Ideal observer analysis for task normalization of pattern classifier performance applied to EEG and fMRI data

Matthew F. Peterson,[1,*] Koel Das,[1] Jocelyn L. Sy,[1] Sheng Li,[3] Barry Giesbrecht,[1,2] Zoe Kourtzi,[3] and Miguel P. Eckstein[1,2]

[1]*Department of Psychology, University of California, Santa Barbara, California 93106, USA*
[2]*Institute for Collaborative Biotechnologies, University of California, Santa Barbara, California 93106, USA*
[3]*School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*
[*]*Corresponding author: peterson@psych.ucsb.edu*

Received April 20, 2010; revised October 8, 2010; accepted October 19, 2010;
posted October 22, 2010 (Doc. ID 127094); published November 24, 2010

The application of multivariate techniques to neuroimaging and electrophysiological data has greatly enhanced the ability to detect where, when, and how functional neural information is processed during a variety of behavioral tasks. With the extension to single-trial analysis, neuroscientists are able to relate brain states to perceptual, cognitive, and motor processes. Using pattern classification methods, the neuroscientist can extract neural performance measures in a manner analogous to human behavioral performance, allowing for a consistent information content metric across measurement modalities. However, as with behavioral psychophysical performance, pattern classifier performances are a product of both the task-relevant information inherent in the brain and in the task/stimuli. Here, we argue for the use of an ideal observer framework with which the researcher can effectively normalize the observed neural performance given the task's inherent objective difficulty. We use data from a face versus car discrimination task and compare classifier performance applied to electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) data with corresponding human behavior through the absolute and relative efficiency metrics. We show that confounding variables that can lead to erroneous interpretations of information content can be accounted for through comparisons to an ideal observer, allowing for more confident interpretation of the neural mechanisms involved in the task of interest. Finally, we discuss limitations of interpretation due to the transduction of indirect measures of neural activity, underlying assumptions in the optimality of the pattern classifiers, and dependence of efficiency results on signal contrast.

OCIS codes: 330.4300, 330.5510, 330.4060.

1. INTRODUCTION

The neuroscientist's use of noninvasive measures of neural activity to infer functional brain mechanisms rests on many assumptions [1]. First and foremost is the idea that the pattern of neuronal activity (i.e., a brain state) directly represents an organism's perceptual-cognitive-motor "experience" (i.e., a mental state; though these states may be inaccessible, unused, or below the organism's conscious experience) [2]. As Barlow succinctly states: "*To understand nervous function one needs to look at interactions at a cellular level … because behaviour depends upon the organized pattern of these intercellular interactions*" [3]. If this reasoning is correct, then it should be possible to decode an organism's neural activity to gain a representation of its mental state. That is, the information that the organism is using to interact with its environment must be instantiated in the neural activity, guided by the brain's architecture of distributed networks. Indeed, this has been accomplished in a variety of studies, lending credence to Barlow's dictum [2,4,5].

Classically, researchers have used univariate methods to analyze imaging data averaged over many observations (trials) in controlled tasks to infer the participation of specific brain areas and time windows. While these methods have led to great advances in determining a coarse structure of functional brain organization they say little about how, or even if, this activation corresponds to useful task-relevant information processing. Indeed, analyzing observable variables in isolation (e.g., single voxels, electrodes, time points; from here on we will refer to these as samples) to some extent ignores Barlow's statements on the interactions between neurons [6].

Recently, researchers have begun to use more advanced statistical methods to implicate brain regions in specific behaviors while allowing for greater signal sensitivity. Perhaps the most popular of these methodologies is pattern classification, a general term referring to a large and diverse suite of multivariate statistical procedures culled from the field of machine learning (see Fig. 1) [5,7–9]. These techniques take the raw imaging data, distributed among many samples, as a high-dimensional multivariate distribution. Each trial, then, results in a noisy sample of task-specific distributed neural activity. The goal of the pattern classifier is to specify the "best" way to segregate the observations such that certain patterns of activity correspond to specific properties of the task or observers' behavior. Importantly, these are predictive methods, whereby the optimal classification criteria are inferred from a set of training data and then tested by applying these criteria to an independent set of observations. Multivariate techniques offer both pragmatic and interpretational benefits over unidimensional analysis. On the prac-

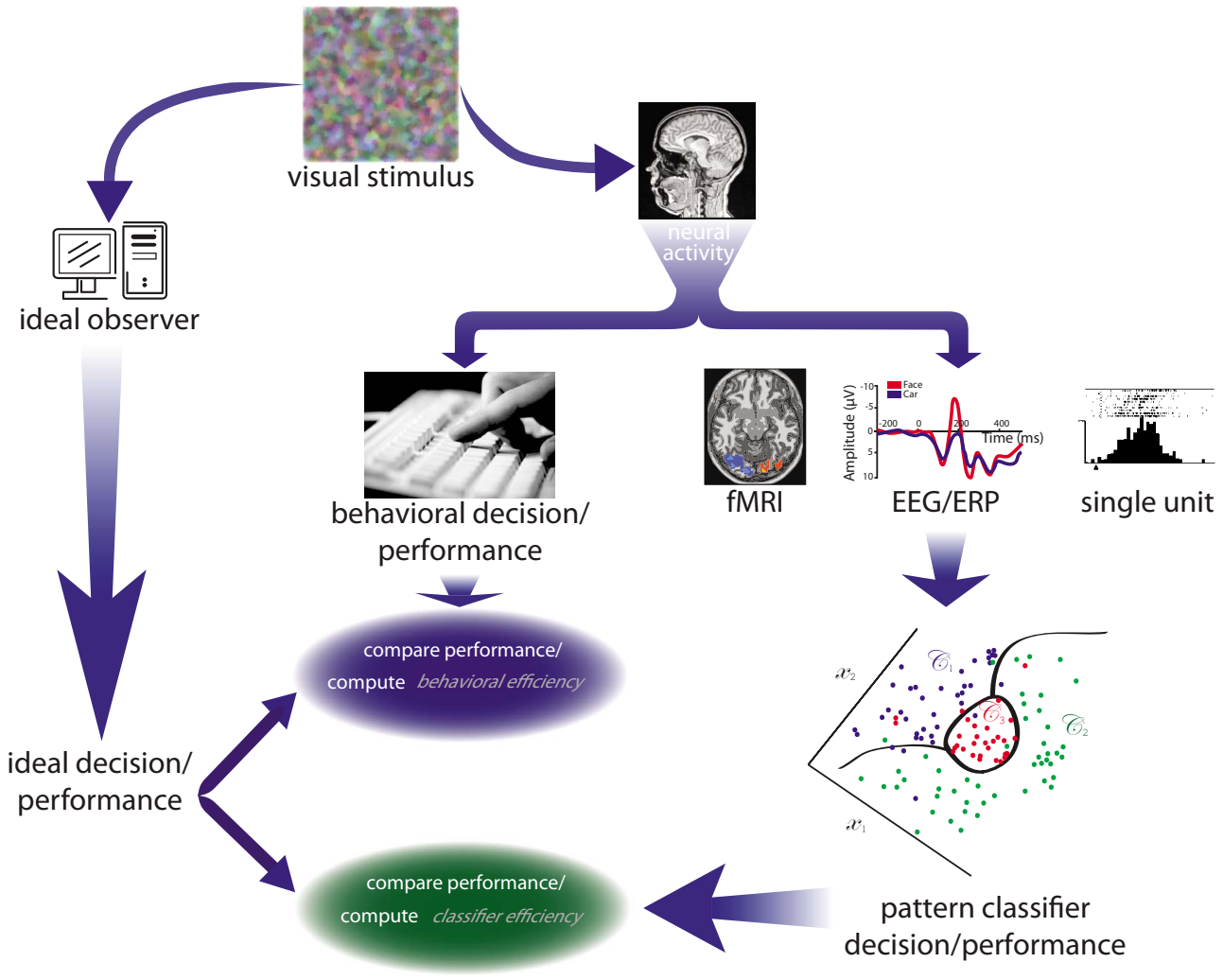

Fig. 1. (Color online) Outline of an ideal observer framework to analyze human behavior and classifier performance based on neural activity. The visual stimulus impinges on the observer's retina, where it is sampled by the photoreceptor lattice. This neural signal is then propagated back to the cortex. Imaging techniques such as fMRI and EEG observe indirect measures of neural firing related to the stimulus and task. Pattern classifiers are used to map patterns of activity to specific postulated brain states that can then be used to make inferences about the state of the world. The human brain also uses the pattern of activity to make decisions. The ideal observer performs the visual task in a mathematically optimal way, giving an upper bound on task performance. The human's behavior and the pattern classifier's decisions can be compared with those of the ideal observer, allowing for a principled comparison of the quality of information being used by each decision mechanism.

tical side, the neuroscientist can take advantage of the increased sensitivity, the ability to detect diffuse information, and utilization of information not necessarily contained in mean activity differences (e.g., covariance) [9]. Thus, information that could be lost in averaging techniques or contained in spatial covariance is now accessible (for a treatment of the benefits of combining information across voxels using multivariate methods see [10,11]). Along with increased observational power comes a clear advantage in the analysis and interpretation of the data's functional utility and its relation to the associated human behavior. In an event-related paradigm the classifier is able to form a decision for each trial (the test set) based on the activity patterns from a mutually exclusive set of trials (the training set, generally from the same subject and task). The output of the classifier is a performance measure analogous to the behavioral performance measure recorded from the human subject. This allows the investigator to infer not only which brain regions and/or time intervals contain task-relevant neural information but also the relative amount of information they possess (e.g., see [12–16]).

