Bayesian data fusion and credit assignment in vision and fMRI data analysis

Paul R. Schrater

[a]Departments of Psychology and Computer Science & Engineering, University of Minnesota, N218 Elliott Hall, 75 E. River Rd., Minneapolis, MN, U.S.A.

ABSTRACT

One of the most important challenges in understanding expert perception is in determining what information in a complex scene is most valuable (reliable) for a particular task, and how experts learn to exploit it. For the task of parameter estimation given multiple independent sources of data, Bayesian data fusion provides a solution to this problem that involves promoting data to a common parameter space and combining cues weighted by their reliabilities. However, for classification tasks this approach needs to be modified to find the information that most reliably *distinguishes* between the categories. In this paper we discuss solutions to the problem of determining the task-dependent reliability of data sources both objectively for a Bayesian decision agent, and in terms of the reliability assigned by a human observer from the performance of the observer. Modeling observers as Bayesian decision agents, solutions can be construed as a process of assigning credit to data sources based on their contribution to task performance. Applications of this approach to human perceptual data and the analysis of fMRI data will be presented.

Keywords: Bayesian data fusion, feature selection, human perception

1. INTRODUCTION

In performing a visual task the human brain needs to integrate and fuse the sensory information available. There are several distinct ways this integration can occur, across space, across time, across qualitatively different cues, and across modalities (e.g. visual vs. touch). A general cue integration model that can successfully predict these variations for a number of experimental tasks have been developed, refined, and extended by several investigators[1–5] based on probabilistic models of perceptual inference. While early data fusion models in perception were not explicitly described in probabilistic terms, all existing models can be recast in a straightforward way as statistical inference problems,[2,4] which has the benefit of allowing us to formulate, test, and relate data fusion models using the common language of Bayesian probability theory.

Bayesian probability theory has a long history of successful applications to a range of problems in computer vision as well as modeling human visual inference (for a review, see[6]). The ideal Bayesian observer bases its decisions on knowledge of the posterior probability $p(S|I)$, where S is the quantity to be estimated (shape, depth, motion direction, etc.), and I is a collection of image data. The posterior is given by:

$$p(S|I) = p(I|S)p(S)/p(I) \tag{1}$$

where $p(I)$ is a constant for a fixed set of image data. Thus the posterior is proportional to two terms,the generative model $p(I|S)$ which is a probabilistic model describing the image ensemble that can be produced from a scene description S, and the prior model $p(S)$ that describes the observers prior knowledge about the relative frequency of occurrence of the scene descriptors. It is important to note that I is assumed to be a high-dimensional vector of image features $\vec{f}$, where the features are functions of the image intensities (e.g. filter outputs, raw pixel intensities, edge estimates, etc.). For detection, discrimination, and recognition problems, S indexes the discrete categories, but S can be high dimensional vectors for estimation problems.

Data fusion rules that produce minimal estimation error arise naturally from the probabilistic combination of information inherent in equation 1. Different rules result depending on how the posterior distribution factors

Send correspondence to: E-mail: schrater@umn.edu, Telephone: (612) 626-8638

over the image features and on the particular form of the probability distributions. One important special case, termed weak fusion, results when the image features are conditionally independent given S:

$$p(I|S) = \prod_{i=1}^{N} p(\vec{f_i}|S)$$

and the image data are deterministic functions of the scene parameters with added gaussian noise. In this case optimal fusion produces an estimate of the scene parameters that is equivalent to a weighted linear combination of the best estimates of the scene parameters, weighted by the uncertainty of the estimates:

$$\hat{S} = w_0 b + \sum_i w_i s_i$$

where

$$w_i = \sigma_i^2 / \sum \sigma_i^2$$

and σ_i^2 denotes the uncertainty in s_i and $w_0 b$ is the bias b weighted by its relative uncertainty w_0 contributed by the prior $p(S)$. It is important to note that the weak fusion model can be a reasonable approximation with non-gaussian distributions and weak dependence. In these cases, σ_i^2 are replaced by the Fisher Information for each feature evaluated at the maximum likelihood estimates for S. In general, optimal fusion rule results in estimates that are non-linear functions of the image data. These fusion rules are termed strong fusion and have been empirically observed.[7,8]

Despite this difference, both kinds of fusion share some qualitative behavior, including the idea that the more reliable image data should provide a larger contribution to the best estimate. Conversely, data that is unrelated to the scene parameters are unreliable and provide no contribution. Moreover, for a given set of image data and generative model $p(I|S)$ the reliabilities will appropriately vary with changes in scene parameters.

While the idea of feature reliability is fairly straightforward in weak fusion models, this notion is limited in two important respects. First, this definition involves the reliability of a feature given a particular value of the parameter. However, in many tasks the most important image information for the task is not necessarily the most reliable for any particular parameter values. Consider a potentially difficult task, like distinguishing identical twins. The most important image features for distinguishing between the twins will be very different than the most important image features for recognizing the twins in a crowded room. A graphical illustration of task dependence is shown in figure 1. Second, notions of feature reliability from weak fusion models provides only prescriptive information. A related but different problem is to determine the effective feature reliability assigned by observers in a task. In this paper we will discuss a general framework that provides solutions to these problems, and discuss applications of these ideas in analyzing motion perception and fMRI data.

