Estimating predictive stimulus features from psychophysical data: The decision image technique applied to human faces

Jakob H. Macke

Max-Planck-Institut für biologische Kybernetik, Tübingen, Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Germany

Felix A. Wichmann

Technische Universität Berlin, Bernstein Center for Computational Neuroscience, Berlin, Germany

One major challenge in the sensory sciences is to identify the stimulus features on which sensory systems base their computations, and which are predictive of a behavioral decision: they are a prerequisite for computational models of perception. We describe a technique (decision images) for extracting predictive stimulus features using logistic regression. A decision image not only defines a region of interest within a stimulus but is a quantitative template which defines a direction in stimulus space. Decision images thus enable the development of predictive models, as well as the generation of optimized stimuli for subsequent psychophysical investigations. Here we describe our method and apply it to data from a human face classification experiment. We show that decision images are able to predict human responses not only in terms of overall percent correct but also in terms of the probabilities with which individual faces are (mis-) classified by individual observers. We show that the most predictive dimension for gender categorization is neither aligned with the axis defined by the two class-means, nor with the first principal component of all faces—two hypotheses frequently entertained in the literature. Our method can be applied to a wide range of binary classification tasks in vision or other psychophysical contexts.

Keywords: classification images, decision images, logistic regression, machine learning, psychophysics, face perception

Citation: Macke, J. H., & Wichmann, F. A. (2010). Estimating predictive stimulus features from psychophysical data: The decision image technique applied to human faces. *Journal of Vision, 10*(5):22, 1–24, http://journalofvision.org/content/10/5/22 doi:10.1167/10.5.22.

Introduction

One of the central goals of visual psychophysics is to identify the aspects of a physical stimulus that determine human behavior. For simple visual patterns the experimenter imposes on the observer which information, or features, she can use to solve a given perceptual task. Thus we can measure perception as a function of the available features and build models based on these features. While this approach has led to the success of many low-level computational models, it does not yield an answer to the question of which features are used by the visual system in more natural tasks. For natural vision tasks, there are typically a multitude of features in a given visual scene which the observer system could potentially exploit—corners, contours, lightness gradients or color gradients to name but a few.

In single-cell neurophysiology an analogous goal would be to characterize the receptive field of a neuron—which physical stimulus properties make the neuron spike. One very successful and general approach is to stimulate the neuron with a sequence of noisy stimuli and then correlate the input with the spike train of the neuron (Marmarelis & Marmarelis, 1978). Many variants of this *reverse correlation* approach exist, for an overview see Wu, David, and Gallant (2006). In psychophysics, noise-based reverse correlation techniques are known as *classification image* methods (Ahumada & Lovell, 1971; Beard & Ahumada, 1998; Eckstein & Ahumada, 2002; Knoblauch & Maloney, 2008; Ross & Cohen, 2009). In cases where classification images adequately predict behavior, they can also give insights into the neural mechanisms involved in a psychophysical task (Nienborg & Cumming, 2009). To stress the similarity of this feature identification approach to reverse correlation in neurophysiology, some authors also refer to this approach as *perceptive field* estimation (Neri & Levi, 2006). In addition to classification image techniques using (white) noise, a variant of reverse correlation termed the *bubbles technique* recently gained some popularity in the vision sciences (Dupuis-Roy, Fortin, Fiset, & Gosselin, 2009; Gosselin & Schyns, 2001; Murray & Gold, 2004; Schyns, Gosselin, & Smith, 2009).

These techniques have been of great importance for identifying the features used by human observers in detection and discrimination tasks. However, both techniques rely on substantial manipulations of the stimulus: In the case of classification images, the target is embedded in

doi: 10.1167/10.5.22 Received June 19, 2009; published May 18, 2010 ISSN 1534-7362

either white or correlated noise (Abbey & Eckstein, 2007), whereas in the bubbles technique, it is windowed on multiple spatial scales. This raises the question of how such noise embeddings affect the strategy used by the observers: The features returned by these methods may not be those which are predictive of human behavior during natural viewing, but only those used by human observers if forced to view an image under the conditions imposed by the techniques. Both techniques gain most information about the features used by human observers close to threshold, whereas under typical every day conditions performance is not close to threshold. In gender classification, for example, we may well make errors in real life but typically not because the face was nearly invisible ("dense fog" as for classification images), but rather because the *clearly visible* features analyzed led us astray. Finally, while the bubbles-technique localizes important regions in a stimulus, it does not yield a predictive model which can be used to predict human responses on novel stimuli.

These limitations have led us to search for feature identification techniques which can be applied during natural viewing, and are thus complementary to noise embedding methods. In particular, we will describe a technique enabling us to extract the features that are important for predicting the responses of human observers in natural classification tasks and apply it to a gender categorization experiment of human faces. Our strategy is to exploit the fact that there is substantial variability both across different faces and human responses, which makes it possible to analyze what features of a faces lead to human (mis) classifications. Therefore, our approach requires tasks with enough variability across stimuli (here faces) and experimental conditions in which performance is not completely at ceiling. The benefit of the method, however, is that stimuli are presented in natural viewing conditions.

Gender categorization of human faces provides an ideal testing ground for our method, as it has been extensively studied in the past both with psychophysical (Dupuis-Roy et al., 2009; Mangini & Biederman, 2004; Sekuler, Gaspar, Gold, & Bennett, 2004; Smith, Gosselin, & Schyns, 2004), neuro-imaging (Ng, Ciaramitaro, Anstis, Boynton, & Fine, 2006; Smith, Fries, Gosselin, Goebel, & Schyns, 2009; Smith et al., 2004) and computational approaches (Abdi, Valentin, Edelman, & O'Toole, 1995; Gray, Lawrence, Golomb, & Sejnowski, 1995; Lu, Plataniotis, & Venetsanopoulos, 2003; O'Toole, Vetter, Troje, & Bulthoff, 1997; O'Toole et al., 1998). Most computational studies have focussed on building algorithms which can be used to determine the gender of a face from its statistical properties alone, and which thus constitute candidate mechanisms for the algorithms that underlie human gender classification. One influential idea is the observation that the first principal component of (unnormalized) faces is informative about gender (Abdi et al., 1995; O'Toole, Abdi, Deffenbacher, & Valentin, 1993; O'Toole et al., 1998; Sirovich & Kirby, 1987; Turk & Pentland, 1991; Valentin, Abdi, Edelman, & O'Toole, 1997). To evaluate the algorithms described above and their plausibility as models of human face processing, their overall performance (in percentage correct) is compared against the performance of human observers (Blackwell et al., 1997; Hancock, Bruce, & Burton, 1998). We extend these studies by trying to build algorithms which not only predict the over-all percentage correct of human observers, but also their responses on a stimulus by stimulus level (Graf, Wichmann, Bulthoff, & Scholkopf, 2006). Secondly, we work with stimuli for which size, mean pixel intensity and variance of pixel intensity has been normalized, and for which we show the first principal component not to be a good predictor of gender (see Discussion).

Studies using classification images (Mangini & Biederman, 2004; Ross & Cohen, 2009; Sekuler et al., 2004), or bubbles (Dupuis-Roy et al., 2009; Smith et al., 2008; Tyler & Chen, 2006) have been employed to identify features which are used by human observers in gender categorization tasks. These studies have generally concluded that the eye-region is important for humans to determine the gender of a presented face. However, these studies did not evaluate the ability of these models to predict the gender-classification decisions of human observers on novel stimuli.

Our approach is an extension of classification image techniques which use generalized linear models (such as logistic regression) (Knoblauch & Maloney, 2008). As we use it to predict decisions in a binary classification task, we refer to the estimated filters as *decision images* (Kienzle, Franz, Scholkopf, & Wichmann, 2009; Wichmann, Graf, Simoncelli, & Schoelkopf, 2005) (Yovel, Franz, Stilz, & Schnitzler, 2008, in the auditory domain). The central idea is to train a statistical classification algorithm on the responses provided by human subjects in a psychophysical task (here: perceived gender) instead of ground truth (real gender). In this way, we will obtain decision images which resemble the internal decision space of individual human subjects, rather than create a decision space that optimally separates the physical stimulus classes, as is done when training learning machines on problems in engineering (Gray et al., 1995; Moghaddam & Yang, 2002). Clearly, if human observers perform above chance-level, the extracted decision boundaries will be correlated with the (physically) optimal decision boundaries. We perform additional analyses to show that the extracted decision models do not merely reflect the optimal features, but rather observer-specific decision boundaries.

We are not limited to predicting an observer's binary response (correct versus incorrect) but we can predict the observer's response probabilities, e.g. when indicating the gender of human faces shown repeatedly during a gender discrimination task. By using repeated stimulus

presentations, we can also quantitatively evaluate and compare decision images corresponding to different categorization models. The decision image not only defines the critical region-of-interest within a stimulus (as the bubbles technique does) but, similar to a classification image it is a quantitative template which defines a direction in stimulus space. Therefore, they can be used for the generation of optimized stimuli (Wang & Simoncelli, 2008) for subsequent psychophysical experiments, as we will show below. In contrast to classification images (which are defined relative to the noise used in the experiment), the decision images can be interpreted relative to the natural variability in the stimulus.

Methods I

Statistical techniques

Estimating informative features

Of the very many potential stimulus features a human observer may use when performing a classification task, we aim to extract those features of the stimulus which are predictive of their decisions. In general, we consider an experiment in which each of n stimuli s_i—typically an image in visual psychophysics, but equally a sound in auditory psychophysics—is presented multiple times (m presentations). At each presentation j, the observer responds with a binary decision $y_{ij} \in \{1, -1\}$, depending on whether they perceive the stimulus to be belonging to class -1 or 1—female or male in case of the experiments reported herein. We assume decisions to be stochastic, i.e. that observers respond with probability $P(Y_i = 1|s_i) = :p(s_i)$. Statistically speaking, we assume each trial of the experiment to be an independent Bernoulli trial, an assumption that is widespread in psychophysics (e.g. Wichmann & Hill, 2001a). Here we estimate the decision probabilities by using the empirical estimate $\hat{p}(s_i) = \frac{1}{m}\#\{j : y_{ij} = 1\}$, but our method can also be applied in situations where the $\hat{p}(s_i)$'s are obtained in a different manner, e.g. by explicit uncertainty estimates provided by the observers. The use of m multiple presentations allows for the quantitative assessment of the resulting model, i.e. how much of the variability in the human responses is accounted for Murray, Bennett, and Sekuler (2002).

Our key interest is to link the relative decision frequencies (referred to as "empirical probabilities" in the following) with the stimulus: Given a stimulus s_i, we want to obtain a prediction of the corresponding $p(s_i)$. This task is analogous to fitting psychometric functions (Wichmann & Hill, 2001a), with the difference that here s is high-dimensional rather than univariate. We assume that the relationship between the stimulus and the decision probability can be represented by the combination of a linear filter ω and a static nonlinearity f, i.e. by a model of the form

$$p(s_i) = f(\omega^\top s_i + \beta). \quad (1)$$

Models of this form are sometimes referred to as linear–nonlinear cascade models in computational neuroscience (Simoncelli, Paninski, & Schwartz, 2004) or Generalized Linear models in statistics (McCullagh & Nelder, 1989). Linear–nonlinear models have been extensively studied in psychophysics (Knoblauch & Maloney, 2008; Neri, 2004) and successfully applied in a variety of tasks. Clearly, there are also many situations in which this model is too constrained to provide a good model of human behavior, for example for the recognition of objects embedded in a natural scene.

In the formula above, ω is a vector of the same dimensionality as s. (If the stimulus s is an image of arbitrary dimension $K \times Z$ (as in visual psychophysics) we simply re-shape s to be a vector of dimension $KZ \times 1$.) The scalar β is a bias term which can be used to shift the decision boundary $p(s) = 0.5$ closer to one class or the other. To calculate the predicted classification probability, we take an inner product between the stimulus s_i and ω, add β, and apply a static nonlinearity f to the result. In our psychophysical application we refer to ω as the *decision image*, in neurophysiology it would be referred to as the receptive field. In psychophysics, the nonlinearity f is a psychometric function which maps the real-valued filter outputs $z_i = \omega^\top s_i + \beta$ to probabilities in the range [0,1]. In neurophysiology, f could be used to model nonlinear spike-generation mechanisms. All stimuli s which lie in the hyperplane given by the equation $\omega^\top s + \beta = z$ for some z will have the same decision probability $p(s) = f(z)$. In particular, the decision boundary $p(s) = 0.5$ is the hyperplane $\omega^\top s + \beta = f^{-1}(\frac{1}{2})$. The decision probability $p(s_i)$ is thus a function of the (signed) distance of s_i to this decision hyper-plane: Moving stimuli perpendicular to the decision boundary will increase (or decrease) the associated decision probabilities, whereas moving in parallel will keep them constant.

Our aim is to find ω—the weight vector or *decision image*—and the function f such that the predicted probabilities $p(s_i)$ are as close as possible to the empirical probabilities $\hat{p}(s_i)$, a task analogous to psychometric function fitting. As optimizing both the decision image and the nonlinearity f together would be a non-convex optimization problem, we will separate the two stages in a way that is analogous to techniques for estimating receptive fields in neurophysiology (Sharpee, Miller, & Stryker, 2008): First, we estimate the decision image ω under the assumption that the nonlinearity is the logistic function, and then optimize the nonlinearity in a second step.