However, inferences about the functionality of brain regions from classifier performance across different visual tasks can be problematic. As an example, assume a study is run in which observers are asked to respond whether they see a face or a house in a brief presentation interval while fMRI or EEG data is recorded. A classifier applied to the imaging observables from a determined brain region in the ventral stream results in a performance level of 0.67. Next, a second similar study is run where the faces have been replaced by letters. The classifier now performs significantly better (0.8) for the same brain region. Are these results evidence that the brain region is "better" at detecting letters than detecting faces? Intuitively, we recognize a problem: the tasks are very different. The distinct possibility remains that the task of discriminating letters and houses is objectively much easier

than faces and houses. If this is the case, is there a way we can account for this discrepancy and effectively control, or normalize, for the differences in task and stimuli?

Here, we propose the use of an ideal observer framework for the realm of neuroimaging and electrophysiology data in order to normalize classifier performance with respect to task difficulty. Just as we must account for the objective difficulty of a task when evaluating and interpreting psychophysical behavioral performance [17–22], we must also frame a neural decoder's performance within the context of the task and the stimuli's available task-relevant perceptual information. The goal of the present work is to use data and analysis from a face versus car discrimination task utilizing two imaging modalities (EEG and fMRI) to illustrate how use of ideal observer analysis allows for direct comparison between tasks, modalities, and subjects. We suggest the use of the efficiency metric, whereby raw observer and classifier performance alike are evaluated relative to the ideal observer yielding a measure of the amount of available information present in the imaging data (see Fig. 1). We hope to show that use of this framework allows the neuroscientist to state where, when, and how critical neural information is propagated, transformed, organized, and, ultimately, used by the observer to make a decision.

2. METHODS

A. Observers

For the EEG sessions, three naïve observers with normal vision from the University of California, Santa Barbara, participated in the study. For the fMRI sessions, three different naïve observers with normal vision from the University of Birmingham participated.

B. Experimental Setup

1. EEG

Observers sat 125 cm from a 19-in. ViewSonic E90f color monitor with a refresh rate of 75 Hz and a resolution of 1024 by 768 pixels. Each image subtended 4.57° of visual angle. The monitor was calibrated with a linear gamma function with a luminance range from 0–50 Cd/m^2.

2. fMRI

Images were displayed using a color and luminance calibrated JVC D-ILA SX21 video projector with a linear gamma function and luminance range from 0–50 Cd/m^2. The stimuli were projected onto a plastic screen situated 65 cm from the observers at the back of the magnet's bore. To view the images, observers looked at a mirror placed above their heads angled 45° to the image surface. Each image spanned 5.13° of visual angle.

C. Stimuli

Both sessions used the same stimuli: 290×290 pixel grayscale images, 12 faces (six frontal view, six 45° profile; courtesy of the Max Planck Institute for Biological Cybernetics face database) and 12 cars (six frontal view, six 45° profile; see Fig. 13 in Appendix A). For the rest of this paper we will designate the set the image belongs to with an i (i=*car* or *face*) and the specific exemplar from the set with a j ($j=1,2,\ldots 12$). Also, vectors will be displayed in boldface. The images are collections of pixel luminance values and thus can be represented as vectors of size P pixels. The images were transformed to the frequency domain using the fast Fourier transform (FFT). The average amplitude spectrum of all images, $\boldsymbol{\rho}$ [Eq. (1)], was used to filter each original image, $\mathbf{s}_{i,j}'$, before transforming back to the spatial domain, resulting in a set of 24 images with identical amplitude spectra but distinct phase information [$\mathbf{s}_{i,j}$; Eq. (2)]:

$$\boldsymbol{\rho} = \frac{\Sigma_{i,j}\|\Im(\mathbf{s}_{i,j}')\|}{24}, \tag{1}$$

$$\mathbf{s}_{i,j} = \Im^{-1}\left(\Im(\mathbf{s}_{i,j}')\frac{\boldsymbol{\rho}}{\|\Im(\mathbf{s}_{i,j}')\|}\right), \tag{2}$$

where $\Im$ and $\Im^{-1}$ are the Fourier and inverse Fourier transforms, respectively. The resulting images had an average RMS contrast of 28.1%, with a minimum of 23.8% and a maximum of 32.7%.

Noise was created by filtering zero-mean white Gaussian noise (σ=3.61Cd/m^2 for an RMS noise contrast of 14.5%) with the common amplitude spectrum [Eq. (3)]and adding the resulting field, $\mathbf{n}$, to the image in the spatial domain:

$$\mathbf{n} = \Im^{-1}[\Im(\mathbf{n}')\boldsymbol{\rho}]. \tag{3}$$

This multiplicative process in the Fourier domain translates to a spatial convolution that introduces correlations between each pixel's noise value. That is, $\mathbf{n} \sim \mathrm{N}(0,\Sigma)$, where Σ is a $P\times P$ covariance matrix. The noise was then added to the image, resulting in a stimulus drawn from a multivariate normal (MVN) distribution, $\mathrm{N}(\mathbf{s},\Sigma)$, where the pixel luminance values of $\mathbf{s}$ shift the mean of the noise distribution.

D. Procedure

1. EEG

On test day, observers completed 100 practice trials immediately preceding the test session. The 1000 trials in the test session (500 car images, 500 face images, order randomly permuted for each observer) were divided into five blocks of 200 trials each, separated by brief breaks for the comfort of the observers. Figure 2 outlines the basic trial sequence. Each trial began with the observer pressing the spacebar while fixating a central cross. After a random amount of time (500–1500 ms) a noisy image of either a face or a car (equally likely) was presented for 40 ms after which a blank screen was shown for 500–1500 ms (see Fig. 2). Mouse movements were prohibited during the stimulus and blank screen presentation. A response screen was then given where the observer had to give a rating on the just-seen stimulus. Using a mouse, the observer could respond with a number from 1–10, with a 1 representing extreme certainty that a face was shown, a 10 representing extreme certainty that a car was shown, and a 5 or 6 representing extreme uncertainty (for a face or a car, respectively). No feedback was provided.

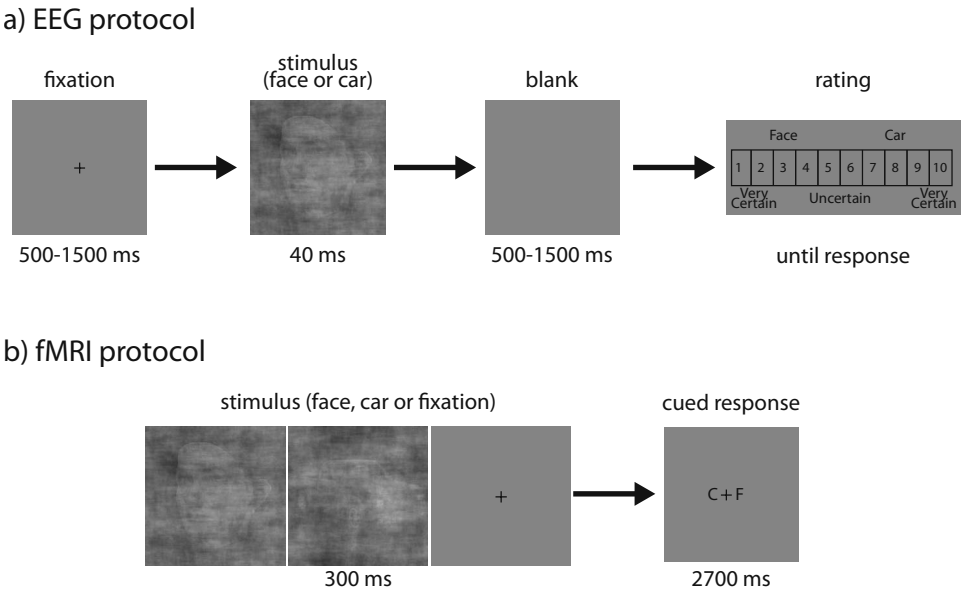

Fig. 2. Task structures for the EEG and fMRI sessions. The images and task were the same, but acquisition concerns required slight modifications to presentation and response. The C and F in the fMRI protocol show the observer which hand (left side of fixation cross means use left hand and vice versa) to use for a car or a face response, respectively.

2. fMRI

The core of the task remained the same, but acquisition considerations prompted some changes to the protocol. The task was still face vs. car discrimination, with each trial matched for history (2-back). A total of 630 images (315 cars, 315 faces) were divided into 7 blocks. Each block consisted of 127 test trials (45 each of cars, faces, and fixations, along with 2 trials at the beginning of the block to equate the history for trials 3 and 4). A 9 s fixation was also added to the start and end of each block. Each trial lasted for 3 s, initiated by a 300 ms stimulus presentation, followed by a 2700 ms response interval. Observers used two button boxes to respond, one for each hand. Each box had four buttons that could be depressed by the observers' corresponding fingers (no thumb). The buttons represented confidence ratings, with the little finger always used for the highest rating and the index finger always used for the lowest rating. During the response time observers were cued as to which hand represented which category: half the time the right hand would be used for a car response while the left hand was used for a face response, and vice versa (see Fig. 2). While rating data were collected for both EEG and fMRI sessions, we collapsed to a binary decision variable for simplicity.

E. Data Acquisition and Preprocessing

1. EEG

Observers were fitted with a 64 channel Ag/AgCl sintered electrode cap in accordance with the International 10/20 System. In addition, external electrodes were placed at the left and right mastoids, 1 cm lateral to the left and right canthii, and above and below each eye. EEG activity was sampled at 512 Hz and preprocessed using the central midline electrode (Cz) as the reference, band-pass filtering with a range of 0.01 to 100 Hz, and excluding trials with blink or eye movement artifacts (as detected by electro-oculogram amplitudes in excess of ±100 mV).