2. PROBLEM FORMULATION

The basic problem is to determine the relative contribution of features to success on a particular *task*. In order to describe the task importance of a feature, we need to expand our previous discussion of Bayesian decision theory to include Loss functions $\lambda(a; s)$ that measure the cost of making an action a when the state of the world is s. For perceptual decisions, actions largely involve decisions about the true state of the world, hence the loss function can be written $\lambda(\hat{s}; s)$, where $\hat{s}$ is the observer's state decision (guess) about S. For the case of discrete S, the optimal Bayesian decision involves finding $\hat{s}$ with the minimal expected cost:

$$\hat{s} = \arg\max_{\hat{s}} \sum_{s} \lambda(\hat{s}; s) p(s|\vec{f_1}, \ldots, \vec{f_N}) \tag{2}$$

A principled solution to determining the relative importance of features for perceptual decisions is to compute the amount of information each feature carries about $\hat{s}$. The amount of information (in nats) each feature $\vec{f_j}$ carries about $\hat{s}$ is given by:

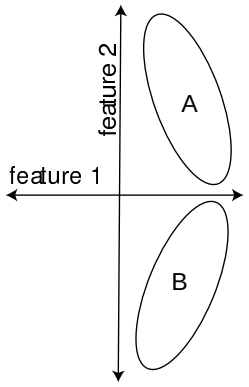


Figure 1. Illustration of the dependence of feature reliability on the task. Figure depicts constant probability ellipsoids for two classes, A & B, in a 2-D feature space. For the task of task of detecting the presence of A or B in background noise, feature 1 is most reliable and feature2 provides no information. For the task of distiguishing A from B, feature 1 provides no information and feature 2 is most reliable.

$$MI(\hat{s};\vec{f_j}) = \sum_{\hat{s}} \int_{\vec{f_j}} p(\hat{s},\vec{f_j}) \ln \frac{p(\hat{s},\vec{f_j})}{p(\hat{s})p(\vec{f_j})} = \left\langle \ln \frac{p(\hat{s}|\vec{f_j})}{p(\hat{s})} \right\rangle_{p(\hat{s},\vec{f_j})} \tag{3}$$

which induces an ordering on the informativeness of each feature for the perceptual decision. The most informative features can also be interpreted as the most important features to consider when guessing an optimal decision agent's choices.

Computing the mutual information between features and class choices is useful both to theoretically determine the objective importance of features for an ideal observer, and to determine the determine the relative impact of features on the recorded decisions of an actual observer. Although it is possible to solve in special cases, equation 3 is typically intractable to solve in the ideal observer context, due to the complexity of computing $p(\hat{s},\vec{f_j})$. However, $\hat{s}$ is a non-linear function of s (equation 2), thus by the Data Processing Theorem,[9] $MI(\hat{s};\vec{f_j}) \leq MI(s;\vec{f_j})$, which suggests a solution to the problem for sets of features. In general, if we pick the best feature set using $MI(s;\vec{f_j})$, then it will also be the best feature set for $MI(\hat{s};\vec{f_j})$, but the overall importance of any given feature (and hence the overall ordering) will change with changes to the loss function. In general, the importance of a feature for $\hat{s}$ depends both on its relative informativeness for inferring particular classes s weighted by the importance of those classes to the decision as specified by the loss function. For simplicity, in the rest of the paper we will restrict our attention to the case in which costs of all errors are equal.

The distinction between $MI(\hat{s};\vec{f_j})$ and $MI(s;\vec{f_j})$ is important to understand the difference between the problem of determining the objectively best features for a task and the problem of determining what information is actually used by an observer in a task. To analyze the objective importance of features for a task, we can focus on computing $MI(s;\vec{f_j})$. To analyze what information is used from human response data, we are given the set of responses, and thus we are forced to consider $MI(\hat{s};\vec{f_j})$. Because the information the observer has about the classes s is reduced when the observer makes a decision $\hat{s}$, an observer's responses on any particular task do not provide a complete picture of the amount of class information extracted by the observer. Thus, the practice of estimating the reliability of a cue from the observer's single cue discrimination performance should result in an underestimate, because it neglects the information lost due to the decision.

3. PROBLEM SOLUTION

In general, we do not know $p(s,\vec{f})$ or $p(\hat{s},\vec{f})$. However, we assume the availability of a *sample distribution* provided by the collection of images labeled by either the true or observer estimated classes. This is clearly true for behavioral experiments, but it also accounts for the case in which we have a ensemble of images produced,

for instance via computer graphics algorithms. The sample distribution of N_{ex} examples $(s_j, \vec{f_j})$ is the sum of Dirac delta functions:

$$p_{sample}(s, \vec{f}) = \frac{1}{N_{ex}} \sum_j \delta(\vec{f} - \vec{f_j}.s - s_j) \tag{4}$$

Let c denote the class labels either objectively assigned (e.g. s) or observer assigned (e.g. $\hat{s}$). For any given sample distribution, we can compute the mutual information between features and class labels as:

$$\begin{aligned} MI(c; \vec{f}) &= \sum_c \int_{\vec{f}} p_{sample}(c, \vec{f}) \ln \frac{p(c, \vec{f})}{p(c)p(\vec{f})} \\ &= \sum_c \int_{\vec{f}} \frac{1}{N_{ex}} \sum_j \delta(\vec{f} - \vec{f_j}.s - s_j) \ln \frac{p(c, \vec{f})}{p(c)p(\vec{f})} \\ &= \frac{1}{N_{ex}} \sum_j \ln p(c_j|\vec{f_j}) - \sum_c p(c) \ln p(c) \end{aligned} \tag{5}$$

The first term in equation 5 is the total likelihood of the data, while the second term is the entropy of the labels.