Estimation of the decision image ω

Finding the direction ω along which the two classes vary most can be regarded as a classification problem, as in Wichmann et al. (2005): Linear algorithms, such as support vector machines, can be used to find the direction of maximal separation between the two sets of stimuli, here female and male faces. Unlike in Wichmann et al. (2005), however, we present each stimulus m times, and thus we have not only the observer's binary decision for each stimulus, but also an estimate of the decision probability $\hat{p}(s_i)$ over the m repetitions. Hence, we can interpret the estimation step as a regression from s_i against $\hat{p}_i$: we move from binary classification as in Wichmann et al. (2005) to probability estimation. One standard algorithm for this task is penalized logistic regression, which finds ω by minimizing the objective function

$$L = \sum_i \hat{p}(s_i)\log(1 + e^{-(\omega^\top s_i+\beta)}) + (1-\hat{p}(s_i))\log(1 + e^{\omega^\top s_i+\beta}) + \eta\|\omega\|^2. \quad (2)$$

This is equivalent to maximizing the regularized log-likelihood under the assumption that the nonlinearity f is the logistic function $l(t) = (1 + \exp(-t))^{-1}$, where t is the signed distance to the separating hyperplane. Although we do not restrict f to be of this form, psychometric functions can often be well approximated by such a shape. Importantly, the cost function of logistic regression is convex. Therefore, the optimization problem has a unique solution, a single, global optimum which can be found efficiently. The loss function in Equation 2 is linear in the probabilities ($\hat{p}$). Thus, for logistic regression, the two views of *classification with repeated stimuli* and *regression onto probabilities* are mathematically equivalent.

The regularization term $\eta\|\omega\|^2$ is used to avoid overfitting, which is important in high-dimensional regression problems (Hofmann, Schölkopf, & Smola, 2008). In a Bayesian setting, maximization of the penalized log-likelihood can be interpreted as finding the maximum-a-posterior estimate (MAP) for ω: In this case, the regularization term is the log of the prior distribution over parameters ω, i. e. a prior of the form $\exp(-\|\omega\|^2\eta)$ is assumed. In any case, choice of the regularizer should be guided by prior knowledge about the problem. Often, however, explicit domain knowledge is unavailable, or unreliable, and then the regularization term is used to enforce smoothness (Knoblauch & Maloney, 2008) or sparseness (Gerwinn, Macke, Seeger, & Bethge, 2008; Mineault, Barthelme, & Pack, 2009) of the filters ω. Note that in this case the regularization term acts very much like the penalty terms in model selection (Jäkel, Schölkopf, & Wichmann, 2007). Applying our decision image estimation technique to gender classification of faces, we simply chose the standard regularization $\|\omega\|^2 = \sum_i \omega_i^2$. It is possible that more informed choices of the regularization term may lead to even better results (Knoblauch & Maloney, 2008). The regularization parameter η which determines the trade-off the goodness-of-fit of the data with the penalty for overly complex models was chosen by ten-fold cross-validation. This means that we repeatedly fitted the model on 90% of the data, and took those parameters which, on average, lead to the best performance on the remaining 10% of the data.

It can be shown that the vector ω which minimizes the function in Equation 2 (where the summation is over the entire training set) will always be in the subspace spanned by the training stimuli s_i, the so-called representer theorem (Hofmann et al., 2008). Thus, rather than working in the full space of $K \times Z$-dimensional stimuli, we can perform an (orthogonal) change of variables and work in a n dimensional space (if there are n stimuli): We apply a PCA to the full set of stimuli, but keep all n principal components, and thus still work in the space spanned by all n stimuli: We use PCA to increase computation efficiency by dimensionality reduction ($n = 428$; $K \times Z = 65536$), but not for it to have any influence on the performance of the algorithms (Abdi et al., 1995). (An orthogonal change of variables does not affect any linear algorithm, as it leaves the inner products between any two vectors unchanged.)

If the stimuli s_i are high-dimensional compared to the number of datapoints ($K \times Z \gg n$) it is always possible to perfectly fit the model to the data. Consequently, mere goodness of fit can not be used as a performance measure. Instead, we directly tested how well the model can *predict* data the model was not trained on. Therefore, to obtain a realistic estimate of the generalization performance of the model, we use cross-validation: The model is fitted to a subset of the data, and then the performance is evaluated on a different and disjoint subset. We used leave-one-out estimates: To predict the output of the model to the stimulus s_i, which we call z_i, we trained it on all other stimuli *but* s_i. All values of z_i in this article, including those in the figures, are cross-validated predictions. Thus, we have two levels of cross-validation: We use 10-fold cross validation to *find* the regularization parameter η, and we use leave-one-out estimates to *report* predictions (and thus evaluate how well our decision images really predict human behavior) (Bishop, 2006).

Fitting the nonlinearity f

Above, we estimated ω under the assumption that the nonlinearity is the logistic function $l(s)$. In practice, however, it often leads to superior performance to readjust the nonlinearity after identification of the filters: Having estimated ω, we can calculate the scalar products $z_i = \omega^\top x_i$ and determine f by fitting a function to the set of points $(z_1, \hat{p}_1)\ldots(z_n, \hat{p}_n)$. Assuming each subject's response

to be the outcome of a Bernoulli trial, we find f by maximizing the log-likelihood

$$l(f,\hat{p},z) = \sum_{i=1}^{n} \hat{p}_i \log(f(z_i)) + (1-\hat{p}_i)\log(1-f(z_i)), \tag{3}$$

(see Wichmann & Hill, 2001a for details).

To each data set, we fitted each of three functions with two free parameter (a thresholded linear, a Weibull (asymmetric) and a logistic (symmetric) function to cover the most often required shapes of the psychometric function). For further analysis we picked the functional form that yielded the best fit. Almost always the data were best fit by the logistic function. Human observers occasionally "lapse" independently of stimulus magnitude, for example because of eye-blinks, attentional fluctuations, or finger-errors. In the current settings this translates into a nonzero probability that an observer will indicate to have seen a female (or male) face even if the opposite gender should have been "crystal clear". Such lapses potentially lead to serious parameter misestimates during maximum-likelihood fitting. This can be overcome if the *asymptotic error probabilities* are fitted simultaneously (Wichmann & Hill, 2001a). Therefore, the fit of the pointwise nonlinearity involves a total of 4 free parameters (two for the function proper, two for the asymptotes), as well as selection of the best fitting function type. Simultaneous fitting of both the filter ω and the asymptotes did not lead to any gain in prediction performance, most likely because such an optimization is non-convex: the theoretical advantage of estimating the decision image ω and the asymptotes directly together is more than offset by the disadvantage of non-convex optimization.[1]

Comparison with alternative models

There are alternative models attempting to predict the decisions of human subjects. For example, one could assume a simple classification model in which each stimulus is assigned to the class whose mean is closer. We refer to this model as the *prototype* classifier, and this is a very popular model for face perception (Burton, Jenkins, Hancock, & White, 2005; Leopold, Bondar, & Giese, 2006; Loffler, Yourganov, Wilkinson, & Wilson, 2005). The prototype classifier can be cast in our linear–nonlinear cascade framework. Its filter ω is simply given by $\omega_{prot} = \frac{1}{nm}\sum_{ij} y_{ij} s_i$, i.e. the difference of the two class means.

The simple prototype classifier ignores the fact that the different dimensions of the stimuli (i. e., the pixel-intensities) are correlated. However, one can correct for these correlations by pre-multiplying the filter ω by the inverse of the within-class covariance C_{within}, yielding the filter $\omega_{lda} = {C_{within}}^{-1}\omega_{prot}$. This algorithm (Lu et al., 2003) is known as (Fisher's) linear discriminant analysis in statistics (LDA in the following). In psychology, it has been noted several times that it is important to take the variances (and co-variances) into account when learning a category (Fried & Holyoak, 1984; Reed, 1972). This effectively results in categorization models which are related to LDA. LDA can be shown to be an optimal classifier, provided that the stimuli in each of the two classes are Gaussian, have the same within-class covariance, and the means and covariances are known exactly. In practice, however, we do not know the exact means and the covariance, and we have to estimate both from the data. Misestimation of the covariance could result in worse classification performance then the simple prototype classifier. On small data sets it is thus advisable to use a regularized covariance estimate $C_{reg} = C_{within} + \eta\mathbf{1}$, where $\mathbf{1}$ is an identity-matrix. We optimize the regularization η by cross-validation, as we did for logistic regression.

Finally, a decision boundary ω can be found by minimizing a different convex cost function in place of the logistic-regression cost function from Equation 2. For example, the support vector machine (SVM) algorithm finds a decision boundary by minimizing the so-called hinge-loss (Hofmann et al., 2008). The SVM is a general purpose machine learning algorithm which has successfully been applied in a variety of domains, and has also been used previously for the estimation of decision images (Wichmann et al., 2005). We modified the original SVM-algorithm (which performs binary classification) in order to also take into account the estimated decision probabilities $\hat{p}(s_i)$ (see Appendix A.1 for derivation and discussion of our modified algorithm).

For each of these models, we first obtained the corresponding filter ω, and then fitted the nonlinearity to the residuals as described above.

Stimulus generation

We presented stimuli in four experimental conditions, which differed in orientation, illumination and texture. Figure 1 shows four example stimuli (two female and two male faces). In condition 1, faces were presented in frontal view, illuminated predominantly by diffuse, ambient light and an additional point source (spotlight) at 65 degrees azimuth and 25 degree eccentricity to create a "natural looking" illumination. Faces were rendered with the same texture obtained from a neutral mean face to eliminate texture cues to gender, e.g. facial hair (Vetter & Troje, 1997). Condition 2 was identical to condition 1, except that faces were shown slightly from the side, at an angle of 15 degrees. Condition 3 differed from condition 2 in that faces were shown with their original texture, i.e. each face had a different texture. For some men with strong

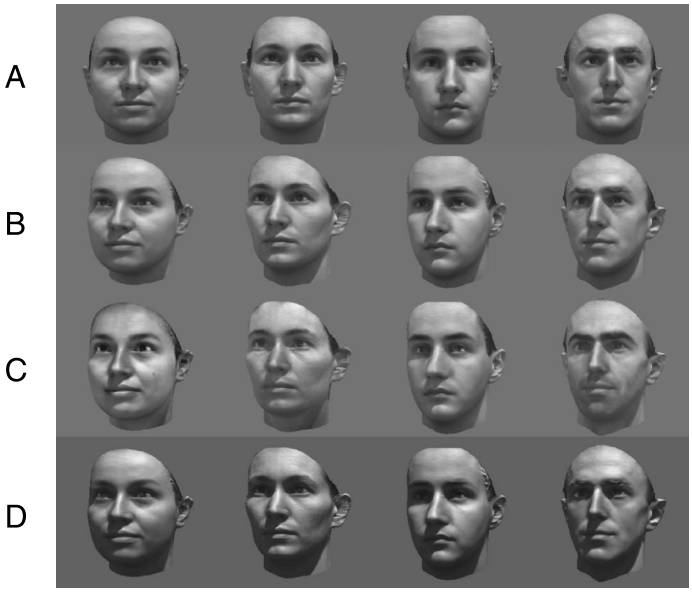

Figure 1. Examples of faces used in the experiments; the two leftmost columns show the same two female, the two rightmost columns the same two male faces. Observers were asked to indicate the gender of a presented face. Conditions differed in orientation, illumination and texture. Top row, condition 1: faces were presented in frontal view, illuminated predominantly by diffuse, ambient light and an additional point source (spotlight) at 65 degrees azimuth and 25 degree eccentricity to create a "natural looking" illumination; faces were rendered with the same texture obtained from a neutral mean face to eliminate texture cues to gender. Second row, condition 2: same as condition 1 except that faces were shown slightly from the side, at an angle of 15 degrees. Third row, condition 3: same as condition 2 except that faces were shown with their original texture, i.e. each face had a different texture. For some men with strong beard growth, like the gentleman in the rightmost column, this meant that there was a slightly darker region around the mouth—at least from an introspective point of view a reasonable cue to gender. Bottom row, condition 4: same as condition 2 (mean facial texture for all faces) except that the illumination had no diffuse, ambient component resulting in a sharp spotlight creating deep shadows.

beard growth this meant that there was a slightly darker region around the mouth. In condition 4, we used the same view and texture as condition 2, but a different illumination: The light source had no diffuse ambient component, resulting in a sharp spotlight creating deep shadows.

For each condition 428 images of faces (256 × 256-pixel resolution; 215 females and 213 males) were generated using the MPI-face database.[2] 173 of these 428 were obtained by morphing pairs of randomly selected faces of the same gender using methods described in Blanz and Vetter (1999). The morphs were linear in texture-coordinates, but not in pixel-space.

All faces were normalized to have the same mean and standard deviation of pixel intensity (Russell, 2009), the same area (i.e. no size cue), and the faces were aligned such that the cross-correlation of each face to a mean face of the database was maximal.

Stimuli were presented using a carefully linearized Siemens SMM21106LS gray-scale monitor driven by a Cambridge Research Systems Visage display controller (spatial resolution 1024 × 768 at a refresh rate of 130 Hz non-interlaced). Neither male nor female faces changed the mean luminance of the display. Subjects viewed the screen binocularly with their head stabilized by a headrest.