2. fMRI

The experiments were conducted at the Birmingham University Imaging Centre (3T Achieva scanner; Philips, Eindhoven, The Netherlands). EPI and T1-weighted anatomical ($1\times1\times1$ mm) data were collected with an eight channel SENSE head coil. EPI data (Gradient echo-pulse sequences) were acquired from 24 slices (whole brain coverage, TR: 1500 ms, TE: 35 ms, flip-angle: 73 degrees, $2.5\times2.5\times4$ mm resolution) for the main experiment and 32 slices (whole brain coverage, TR: 2000 ms, TE: 35 ms, flip-angle: 80 deg, $2.5\times2.5\times3$ mm resolution) for the localizer scans. MRI data were processed using Brain Voyager QX (Brain Innovations, Maastricht, The Netherlands). T1-weighted anatomical data were used for 3D cortex reconstruction, inflation and flattening. Preprocessing of functional data included slice-scan time correction, head movement correction, temporal high-pass filtering (3 cycles) and removal of linear trends. No spatial smoothing was performed on the functional data. The functional images were aligned to anatomical data and the complete data were transformed into Talairach space at the standard resolution of $3\times3\times3$ mm.

F. Pattern Classification

We used linear discriminant analysis (LDA; specifically, we used the Fisher linear discriminant (FLD); see Appendix A for more information) for classification analysis on both modalities' data sets. Below are details specific to each set.

1. EEG

Input data to the classifier were taken for the time epoch beginning at stimulus presentation through 512 ms poststimulus, yielding 256 time points (see Fig. 3(a)). Each trial thus provided 16,128 independent inputs (256 time points for 63 electrodes) rendering traditional FLD unfeasible. To deal with the high-dimensional data we implemented a Fukunaga–Koontz transform on the Fisher

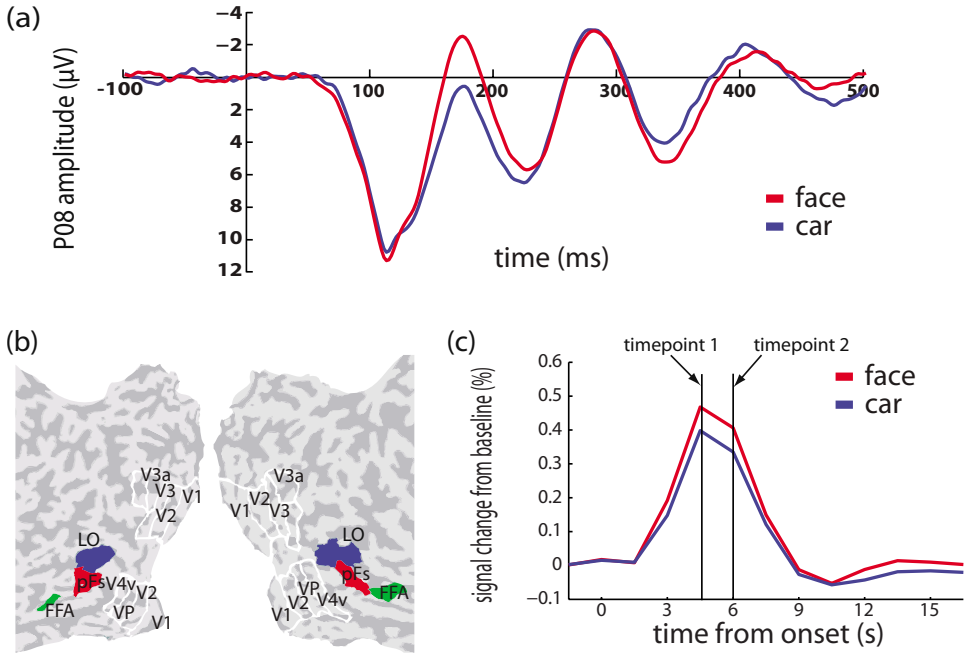

Fig. 3. (Color online) (a) Averaged EEG signal time-locked to stimulus presentation across the three observers for the time epoch of interest showing a classic large negative deflection for faces around 170 ms. (b) Localized regions of interest for one of the observers in the fMRI session. Here, the FFA was selected for further analysis. (c) Prototypical hemodynamic response functions (HRF) from the FFA from one of the observers in the fMRI sessions showing increased response to faces compared to cars. The two time points of interest are provided for reference.

scatter matrices (FLD/FKT) [23]. FKT decomposes the high-dimensional data to four lower-dimensional subspaces based on the eigenvalues of the between-class and within-class scatter matrices (see Appendix A for details). This allowed us to both formulate responses and optimize performance despite singularities in the scatter matrices due to the under-sampling problem that makes simple matrix inversion impossible. We used a ten-fold stratified cross validation, leading to 900 trial training sets and 100 trial test sets. Performance was taken as the average of these ten computations, with the standard error taken across the ten observations.

2. fMRI

We focused our analysis on the fusiform face area (FFA) [24]. We mapped the FFA by selecting voxels in the fusiform gyrus that responded significantly more to faces than images of objects, bodies and scenes ($p<10^{-4}$; see Fig. 3(b)). Observers were shown blocks of 20 images from single categories during which they performed a one-back matching task while maintaining a central fixation. Each image was displayed for 300 ms followed by a 500 ms blank interval. Due to individual differences in FFA size each observer's data set consisted of a unique number of voxels (87, 118, and 114 for observers 1, 2, and 3, respectively). Classification analysis was run on these voxel subsets using the same FLD/FKT algorithm [23]. Input data to the classifier consisted of the raw blood-oxygen-level dependent (BOLD) signal sampled at two time points (3 and 4.5 s post-stimulus in order to account for the hemodynamic response lag; see Fig. 3(c)). We implemented a leave-one-session-out validation. Each 127-trial session was tested by training on the remaining 6 sessions. Performance was taken as the average of these 7 computations, with the standard error taken across the 7 observations.

G. Ideal Observer Analysis

Use of ideal observer analysis entails specifying the task, the set of stimuli, and the external noise properties. In the general case, the stimulus space is defined by a set of possible signals, S, perturbed by a stochastic noise process, n, leading to a limited set of distributions. The ideal observer has full knowledge of the multivariate probability density functions, allowing for an optimal decision rule based on Bayesian inference.

In the task reported here there are 2 possible states, x_i (i=face or car present), with each state comprising 12 possible signals (the individual car and face templates, each with a dimensionality of 84,100 equal to the total number of pixels; $j=1,2,\ldots,12$). Each trial, t, produces an independent observation, the data vector $\mathbf{g}_t$, which is a randomly sampled signal, $\mathbf{s}_{i,j}$, perturbed by a stochastic noise process, $\mathbf{n}$. In this case the noise is additive, such that $\mathbf{g}_t=\mathbf{s}_{i,j}+\mathbf{n}$. The task of the ideal observer is to decide which state of the world is most likely to be true (x_{face} or x_{car}) given the observation. This is accomplished through use of a Bayesian decision rule, whereby the posterior probability for each state, $P(x_i|\mathbf{g}_t)$, is computed and the maximum value is selected as the response [Eq. (4)]:

$$P(x_i|\mathbf{g}_t)=\frac{P(x_i)P(\mathbf{g}_t|x_i)}{P(\mathbf{g}_t)}\Rightarrow P(\mathbf{g}_t|x_i)=\ell_{i,t}. \tag{4}$$

Here, the prior probability, $P(x_i)$, is the same for each state (there is a 0.5 probability for either a car or a face being sampled), and so it can be ignored. The probability

of observing the data, $P(\mathbf{g}_t)$, is also the same and can be discarded. The posterior calculation thus simplifies to finding the likelihood of observing the data given the possible states, $P(\mathbf{g}_t|x_i)$, which we will refer to as $\ell_{i,t}$. Due to signal uncertainty (each state can be represented by any of their respective 12 signals) the ideal observer must compute individual likelihoods for each possible signal, $\mathbf{s}_{i,j}$and sum these values [Eq. (5)]. Again, each signal is equally likely, allowing for the removal of the priors:

$$\ell_{i,t} = \sum_j P(\mathbf{s}_{i,j})P(\mathbf{g}_t|\mathbf{s}_{i,j}) \Rightarrow \sum_j P(\mathbf{g}_t|\mathbf{s}_{i,j}) = \sum_j \ell_{i,j,t}. \quad (5)$$

The likelihoods are calculated in an optimal Bayesian manner. The likelihood function for a MVN distribution is shown in Eq. (6) [25]:

$$\ell_{i,j,t} = \frac{1}{(2\pi)^{P/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{g}_t - \mathbf{s}_{i,j})^T \Sigma^{-1} (\mathbf{g}_t - \mathbf{s}_{i,j})\right). \quad (6)$$

In order for a decision to be ideal, the filter, or template, that we are matching to the observation must be optimal. If the noise were white, the optimal algorithm would be a simple approach where the observed data is cross-correlated with each possible signal (the values from corresponding pixels between the data and a template are multiplied and the resulting values are summed via a matched-filter method) [7]. Filtering the noise in the Fourier domain introduced correlations between the noise values in each pixel, represented by the covariance matrix, Σ. In order to account for these spatial correlations we need to use the correct template. To this end we use a pre-whitened match-filter, which is essentially the inverse of the noise-filtering process (Eq. (7)) [26]. We transform the templates to the Fourier domain, divide by the normalizing filter, and transform back to the spatial domain [27]:

$$\hat{\mathbf{s}}_{i,j} = \Im^{-1}\left(\frac{\Im(\mathbf{s}_{i,j})}{\boldsymbol{\rho}}\right), \quad (7)$$

where the symbol definitions are as described under Methods (Subsection 2.B).