Unless the distribution factors over the components of $\vec{f}$, the most important features are still difficult to compute. Although generally suboptimal a simple, useful and interpretable way of assessing feature importance is to find a set of r linear weightings $\vec{w}_k$ of the feature vector $\vec{f}$ that capture the majority of the information that the m-dimensional vector $\vec{f}$ conveys about c. To find the best feature, we simply maximize the mutual information over w. When the mutual information in computed using a sample distribution, finding the best $\hat{w}$ is equivalent to maximizing the log-likelihood in equation 5 over w:

$$\begin{aligned} \hat{w} &= \arg\max_w \frac{1}{N_{ex}} \sum_j \ln p(c_j|w^T \vec{f_j}) - \sum_c p(c) \ln p(c) \\ &= \arg\max_w \frac{1}{N_{ex}} \sum_j \ln p(c_j|w^T \vec{f_j}) \end{aligned} \tag{6}$$

In general, this maximization needs to be performed numerically, however, explicit solutions are available for special cases (see below). Clearly, in order to evaluate the objectively best features and to avoid overfitting we need to evaluate equation 6 over many sample distribution. Nevertheless, the sampling approach produces a major reduction in the complexity of the problem and makes it applicable to both theoretical and empirical problems. Although equation 6 is expressed in terms of a single maximization, it is possible to extend this approach to multiple features using methods like exploratory projection pursuit[10] in which the data is transformed to remove any difference in the $\hat{w}$ direction or subspace removal (see below) in which the $\hat{w}$ direction in the feature space is removed via projection.

3.1. Two class discrimination

Consider the case of a two-class discrimination. Then the optimal decision given the feature weighting is to choose :

$$\hat{s} = s_1 \qquad \text{if} \qquad \frac{p(w^T \vec{f}|s_1)}{p(w^T \vec{f}|s_2)} > \frac{\lambda(\hat{s}_1; s_2) - \lambda(\hat{s}_2; s_2)}{\lambda(\hat{s}_2; s_1) - \lambda(\hat{s}_1; s_1)} \frac{p(s_2)}{p(s_1)} \tag{7}$$

$$\hat{s} = s_2 \quad \text{otherwise} \tag{8}$$

In this case, irrespective of the priors and loss function, maximizing the mutual information between $w^T \vec{f}$ and $\hat{s}$ is equivalent to maximizing:

$$\frac{1}{N_{ex}} \sum_j \ln \frac{p(w^T \vec{f_j}|s_1)}{p(w^T \vec{f_j}|s_2)} \tag{9}$$

over w.

When the two class densities are (or can be approximated as) m-dimensional Gaussian distributions, with different means $\{\vec{\mu}_1, \vec{\mu}_2\}$ but a common covariance, C, then finding a set of r $\vec{w}_k$ with $r < m$ that maximize equation 9 is equivalent to Fisher's Linear Discriminant Analysis (LDA).[11–13] In LDA, the p-best feature weights $\{\vec{w}_1, \ldots, \vec{w}_p\}$ are given by the top p eigenvectors of the matrix $W^{-1}B$, where W is the within class and B the between class covariance matrices. The best w is given by:

$$w = C^{-1}(\vec{\mu}_1 - \vec{\mu}_2) \tag{10}$$

This result has a simple relationship to the weak data fusion result of combining features according to the inverse variance of the estimate derived from s. In this context, s can be identified as the decision variable that will be compared to a threshold. For a given feature vector $\vec{f}$, the decision variable $d = \vec{w}^T\vec{f} = \vec{\mu}_1^T C^{-T}\vec{f} - \vec{\mu}_2^T C^{-T}\vec{f}$. When C is diagonal, this becomes:

$$\begin{aligned} d &= \vec{\mu}_1^T C^{-T}\vec{f} - \vec{\mu}_2^T C^{-T}\vec{f} && (11)\\ &= \sum_{j=1}^{m}(\frac{1}{\sigma_j^2}f_j)\mu_{1j} - (\frac{1}{\sigma_j^2}f_j)\mu_{2j} && (12)\\ &= d_1 - d_2 && (13) \end{aligned}$$

Thus the optimal discriminant can be thought of as combining the features into d_i separately for each class according to their reliability and then computing the difference between the d_i. This result fits the intuition given above. The solution for the optimal discriminant features will involve re-weighting the features that are most informative for each class. However, notice that the shared features will cancel in the above expression, demonstrating that the best features for distinguishing classes are not necessarily the best features for a class considered alone.

The case of unequal class covariance is more difficult. The optimal Bayesian decision results in decision boundaries that are hyperquadrics,[14] thus any single linear weighting is suboptimal. Nevertheless, a collection of $\vec{w}_j$ are a useful reduced dimensional summary of the relevant feature information. Recently, approaches to finding $\vec{w}_j$ for more general class probabilities have appeared.[11, 12, 15]

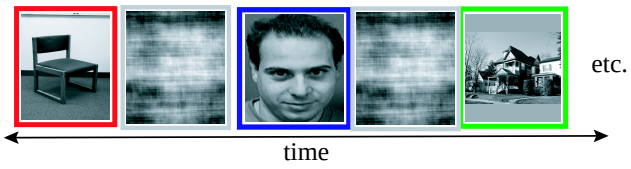
In the next two sections, we apply special cases of this approach to the analysis of fMRI data and to the analysis of human response data in a motion detection task.

4. FMRI DATA ANALYSIS

Recently, my colleagues and I have investigated the problem of finding the most the patterns of brain activity that are most informative about the observer's perceptual state.[16] One of the key problems in the analysis of brain imaging data is a version of the credit assignment problem: what brain activity deserves credit for supporting some behavior or brain function? The difficulty is that whenever a task like object recognition is performed, many areas of the brain that are not necessary for the task are active. Moreover, the observed activity results from many processes not critical to co-occurring but not directly tied to object recognition, including the category independent responses to the individual image features, the decision process, the motor response, and non task-specific activity resulting from daydreaming, bodily discomfort, etc.