In the first set of experiments, the mean luminance of the screen was 213 cd/m^2 for conditions 1 to 3 and 181 cd/m^2 for condition 4. The temporal envelope of stimulus presentation was a modified Hanning window (a raised cosine function with rise and fall times of 200 ms and a plateau time of 800 ms). Faces were presented in random order and subjects were instructed to respond quickly; the median response time was 647 ms with a standard deviation of 190 ms after stimulus onset, i.e. on the vast majority of trials subjects responded whilst the stimuli were still on screen. No feedback as to whether the response was correct was provided. At the viewing distance of 60 cm the stimuli (nominally) subtended 9.5 degrees of visual angle. Seven observers with normal or corrected-to-normal vision who were naive to the purpose of the experiment acted as experimental subjects; they were paid for their participation. Each subject categorized each of the 428 faces per condition 10 times, for a total of 4280 trials per subject and condition, i.e. 17 120 trials per subject in total and a grand total of 119 840 trials reported in this study. Stimuli were presented in blocks of (about) 500 stimuli from each condition, and the order of blocks and stimuli within each block was randomized.

Results I

Performance of observers in the gender discrimination task

In each condition, each of the 428 stimuli was presented 10 times to the observers in random order. Subjects indicated the perceived gender by pressing one of two buttons. Overall, observers performed accurately, but not at ceiling, with an average percentage correct (pc) of 85.0%, (*SEM* 1.80%).

There was a bias towards responding "male," which was indicated 59.0% of the time (*SEM* 2.9%), on a balanced stimulus set. Consistent with previous studies (Troje & Bulthoff, 1996, and references therein), observers were slightly better at discriminating the faces which were shown slightly from the side (condition 2, pc = 84%, *SEM* 2.0% than when presented in frontal view (condition 1, pc = 81%, *SEM* 2.3%). Each of the observers performed better on condition 2 than on condition 1 (binomial test, 7

out of 7, $p < 0.01$). The stimuli in conditions 2 and 3 were identical except for the fact that the faces in condition 3 were rendered with their original texture, whereas an uninformative texture was used in condition 2. Not surprisingly, the texture-cue helped gender-discrimination performance, with the average up to pc = 90%, *SEM* 1.2%, in this condition, and all observers performing better in this condition than in condition 2 (7 out of 7, $p < 0.01$). Condition 4 consisted of faces which where presented slightly from the side, had a neutral texture, and in which the shadows caused by the spotlight illumination removed some potential cues. This condition proved to be of intermediate difficulty, with an average performance level between those of conditions 2 and 3 (pc = 85%, *SEM* 1.7%).

The errors made by the observers in determining the gender of a face cannot (solely) be attributed to stimulus-independent lapses, but depended on the stimuli in a systematic fashion: First, within each observer, the probability of an error depended strongly on the stimulus. On both female and male faces separately, the errors per stimulus were inconsistent with the assumption that occurrence-probability of an error was independent of stimulus identity. The mutual information between "stimulus" and "error" was 0.14 bits (average across observers and gender of stimulus, no bias correction, *SEM* 0.007). The hypothesis of independence could thus be rejected (G-test (Kullback, 1997), $p < 0.01$ for each observer). Second, observers had a tendency to make errors on the same stimuli, indicating that some faces were more difficult to categorize than others. We correlated the response probabilities between observers for female and male faces separately. The mean rank-correlation between the error-probabilities of any two observers was 0.27 for the female, and 0.57 for the male faces. The correlations between any two observers were significant ($p < 0.05$) on the male faces for each pair of observers in each condition, and on 95% of pairs for the female faces (158/168 pairs). Thus, it is clear that some faces are more difficult to categorize than others, and that there is a significant amount of agreement amongst observers about which of the stimuli are difficult. However, we will show below that there are also significant differences between observers, and that the decision image technique's sensitivity is high enough to identify those differences.

Modeling decision probabilities of a single observer

To capture the relationship between the face stimuli s and decision probabilities $p(s)$, we fitted a (penalized) logistic regression to the psychophysical data of each observer and condition. We first describe the data of observer CGF for condition 3 fitted with the logistic-regression model (Logreg), and compare it to the simple prototype model (Prot).

Observer CGF chose the correct gender with probability 0.88. Our decision model fitted with logistic regression chose the correct gender with probability 0.87, whereas the prototype classifier chose the correct gender only with probability 0.75. For a model to be a good predictor of the decisions of the observer, we would, at the very least, expect that they can also predict *which* of the stimuli the observer perceived as female. Logreg predicted the class (i.e. gender) chosen by the observer in 93% of cases. Prot, on the other hand, was correct only 83.0% of the time. Figure 2A depicts the psychometric function of observer CGF for condition 3 predicted by the logistic regression model. The empirically estimated gender assignment probability ($\hat{p}(s)$) for each face (y-axis) is plotted against the distance of that face from the decision boundary (filled yellow circles; $B = 428$)). The black line is the decision probability $p(s)$ predicted by the model. To help judge the quality of the fit—given that N is large and that many data-points overlap—the gray histograms show the distribution of classification responses. The top histogram consists of all the stimuli which are predominately classified by the observer as female $\hat{p}(♀) \geq 0.5$), the bottom histogram those classified as male. Stimuli which fall outside two standard deviations of the model-prediction are colored in red.

Clearly, for any model with good predictive power the data-points (filled yellow circles) should be scattered around the model prediction. For a perfect model fit the only source of deviation between model prediction and data is the (unavoidable) binomial noise resulting from the Bernoulli decision process. For a perfect model one requirement is that only 5% of the data-points should fall outside two standard deviations from the model prediction (gray lines). For the logistic regression model $Q = 12.0\%$ of points fall outside two standard deviations—not perfect, but arguably reasonable (88% versus 95% fall within two standard deviations). In addition, we quantify for each face its "deviation" from the model-prediction by the deviance residual (Wichmann & Hill, 2001a). For a perfect model the average deviance $\bar{D}$—defined as the (total) deviance divided by the number of data-points—is expected to be, on average, 1.0 (Wichmann & Hill, 2001a, 2001b). In practice in low-level psychophysical tasks average deviance values around 1.5 still are regarded as a "good" fit of the model (e.g. Goris, Wagemans, & Wichmann, 2008; Wichmann, 1999). For the logistic regression model shown in Figure 2 the average deviance equaled 1.81—again not perfect, but reasonably close given the simplicity of our model: a linear filter, the decision image, followed by a static nonlinearity (the fitted psychometric function).

To see whether the inflated average deviance $\bar{D}$ arises from systematic errors in the logistic regression model, we performed a Monte Carlo simulation of a "synthetic"

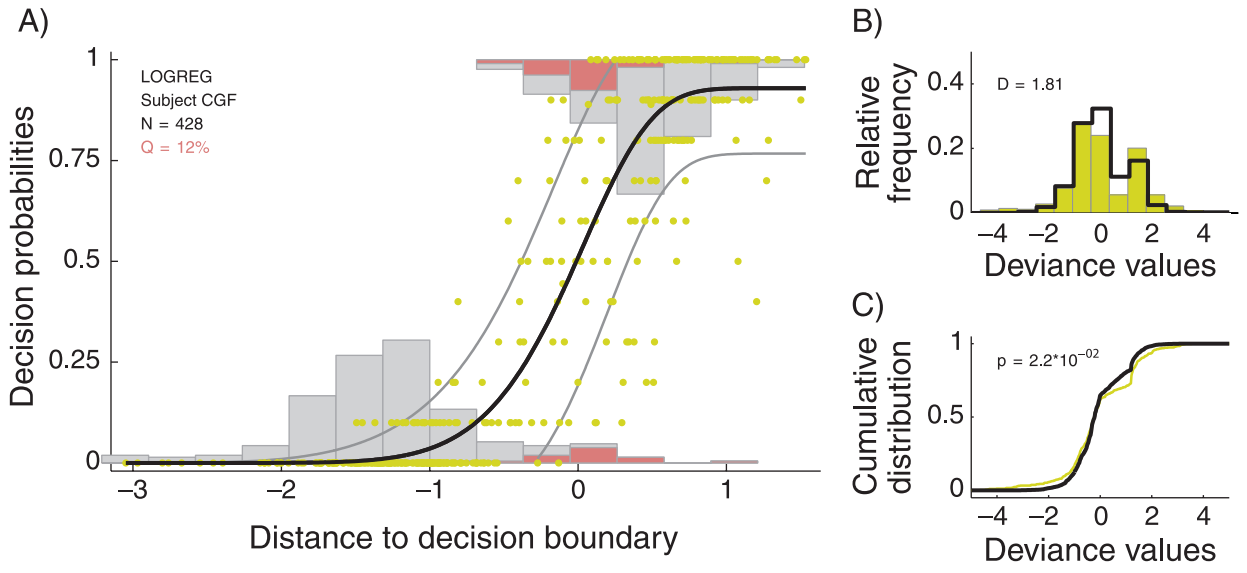

Figure 2. Data for observer CGF in condition 3 using logistic regression to estimate the decision image. A) Gender assignment probability for a face on the y-axis is plotted against the distance of that face from the decision boundary in the decision image model (filled yellow circles; N = 428). For a model with good predictive power the data-points (filled yellow circles) should be scattered around the model prediction (black line) with deviations consistent with a binomial decision process; for a perfectly correct model 95% of the data-points should be within two standard deviations from the model prediction (gray lines). To help judge the quality of the fit—given that N is large and many data-points fall on top another—the gray histograms show the distribution of classification responses (top histogram: $\hat{p}(♀) \geq 0.5$); bottom histogram: ($\hat{p}(♂) > 0.5$). Stimuli falling outside the 95%-confidence intervals of the model-prediction are colored in red in the histograms. (For details on Q see text.) B) Histogram of deviance residuals (yellow): The deviance of each data-point is the signed distance from the model prediction rescaled by its (binomial) standard deviation; for comparison the histogram of expected deviance residuals for an exactly correct model is shown (black line). (For details on D see text.) C) Cumulative distribution of deviances (yellow line): Cumulative distribution of deviance residuals rather than the histogram. Again, for comparison the cumulative distribution of expected deviance residuals for an exactly correct model is shown (black line). (p is the p-value of a KS-test on the two distributions, for details see text.)

observer which is described exactly by the logistic regression model. We performed 500 such experiments, and compared the simulated deviance residuals to those observed experimentally. Figure 2B shows the histogram of deviance residuals for observer CGF and condition one as well as simulated deviance residuals for the logistic-regression model. The close similarity of the histograms indicates that the fitted regression model and empirically obtained data from observer CGF match well. To quantify the similarity, in Figure 2C we plot the cumulative distribution of empirical and simulated deviance residuals, and perform a Kolmogorov–Smirnov test for the difference of cumulative distributions. For each of the 500 simulated experiments, we calculated the p-value for rejecting the null-hypothesis of no difference. In this case, the median p-value was 0.02. Thus, in most cases, the null-hypothesis of no difference could be rejected at a level of 5%, but not at 1%. Again, this indicates that the logistic regression decision image accounts for a large part, but not all of CGF's gender discrimination probabilities.

For comparison, Figure 3 shows the same psychophysical data, but using the prototype classifier as predictive model. In this case, the deviations from the model prediction are much larger, the average deviance was 4.92. This indicates the clear failure of the simple Prot model for which Q = 36.0% of the data points are more than two standard deviations away from the model prediction. Figures 3B and 3C show the histogram and cumulative distributions of residuals for the prototype model. The differences between the simulated and real data are evident, and the median p-value of the Kolmogorov–Smirnov test was $p = 6.1 \times 10^{-10}$, and less than 0.01 for each simulation. Thus, while the overall percentage correct of this model was seemingly high—the Prot model predicts the observer's gender assignment 83.0% of the time—evaluating its model performance on a stimulus-by-stimulus basis shows that the prototype classifier predicts the *decision probabilities* of the observer only poorly.

The discrepancies on a stimulus by stimulus level clearly underscore the importance of analyzing a model in detail: predicting the overall categorization performance does *not* imply that the model is a good description of the underlying psychophysical process. To further test and quantify the difference between the different ways to obtain the decision image, i.e. logistic regression versus mean-of-class (prototype), we calculated rank-correlations between the measured responses $\hat{p}(s_i)$ and the distances to the decision boundary in the model, $z(s_i)$—this statistic is

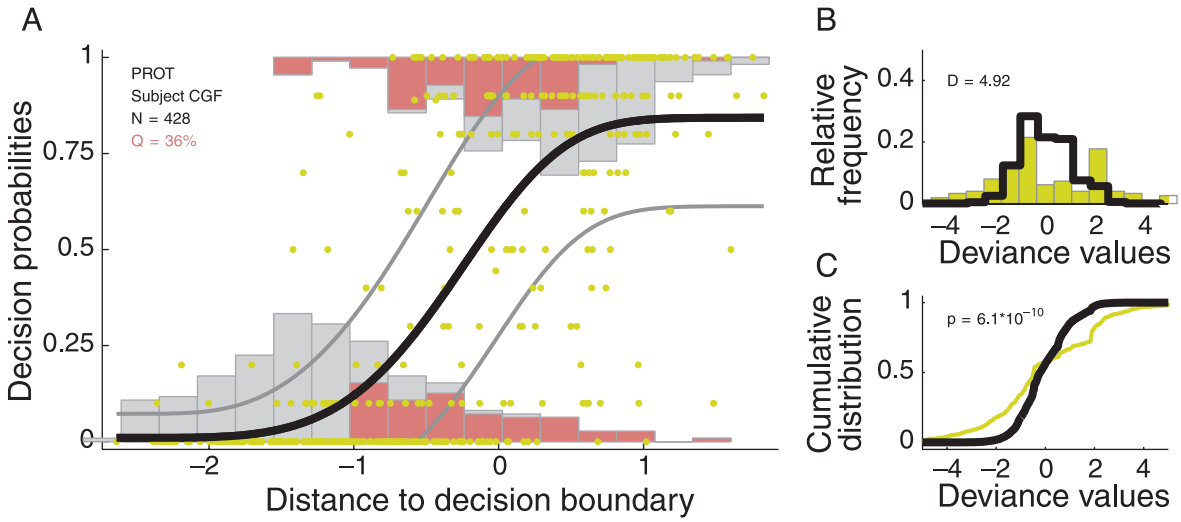

Figure 3. Data for observer CGF in condition 3 re-plotted from Figure 2 but using the prototype classifier as predictive model.

independent of the fitted (monotonic) nonlinearity. Due to the binomial scatter, the median expected rank correlation even of a perfect model would not be 1.0, but 0.915. (We obtained 0.915 from Monte Carlo simulations of a synthetic observer which is perfectly described by the model, see above.) The rank correlation for Logreg was 0.88. In stark contrast, the rank-correlation for Prot was only 0.74 (simulated: 0.92).