We can now use this pre-whitened template as a match-filter. After some algebraic simplification we arrive at Eq. (8) and Fig. 4,

$$\ell_{i,j,t} = \exp(\mathbf{g}_t{}^T \hat{\mathbf{s}}_{i,j} - 0.5\hat{E}_{i,j}), \quad (8)$$

where $\hat{E}_{i,j} = \hat{\mathbf{s}}_{i,j}{}^T \mathbf{s}_{i,j}$ is the pre-whitened template energy.

H. Calculating Absolute and Relative Efficiency

Ultimately, the ideal observer allows meaningful comparison of human performance across a broad range of psychophysical tasks. The ideal observer makes errors due to the stochastic nature of the noise. In addition to being limited by the same external noise, the assumption is that various inefficiencies inherent to the observer further impair human performance (e.g., retinal sampling density, the use of suboptimal templates, internal neuronal noise, bias, spatial uncertainty, etc.). It is these observer-specific inefficiencies with which we are conerned. A useful metric for analyzing and summarizing performance is absolute efficiency, η_0, defined per Eq. (9) as the squared ratio of the signal contrast used on human trials, C_{human}, to the contrast required to equate the ideal observer's performance with the human's, $C_{IO|human}$ [28]. Here, the signal

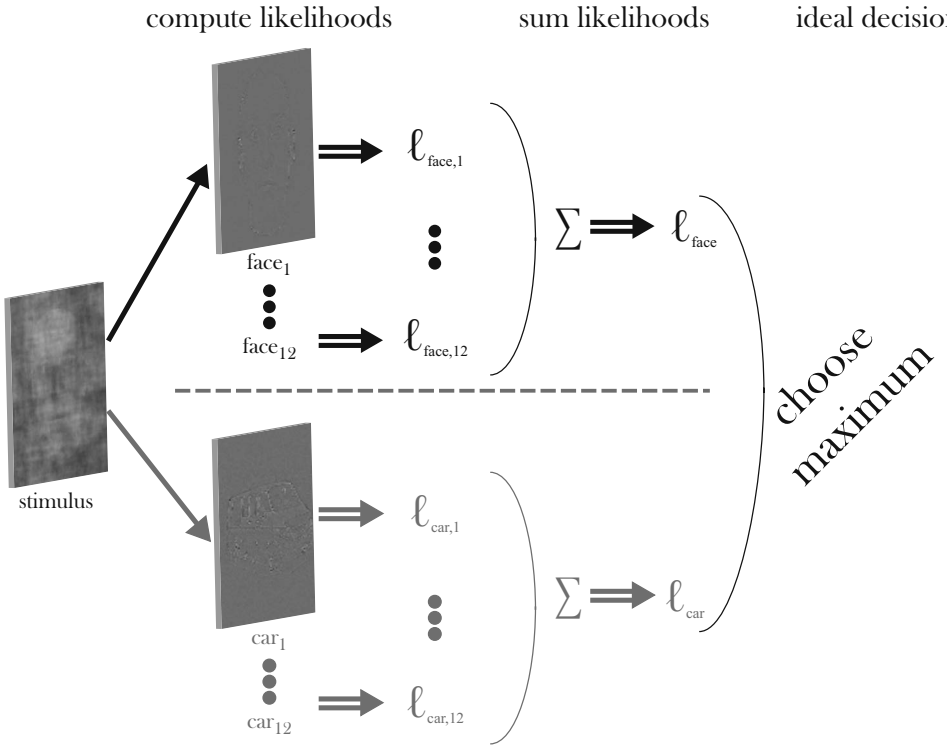

Fig. 4. Ideal observer algorithm. Likelihoods are computed by cross-correlating the stimulus with the possible prewhitened templates and then summing across within category products. The maximum summed likelihood is taken as the decision.

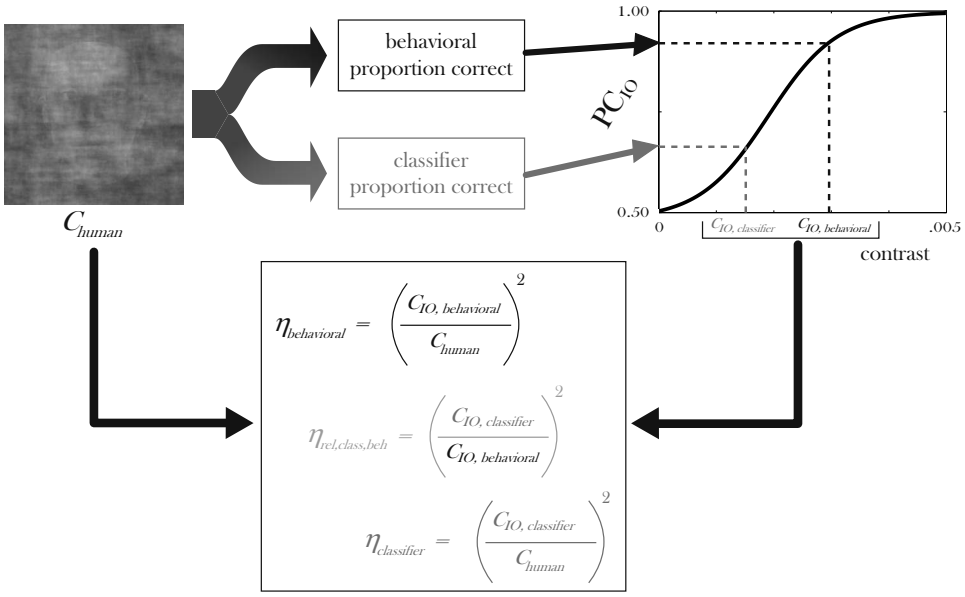

Fig. 5. Absolute and relative efficiencies are calculated by determining the signal contrast that causes the ideal observer to perform at the level of the human or the classifier and then taking the squared ratio of the contrast values.

contrast C is a scalar multiplier applied to the signal before the addition of the filtered noise with a range between 0 and 1. For this purpose we can simulate the ideal observer's performance over a range of signal contrasts and then look up the contrast level that sets the ideal observer's performance to that of the human (Fig. 5). The same analysis can be applied to a pattern classifier's performance [29], providing a measure of how much extractable task-related information relative to the ideal observer is present in the neural data:

$$\eta_{0,human} = \frac{C^2_{IO|human}}{C^2_{human}} \quad \text{or} \quad \eta_{0,classifier} = \frac{C^2_{IO|classifier}}{C^2_{classifier}}. \tag{9}$$

We may ask the question, How well does a pattern classifier, acting on a human's neural activity, perform a task compared to the human's explicit behavioral response? Does this result hold across tasks? To answer these questions we use the relative efficiency metric, η_{rel}, defined as the ratio of the absolute efficiencies for the two situations of interest [Eq. (10) and Fig. 5; see [30] for relative efficiency of saccadic vs. perceptual decisions). In this paper we compute relative efficiencies for a pattern classifier acting on two imaging modalities relative to human behavioral performance:

$$\eta_{rel,cond1,cond2} = \frac{\eta_{0,cond1}}{\eta_{0,cond2}}. \tag{10}$$

3. RESULTS

A. Proportion Correct

Figure 6 shows task performance for three observers in terms of proportion correct. There were some differences in behavioral performance: observer performance was higher on average in the EEG sessions than in the fMRI ($t(15)=2.92$, $p=.01$, two-tailed). This is not surprising, given that the studies utilized two different sets of human observers [31]. Classifier performance for the EEG data were not significantly higher on average than the fMRI [$t(15)=1.72$, $p=0.11$, two-tailed]. Note that observer 3 in the EEG condition performed very well behaviorally, yet the classifier performed very poorly on the associated imaging data. Observer 3 in the fMRI condition clearly outperformed the other two observers behaviorally. The classifier also did well with this observer's imaging data, but the performance gap across observers seems to have diminished. These comparisons motivate the usefulness of a well-defined metric to quantify the extent to which the classifier performance corresponds to the behavioral performance. Observer 3's neural activity (fMRI) contains more information about the stimulus than that from observers 1 and 2, but is it as informative as we would expect given the drastic difference in behavioral performance?

B. Absolute Efficiency

Figure 7 shows the proportion correct performance measures transformed into absolute efficiencies (see Methods, Subsection 2.H). In this case, the ideal observer is identical for both sessions, and thus, at first glance, the efficiency metric seems to offer few benefits over performance

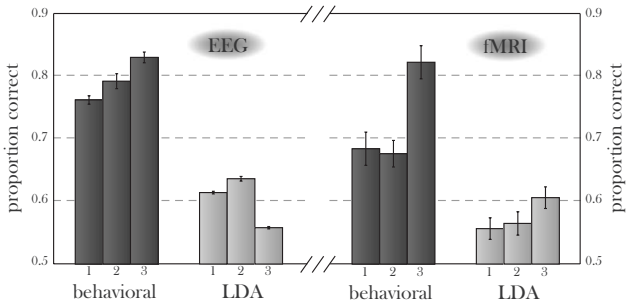

Fig. 6. Task performance, in terms of proportion correct, for each observer's (1, 2, and 3) behavioral and neural classifier decisions. Error bars represent one SEM.

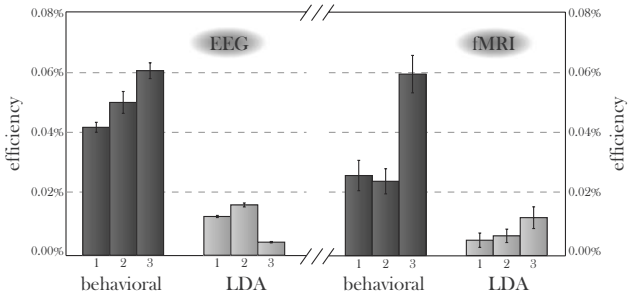

Fig. 7. Performance transformed into efficiency units using ideal observer analysis. Error bars represent one SEM.

measured with proportion correct. Here, the absolute efficiency is mainly used as a step toward the final metric of interest, the relative efficiency. Note, however, that if we were comparing different tasks the absolute efficiency would be useful in itself to disambiguate performance of the classifier from task difficulty.