Our solution to this problem is to treat the volume of BOLD response data at each scan time as a point in a high-dimensional feature space we term *activity space* that is an indirect measure of the neuronal activity in the brain at each time. Given this feature space, we solve the credit assignment problem for object recognition by finding weighted combinations of features (i.e. patterns of activity) that are most informative about the observer's object recognition state by maximizing the mutual information between the brain activation feature vectors and the object category labels. If the experimental paradigm involves stimuli with large variations in the images from each object category, then we have a reasonable assurance of extracting patterns of activity that are reliably related to the abstract categories rather than particular exemplars of that category.

We modeled the distribution of brain activity given an object category s as a multi-dimensional Gaussian distribution: $p(\vec{b}|s) = N(\vec{b}; \mu_s, C_s)$ with mean μ_s and covariance C_s. If we assume the class covariance matrices



Object vs. Object Classification Performance

	Chairs vs. Faces	Chairs vs. Houses	Faces vs. Houses
Delayed Matching	80.3% (2.4)	73.2% (2.5)	76.3% (2.5)
Passive Viewing	79.2% (2.4)	74.9% (2.4)	73.6% (2.4)

Object vs. Noise Classification Performance

	Chairs vs. Noise	Faces vs. Noise	Houses vs. Noise
Delayed Matching	98.7% (1.3)	97.4% (1.4)	99.3 (1.0)
Passive Viewing	86.3% (2.2)	84.3% (2.3)	84.6 (2.3)

Figure 2. Top: Illustration of the paradigm. Data from.[17] Observers observed photographs of three object categories, faces, houses, and chairs while their brains were imaged using fMRI. Photographs of one category were shown for 21 sec druing which time 7 complete brain volumes (time slices) were acquired. Subjects performed one of two tasks, a delayed matching and passive viewing task. Stimuli presented during passive viewing were presented at a rate of 2 photograpghs/second. In the delayed matching task, target stimuli were presented for 1.5 seconds, then after delay of .5 seconds, were presented two alternative stimuli for 2 seconds. **Table:** The table shows the prediction error (error for Bayesian classification performed on excluded data points along the best pairwise axis in activity space derived from LDA). Prediction error was evaluated via the 632+ bootstrap procedure[18]

are similar, then as mentioned previously, the solution is LDA. Direct application of LDA to fMRI is not feasible, however. The problem is that LDA requires there to be more data points than the number of feature dimensions minus the number of classes. Unfortunately, fMRI data typically involves 10-12k voxels, and only a few hundred data points. Thus, we first performed dimensionality reduction using a robust form of PCA, and performed LDA in this reduced dimensional space.

We used LDA to re-analyze a set of data originally published by Ishai and her colleagues (Ishai, A., Ungerleider, L. G., Martin, A., Haxby, J. V. (2000) The Representation of Objects in the Human Occipital and Temporal Cortex. Journal of Cognitive Neuroscience, U.S.A., 12 Supplement 2, pp. 35-51), publicly available from fMRIDC (accession no. 2-2000-1113D). The data set consists of fMRI recordings from 6 subjects while they viewed photographs of objects from 3 general categories of objects (faces, houses, and chairs) while performing two different tasks, passive viewing, and a delayed matching task that forced observers to attend to the stimuli presented (see figure 3). We derived activation patterns that best predict the object category viewed at each time point, both relative a particular other object category (e.g. faces vs. houses) or relative to all the other categories (e.g. houses vs. faces & chairs).

In order to avoid overfitting, the generality of the brain activation patterns was assessed by computing the predictive error using a cross-validation resampling procedure. Prediction error for several types of classification are shown in the table in figure 3. One of the interesting aspects of these results is the large effect of task demand (delayed matching task vs. passive viewing) for all three categories of objects. This is contrasted with the lack of a task demand effect on the object vs. object predictions. In addition, unlike object vs. object activation patterns, object vs. noise activation patterns were essentially exchangeable: any object vs. noise activation pattern could be used to successfully distinguish the other objects from noise. Furthermore, the locus of the activation patterns was in the earliest visual areas, like V1. These and other results from this analysis strongly

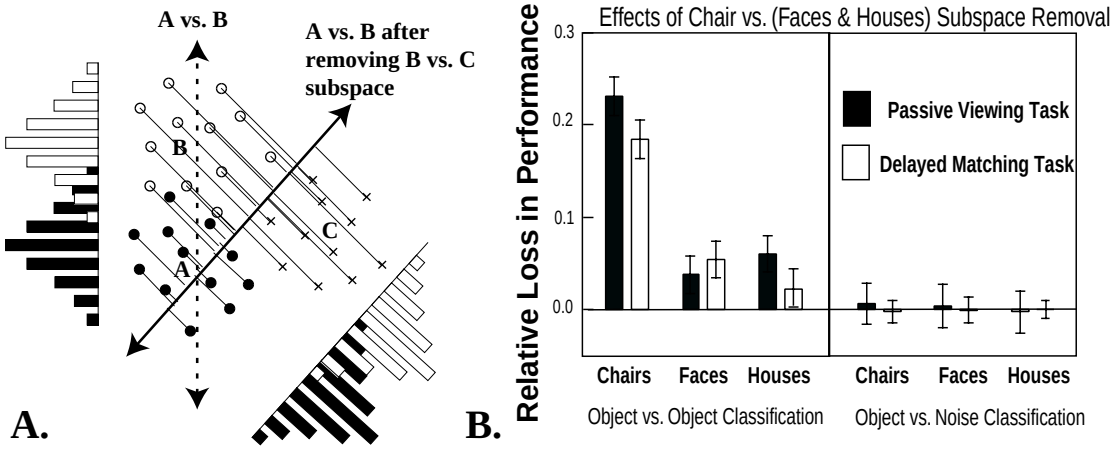


Figure 3. Top: Effects of subspace removal on other classification. **A.** Illustration of the effects of subspace removal. **B.** The effect of removing the chair-class specific brain activity subspace on other classifications, measured in terms of the increase in prediction error after removal.

suggest that the only effect of task demand is on visual processing in early visual areas.