Conditional correlations: Predictions beyond class-structure

Some of the correlation between our model and the experimental data is, however, a consequence of the fact that the stimuli fall into distinct classes. If a classification algorithm and an observer classify most of the stimuli correctly, their decisions will be correlated even if the algorithm was not fit to the psychophysical data of the human observer at all and may use entirely different features to perform the task. Thus, we need to perform additional analyses to determine whether the model is capturing the decisions of the observers, and not simply the class-structure of the stimulus.

Thus we need to determine whether the algorithm can predict the *perceived* gender of each face (by each observer), and not just the *true* gender of each face. This can be achieved by calculating correlations conditioned on the stimulus class, here gender. Correlations between human responses and the model that we find using only stimuli in one class indicate that the model is indeed capturing the observers decisions, and not just re-discovering the class-structure of the stimulus set. For a perfectly correct observer—i.e. an observer for which we know the decision features *exactly*—the median correlations were 0.75 on the female faces and 0.58 on the male faces. The overall conditional correlation, taken as the mean of the two correlations, was 0.68. For the Logreg model the rank correlations between decisions and model predictions were 0.56 on the female faces, and 0.45 on the male faces, conditional correlation thus 0.51 ($p < 10^{-10}$ in both cases). In contrast, the conditional correlations for Prot was only 0.24 (female: 0.27, male: 0.22). This provides further evidence that the prototype model is inadequate for modeling the observer's decisions.

Summary statistics across observers

So far, we performed a detailed comparison of the performance of the logistic-regression model with the simple prototype classifier for a single subject, and showed that the prototype classifier is unable to predict his responses. In contrast, the logistic regression model is a good predictor, and is statistically similar to a perfectly correct model. In the following, we show that these results also hold for the other observers and conditions. We report a summary of performance statistics in Table 1).

Across observers and conditions, the average deviance of the logistic regression model was 1.94 (*SEM* 0.0578, median 1.8978), and was smaller than 2.39 for each subject and condition. For single subjects, the values (average across conditions) ranged from 1.80 to 2.2. Performance was slightly better on conditions 3 and 1 (1.85 and 1.87) than on 2 and 4 (1.99, 2.02). For the prototype classifier, the mean average deviance was 4.38 (*SEM* 0.30, median 3.99.), and the average deviance was larger than 3.25 in each experiment (see also Figure 6, top left).

The rank-correlation between predictions and actual responses was 0.86 (*SEM* 0.0035, median 0.86) for the logistic regression model, which was very close to the corresponding value for the simulated observers, which had an average correlation of 0.91. For the prototype

classifier, the correlation was 0.72 on average (simulated: 0.90). Additionally, we want to evaluate the ability of the classifiers to predict the observers decisions (and not merely the class-structure of the stimulus) by calculating conditional correlations: The average conditional correlation for the logistic regression model was 0.46 (*SEM* 0.021, median 0.46, simulated: 0.65), whereas for the prototype classifier it was only 0.20 (*SEM* 0.0064, median 0.20, simulated 0.70). Thus, while the logistic regression model gets reasonably close to "optimal" (as estimated from the simulated data assuming an exact model fit), this is not the case for the prototype classifier. Figure 6 (first column) shows these comparisons for the two models and for each subject and condition.

As specified above, 173 stimuli were created by morphing two faces of the same gender, with methods as described in Blanz and Vetter (1999). We also analyzed prediction performance of the decision image models separately on the 255 genuine faces. On this stimulus set, the overall prediction performance was reduced slightly: The average correlation between responses and prediction dropped from 0.86 to 0.85, the class conditional correlation stayed at 0.46, and the average deviance was increased to 2.23.

We also analyzed model performance by pooling the data across conditions and observers: For each stimulus s_i, we estimated the average probability of it being classified as female $\hat{p}_{pool}(s_i)$, where the average is now across all experiments. Similarly, we calculate the predicted probability $p_{pool}(s_i)$ by averaging the predictions on different experiments. Figures 4A and 5A show scatter plots of the decision probabilities plotted against the model predictions. For the logistic regression model (Figure 4), the decision image model predictions and those of the observers are highly correlated $c = 0.94$, whereas for the prototype classifier (Figure 5), the correlation is 0.80. The conditional correlation is 0.76 for Logreg (higher than for the typical single subject, as the noise can average out across observers), but only 0.30 for Prot. Thus, while the performance of the prototype classifier can be attributed mainly to its ability to reproduce the class-structure of the stimulus, the logistic regression model is highly correlated with the decisions of the human observers even within each class.

Finally, we calculated pooled deviance residuals for every stimulus by averaging residuals across experiments. Figures 4C and 5C compare the histograms of the actual pooled residuals against those from simulated observers: For the logistic regression model, the histograms are highly overlapping, whereas for the prototype classifier, many residuals are larger than they should be for a model fitting the data well.

Comparison with other models: Fisher's linear discriminant and SVM

Next, we will compare the logistic-regression model and the simple prototype classifier with alternative models, namely Fisher's linear discriminant (LDA) and linear support vector machines (SVM). The performance of LDA is of interest, as it can be interpreted as a more

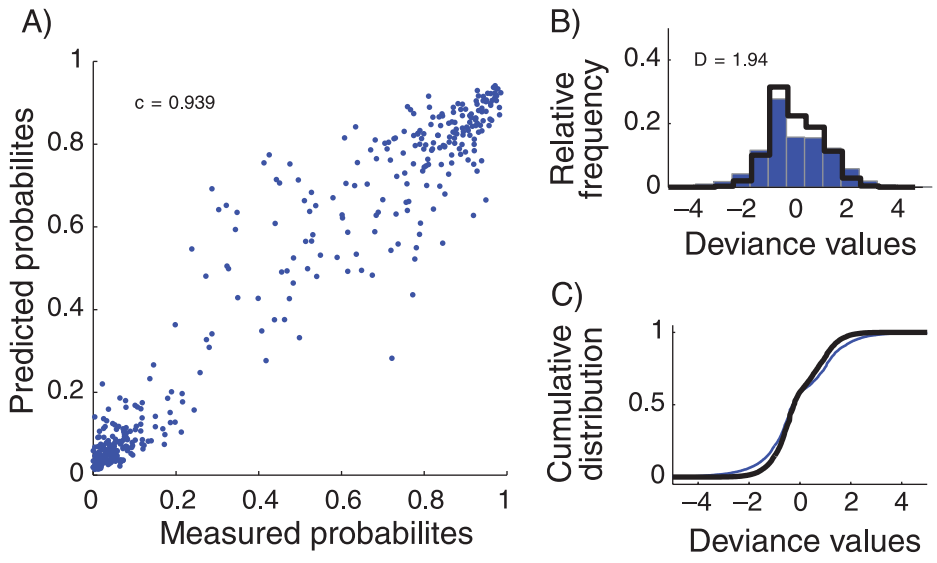

Figure 4. Summary plot for the logistic regression decision image model across all 7 subjects and 4 conditions for each of the $N = 428$ face stimuli. (Each of the 428 data-points is based on $10 \times 7 \times 4 = 280$ observations.) A) The (average) male-decision probability p(♂) on the x-axis is plotted against model predictions on the y-axis. Decision image model predictions and those of the observers are highly correlated, $c = 0.97$. B) Histograms of expected (black) and observed (blue) deviance residuals as in Figure 2B but averaged across all 7 subjects and 4 conditions. C) Cumulative distribution of expected and observed deviance residuals as in Figure 2C but averaged across all 7 subjects and 4 conditions.

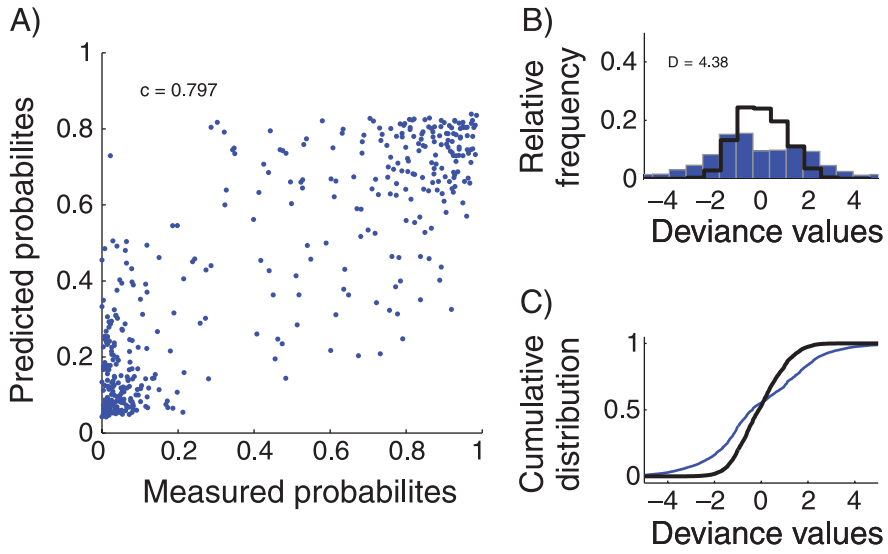

Figure 5. Summary plot with data re-plotted from Figure 4 but using the prototype classifier as predictive model.

principled variant of the prototype learner which also takes into account the covariance structure of the stimulus (Ashby & Gott, 1988; Fried & Holyoak, 1984; Reed, 1972). Support vector machines have been used in a variety of technical applications, as well as for the estimation of decision images in previous studies (Graf et al., 2006). We used a modified version of the SVM algorithm which also take the decision probabilities into account (see Appendix A). Figure 6 compares the performance statistics of these models with our decision image model across conditions and observers. Summary statistics can be found in Table 1.

The decision image model estimated with LDA was much better than the prototype classifier, and had an average deviance of 2.24 (*SEM* 0.11, median 2.18). Thus, taking into account the covariance structure of each class can considerably improve the ability of means-of-class classifier to model our psychophysical data. However, LDA was markedly worse than the logistic regression model: For each of the 28 conditions (7 observers, 4 stimulus types), the average deviance across conditions was higher than that of the logistic regression model ($p < 10^{-9}$). In addition, the mean deviance was also significantly higher (paired-sample t-test, $p = 0.0003$). Qualitatively, the same also holds true for our other performance measures, as can be seen from Figure 6, second column.

Finally, using the SVM resulted in an average deviance of 2.38 (*SEM* 0.12, median 2.29), and was outperformed by both the logistic regression (see Figure 6, right column) and the LDA model: It had a lower average deviance than LDA in only 3 out of 28 experiments ($p = 1.4 \cdot 10^{-5}$). In addition, both its rank correlation and class-conditional correlation with the observer's decisions were lower in 27 out of 28 comparisons ($p < 10^{-6}$). Thus, while the LDA and the SVM performed much better than the simple prototype classifier, they did not reach the same level of accuracy as the decision image model estimated by logistic regression. This was also the case when we excluded the morph faces. The logistic regression model performed better than linear discriminant analysis (26 out of 28 conditions and observers), which in turn performed better than the SVM (28 out of 28) and the prototype classifier (average deviance 4.87). For modeling gender categorization of human faces logistic regression thus appears to be the method of choice to estimate the decision image (at least superior to LDA, SVM and the simple prototype).

Analyzing predictions across observers

In a reasonably easy task, there is necessarily a lot of agreement between observers in which stimuli are perceived as male or female. As a result, the average rank correlation between the responses of any two observers is high, and we found it was 0.82 for our observers. Thus, the responses of any one observer can be used as a predictor for predicting the responses of any other observer. We investigated whether our decision models were better predictors than these "cross-observer" predictions. If so, this implies that the fitted logistic regression model is sensitive to an individual's idiosyncratic features used to assign gender (Peterson & Eckstein, 2009), i.e. that the individual differences between the decision images are not merely "noise" but contain crucial inter-observer differences.

