as independent component analysis, nonnegative matrix factorization, or Gabor wavelet filters.

## Acknowledgments

We thank C. Wallraven, M. Giese, A. Kohn, M. Jazayeri, J. A. Movshon, and I. Bülthoff for helpful comments and suggestions. A. B. A. G. was supported by a grant from the European Union (IST 2000-29375 COGVIS). In addition, part of this work was supported by the German Research Council (DFG) grant Wi-2103 awarded to F. A. W.

## References

Ashby, F., Boynton, G., & Lee, W. (1994). Categorization response time with multidimensional stimuli. *Perception and Psychophysics*, *55*(1), 11–27.

Ashby, F., & Ell, S. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, *5*(5), 204–210.

Bartlett, M., Movellan, J., & Sejnowski, T. (2002). Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, *13*(6), 1450–1464.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Siggraph99* (pp. 187–194). New York: ACM Press.

Calder, A., Burton, A., Miller, P., Young, A., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research, 41*, 1179–1208.

Churchland, P., & Sejnowski, T. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Cottrell, G., & Metcalfe, J. (1991). EMPATH: Face, emotion, and gender recognition using holons. In D. Touretzky & R. Lippmann (Eds.), *Advances in neural information processing systems, 3* (pp. 654–571). San Mateo, CA: Morgan Kaufmann.

Dailey, M., Cottrell, G., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience, 14*(8), 1158–1173.

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience.* Cambridge, MA: MIT Press.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines*, *5*, 45–68.

Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188.

Földiák, P. (1998). What is wrong with prototypes. *Behavioral and Brain Sciences*, *21*(4), 471–472.

Freund, Y., & Schapire, R. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In P. M. B. Vitányi (Ed.), *Second European Conference on Computational Learning Theory* (pp. 23–37). New York: Springer.

Furl, N., Phillips, P., & O'Toole, A. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, *26*, 797–815.

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge: Cambridge University Press.

Golomb, B., Lawrence, D., & Sejnowski, T. (1991). SEXNET: A neural network identifies sex from human faces. In D. Touretzky & R. Lippmann (Eds.), *Advances in neural information processing systems, 3* (pp. 572–577). San Mateo, CA: Morgan Kaufmann.

Graepel, T., Herbrich, R., & Williamson, R. (2001). From margin to sparsity. In T. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems*, *13* (pp. 210–216). Cambridge, MA: MIT Press.

Graf, A. (2004). *Classification and feature extraction in man and machine*. Unpublished doctoral dissertation, Max Planck Institute for Biological Cybernetics.

Graf, A., Smola, A., & Borer, S. (2003). Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, *14*(3), 597–605.

Graf, A., & Wichmann, F. (2002). Gender classification of human faces. In H. H. Bülthoff, S.-W. Lee, T. A. Poggio, & C. Wallraven (Eds.), *Biologically Motivated Computer Vision*, LNCS 2525 (pp. 491–501). New York: Springer.

Graf. A., & Wichmann, F. (2004). Insights from machine learning applied to human visual classification. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems, 16* (pp. 905–912). Cambridge, MA: MIT Press.

Gray, M., Lawrence, D., Golomb, B., & Sejnowski, T. (1995). A perceptron reveals the face of sex. *Neural Computation, 7*(6), 1160–1164.

Hancock, P., Bruce, V., & Burton, A. (1998). A comparison of two computer-based face recognition systems with human perceptions of faces. *Vision Research, 38*, 2277–2288.

Haykin, S. (1999). *Neural networks: A comprehensive approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

LeCun, Y., Bottou, L., Orr, G., & Müller, K.-R. (1998). Efficient backprop. In G. B. Orr & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade*, LNCS 1524. New York: Springer-Verlag.

Leopold, D., O'Toole, A., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*(1), 89–94.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

Mel, B. (1997). SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation, 9*(4), 777–804.

Nosofsky, R. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 3–27.

O'Toole, A., Abdi, H., Deffenbacher, K., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A, 10*(3), 405–411.

O'Toole, A., Deffenbacher, K., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and Cognition, 26*, 146–160.

O'Toole, A., Phillips, P., Cheng, Y., Ross, B., & Wild, H. (2000). Face recognition algorithms as models of human face processing. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition.* Piscataway, NJ: IEEE Computer Society Press.

O'Toole, A., Vetter, T., & Blanz, V. (1999). Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. *Vision Research, 39*, 3145–3155.

Palmeri, T. (2001). The time course of perceptual categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization.* New York: Oxford University Press.

Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature, 428*, 419–422.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382–407.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*, 162–168.

Rolls, E., & Deco, G. (2002). *Computational neuroscience of vision.* New York: Oxford University Press.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Rosenblatt, F. (1958). The perception: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408.

Schapire, R., Freund, Y., Bartlett, P., & Lee, W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics, 26*(5), 1651–1686.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A, 4*(3), 519–524.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*, 211–214.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Valentin, D., Abdi, H., Edelman, B., & O'Toole, A. (1997). Principal component and neural network analyses of face images: What can be generalized in gender classification? *Journal of Mathematical Psychology, 41*, 398–413.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. (2000). *The Nature of statistical learning theory* (2nd ed.). New York: Springer.
Wichmann, F., & Hill, N. (2001). The psychometric function: I. Fitting, sampling and goodness-of-fit. *Perception and Psychophysics, 63*(8), 1293–1313.
Wickens, T. (2002). *Elementary signal detection theory*. New York: Oxford University Press.
Williams, C., & Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(12), 1342–1351.

---

Received December 27, 2004; accepted June 1, 2005.