Given two separately derived feature weights, one best for distinguishing category A from B, and the other for distinguishing B from C, how can we determine the effective overlap of these weight vectors? One way is to assess the impact of removing a feature axis best for classifying one stimulus pair on the prediction error for the other (see figure3). We performed removal by projecting out an informative feature direction and measuring increases in prediction error. We found a small but significant overlap between the best activation patterns for different object categories. For further results and discussion see Carlson et. al..[16]

5. ANALYZING HUMAN BEHAVIOR

Recent years have seen the development of psychophysical methods that attempt to assign credit for the observers successes and errors to particular stimulus features by correlating the observers responses against the values of the stimulus features on each trial (e.g.[19–23]). This analysis is only appropriate for problems in which stimuli are weighted linear combinations of the stimulus features with added gaussian noise,[22,24] and requires adding gaussian noise to the images. However, generalizations of this method have been proposed that relax feature additivity and the addition of noise. Knill[25] in slant from texture and the author[26] in motion detection developed methods for estimating the weights on independent cues by fitting a probabilistic cue combination model to the observers responses. Recently, Gosselin & Schyns[19] assessed the spatial dependence of object recognition using sets of small randomly placed smooth windows (bubbles) to limit the visible area of an image. The trial-by-trial relation between the observers responses and the visible areas was used to estimate the image regions used by the observer. The key idea in all these analyses is to construct an ensemble of images using either added noise (for example,[20,21,23] a probabilistic rendering procedure,[25,26] or via random windowing,[19,27] and to use the fluctuations in the image features across members of the ensemble to assess an observers relative use of the stimulus information. Below I briefly demonstrate the utility of the approach described above to estimating the features used by human observers in a motion detection task.

Previous work by the author[28] and others (for example [29–31]) have shown that the estimation of local image velocities in humans and monkeys can be described in terms of weighted sums of the outputs of local motion energy mechanisms.[32] Consider an image sequence $I(x, y, t)$. If in a window of space and time $W(\vec{x}, t)$ (e.g. gaussian) the image motion can be described as a translation, then $I(x, y, t) \approx \sum_{ij} w_{ij}(\vec{x} - \vec{x}_i, t - t_j) I(\vec{x} - \vec{v}_{ij} t, t)$. It is easy to show that the spatio-temporal (3-D) Fourier transform of one windowed region is given by

$$\mathcal{F}\{w_{ij}(\vec{x} - \vec{x}_i, t - t_j) I(\vec{x} - \vec{v}_{ij} t, t)\} = S(\vec{\omega_x}, \omega_t) = W(\vec{\omega_x}, \omega_t) \otimes (S_I(\vec{\omega_x}) \delta([\vec{\omega_x} \;\; \omega_t]^T [\vec{v}_{ij} \;\; 1]))$$

Note that the delta function term is an equation for a plane in the Fourier domain specified by $[\vec{\omega_x}\ \omega_t]^T[\vec{v}_{ij}\ 1] = 0$. Thus the equation says that local image translations in the Fourier domain are characterized by the spatial spectrum of the image projected onto a plane whose orientation is uniquely specified by the velocity of the translation, which is convolved by the Fourier transform of the windowing function. Given this description, a simple velocity detector can be constructed by pooling the squared outputs of spatio-temporal filters whose peak frequencies lie on a common plane. For a windowed signal S, the output R of the detector is given by

$$R = \sum_j \sum_{\vec{\omega_x},\omega_t} a_j |F_j(\vec{\omega_x},\omega_t)|^2 |S(\vec{\omega_x},\omega_t)|^2$$

where F_j denotes whose peak frequency lies on the plane specified by the signal. Within this simple theory, we have a choice of the weights a_j.

In order to investigate the plausibility of such a model for human motion processing, Schrater et. al.[28] have recently shown that the putative motion detectors are ideal observers for detecting a class of novel stochastic signals added to Gaussian white noise. The stochastic signals are produced by passing Gaussian white noise through the filters used to construct the motion detector. In general, a detector which computes the Fourier energy within a filter is an ideal observer for stochastic signals generated by passing Gaussian white noise through the filter. Thus, by varying the number and placement of filters, we can produce motion stimuli that are consistent with a single translational velocity and have various spatial frequency spectra, or even produce stimuli that are consistent with multiple velocities. Examples of two filters and the resultant stimuli are shown in figure 4.

Observers detected the presence of these stochastic signals buried in noise vs. noise alone in a 2AFC detection task. In previously published only in my dissertation, I estimated the weighting observers assign to different frequency bands by maximizing the mutual information between observer's responses and the energies in a small set of frequency bands by numerically optimizing equation 6 adapted to this problem domain. Because our method required numerical optimization, I reduced the dimensionality of the problem dramatically by decomposing frequency space into a set of 13 non-overlapping bands arranged on a sphere. The decomposition is illustrated in figure 4. The choice of decomposition was subject to several constraints. The most important constraints were to ensure that the signal spectra were completely contained within a small number of bands and that the numerical optimization was converging to similar answers. The number of bands was determined by a trade off between the reliability of the weight estimates (which requires a smaller number of bands) and reducing biases in the weights due to discretizing the subject's spectral weighting function (which requires a larger number of bands). For details concerning stimulus construction, filter parameters,etc., see.[26]

5.1. Theory

The method relies on the stochastic nature of the stimuli. Recall that signal stimuli are filtered noises which have mean power spectra given by the filter which produced them, while the backgrounds are white noise samples which have expected flat power spectra. Because the stimuli are noises, their spectral power fluctuates around the mean. In figure 5 we show the total power (energy) in a set of seven different nonoverlapping frequency bands plotted above the observer's binary response as a function of trial number. Depending on how an observer weights frequency space, the fluctuations in power within each band will cause different patterns of correct and incorrect decisions.