The logistic regression models were better at predicting responses (as measured by rank-correlation) than another human observer in 98% out of $4 \times 6 \times 7 = 168$ comparisons (95% when we excluded the morphed faces), and in 82% if only considering percent correct. For each of the 28 experiments, the model-prediction was better

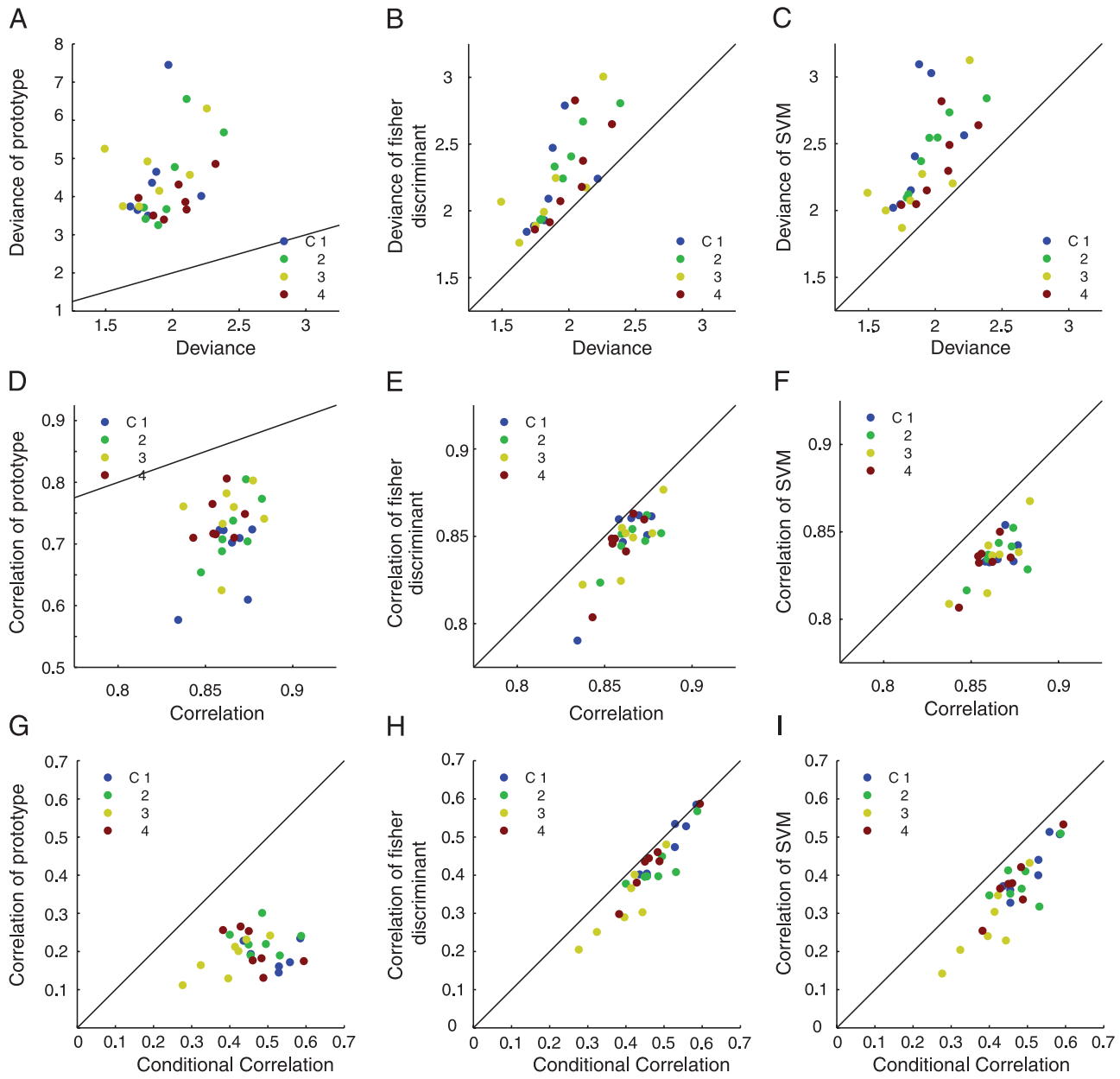

Figure 6. Comparing the performance of the decision model estimated using logistic regression with other estimation methods; the diagonal lines in each of the subplots mark the equal performance contour. Top row: Comparison of the average deviance of the logistic regression model with those of the prototype model, the Fisher discriminant model, and the support vector machine (from left to right). For each of the four conditions and seven subjects, the logistic regression model outperforms the three alternative models. Second row: Scatter plots of the rank-correlations of the model-predictions with the subjects' predictions. Bottom row: Scatter plots of the partial rank-correlation between the predicted and observed probabilities after conditioning on the true gender of each stimulus. Partial rank-correlation conditioned on the true gender show how well a model predicts the *exact* decision probabilities and not just the gender; see text for details.

then the (averaged) performance of the other 6 observers (see Figure 7). For the Prot, the cross-observer predictions were better in 98% of all comparisons. LDA and SVM showed better performance for 88%, resp. 76% of comparisons, but neither reached the level of performance of the logistic regression model.

Correlating reaction times with predictions

Above we showed that responses of the observers in our experiments could be predicted fairly well by a simple generalized linear model. However, the model was not perfect, and a small number of faces were not predicted

Model	Average Deviance	Percent Correct	Rank Corr	Corr(♀)	Corr(♂)
Logreg	1.935 ± 0.058	0.922 ± 0.008	0.863 ± 0.003	0.577 ± 0.031	0.352 ± 0.045
Prot	4.382 ± 0.297	0.842 ± 0.013	0.722 ± 0.015	0.279 ± 0.013	0.125 ± 0.011
LDA	2.236 ± 0.114	0.913 ± 0.010	0.847 ± 0.006	0.520 ± 0.033	0.312 ± 0.040
SVM	2.383 ± 0.118	0.910 ± 0.011	0.832 ± 0.007	0.465 ± 0.039	0.264 ± 0.031

Table 1. Summary of results: For each of the four estimation methods, we show different performance metrics averaged over all observers and conditions, with standard errors calculated across observers. Corr(♀) and Corr(♂) are the rank-correlations between the predictions of the algorithm and the observers' decisions evaluated on the female and male faces separately.

well, i.e. their deviance residuals were clearly too large. What makes these stimuli harder to predict? One hypothesis is that our decision image model can predict human responses better if the decision is mainly driven by bottom-up cues from the stimulus and is made quickly. If observers spend more time cogitating about their decision, perhaps comparing the current face to ones that she is familiar with, our model based purely on a linear stimulus filter is bound to fail. If this were correct, this would imply that stimuli for which reaction times are low have small deviance residuals, and vice versa.

For each observer and condition, we separated all face stimuli into 10 bins based on their reaction time (RT), each containing 10% of all stimuli, and calculated the average squared deviance residual in each bin (Wichmann

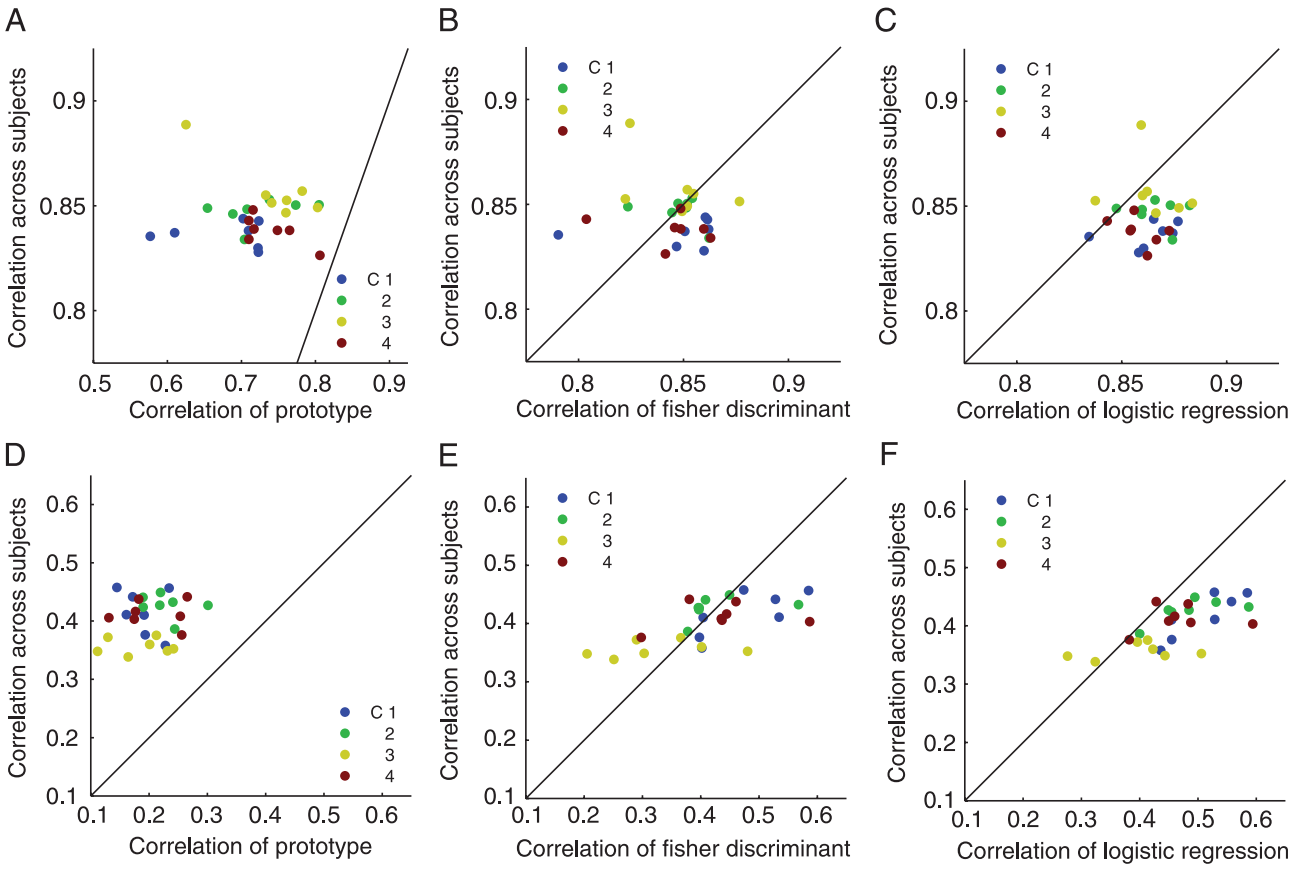

Figure 7. Comparing the performance of the decision model estimated using the prototype classifier, Fisher discriminant analysis and logistic regression (left, middle, and rightmost columns, respectively) with cross-subject predictions; the diagonal lines in each of the subplots mark the equal performance contour. Top row: Rank-correlations of the cross-subject predictions plotted against model-predictions. (Rank-correlations of model predictions on the x-axes re-plotted from Figure 6, middle row; see text for details.) Bottom row: Same as top row except that partial rank-correlations conditioned on the true gender are shown. (Partial rank-correlation of model predictions on the x-axes re-plotted from Figure 6, bottom row; see text for details.)

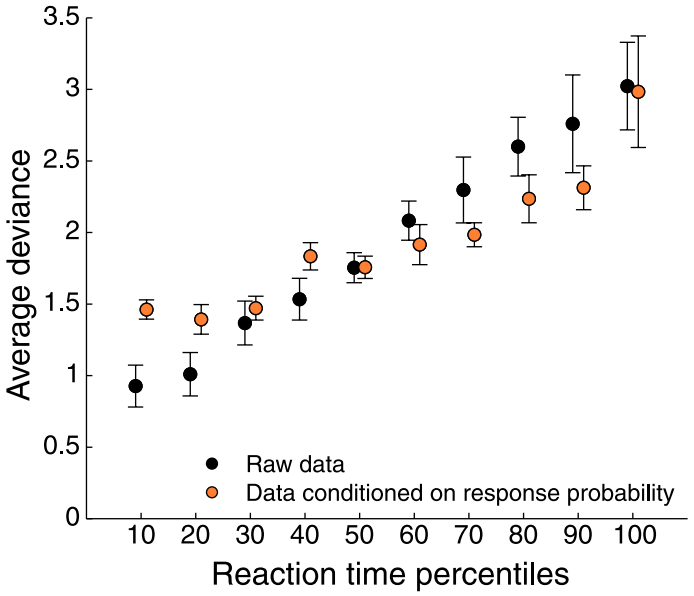

Figure 8. Correlating reaction times (RTs) with the average deviance for the decision images estimated using logistic regression. For each subject and condition, faces were assigned into ten RT-bins and the deviance residuals associated with each face were averaged within a bin (see text for details on RT-bin boundary assignment). Such averaged deviance is shown against RT-bin percentiles (filled black circles; error bars correspond to standard errors of the mean (*SEM*) across subjects). RTs are likely a function of the observers' gender assignment probabilities; to remove this potentially confounding influence we conditioned on the entropy of the observers' responses (see text for details). The correlation between RTs and average deviance persists after conditioning (filled orange circles; error bars correspond to *SEM* across subjects).

& Hill, 2001a). From Figure 8, one can see that the RTs increases monotonically with binindex. In fact, the rank correlation between the percentile and mean deviance is perfect, $c = 1$. When we excluded the morphed faces from the analysis, the correlation was 0.9. For very fast decisions, the average deviance residual is as expected from a perfect model, around 1.0, whereas for slow decisions, they are substantially bigger. We repeated the same analysis for each subject individually, and found a significant correlation (at $p < 0.01$) for 6 out of the 7 subjects, with a median correlation of 0.94. The correlation was also significant for each of the four conditions separately, and was at least 0.96 in each case.