C. Relative Efficiency

Figure 8 shows relative efficiencies of the two classifiers compared with behavioral performance. This metric quantifies how well a classifier performs with an observer's neural data relative to what we might expect given the behavioral performance. Indeed, some interesting, and perhaps surprising, results become apparent. Next, we will discuss a few situations more closely and see how our interpretations of the results are clarified through the relative efficiency metric.

4. DISCUSSION

A. Why Use Absolute and Relative Efficiency over Proportion Correct?

What knowledge has been gained through these extra computational steps that was not immediately obvious with the basic performance measures based on proportion correct? To answer this question one can compare the efficiency measures across observers in the fMRI sessions. Observer 3 shows the highest behavioral performance. Observer 3 also yields the best classifier results (Fig. 6). Proportion correct measures show a larger gap between observer 3 and the other two participants in behavioral performance relative to the gap in their classifier performances. On the surface this implies that information extracted from the neural activity of observer 3 does not reflect the information used in the behavioral performance as well as with observers 1 and 2. However, use of the relative efficiency (Fig. 8) shows the classifier is able to extract information approximately equally for all three observers.

A more subtle example of the extra information gained from efficiency measures can be seen when comparing observers 1 and 2 in the EEG sessions. Observer 2 performs at a higher level both behaviorally and neurally (Fig. 6). The relative efficiencies show that the higher classifier accuracy can be mostly explained through the higher behavioral performance. Indeed, the ability of the classifier to cull relevant neural information is essentially equivalent, a result that would be hard to notice, much less quantify, without a well-defined metric (Fig. 8).

While we argue that efficiency measures provide more information than can be gained from proportion correct alone, sometimes these metrics point toward equivalent conclusions. Indeed, observer 3 in the EEG condition shows the most striking discrepancy between performance levels, yielding the best behavioral results and the worst classifier accuracy (Fig. 6). Clearly, observer 3's behavior reflects that high-quality visual information (e.g., signal to noise ratio) is present in the brain, yet it is not well-represented in the neural observables. This example illustrates a case in which the same conclusion can be drawn looking solely at basic proportion correct performance. Observer 3's low relative efficiency validates this result (Figs. 6 and 8).

B. Cross-Modality Comparison

Looking at only basic performance measures such as proportion correct, we cannot confidently attribute differences in classifier performance between the EEG and fMRI conditions to the same factors that influenced the behavioral performance disparity. The relative efficiency helps control for factors that influence the behavioral and neural results in a similar manner but are not entirely captured by the ideal observer. For instance, the presentation time in the fMRI sessions was much longer than in the EEG. It would be surprising if this did not affect performance levels. The important point to realize is that this difference in the amount of information entering the brain is common to both the behavioral and classifier decision processes. When classifier performance is normalized for task difficulty, the gap in neural information extraction ability between EEG and fMRI is mitigated dramatically (Fig. 9).

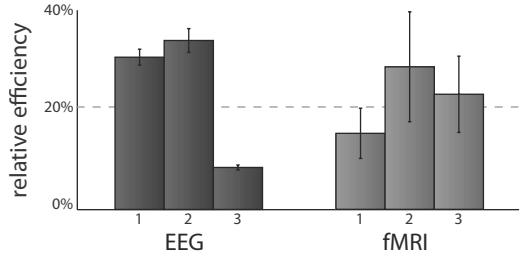

Fig. 8. Relative efficiencies of the classifier compared to behavioral performance. Of note is the especially poor extraction of neural information from observer 3 in the EEG session compared with the behavioral performance. Error bars represent one SEM.

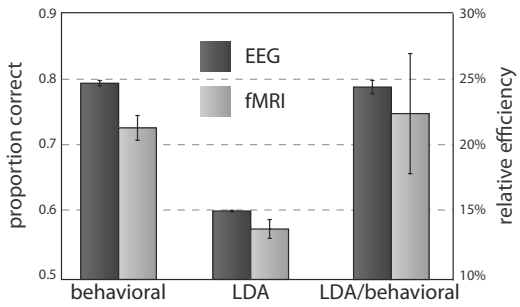

Fig. 9. Mean performances across observers. On the left are the behavioral and classifier proportion correct scores for each modality. On the right are the corresponding relative efficiencies. Error bars represent one SEM.

The conditions also differ with respect to acquisition fidelity, limited by their various sources of noise and signal loss such as aberrations in skin conductance, electrical shielding, head motion, etc., not to mention the distinct relationships between the observables and underlying neural activity. These factors are not accounted for in this treatment, though this should not be viewed as a fundamental limitation of ideal observer analysis. With rigorous environmental, physical, and physiological measurements, many of these inefficiencies could be modeled and implemented [32].

Accounting for a greater proportion of these inefficiencies allows the neuroscientist to focus on the core of many research inquiries: How is information represented, transformed, and organized in the brain? With EEG the classifier has access to data with good temporal resolution at the cost of coarse spatial sampling. Each trial provides 63 data samples, one for each electrode, every 2 ms. We might expect that information carried in the temporal structure of the brain's data processing is represented quite well. fMRI on the other hand has very poor temporal resolution (each trial encompasses only two time points) but superior spatial resolution, with hundreds of voxels providing data. Information contained in relatively small spatial structure can be extracted more easily.

C. Further Applications

Finally, there is a common situation that we do not address here yet is arguably where ideal observer analysis has its most powerful impact. In Figs. 6 and 7 we saw that absolute efficiency did not offer a strong interpretational advantage over raw performance. This is because the ideal observer is exactly the same for both conditions, meaning that the contrast look-up curve in Fig. 5 is used for both data sets. If instead we were to compare observations across different tasks, each with its own unique ideal observer (i.e., to compare performance between this task and a within-category face identification task), the efficiency metric could assist interpretation by normalizing for the differential task requirements and difficulty.

D. Limitations

We conclude this treatment with cautionary discussions of two limitations to the proposed method. The first deals with the possible non-constancy of the efficiency metric across signal contrasts. The second concerns the ability of pattern classifiers acting on imaging data to optimally extract information from neural activity.

1. Dependence of Efficiency on Signal Contrast

In this paper we calculated behavioral and pattern classifier efficiencies for the stimuli at a single contrast level. But how well does this efficiency generalize across other signal strengths? The constancy of efficiency will depend on how human behavioral performance and pattern classifier performance vary across signal contrasts. We can rely on previous work measuring and modeling behavioral performance as a function of signal contrast [30] to elaborate on the dependence of pattern classifier efficiency on signal contrast. Figure 10 shows model simulation results illustrating a situation where efficiency changes as a function of signal contrast, consistent with

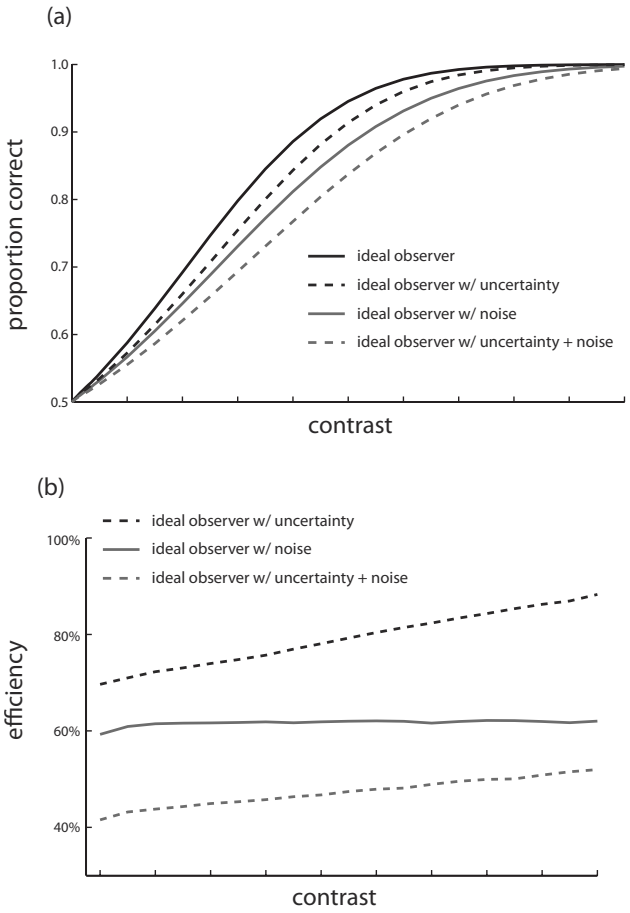

Fig. 10. Ideal observer simulation results for hypothetical situations. (a) Ideal observer performance can be degraded with either additional noise or intrinsic uncertainty as seen in these simulated psychometric curves. Here, uncertainty was created by adding modified versions of the original car and face templates to the signal sets through shifting the images left, right, up, and down by 0.25° visual angle. Noise was zero-mean white Gaussian with an RMS noise contrast equivalent to the image noise at 14.5%. (b) When converted into efficiency units (by comparing the sub-optimal models with the ideal) it is shown that additional noise lowers efficiency but by a constant amount across signal strengths. Meanwhile, non-linear effects, such as those given by added signal uncertainty, lead to an efficiency that varies monotonically with signal strength.

results for human perceptual decisions and saccadic eye movements [30]. Shown are four hypothetical psychometric functions for our car-versus-face discrimination task. The solid black curve represents the familiar ideal observer from Fig. 5. The solid gray curve is the performance of an ideal observer with additional internal noise (white Gaussian noise added to the original noisy image data of equivalent power, 14.5% RMS noise contrast). The dotted black line represents the ideal observer given greater signal uncertainty (in this case, we added spatial uncertainty by creating a model with new signal templates that were the original templates jittered left, right, up and down by 0.25° visual angle). This is meant to mimic human observers' intrinsic spatial uncertainty (e.g., [33,34],). This uncertainty lowers performance, especially at low signal contrasts. The dotted gray curve shows the performance of an ideal observer perturbed by both spatial uncertainty and additional internal noise.