The approach taken here is to explicitly model the observer's decisions as a near optimal Bayesian decision agent. To estimate the observer's weights, we assume observers use weighted combinations of energies from different Fourier energy bands corrupted by internal noise to perform the task. In this model the observer is assumed to compute the energies e_i within each band for both signal plus noise and noise alone intervals. Each of the energy estimates are corrupted by an independent additive noise N_i. Subsequently the energies are weighted by the scalars w_i and passed through an unknown function g. The weighted, transformed energies are then summed within each interval, and finally the difference of these sums is used as a decision variable dv, which is corrupted by central noise $N_{central}$.

$$dv = \sum_i \Delta g(w_i \cdot (e_i + N_i)) + N_{central} \tag{14}$$

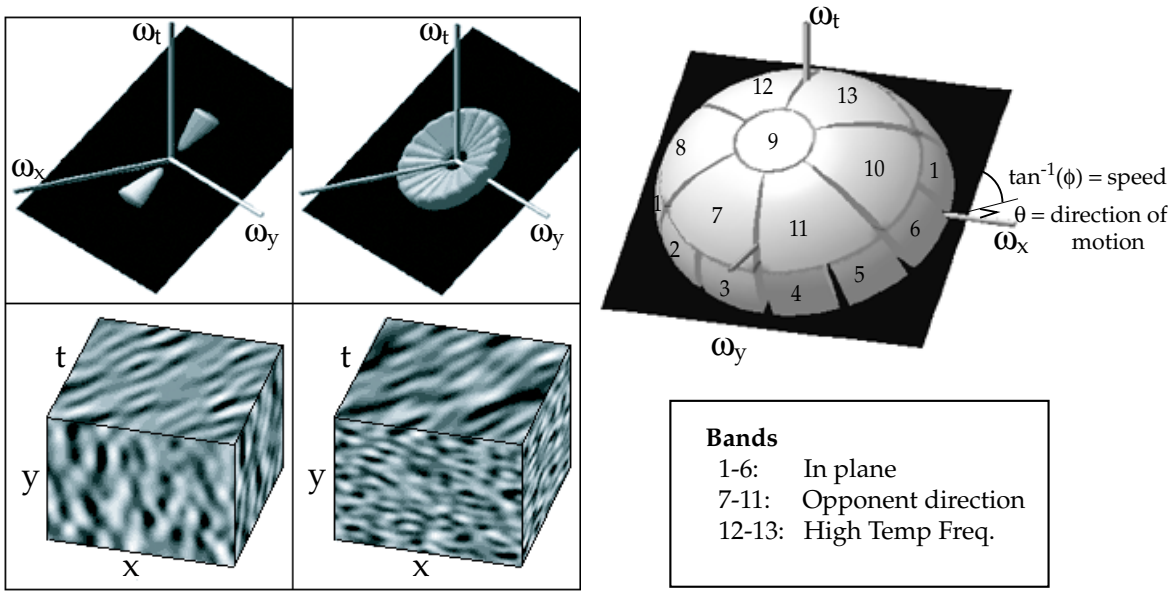


Figure 4. Top: Left: Observers detected stochastic motion stimuli buried in spatio-temporal Gaussian white noise in a 2AFC task. Stochastic stimuli are produced by passing spatio-temporal Gaussian white noise through a sum of spatio-temporal filters. Top-left, Level-set of the amplitude spectrum of a single filter in shown in spatio-temporal frequency space. Bottom-left, Depiction of the movie that results when noise is passed through the above filter shown as an intensity cube. x-y face shows shows that static frames look like oriented band-pass textures. y-t face orientation shows the left-to-right motion. Top-right, 'Planar' stimuli: A set of filters arranged to tile a plane in frequency produce (Bottom right) a spatially isotropic band-pass static frames (x-y face) but motion identical to 'Component stimuli'. **Right** The ideal observer for the task of detecting these stimuli buried in Gaussian white noise uses a weighted sum of energies from Fourier bands that contain the signal. Frequency space was divided into a set of 13 bands, centered around the plane containing the signal energy such that all the signal energy for both stimulus types is contained in bands 1-6. Weights effectively assigned by observers for these bands were estimated via maximum likelihood.

Subject's responses are determined by the decision variable dv, with $dv > 0$ producing a correct response and $dv < 0$ an incorrect response. The unknown function g is used to represent the effects of pointwise non-linearities and/or non-gaussian noise to the pooling process* This model can be linearized to yield the simpler expression:

$$dv = \sum_i w_i' \cdot \Delta e_i + N_{total} \tag{15}$$

where w_i' represent the linearized weights and N_{total} is the total noise at the decision variable. This model essentially lumps the effects of nonlinearities and non-gaussian noises into the noise term, whose distribution then ultimately determines the shape of the psychometric function in the model. Observer's psychometric functions were significantly skewed, but could be modeled by assuming N_{total} is a gaussian random variable with mean and variance functions linear in the input energies.

Using this model, the probability of a correct response is given by:

$$p(R_j = 1) = p(dv > 0) = 1 - \Phi(0, \mu, \sigma^2_{N_{total}}) \tag{16}$$

$$\mu = \mu_{bias} + (w_i' \cdot \Delta e_i) \tag{17}$$

$$\sigma^2_{N_{total}} = c + d(w_i' \cdot (e_{i_{signal}} + e_{i_{noiseonly}}))$$

*Note that the weight estimation procedure is only guaranteed to be unbiased if g is linear. For non-linear g, the linearization may be different at different signal levels, causing the estimated weights to be compromises across signal level.