This analysis could be confounded if the deviance and the reaction times both depend on the empirical decision probabilities $\hat{p}(s)$: If a face is difficult to categorize, then subjects might spend more time thinking about it. But if the algorithms also have trouble in categorizing such faces correctly, then the correlation between reaction times and deviance residuals would merely be a consequence of the common correlation with the factor "stimulus difficulty." To rule out this possibility, we computed a partial correlation between the reaction times and deviance residuals, conditioning on the decision probabilities. We binned the decision probabilities into 5 bins, and normalized the reaction times within each bin by subtracting their mean and dividing by the standard deviation. This processing step did not eliminate the correlation between reaction time percentile and deviance, which was still 0.95, as can be seen from Figure 8. The conditional correlation was also significant for each of the four conditions (minimum 0.81) and for 6 out of 7 observers (median 0.84). Similarly, for 21 out of 28 observers and conditions separately, reaction times and deviance residuals were significantly correlated (14 out of 28 conditioning on response probability).

This analysis lends credibility to our claim that the logistic regression model really captures the features used by human observers when categorizing faces according to gender if they respond quickly, and treat this as a "sensory" decision. For longer decision times—not more difficult decisions—they appear to use other strategies, perhaps memory, to arrive at their gender judgement.

Methods II

Using decision images to generate new stimuli

We can use the decision images to generate optimally discriminable stimuli (Wang & Simoncelli, 2008). According to our model, the probability that a stimulus is classified into a particular class, e.g. female, is monotonically related to its distance from the decision boundary. Therefore, if we add a multiple of the normal vector $\lambda\omega$ ($\lambda > 0$) to any stimulus s_0, we can create a stimulus $s_0 + \lambda\omega$ with $p(s_0 + \lambda\omega) > p(s_0)$. For simplicity, we take s_0 to be the mean face in each condition.

Similarly, we can create a pair of stimuli symmetric about s_0 by adding or subtracting equal amounts of ω to/from it: $p(s + \lambda\omega) > p(s_0 - \lambda\omega)$ Moreover, the probabilities vary fastest when we go perpendicular to the decision boundary: For any other vector with $\|\omega\| = \|\xi\|$ we have that $p(s_0 - \lambda\omega) \leq p(s_0 - \lambda\xi) < p(s_0 + \lambda\xi) \leq p(s_0 + \lambda\omega)$. The reason is that $\omega^\top\xi < \omega^\top\omega = \|\omega\|^2$ for any vector ξ with $\|\xi\| = \|\omega\|$. As the absolute "length" $\|\omega_i\|$ is arbitrary, we normalized the decision faces to have norm $\|\omega\| = 1.0$. In other words, adding some multiple of ω to a stimulus should lead to stimuli which are more discriminable than when any other vector (of same length) is added to s_0. If ω really captures the direction in feature space along which the male–female perception varies the most, then even small changes of λ should lead to measurable changes in

the decision probability $p(s + (\lambda\omega))$. On the other hand, the model predicts that addition of any vector γ which is orthogonal to ω should not result in a stimulus which looks more female or male than s_0: $p(s_0 \pm \lambda\gamma) = f(\omega^\top(s \pm \lambda\gamma)) = f(\omega^\top s_0) = p(s_0)$, if $\omega^\top\gamma = 0$.

Our approach of generating stimuli which are either maximally discriminative or invariant with respect to the stimulus class is related to previous studies of linear mechanisms in psychophysics. For example, analogous ideas based on the duality between the most-discriminable directions and constant-response contours have been used in color vision (Knoblauch, 1995).

Testing the models with optimized stimuli

We tested these predictions by conducting a second set of experiments, in which three of the subjects of the first experiment took part. The subjects performed a standard spatial 2 alternative forced choice task in which they were asked to indicate which of the two artificial faces $s + \lambda\omega$ and $s - \lambda\omega$ they perceived as looking more female. If the values of λ which lead to visible gender differences for the filter ω are smaller than the ones for an alternative filter ξ, we can conclude that the model ω captures the decision boundary better than ξ. For each choice of ω, we tested 10 different values of λ in order to map out psychometric functions and thus test the sensitivity of the subjects to changes in ω. The pairs of faces $s_\pm(\lambda\omega)$ were generated using the decision images ω obtained from the first experiment. We used the ωs corresponding to the weight vectors of logistic regression, the prototype classifier and the support vector machine for conditions 1, 2 and 3. In addition, we used a slightly low-passed version of the Logreg filter ω_2. Using a prior which explicitly enforces smooth decision boundaries would make this smoothing step unnecessary (see Discussion) (Knoblauch & Maloney, 2008; Ross & Cohen, 2009).

In addition, we wanted to test the hypothesis that subjects are insensitive to changes along a direction which is thought to contain no information about gender, i.e. directions within the decision boundary. This should lead to much flatter, perhaps even degenerate psychometric functions with $p(s_0 \pm \lambda\gamma) = p(s_0)$ for all values of λ. We therefore generated two additional filters γ for each subject which were chosen to be orthogonal to the filter of logistic regression, i.e. such that $\omega^\top\gamma = 0$. We constrained γ to be a linear combination of the eigenvectors of the data set of all faces, where the weights were chosen to be of similar magnitude as the decision images. This procedure was used with the aim of making the statistics of the non-directions as similar to the decision images as possible. However, as this constraint does not uniquely determine γ, we picked a direction at random for each subject.

Each subject was tested on each of the five choices of filters (corresponding to the prototype classifier, logistic regression, low-pass logistic regression, two neutral directions γ), and on ten different values of λ. Combinations were presented intermixed and in random order, and each combination was presented 225 times. This resulted in a total number of 11250 trials for each subject.

Pairs of faces $s \pm \lambda\omega$ were presented next to each other in the same experimental setup as described above. We randomized whether the face on the left side corresponded to $s + \lambda\omega$ or $s - \lambda\omega$. Each pair was presented using a modified Hanning window with a rise time of 300 ms, a plateau of 300 ms and a fall time of 300 ms. The interstimulus interval was 200 ms. On each trial, the subjects indicated whether the left face or the right face looked more female by pressing a button on a touchpad. Subjects were asked to respond quickly, and were not given feedback on their performance.

Results II

Interpreting the decision images

The filter ω of each model is a vector which has exactly the same dimensionality as the stimuli (faces). Therefore, it can be visualized as an image (see Figure 9), and this is the reason we refer to these filters as *decision images,* or "decision faces" in the particular experimental context considered here. Both the decision image of logistic regression and that of the prototype classifier place power at the eye-region, indicating that this region is important for gender categorization of human faces. This finding is consistent with previous studies using bubbles (Dupuis-Roy et al., 2009), classification images (Mangini & Biederman, 2004; Sekuler et al., 2004), or analysis of photographs and gender ratings (Russell, 2009). However, while both models place emphasis on the eye region, they actually have vastly different prediction performance: Thus, mere localization of important regions is not sufficient to predict human behavior, but the exact filter shapes also matter.

Secondly, one can see that the decision image for the prototype is dominated by low spatial frequencies, which is not the case for the optimal decision boundary. The reason for that is that, as the prototype is obtained by averaging faces (which are dominated by low spatial frequencies), it is bound to also be of this form, even if low spatial frequencies contain no class-specific information at all.

The decision faces can be interpreted more readily if they are used to generate female or male looking faces by adding them to a neutral looking mean face s_0, as described above. From Figure 10, we can see that subtracting ω

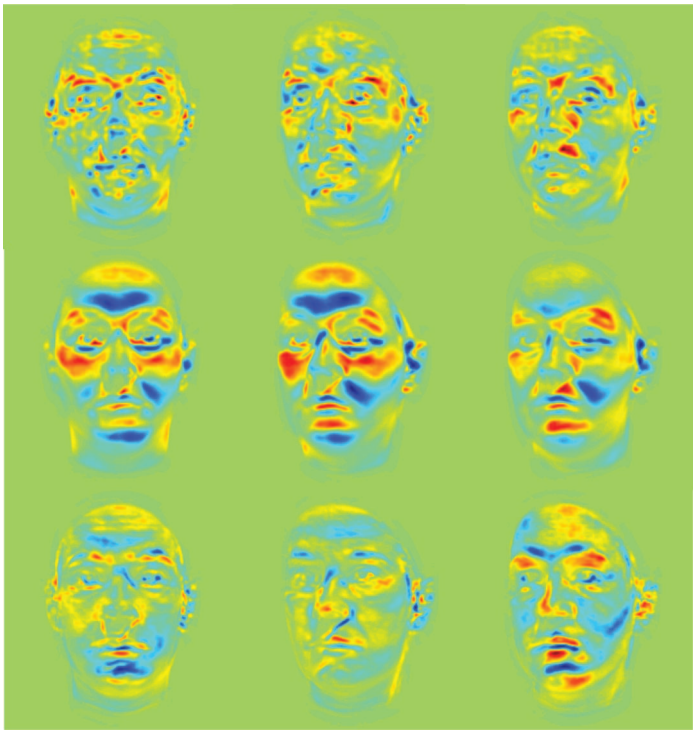

Figure 9. Decision images ("decision faces") for conditions 1, 2 and 3 (left, middle and right columns, respectively; see Figure 1, topmost three rows) averaged across observers. Different methods to estimate the decision images are shown in different rows. Top row: Using logistic regression to estimate the decision images (good fit to psychophysical classification data). Middle row: Using the prototype classifier as predictive model (poor fit to psychophysical classification data). Bottom row: "Synthetic" decision image orthogonal to that estimated using logistic regression in the top row (see text for details).

leads to a male looking face (left column) whereas adding ω results in a female looking face (right column). This is true both for the filter for Logreg (first row) as well as for Prot (second row), but the effect is stronger for the Logreg-filter. In contrast, using a direction along the decision boundary does not result in faces which look either male or female (last row).

Evaluating the experiments with optimized stimuli

For each observer, experimental condition and choice of the filter ω, we fitted psychometric functions to the responses of the subject against the weight λ, which determines 'how much' of the decision image is added to a neutral face. We quantified the performance of each model by finding the value of λ at which subjects were 90% correct. Averaged across observers and conditions, the filter ω corresponding to the decision image of logistic regression performed best, and reached the performance criterion for values of λ less than 0.1 (see Figure 11D). The filter of logistic regression with low-pass filtering slightly outperformed the original Logreg filter (mean-λ at 90% correct: 0.081 vs. 0.0947). This is an indication that our regularization procedure is not optimal yet (see Discussion). Both of these decision images consistently outperformed the SVM (mean-λ = 0.11) and the prototype classifier (mean-λ = 0.14).

Similar results were obtained for each of the three experimental conditions, and each observer individually: In each of the 9 experiments, the filter corresponding to logistic regression (without smoothing) outperformed both the SVM and the Prot filters (Binomial test, 9/9, p = 0.002). In 7 out of 9 experiments, smoothing of the decision image did lead to a performance increase (p = 0.089). As we mentioned previously, the absolute scaling of λ is arbitrary, and it is really the relative differences which are meaningful. Therefore, for each classifier and condition, we calculated the percentage correct attained at the value of λ at which the best algorithm achieved 90% correct. On average, when logistic regression (lowpass) was at 90%, Logreg was at 84% ± 2.3, SVM at 78% ± 2.6 and Prot at only 74% ± 3.0.

The filter γ which was parallel to the decision boundary did not result in any differences in perceived gender; in fact, the psychometric function was flat across the whole range of λ's considered, and never achieved 90% correct. Thus, at least for the particular directions that we tested,

Figure 10. Synthesising maximally discriminative faces with respect to gender using the decision images for condition 3 (see Figure 9, middle column), averaged across observers. Comparable distances along the respective axes are visualized, from $\lambda = -0.5$ to $\lambda = 0.5$ (see text for details). Top row: Synthesized using the logistic regression-estimated decision image, for which observers were typically at 90% correct for $\lambda = \pm 0.08$ Middle row: Synthesized using the prototype-based decision image, 90% correct for $\lambda \approx \pm 0.15$. Bottom row: Synthesized using the synthetic decision image orthogonal to that estimated using logistic regression, no clear gender change even for $\lambda > 3$ as predicted by our model.

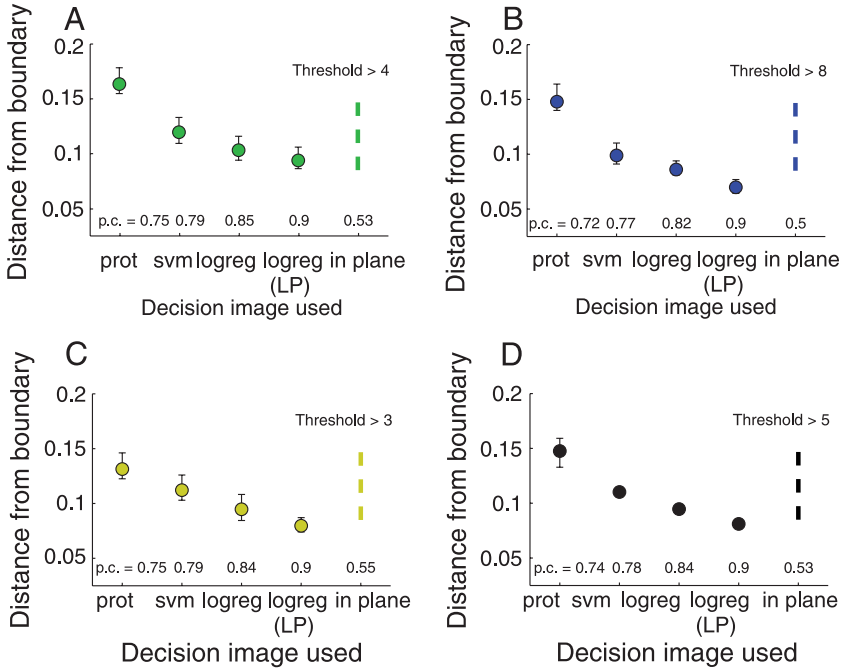

Figure 11. Results of the gender discrimination experiment using synthetic faces symmetrically placed around a neutral face in "face-space". A) Thresholds for 90% correct gender discrimination shown for synthetic face images synthesized using various decision images for faces of condition 1 (see Figure 1, top row). Results are pooled across observers and error bars indicate the *SEM*. B) Same as A except for faces from condition 2 (see Figure 1, second row, for the face condition and Figure 10 for example stimuli). C) Same as A except for faces from condition 3 (see Figure 1, third row). D) Same as A except pooled not only across observers but conditions 1, 2, and 3, too.

we can conclude that moving along the decision boundary does not result in stimuli which appear male or female (see Figures 10A–10D), which is in accordance with our predictions.