Performance once again is decreased. While uncertainty is used in our model simulations to exemplify its detrimental effect on performance at lower signal contrasts, other non-linearities can have similar effects. The way in which performance of the pattern classifier varies with signal contrast (Fig. 10(a)) will have different consequences on the efficiency measure. Figure 10(b) displays how the shape of the psychometric/neurometric function influences the efficiency metric. A neurometric function that can be modeled as an ideal observer with additive noise leads to a constant efficiency. If the neurometric function can be fit only by the inclusion of non-linearity such as uncertainty, then it will lead to a non-constant loss of efficiency, especially at lower signal strengths. In these cases, pattern classifier efficiency will vary across signal contrasts. Similar arguments can be raised about the relative efficiency: if the psychometric and neurometric functions are modeled with varying degrees of nonlinearities, then their relationship (relative efficiency) will vary across signal contrasts (e.g., see [30] for relative efficiency of saccades vs. eye movements across signal contrasts). In this respect, one thorough approach is to measure absolute and relative efficiencies at various signal contrasts. Another recommendation is to not choose very small signal contrasts at which the non-linearities potentially have the strongest effects on the psychometric and neurometric functions. In the current study observers operated at fairly high signal contrasts and thus theoretically in the regions where efficiencies should be less variant with contrast. Finally, note also that the absolute efficiency measure will also change with external noise amplitude due to the effects of the constant additive internal noise being greatest with low external noise.

2. Classifier Optimality and Information Transduction

We use pattern classifier performance as a measure of the task-specific information content carried in the brain's neural activity. However, at least two factors might limit this interpretation of the pattern classifier performance.

First, there is an underlying assumption that our pattern classifier is optimal at extracting the measured neural activity; however, it may be suboptimal for the data set under study. For instance, linear discriminant analysis (LDA) relies on the data being normally distributed. Yet if the measurements (EEG/fMRI BOLD) yield highly non-Gaussian data, then LDA will result in suboptimal performance that does not entirely capture the task-related information inherent in the neural activity [35,36]. This problem is not restricted to LDA. All pattern classification techniques employ their own sets of assumptions about the form of the data. For instance, imaging data analysis using classifiers is mostly limited to linear mappings, although it can result in optimal separation only if the data distribution follows strict assumptions. Nonlinear transformation should, in principle, enhance the class separability of complex, large multidimensional datasets; however, in practice nonlinear mapping is not widely explored due to the inherent difficulty of solving a nonlinear problem. Nonlinear classification typically results in complex decision boundaries and a high-dimensional decision variable, where low-dimensional features are extracted from the original dataset using a nonlinear mapping. Unfortunately, there is no optimal method to find the intrinsic dimensionality of the dataset that accounts for the underlying data manifold and enhances class separability; hence, various assumptions are imposed on the dataset that involve free parameters requiring computationally intensive parameter searches. Thus, "simple" linear mapping techniques are generally preferred over complex nonlinear algorithms in accordance with the general principle of "Occam's razor" in machine learning [37] which can be informally interpreted as "the simplest explanation is best" [38].

One possible way to mitigate this problem and check the reliability of pattern classifier results is to run multiple classification algorithms and test whether the conclusions change depending on choice of classifier. Figure 11 shows the results of various classifiers (see Appendix A for brief descriptions of each algorithm) including the Fukunaga–Koontz transform (FKT) approach that has been used throughout this study, regularized LDA, support vector machine (SVM), and classwise principal component analysis (CPCA) all acting on the same data. To note, LDA, SVM, and FKT assume linear separability between the two classes (face/car) while CPCA has a piecewise linear class boundary making it a simple nonlinear classifier. Not surprisingly, there are differences in performance across classifiers. The important point is that the pattern of results across the variables being compared (in

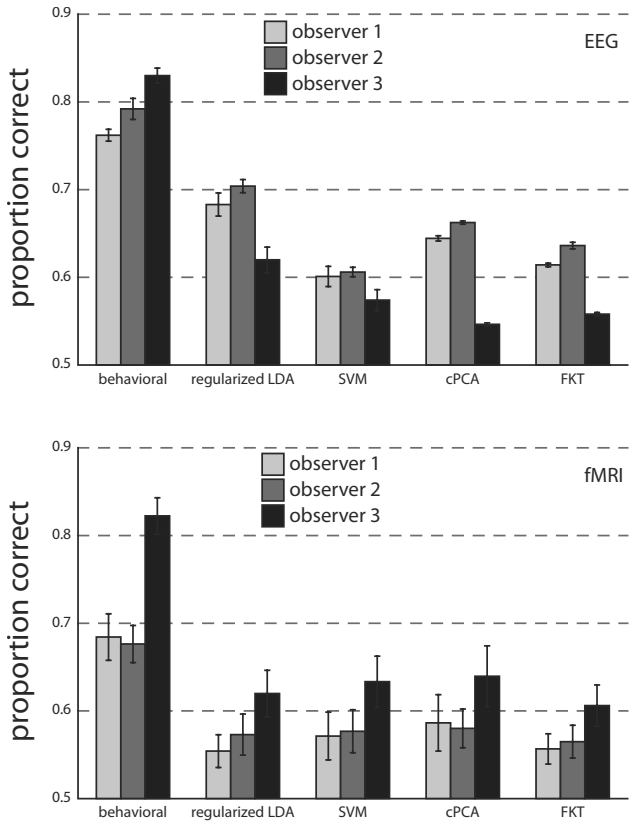

Fig. 11. Behavioral performance is shown alongside four different classifiers (including FKT, which has been used throughout the paper). Overall performance levels vary, but relative patterns remain intact. Error bars represent one SEM.

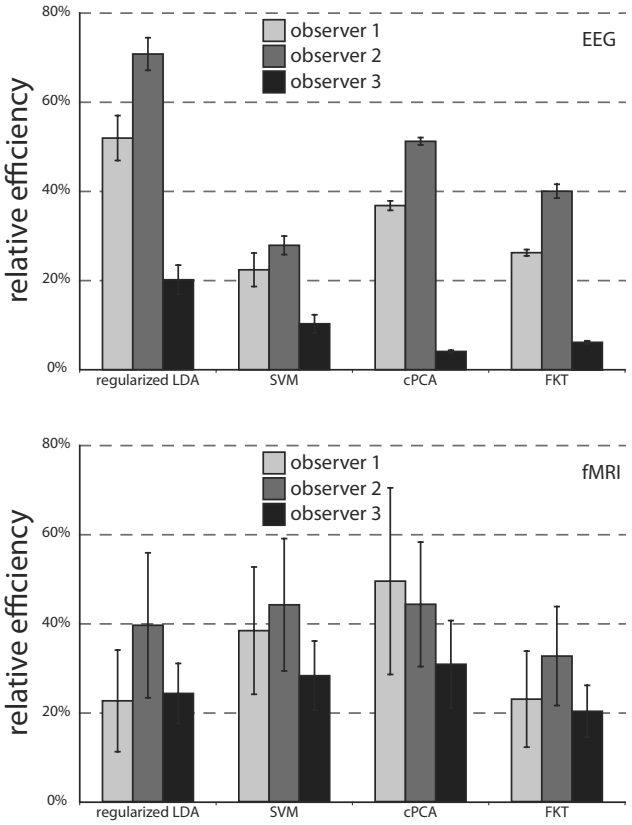

Fig. 12. Relative efficiency of the four classifiers tested. Once again, patterns remain clear except possibly with the CPCA in the fMRI session. This lends strength to the argument that the results are not a product of the choice of classifier. Error bars represent one SEM.

this case, between individual observers) does not change with classifier choice (as can be seen in the relative efficiency metric in Fig. 12). This may not always be the case, and the neuroscientist would be well advised to test multiple algorithms for consistency of results.

The second factor that can invalidate the interpretation that the classifier is measuring an upper bound of task-related information in the brain or brain region is the transduction of neural activity to the imaging observable. The classifier is acting on coarse, indirect measures of neural activity. Each voxel/electrode incorporates activity from many thousands of neurons. While there is a general relationship between neural activity and the imaging observable [39,40], the transduction remains complex and not completely understood [41]. It is very likely that indirect measures of neural activity such as EEG and fMRI do not comprise all the task-related activity in a brain. Furthermore, it is possible that the transduction of the task-related information might vary across brain regions. For example, for fMRI the BOLD responses depend on local hemodynamic properties of the brain as well as the spatial distribution of neurons coding the task-relevant information. Both of these limit interpretations of pattern classifier performance when making comparisons across brain regions. Arguably, however, this is not a limitation that is restricted to pattern classifier analysis but rather to all fMRI and EEG research.

5. CONCLUSION

We have illustrated how use of ideal observer analysis can provide a framework for the comparison of pattern classifier performance across tasks, observers, brain areas, and imaging modalities. Through a series of examples we have shown how pattern classifier performance can be related to individual behavioral performance, using the relative efficiency metric, to compare information content between observers and across imaging modalities (EEG and fMRI) in a meaningful way. Finally, we described current limitations to the approach, including the possible non-constancy of the efficiency metric across signal strengths and a potentially imperfect relationship between classifier performance and underlying neural information content.