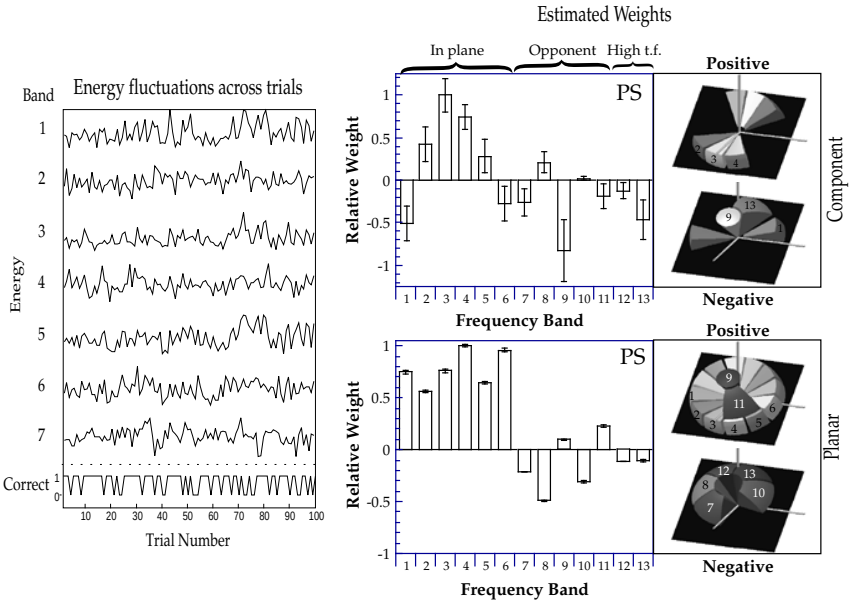


Figure 5. Top: Left: Stimulus energy in each band fluctuates across trials. Fluctuations are shown for 7 bands. We seek a weight vector that best explains the observer's performance across trials. **Right** Estimated weights for one observer for the two different types of stochastic stimuli given above. The data for more observers and stimulus types is presented by Schrater in.[26]

where R_j denotes the observer's response on the j^{th} trial, and Φ is the cumulative gaussian function with the first argument giving the upper integrand. The free parameters in this model are the weights w_i', the mean bias term μ_{bias}, and the variance constant c. The separate scale factor d allows the mean and variance functions to grow at different rates.

The total likelihood of the weights and parameters given the sample distribution of responses and stimulus energies can be computed as:

$$L(w_i, \mu_{bias}, c, d|\{R_j\}) = \prod_{j=1}^{n} R_j p(R_j = 1) + (1 - R_j)p(R_j = 0) \quad (18)$$

$$\begin{aligned} L(w_i, \mu_{bias}, c, d|\{R_j\}) = & \prod_{j=1}^{n} R_j(1 - \Phi(0, \mu_{bias} + w_i' \cdot \Delta e_i, c + d(w_i' \cdot (e_{i_{signal}} + e_{i_{noiseonly}})) \\ & + (1 - R_j)\Phi(0, \mu_{bias} + w_i' \cdot \Delta e_i, c + d(w_i' \cdot (e_{i_{signal}} + e_{i_{noiseonly}})) \end{aligned} \quad (19)$$

This expression was numerically maximized over the model parameters using standard optimization methods find the best weights w.

This method was applied to estimate the weights assigned by 3 observers to 5 different types of stochastic stimuli. Due to space limitations in this paper, the weights resulting of this optimization are only shown for one observer in figure 5. The weights show that the observer assigns close to the ideal positive weights for these two stimulus types. One of the most interesting aspects of the larger analysis is that these are the only two stimulus types out of the 5 studies that produce close to optimal weights. However, the top (Component weights) are too diffuse and slightly suboptimal given the signal is contained in band 4. This orientation bandwidth is similar to what we would expect blurred 2nd derivative filters. The presence of negative weights is more interesting, because the ideal weights for this detection task are all positive. However, the negative weights occur in interpretable locations, corresponding in part to signals with the same dominant orientations but opposite

in motion direction for both stimulus types. In addition, for Component stimuli, all observers showed negative weights in bands 1 and/'or 6, which correspond to *stationary* signals that are *orthogonal* to the Component signal's spatial orientation. This suggests that in detecting Component signals observers are implicitly discriminating the Component's direction of motion against noise fluctuations that are spatially similar and opposite moving and/or stationary and spatially orthogonal.

6. SUMMARY

A general framework for finding the most informative features for classification tasks by maximizing the mutual information between class decisions and features was presented. When a sample distribution of class decisions/labels and features is available, finding optimal weights (axes) in feature space can be treated as a maximum likelihood problem. For the special case of class-conditional Gaussian feature distributions with equal covariance, finding the most informative features is equivalent to Fisher's LDA. We presented two applications of the approach to very different domains, fMRI data analysis, where we derived patterns of brain activity that best distinguish (predict) when the observer is viewing images from different object categories, and human motion detection, where we found weights in spatio-temporal frequency space that best predicted an observer's responses in a motion detection task.

The ability to determine the most informative image data both objectively, and for particular observers will provide a quantitative framework for *information design* that: is critical for the effective design of virtual reality simulators and human/machine interfaces; provides a normative standard and measure of perceptual expertise useful to assess competence in fields like medicine; enables efficient training by providing feedback about the information important for the task. However, for their fruition, these possibilities are contingent on finding efficient algorithms for computing and optimizing the required mutual information.