Discussion

Stimulus correlations and stimulus bias

We presented a method to derive a predictive model of psychophysical decisions during a binary discrimination task, and applied it to gender perception of human faces. As our method derives the predictive features—the decision image—from the example stimuli shown, it is inevitable that the decision images depend on the stimulus class used, as they define the classification task used. For example, our model's decision boundary will be similar to the decision boundary of an optimal classifier, unless subjects are at chance level, and there will be a bias towards the more physically discriminative features. Therefore, it is of critical importance to show that the prediction ability of the decision image goes beyond 'rediscovering' the physical structure of the stimulus, and care is to be taken when comparing the features returned by experiments based on different stimuli.

In addition, the individual entries ω_i of the decision image will depend on the covariance structure of the stimuli. For simplicity, we will discuss this dependence for Fisher's linear discriminant: In this case, the decision boundary is $\omega = C^{-1}(\mu_+ + \mu_-)$, where C is the within-class covariance, and $\mu_\pm$ are the means of the two (perceived) classes. For every pixel k, the difference between the two classes is inversely weighted by the variance within each class. An entry ω_k will be large if either the pixel $s_{(k)}$ varies a lot between classes or varies little within the classes. In terms of spatial frequency, the effect of pre-multiplying by C^{-1} can be interpreted as dampening of those spatial frequencies which are more dominant in the stimulus. For faces and for most natural stimuli low spatial frequencies are more powerful, this can explain why the decision images of the Prot classifier (which ignores the covariance structure) have much power in the low spatial frequencies than those which do take the covariance structure into account.

The discussion above also implies that the generation of stimuli by adding a multiple λ of the decision image ω only makes sense for small values of λ, i.e. for local perturbations: Suppose that for some pixel k, the variance C_{kk} is small, and therefore that ω_i is large. This means that

even small changes in this pixel value will lead to noticeably changes in its perceived class. However, the fact that C_{kk} is small also implies that, for most stimuli, the pixel value $s_{(k)}$ will be small. Thus, if λ is chosen too large, the corresponding pixel entry $s + \lambda\omega_k$ will be large, and is likely to be larger than it would be in any real face, creating artifacts and unnaturally looking stimuli. In our experiments we easily found values of λ which appeared distinctly male or female, and at the same time still looked very 'face-like'. However, larger values of λ resulted in noticeable artifacts.

The fact that the prototype classifier (which entirely ignores C_{kk}) performs so poorly implies that human observers must implicitly take the covariance structure of faces into account for making the decisions about the gender of a face. A related (but not equivalent) idea has been formulated by studies (e.g. O'Toole et al., 1993) using principal component analyses (PCA) on face images and arguing that the first principal component was discriminative between male and female faces. In our experiments, all faces were normalized to have the same mean luminance, and standard deviation of luminance values. Consequently, the first principal component to determine gender was not a good predictor of gender: For example, on the stimuli condition 1, it resulted in a percentage correct of only 56.9%. This suggests that the separation of the two classes by the first principal component is a consequence of an overall brightness difference between male and female faces (Russell, 2009), and therefore only works for un-normalized stimuli. Given that the absolute light level—and hence the light intensity of faces—changes under real-life conditions, reliance on the first principal component for gender discrimination is clearly a sub-optimal strategy. The performance of observers in our experiments with normalized stimuli also shows that absolute luminance is not necessary for gender classification.

We discussed the dependence of the entries ω_i on the covariance structure of the stimulus for the case of linear discriminant analysis, as its decision boundary is solely determined by the covariance and means of the two classes. However, the conclusions qualitatively also apply to algorithms such as logistic regression or SVMs. These algorithms—unlike LDA—are not motivated by specific assumptions about the statistical structure of the stimuli and are therefore likely to work better for stimuli for which simple distributions such as Gaussians are inappropriate: This is the case for many classes of naturalistic stimuli, e.g. natural images.

Limitations and possible improvements

Applicability of the method

We want to estimate predictive features in classification tasks, without having to embed the stimuli in noise. As our strategy is to exploit the variability in the stimulus classes, we require stimuli which are sufficiently heterogeneous within each class to 'sample' different features. In addition, the task has to be such that the performance of human observers is it not completely at ceiling, as in this case, it would be impossible to learn anything about their strategies that goes beyond re-discovering the class structure of the stimulus, as outlined above. If it is not possible to design the experiment such that these requirements are met, then noise embedding techniques would be the method of choice. In this respect, our method is complementary to classification images and bubbles. In addition, the applicability of this method is limited to binary classification tasks, although it might be possible to extend it to situations in which subjects are asked to discriminate between multiple alternatives.

Better feature spaces and model specifications

We applied our methods to face images that were parameterized by their pixel intensities. One shortcoming of the approach presented here is the construction of new stimuli by addition: It is by no means guaranteed that the sum of two faces is another face, or more generally, that the (weighted) sum of two stimuli from a class is itself a member of the class. For faces this problem could be alleviated by not working in pixel-space, but rather in a parameterized morph-space like the one introduced by Blanz and Vetter (1999), or by projecting into a lower-dimensional face-space (Sirovich & Meytlis, 2009). If we applied our decision image technique not to the pixel images but to their standardized versions in morph-space, we would be guaranteed that our synthetically generated stimuli look like faces, at least for reasonable perturbations of λ (see above). On the other hand, interpretation of the decision image might be harder, as the identified features have to be "projected-out" of the morph-space into pixel space to be interpretable (c.f. Kienzle et al., 2009). Alternatively, one could parameterize the stimuli not by their pixel values, but use a set of basis functions which are of the kind that is thought to be implemented in early stages of the visual pathway. For example, one could project the stimuli onto a filter bank of oriented filters, and apply the algorithm to the filter-responses.

The fact that our model is based on only one linear template makes it useful only for stimuli which are aligned, i.e. in which there is no uncertainty about the location of the target. If more general stimulus classes are to be used, models based on multiple templates will need to be used (Cohen, Shiffrin, Gold, Ross, & Ross, 2007; Ross & Cohen, 2009; Tjan & Nandy, 2006).

Better priors

Similarly, it is to be expected that more informed choices of the regularization term will lead to better results both in terms of performance and interpretability of the decision images. As we typically have little data,

the regularizer is critically important during optimization. Currently we use the L2-norm which enforces distributed weights, and is used in popular classification algorithms such as the SVM. The L1-norm leads to sparser solutions, however, and for feature identification we want sparseness (Mineault et al., 2009). In particular, if a feature is correlated with a very predictive one (but not predictive in isolation) we would want such a feature to be set to 0, rather than both features being included in the decision image. For the face stimuli used in this study, using a regularizer which explicitly favors solutions which are dominated by low spatial frequencies would eliminate any need for smoothing decision images post-hoc (Knoblauch & Maloney, 2008). In our experiments, we found that a gentle smoothing of the images helps the prediction performance in our second round of experiments—this can be regarded as evidence that our regularization is not optimal yet (Mineault et al., 2009).

Fully Bayesian models: Iso-probability contours not parallel to the boundary

We have used a nonlinear–linear model which was fitted using logistic regression with a regularization term. Our decision image (filter) ω_{opt} is a maximum a posteriori (MAP) solution, i.e. the most likely filter given the data and the prior. Consequently, $p(s) = p({\omega_{opt}}^{\top} s)$ is the "most likely" value of the decision probability for stimulus s. The MAP-predictor is a very popular estimator in statistics, and is often even referred to as "Bayes-optimal." However, the MAP-estimator is only optimal under very particular conditions (under the assumption of a 0/1 loss (Bishop, 2006)), and a better predictor is the posterior mean probability $p(s) = \int_{\omega} f(\omega^{\top} s)\pi(\omega)d\omega$. Not a single filter is used for prediction, but rather an integral over all possible filters, weighted by their relative posterior probabilities, $\pi(\omega)$. In such a fully Bayesian model, the decision boundary is still linear, and the direction along which the probabilities vary the fastest is still normal to the decision boundary. However, in general, the iso-probability contours (i.e. the lines along which $p(s)$ is constant) are no longer hyper-planes parallel to the decision boundary (Bishop, 2006). Rather, $p(s)$ depends both on the distance of s from the decision boundary, as well as on the distance of s to the mean of the prior. The closer s is to the prior mean, the faster the probabilities change when going away from the boundary. It will be interesting to see in the future whether such fully Bayesian models can be differentiated from our MAP model based on experimental data.

Summary and conclusions

For human beings the accurate perception of faces is essential: Typically we recognize individuals, determine their gender, and even decode their emotional state simply by looking at their face. How this crucial ability is brought about in terms of visual information processing is, however, still unclear. One of the main challenges is to infer the stimulus features on which the human visual system bases its computations—a pre-requisite for successful computational models of visual perception in general, and face perception in particular.

Here we described a technique for extracting the critical stimulus features predictive of the responses of humans observers during natural viewing of faces and use it to model data from a psychophysical gender discrimination experiment. We were able to predict human responses with remarkable accuracy—remarkable, as gender discrimination is a high-level visual task, and thus believed to be complex, whereas our model is comparatively simple: a decision image (linear filter) followed by a static nonlinearity. The linear filter was found by searching for features which are informative of the psychophysical decisions of observers. The good performance of the prediction models suggest that these features do not only correlate with the observers' responses, but rather play a more fundamental role for face categorization, namely that these are the features that are actually used by the human observers. Indeed, our second round of experiments shows that these features alone can be used to generate faces which were perceived as distinctly male or female.

Our findings are consistent with studies postulating "norm based" encoding of faces (Loffler et al., 2005): In our experiments, the distance of a face to a separating hyperplane predicted the probability of (mis-) categorizing a face. However, the optimal gender-separating axis was not aligned with the axis defined by the average male and female face. Therefore, the simple prototype model popular for modeling faces (Leopold et al., 2006; Loffler et al., 2005), which uses the class means or averages (but not the class covariance) is not sufficient for explaining responses in our experiments despite its popularity in the literature. (For a different critique of such prototype face models see Jiang et al., 2006.) Rather, the superior performance of LDA shows that it is at least necessary to model the class covariances. Finally, the small but consistent advantage of logistic regression (which takes into account individual stimuli) over LDA (which takes into account only the covariance across stimuli) suggests that the mechanism subserving gender classification decisions in humans is indeed sensitive to individual stimuli and not only class covariance structure.

We used our method to derive a predictive model of psychophysical decisions during a binary gender classification task with human faces. The method per se is not limited to this task, but offers numerous applications. In the domain of face perception, for example, it could be used to derive objectively the features underlying perceived beauty, health, emotions, or kinship (Dal Martello & Maloney, 2006), to name but a few. Decision

images could also be useful for understanding medical conditions in which feature-processing is impaired, for example by computing the decision images of observers which suffer from face processing deficits such as prosopagnosia. On the other hand, one could derive decision images of trained expert on, e.g. a medical diagnosis task, to see what features they may be using, and how they differ from those used by non-experts. Subsequently, one could then train non-experts on the 'expert features'.

Appendix A

Support vector machines for repeated stimulus representations

The support vector machine (SVM) (Hofmann et al., 2008) is a classification algorithm popular in machine learning, and has successfully been applied in a variety of domains. The SVM separates the two classes by a hyperplane which is chosen such that the margin between the decision boundary and each of the two classes is maximal. Although the algorithm is linear in conception, it can easily be generalized to nonlinear settings by using nonlinear kernel functions. We used a linear kernel for our experiments, but the calculations in this appendix are valid for any choice of kernel. A SVM for binary classification separates two classes of data points by a decision boundary of the form $f(x) = \sum_i y_i \alpha_i k(x_i, x) + \beta$, where the *kernel function* k specifies the dot-product of any two data points in feature-space. Intuitively, the *kernel function* captures the similarity between any two points.

The optimal decision function f is found by minimizing the cost-function

$$L(\alpha) = C\sum_{i=1}^{n} |1 - y_i f(x_i)|_+ + \frac{1}{2}\sum_{ij} y_i y_j \alpha_i \alpha_j k(x_i, x_j), \tag{A1}$$

where $|1 - y_i f(x_i)|_+ = \max(0, 1 - y_i f(x_i))$. This optimization problem can be rewritten as

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \tag{A2}$$

$$\text{subject to } 0 \le \alpha \le C \ \forall i, \tag{A3}$$

$$\text{and } \sum_i \alpha_i y_i = 0. \tag{A4}$$

Thus, the optimal decision boundary is found by optimizing over the weights α_i on the data samples. This implies that the number of parameters in the optimization is independent of the dimensionality of the feature space, which makes it possible to work in high (or infinite dimensional) feature spaces. However, it also implies that the computational requirements of the algorithm scale with the size of the data set considered. Thus, although in principle, repeated presentations of the stimulus can be handled by simply inserting each presentation as a separate data point, this procedure can quickly result in very large data sets which become computationally inconvenient or even infeasible. Furthermore, we also want to be able to handle situations in which the decision probabilities p_i are arbitrary continuous quantities in [0, 1], and not estimated over multiple stimulus presentations.