APPENDIX A: MATHEMATICAL DESCRIPTION OF PATTERN CLASSIFICATION ALGORITHMS

1. Fisher Linear Discriminant

The Fisher linear discriminant (FLD) is commonly used in classification to find a subspace that maximally separates class patterns according to the Fisher criterion [7]. FLD tries to find a linear transformation matrix $\mathbf{W} \in \Re^{D\times d}$, where D is the size of each data vector and $d < D$, mapping the original high dimensional data into a low-dimensional subspace. From the perspective of pattern classification, FLD aims to find the optimal transformation $\mathbf{W}$ such that the projected data are well separated. This is achieved by simultaneously minimizing the within-class distance and maximizing the between-class distance. In terms of the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter matrix $\mathbf{S}_w$, the Fisher criterion can be written as $J_F = trace[(\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})]$, where

$$\mathbf{S}_b = \sum_{k=1}^{C} n_k \mathbf{m}_k \mathbf{m}_k^T, \tag{A.1}$$

$$\mathbf{S}_w = \sum_{k=1}^{C} \sum_{i} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T. \tag{A.2}$$

Here we assume that the kth class (of C total classes) contains n_k data samples $\{\mathbf{x}_i\}$, and $\mathbf{m}_k$ denotes the class mean.

2. Fukunaga–Koontz Transform

FLD/FKT [41] is a method proposed to optimize the Fisher criterion in a proper way. This is achieved by decomposing the whole data space into four subspaces with different discriminabilities, as measured by eigenvalue ratios. Observe that the scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$ are symmetric and positive semi-definite. Performing eigendecomposition on their sum, we have $\mathbf{S}_b + \mathbf{S}_w = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$, where we retain only the nonzero eigenvalues in the diagonal matrix $\mathbf{D}$ and their corresponding eigenvectors in matrix $\mathbf{U}$. We can now transform the data by using a whitening operator, $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1}$:

$$\mathbf{P}^T(\mathbf{S}_b + \mathbf{S}_w)\mathbf{P} = \tilde{\mathbf{S}}_b + \tilde{\mathbf{S}}_w = \mathbf{I}, \tag{A.3}$$

where $\mathbf{I}$ represents the identity matrix. $\tilde{\mathbf{S}}_b = \mathbf{P}^T\mathbf{S}_b\mathbf{P}$ and $\tilde{\mathbf{S}}_w = \mathbf{P}^T\mathbf{S}_w\mathbf{P}$ share the same set of eigenvectors, and the sum of their associated eigenvalues is one ($\lambda_b + \lambda_w = 1$). By connecting the eigenvalue ratio λ_b/λ_w with the generalized eigenvalue of the Fisher criterion, FLD/FKT shows where the Fisher criterion is maximally satisfied. A more detailed derivation can be found in [41].

3. Regularized Linear Discriminant Analysis

Linear discriminant analysis (LDA) assumes the input data to be normally distributed and homoscedastic (equivalent covariances across classes). The linear weighting vector, $\mathbf{w}$, can be found by inverting the data's covariance matrix, $\mathbf{K}$, and multiplying by the difference of the class means, $\boldsymbol{\mu}_\mathbf{f}$ and $\boldsymbol{\mu}_\mathbf{c}$ for face and car trials, respectively:

$$\mathbf{w} = \mathbf{K}^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_c), \tag{A.4}$$

$$\text{where } \mathbf{K} = \frac{1}{N_t}\sum_{i=1}^{N_t}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \tag{A.5}$$

Here N_t is the number of trials, $\boldsymbol{\mu}$ is the overall mean of the data, and T is the transpose operator. The main difficulty with this technique is operating on the covariance matrix which is a $D \times D$ matrix where D is the dimension of the input data. For the EEG sessions D would be equal to 16,128 (number of time points times the number of electrodes). Furthermore, $N_t \ll D$, such that the data space is sparse compared with its dimensionality. This leads to singularities in the covariance matrix. Both of these factors preclude matrix inversion. In order to reduce the dimensionality and thus the size of the covariance matrix we selected a subset of the 63 electrodes (the 17 that had the highest t-values for one observer). To avoid singularities we then regularized the covariance matrix by adding a diagonal matrix with values proportional to the diagonal elements of $\mathbf{K}$:

$$\mathbf{K}_{\text{reg}} = \mathbf{K} + \text{p}\, Diag(\mathbf{K}), \tag{A.6}$$

where p is a proportional constant set to 2 and the function *Diag* zeros all off-diagonal elements of $\mathbf{K}$. Results did not vary greatly with different values of p (1, 2, 3 or 4). The same procedure was done for the fMRI sessions as well, with dimensionality reduction achieved through selecting only the FFA voxels.

4. Linear Support Vector Machine

Support vector machine (SVM) is a widely used nonparametric machine learning technique [42]. The goal of SVM is to maximize the margin between two classes (e.g., face/car). This is achieved by choosing the hyperplane that maximizes the distance from the hyperplane to the nearest data point. For a dataset of N_t trials $\{x_1, \ldots, x_{Nt}\}$ with class labels $\{L_i = 1, -1\}$, the hyperplane can be expressed as $w^T - b = 0$, subject to the constraint

$$w^T x_i - b \geqslant +1 \quad \text{for } \mathrm{L_i} = +1, \tag{A.7}$$

$$w^T - b \leqslant -1 \quad \text{for } \mathrm{L_i} = -1. \tag{A.8}$$

Equations (A.7) and (A.8) can be rewritten as

$$L_i(w^T x_i - b) \geqslant 1. \tag{A.9}$$

Let H_1 and H_2 be the two hyperplanes satisfying constraints (A.7) and (A.8), and points lying on these hyperplanes are given by H_1: $w^T - b = 1$; H_2: $w^T - b = -1$ respectively. The distance between the two hyperplanes can be computed as $2/\|w\|$. The goal of SVM is reduced to minimizing $2/\|w\|$ subject to the constraint (A.9)

$$\min\left(\frac{2}{\|w\|}\right) \quad \text{s.t. } L_i(w^T x_i - b) \geqslant 1, \quad i = 1 \ldots N_t. \tag{A.10}$$

Equation. (A.10) is an optimization problem and can be solved by using the Lagrange method. Importantly, unlike LDA, SVM makes no assumptions about the data's underlying distribution. Deviations from normality may be well accommodated by this method.

5. Classwise Principal Component Analysis

Classwise principal component analysis (CPCA) is a recently proposed simple, computationally efficient pattern recognition algorithm that can systematically extract useful information from any large-dimensional neural dataset [43]. The technique is based on a class-by-class principal component analysis (PCA), which employs the distribution characteristics of each class to discard noninformative subspace. Feature extraction in CPCA is a two-step procedure, comprising the removal of sparse, non-informative subspaces of the large-dimensional data, followed by a linear combination of the data in the remaining subspace to extract meaningful features for efficient classification. Let ω_i ($i = 1, 2$) denote 2 classes with mean μ_i and covariance Σ_i and let $x^* \in \Re^n$ (n being the data dimension) be the unknown test data to be classified.

In the first step, x^* is projected to two PCA subspaces, S_1 and S_2, by the following transformation:

$$x^*_\mathrm{i} = \mathrm{F_i^T}(x^* - \mu_\mathrm{i}), \mathrm{i} = 1, 2. \tag{A.11}$$

Columns of F_i are taken as basis vectors of S_i. The two classes are also transformed in the similar manner. In the second step, linear feature extraction techniques like LDA can now be used on the already reduced dataspace, S_i, to extract meaningful features without changing the mathematical formalism of Eq. (A.11) and by modifying the definition of F_i. The new F_i can be expressed as $F_i = F_i T_i$, where T_i is the feature extraction matrix of the chosen linear method. We have used a mutual information-based feature extraction technique called approximate information discriminant analysis (AIDA, [44]), whose advantages over LDA are discussed in [44]. In this technique, the strength of PCA as a dimensionality reduction technique is exploited while preserving the class-specific information to facilitate subsequent classification.

6. Multivariate Classifiers Versus Univariate Analysis

We investigated the performance advantage conferred upon the multivariate classification techniques employed

Table 1. Performance Values, in Terms of Proportion Correct, for a Linear Univariate Bayesian Analysis Using the Average BOLD Signal from All Voxels in the ROI as Input and the Four Classifiers Implemented for This Study for the Three fMRI Observers [a]

Observer	Average V oxel	FKT	rLDA	SVM	CPCA
1	0.552±0.009	0.557±0.017	0.554±0.019	0.571±0.027	0.586±0.032
2	0.516±0.018	0.565±0.019*	0.573±0.023*	0.577±0.024*	0.580±0.022*
3	0.556±0.018	0.606±0.024	0.620±0.027*	0.633±0.029*	0.639±0.035*

[a]All performance values are shown alongside their standard errors of the mean. Stars indicate significance at the 0.05 level using a paired t-test.

in this study versus more traditional univariate analyses. In order to quantify the gain in sensitivity we conducted a simple linear, univariate Bayesian analysis on the fMRI data using the same FFA ROI voxels. For each trial we averaged the BOLD activation across all voxels in the ROI. For each session we then estimated the data's mean and variance and calculated expected proportion correct assuming an optimal criterion. The results can be seen in Table 1. Observer 1's results show no significant advantages for any method. Observer 2 shows a strong advantage for all multivariate methods. Observer 3 shows significant advantages for rLDA, SVM and CPCA, with FKT showing a trend toward higher performance.