ACKNOWLEDGMENTS

The fMRI analysis summarized here was performed in collaboration with Thomas Carlson and Sheng He at the University of Minnesota. Part of this work was supported by an NEI training grant when the author was at the University of Pennsylvania.

REFERENCES

1. J. J. Clark and A. L. Yuille, *Data Fusion for Sensory Information Processing*, Kluwer Academic Publishers, Boston, 1990.
2. A. L. Yuille and H. H. Blthoff, "Bayesian decision theory and psychophysics," in *Perception as Bayesian Inference*, K. D.C. and R. W., eds., Cambridge University Press, Cambridge, U.K., 1996.
3. A. Blake, H. Bulthoff, and D. Sheinberg, "Shape from texture: Ideal observers and human psychophysics," in *Perception as Bayesian Inference*, D. Knill and W. Richards, eds., pp. 287–321, Cambridge University Press, Cambridge, 1996.
4. M. S. Landy, L. T. Maloney, E. B. Johnston, and M. J. Young, "Measurement and modeling of depth cue combination: In defense of weak fusion," *Vision Research* **35**, pp. 389–412, 1995.
5.
6. D. Kersten and P. Schrater, "Pattern inference theory: A probabilistic approach to vision," in *Perception and the Physical World*, R. Mausfeld and D. Heyer, eds., John Wiley & Sons, Chichester, 2002.
7. D. C. Knill and D. Kersten, "Apparent surface curvature affects lightness perception.," *Nature* **351**, pp. 228–230, 1991.
8. D. C. Knill, "Perception of surface contours and surface shape: from computation to psychophysics," *J Opt Soc Am [A]* **9**(9), pp. 1449–64, 1992.
9. T. M. Cover and A. T. Joy, *Elements of Information Theory*, Wiley Series in Telecommunications, John Wiley & Sons, Inc., New York, 1991.
10. J. Friedman, "Exploratory projection pursuit," *Journal of the American Statistical Association* **82**, pp. 249–266, 1987.

11. M. Zhu and T. Hastie, "Feature extraction for non-parametric discriminant analysis," tech. rep., Department of Statistics and Actuarial Science, University of Waterloo, 2002.
12. N. Kumar and A. Andreou, "On generalizations of linear discriminant analysis," tech. rep., Johns Hopkins University, 1996.
13. N. Campbell, "Canonical variate analysis - a general model formulation," *Australian Journal of Statistics* , pp. 86–96, 1984.
14. R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley Interscience, New York, 2001.
15. M. Gales, "Maximum likelihood multiple projection schemes for hidden markov models.," *IEEE Transactions on Speech and Audio Processing* , 2002.
16. T. Carlson, P. Schrater, and S. He, "Patterns of activation in the categorical representations of objects: A new perspective in functional imaging analysis.," *J Cogn Neurosci* **To appear**, 2002.
17. A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby, "Distributed representation of objects in the human ventral visual pathway," *Proc Natl Acad Sci U S A* **96**(16), pp. 9379–84, 1999. 0027-8424 Journal Article.
18. R. Tibshirani, "Bias, variance, and prediction error for classification rules.," Technical Report Technical Report, Technical Report, University of Toronto, 1996.
19. F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Res* **41**(17), pp. 2261–71, 2001. 0042-6989 Journal Article.
20. J. Gold, P. J. Bennett, and A. B. Sekuler, "Identification of band-pass filtered letters and faces by human and ideal observers," *Vision Res* **39**(21), pp. 3537–60, 1999. 0042-6989 Journal Article.
21. J. Ahumada, A. and R. Marken, "Time and frequency analyses of auditory signal detection," *J Acoust Soc Am* **57**(2), pp. 385–90, 1975. 0001-4966 Journal Article.
22. A. Ahumada and J. Lovell, "Stimulus features in signal detection," *J. Acoust. Soc. Am.* **49**, pp. 1751–1756, 1971.
23. C. Abbey, M. Eckstein, and F. Bochud, "Estimation of humanobserver templates in two-alternative forced-choice experiments.," in *Proceedings of the Society of Photo-optical Instrumentation Engineers*, E. Krupinski, ed., pp. 284–295, SPIE, (San Diego, CA), 1999.
24. V. M. Richards and S. Zhu, "Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients," *J Acoust Soc Am* **95**(1), pp. 423–34, 1994. 0001-4966 Journal Article.
25. D. C. Knill, "Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture," *Vision Research* **40**, 1998.
26. P. R. Schrater, *Local Motion Detection: Comparison of Human and Model Observers.* Neuroscience, University of Pennsylvania, 1999.
27. L. Bonnar, F. Gosselin, and P. G. Schyns, "Understanding dali's slave market with the disappearing bust of voltaire: a case study in the scale information driving perception," *Perception* **31**(6), pp. 683–91, 2002. 0301-0066 Journal Article.
28. P. R. Schrater, D. C. Knill, and E. P. Simoncelli, "Mechanisms of visual motion detection," *Nature Neuroscience* **1**, pp. 64 – 68, 2000.
29. D. J. Heeger, "Model for the extraction of image flow," *Journal of the Optical Society of America* , pp. 1455–1471, 1987.
30. N. M. Grzywacz, S. N. Watamaniuk, and S. P. McKee, "Temporal coherence theory for the detection and measurement of visual motion," *Vision Res* **35**(22), pp. 3183–203, 1995. 0042-6989 Journal Article.
31. E. P. Simoncelli and D. J. Heeger, "A model of neural responses in visual area MT," *Visual Neuroscience* , 1996.
32. E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion.," *Journal of the Optical Society of America* **2**((2)), pp. 284–299, 1985.