Here, we will briefly show how the SVM can be generalized to a setting in which each object x_i does not only have a class-label y_i, but also a corresponding confidence or decision probability p_i. p_i determines our confidence as to whether the class-label y_i is correct. For example, if $p_i = 0.5$, we know that x_i is an entirely ambiguous stimulus. Furthermore, the case $p_i = p$, $y_i = -1$ is equivalent to $p_i = 1 - p$ and $y_i = 1$. To resolve this ambiguity, for every stimulus with labels (y_i, p_i), we add a second stimulus $p_{i+n} = 1 - p_i$, $y_{i+n} = -y_i$. Then, the loss function can be generalized to

$$L(\alpha) = C\sum_{i=1}^{2n} p_i |1 - y_i f(x_i)|_+ + \frac{1}{2}\sum_{ij} y_i y_j \alpha_i \alpha_j k(x_i, x_j). \tag{A5}$$

The loss of every training point is re-weighted by its certainty p_i. The second part of the loss function, which regularizes the shape of the decision function, is unchanged. This is equivalent to

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{2n} \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \tag{A6}$$

$$\text{subject to } 0 \le \alpha \le C p_i \ \forall i, \tag{A7}$$

$$\text{and } \sum_i \alpha_i y_i = 0. \tag{A8}$$

Thus, the only things that have changed is that the summation is now over (at most) $2n$ rather than n weights α, and that the upper bound on each α_i is Cp_i rather than C. If any p_i is zero, we know that $\alpha_i = 0$, i.e. the corresponding stimulus can be eliminated from the optimization problem, and the sum is over less than $2n$ terms.

Importantly, the algorithm does not simply ignore data points with $p_i = 0.5$, but rather aims to place them close to the decision boundary. Even stimuli which are classified very inconsistently can provide important information about the position of the decision boundary. A similar idea has been studied in the field of machine learning under the name of 'Universum SVM' (Weston, Collobert, Sinz, Bottou, & Vapnik, 2006), which uses a 'Universum' of 'non-examples' in addition to the labeled data points. While the derivation and motivation of the USVM is very different to our approach, it is mathematically equivalent to a SVM with confidences p_i which are either 0, 1 or exactly 0.5: The 'Universum' consists of those data points which have completely ambiguous class labels.

Acknowledgments

We are indebted to Eero Simoncelli, Alan Allport and Bruce Henning who independently suggested to explore faces orthogonal to or within the decision boundary. In addition, we benefited from discussions with Matthias Bethge, Frank Jäkel, Larry Maloney and Bernhard Schölkopf, and we would like to thank Roland Fleming, Ingo Fründ, Frank Jäkel, Stefanie Liebe and Vinzenz Schönfelder for carefully reading the manuscript. Last but certainly not least we would like to thank Nadja Althaus and Silke Dutz for their help conducting experiments and for dedicated observing. This research was funded, in part, by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research.

Commercial relationships: none.
Corresponding author: Jakob H. Macke.
Email: jakob@gatsby.ucl.ac.uk.
Address: Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London WC1N3AR, United Kingdom.

Footnotes

[1]The MPI face database is located at http://faces.kyb.tuebingen.mpg.de.

[2]A Generalized Linear Model fit with a distribution-model that has heavy tails (such as the Cauchy for symmetric lapses), can be used to be more robust to lapses, and possibly avoid an explicit, non-convex fit of the lapse-rates. We would like to thank the reviewer for this suggestion.

References

Abbey, C. K., & Eckstein, M. P. (2007). Classification images for simple detection and discrimination tasks in correlated noise. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 24,* B110–B124. [PubMed]

Abdi, H., Valentin, D., Edelman, B., & O'Toole, A. J. (1995). More about the difference between men and women: Evidence from linear neural networks and the principal-component approach. *Perception, 24,* 539–562. [PubMed]

Ahumada, A., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America, 49,* 1751–1976.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology-Learning Memory and Cognition, 14,* 33–53. [PubMed]

Beard, B. L., & Ahumada, A. J. (1998). A technique to extract relevant image features for visual tasks. *Human Vision and Electronic Imaging III, 3299,* 79–85.

Bishop, C. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blackwell, K. T., Vogl, T. P., Dettmar, H. P., Brown, M. A., Barbour, G. S., & Alkon, D. L. (1997). Identification of faces obscured by noise: Comparison of an artificial neural network with human observers. *Journal of Experimental & Theoretical Artificial Intelligence, 9,* 491–508.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Siggraph 99 Conference Proceedings,* 187–194.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51,* 256–284. [PubMed]

Cohen, A. L., Shiffrin, R. M., Gold, J. M., Ross, D. A., & Ross, M. G. (2007). Inducing features from visual noise. *Journal of Vision, 7*(8):15, 1–14, http://journalofvision.org/content/7/8/15, doi:10.1167/7.8.15. [PubMed] [Article]

Dal Martello, M. F., & Maloney, L. T. (2006). Where are kin recognition signals in the human face? *Journal of Vision, 6*(12):2, 1356–1366, http://journalofvision.org/content/6/12/2, doi:10.1167/6.12.2. [PubMed] [Article]

Dupuis-Roy, N., Fortin, I., Fiset, D., & Gosselin, F. (2009). Uncovering gender discrimination cues in a realistic setting. *Journal of Vision, 9*(2):10, 1–8, http://journalofvision.org/content/9/2/10, doi:10.1167/9.2.10. [PubMed] [Article]

Eckstein, M. P., & Ahumada, A. J., Jr. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision, 2*(1):i, i, http://journalofvision.org/content/2/1/i, doi:10.1167/2.1.i. [PubMed] [Article]

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology, Learning Memory Cognitive, 10,* 234–257. [PubMed]

Gerwinn, S., Macke, J. H., Seeger, M., & Bethge, M. (2008). *Bayesian inference for spiking neuron models with a sparsity prior*. Paper presented at the Advances in Neural Information Processing Systems.

Goris, R. L., Wagemans, J., & Wichmann, F. A. (2008). Modelling contrast discrimination data suggest both the pedestal effect and stochastic resonance to be caused by the same mechanism. *Journal of Vision, 8*(15):17, 1–21, http://journalofvision.org/content/8/15/17, doi:10.1167/8.15.17. [PubMed] [Article]

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41,* 2261–2271. [PubMed]

Graf, A. B., Wichmann, F. A., Bulthoff, H. H., & Scholkopf, B. (2006). Classification of faces in man and machine. *Neural Computation, 18,* 143–165. [PubMed]

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

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

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics, 36,* 1171–1220.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology, 51,* 343–358.

Jiang, X., Rosen, E., Zeffiro, T., Vanmeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron, 50,* 159–172. [PubMed]

Kienzle, W., Franz, M. O., Scholkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision, 9*(5):7, 1–15, http://journalofvision.org/content/9/5/7, doi:10.1167/9.5.7. [PubMed] [Article]

Knoblauch, K. (1995). Dual bases in dichromatic color space. *Documenta Ophtalmologica Proceedings Series, 57,* 165–176.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision, 8*(16):10, 1–19, http://journalofvision.org/content/8/16/10, doi:10.1167/8.16.10. [PubMed] [Article]

Kullback, S. (1997). *Information theory and statistics*. Mineloa, MN: Dover.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442,* 572–575. [PubMed]

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience, 8,* 1386–1390. [PubMed]

Lu, J., Plataniotis, K. N., & Venetsanopoulos, A. N. (2003). Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks, 14,* 117–126. [PubMed]

Mangini, M., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28,* 209–226.

Marmarelis, P. Z., & Marmarelis, V. Z. (1978). *Analysis of physiological systems: The white-noise approach*. New York: Plenum Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London and New York: Chapman and Hall.

Mineault, P. J., Barthelme, S., & Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of Vision, 9*(10):17, 1–24, http://journalofvision.org/content/9/10/17, doi:10.1167/9.10.17. [PubMed] [Article]

Moghaddam, B., & Yang, M. H. (2002). Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24,* 707–711.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2*(1):6, 79–104, http://journalofvision.org/content/2/1/6, doi:10.1167/2.1.6. [PubMed] [Article]

Murray, R. F., & Gold, J. M. (2004). Troubles with bubbles. *Vision Research, 44,* 461–470. [PubMed]

Neri, P. (2004). Estimation of nonlinear psychophysical kernels. *Journal of Vision, 4*(2):2, 82–91, http://journalofvision.org/content/4/2/2, doi:10.1167/4.2.2. [PubMed] [Article]

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research, 46,* 2465–2474. [PubMed]

Ng, M., Ciaramitaro, V. M., Anstis, S., Boynton, G. M., & Fine, I. (2006). Selectivity for the configural cues that identify the gender, ethnicity, and identity of

faces in human cortex. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 19552–19557. [PubMed] [Article]

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature, 459,* 89–92. [PubMed]

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

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

O'Toole, A. J., Vetter, T., Troje, N. F., & Bulthoff, H. H. (1997). Sex classification is better with three-dimensional head structure than with image intensity information. *Perception, 26,* 75–84. [PubMed]

Peterson, M. F., & Eckstein, M. P. (2009). Eye movement strategies adapted to individual differences in the loci of performance-maximizing fixations during face recognition [Abstract]. *Journal of Vision, 9*(8):543, 543a, http://journalofvision.org/content/9/8/543, doi:10.1167/9.8.543.

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

Ross, M. G., & Cohen, A. L. (2009). Using graphical models to infer multiple visual classification features. *Journal of Vision, 9*(3):23, 1–24, http://journalofvision.org/content/9/3/23, doi:10.1167/9.3.23. [PubMed] [Article]

Russell, R. (2009). A sex difference in facial contrast and its exaggeration by cosmetics. *Perception, 38,* 1211–1219. [PubMed]

Schyns, P. G., Gosselin, F., & Smith, M. L. (2009). Information processing algorithms in the brain. *Trends in Cognitive Sciences, 13*, 20–26.

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology, 14,* 391–396.

Sharpee, T. O., Miller, K. D., & Stryker, M. P. (2008). On the importance of static nonlinearity in estimating spatiotemporal neural filters with natural stimuli. *Journal of Neurophysiology, 99,* 2496–2509. [PubMed] [Article]

Simoncelli, E. P., J., P., Paninski, L., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences* (vol. III, pp. 327–338). Cambridge, MA, USA: MIT Press.

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 4,* 519–524. [PubMed]

Sirovich, L., & Meytlis, M. (2009). Symmetry, probability, and recognition in face space. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 6895–6899.

Smith, F. W., Muckli, L., Brennan, D., Pernet, C., Smith, M. L., Belin, P., et al. (2008). Classification images reveal the information sensitivity of brain voxels in fMRI. *Neuroimage, 40,* 1643–1654. [PubMed]

Smith, M. L., Fries, P., Gosselin, F., Goebel, R., & Schyns, P. G. (2009). Inverse mapping the neuronal substrates of face categorizations. *Cerebral Cortex, 19,* 2428–2438. [PubMed]

Smith, M. L., Gosselin, F., & Schyns, P. G. (2004). Receptive fields for flexible face categorizations. *Psychology Science, 15,* 753–761. [PubMed]

Tjan, B. S., & Nandy, A. S. (2006). Classification images with uncertainty. *Journal of Vision, 6*(4):8, 387–413, http://journalofvision.org/content/6/4/8, doi:10.1167/6.4.8. [PubMed] [Article]

Troje, N. F., & Bulthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research, 36,* 1761–1771. [PubMed]

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

Tyler, C. W., & Chen, C. C. (2006). Spatial summation of face information. *Journal of Vision, 6*(10):11, 1117–1125, http://journalofvision.org/content/6/10/11, doi:10.1167/6.10.11. [PubMed] [Article]

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

Vetter, T., & Troje, N. F. (1997). Separation of texture and shape in images of faces for image coding and synthesis. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 14,* 2152–2161.

Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision, 8*(12):8, 1–13, http://journalofvision.org/content/8/12/8, doi:10.1167/8.12.8.[PubMed] [Article]

Weston, J., Collobert, R., Sinz, F., Bottou, L., & Vapnik, V. (2006). *Inference with the universum.* Paper presented at the Proceedings of the 23rd international conference on machine learning.

Wichmann, F. A. (1999). *Some aspects of modelling early spatial vision: Contrast discrimination*. University of Oxford.

Wichmann, F. A., Graf, A. B. A., Simoncelli, E. P., & Schoelkopf, B. (2005). *Machine learning applied to perception: Decision-images for classification*. Paper presented at the Advances in Neural Information Processing Systems.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63,* 1293–1313. [PubMed] [Article]

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63,* 1314–1329. [PubMed]

Wu, M. C., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review Neuroscience, 29,* 477–505. [PubMed]

Yovel, Y., Franz, M. O., Stilz, P., & Schnitzler, H. U. (2008). Plant classification from bat-like echolocation signals. *PLoS Computer Biology, 4,* e1000032. [PubMed] [Article]