7. Amplitude Normalized Stimuli
(See Fig. 13.)

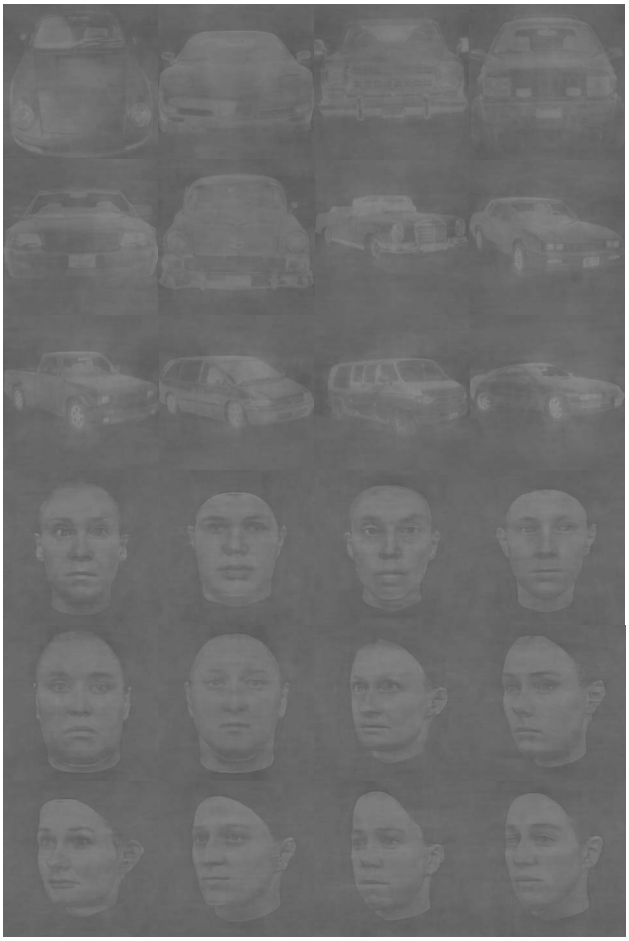

Fig. 13. Complete set of amplitude spectrum-normalized stimuli before addition of the filtered white noise.

ACKNOWLEDGMENTS

Funding for this project was graciously provided by U.S. Army Research Office Institute for Collaborative Strategies (USARO ICB) grant W911NF-09-D-0001.

REFERENCES AND NOTES

1. Further assumptions are well-documented and not the focus of this paper, such as the exact relation between neuronal firing and the observable measurement provided by the specific imaging modality. For our purposes we will assume (somewhat safely, given the large and growing body of evidence) that the imaging observables represent some direct transformation of underlying neural activity (fMRI: [45] below; EEG: [46] below; MEG: [47] below).
2. J. Haynes and G. Rees, "Predicting the orientation of invisible stimuli from activity in human primary visual cortex," Nat. Neurosci. **8**, 686–691 (2005).
3. H. Barlow, "Single units and sensation: A neuron doctrine for perceptual psychology?" Perception **1**, 371–394 (1972).
4. K. Britten, M. Shadlen, W. Newsome, and J. Movshon, "The analysis of visual motion: a comparison of neuronal and psychophysical performance," J. Neurosci. **12**, 4745–4765 (1992).
5. J. Haxby, M. Gobbini, M. Furey, A. Ishai, J. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," Science **293**, 2425–2430 (2001).
6. With spatially coarse measures such as fMRI and ERP the local interactions are pooled together and lost. However, long-range connections can be monitored between the large groups of neurons represented in each voxel or electrode.
7. R. Duda, P. Hart, and D. Storke, *Pattern Classification* (Wiley, 2001).
8. Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," Nat. Neurosci. **8**, 679–685 (2005).
9. J. Haynes and G. Rees, "Decoding mental states from brain activity in humans," Nat. Rev. Neurosci. **7**, 523–534 (2006).
10. D. Ostwald, J. Lam, S. Li, and Z. Kourtzi, "Neural coding of global form in the human visual cortex," J. Neurophysiol. **99**, 2456–2469 (2008).
11. T. Preston, S. Li, Z. Kourtzi, and A. Welchman, "Multivoxel pattern selectivity for perceptually relevant binocular disparities in the human brain," J. Neurosci. **28**, 11315–11327 (2008).
12. L. Pessoa and S. Padmala, "Quantitative prediction of perceptual decisions during near-threshold fear detection," Proc. Natl. Acad. Sci. U.S.A. **102**, 5612–5617 (2005).
13. M. Philiastides and P. Sajda, "Temporal characterization of the neural correlates of perceptual decision making in the human brain," Cereb. Cortex **16**, 509–518 (2006).
14. T. Donner, M. Siegel, R. Oostenveld, P. Fries, M. Bauer, and A. Engel, "Population activity in the human dorsal pathway predicts the accuracy of visual motion detection," J. Neurophysiol. **98**, 345–359 (2007).
15. R. Ratcliff, M. Philiastides, and P. Sajda, "Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG," Proc. Natl. Acad. Sci. U.S.A. **106**, 6539–6544 (2009).

16. S. Li, S. Mayhew, and Z. Kourtzi, "Learning shapes the representation of behavioral choice in the human brain," Neuron **62**, 441–452 (2009).
17. D. Green and J. Swets, *Signal Detection Theory and Psychophysics* (Wiley, 1966).
18. J. Solomon and D. Pelli, "The visual filter mediating letter identification," Nature (London) **369**, 395–397 (1994).
19. B. Tjan, W. Braje, G. Legge, and D. Kersten, "Human efficiency for recognizing 3-D objects in luminance noise," Vision Res. **35**, 3053–3068 (1995).
20. J. Gold, P. Bennett, and A. Sekuler, "Identification of bandpass filtered letters and faces by human and ideal observers," Vision Res. **39**, 3537–3560 (1999).
21. W. Geisler, "Ideal observer analysis," in *The Visual Neurosciences*, L. Chalupa and J. Werner, eds. (MIT Press, 2003), pp. 825–837.
22. J. Gold, D. Tadin, S. Cook, and R. Blake, "The efficiency of biological motion perception," Percept. Psychophys. **70**, 88–95 (2008).
23. Z. Shang and T. Sim, "When Fisher meets Fukunaga-Koontz: A new look at linear discriminants," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2006), pp. 323–329.
24. N. Kanwisher, J. McDermott, and M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," J. Neurosci. **17**, 4302–4311 (1997).
25. W. Peterson, T. Birdsall, and W. Fox, "The theory of signal detectability," IEEE Trans. Inf. Theory **4**, 171–212 (1954).
26. R. McDonough and A. Whalen, *Detection of Signals in Noise* (Academic, 1995).
27. This amounts to a spatial deconvolution, or, equivalently, multiplying by the inverse of the covariance matrix per Eq. (6).
28. H. Barlow, "The absolute efficiency of perceptual decisions," Philos. Trans. R. Soc. London Ser. B **290**, 71–91 (1980).
29. $C_{classifier}$ is equivalent to C_{human} as the human's behavioral response and brain activation are derived from the same displayed stimulus.
30. M. Eckstein, B. Beutter, and L. Stone, "Quantifying the performance limits of human saccadic targeting during visual search," Perception **30**, 1389–1401 (2001).
31. In addition, there were also procedural differences: stimulus presentation time, display luminance, image size on the retina (4.57° vs. 5.13°). However, given that the task is limited by external noise, it is likely that these procedural differences might result in a smaller performance difference than one might expect from noiseless displays.
32. W. Geisler, "Sequential ideal-observer analysis of visual discriminations," Psychol. Rev. **96**, 267–314 (1989).
33. Y. Zhang, B. Pham, and M. Eckstein, "The effect of nonlinear human visual system components on performance of a channelized Hotelling observer in structured backgrounds," IEEE Trans. Med. Imaging **25**, 1348–1362 (2006).
34. B. Tjan and A. Nandy, "Classification images with uncertainty," J. Vision **6**, 387–413 (2006).
35. Here, the imaging data can be thought of as coming from two classes, face present and car present. On top of the category-specific activity is the imaging noise which has been shown to be highly Gaussian for fMRI data [36]. Thus, the data themselves can be thought of as a Gaussian noise process with a mean displaced by the category-specific activation.
36. C. Chen, C. Tyler, and H. Baseler, "Statistical properties of BOLD magnetic resonance activity in the human brain," Neuroimage **20**, 1096–1109 (2003).
37. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Occam's Razor," Inf. Process. Lett. **24**, 377–380 (1987).
38. T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, 1991).
39. D. Heeger and D. Ress, "What does fMRI tell us about neuronal activity?" Nat. Rev. Neurosci. **3**, 142–151 (2002).
40. G. Boynton, S. Engel, G. Glover, and D. Heeger, "Linear systems analysis of functional magnetic resonance imaging in Human V1," J. Neurosci. **16**, 4207–4221 (1996).
41. S. Zhang and T. Sim, "Discriminant subspace analysis: a Fukunaga-Koontz approach," IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1732–1745 (2007).
42. V. Vapnik, *Statistical Learning Theory* (Wiley, 1998).
43. K. Das and Z. Nenadic, "An efficient discriminant-based solution for small sample size problem," Pattern Recogn. Lett. **42**, 857–866 (2009).
44. K. Das and Z. Nenadic, "Approximate information discriminant analysis: a computationally simple heteroscedastic feature extraction technique," Pattern Recogn. Lett. **41**, 1565–1574 (2008).
45. N. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," Nature (London) **412**, 150–157 (2001).
46. P. Nunez, *Electrical Fields of the Brain: the Neurophysics of EEG* (Oxford University Press, 1981).
47. Y. Okada, "Neurogenesis of evoked magnetic fields," in *Biomagnetism: an Interdisciplinary Approach*, S. Williamson, ed. (Plenum, 1983), pp. 399–408.