An Efficient Image Search Algorithm For Object Feature Identification in Biological Vision

Alexander Franks
Brown University

April 28, 2009

1 Introduction

While the functional architecture for low level vision in the primary visual cortex is relatively well understood, very little is known about how higher level vision functions in the ventral temporal cortex. For researchers, the ventral temporal (VT) cortex, which is associated with visual object perception and recognition, is still essentially a black box. The VT cortex performs several impressive computations related to object recognition. Importantly, as Haxby *et al.* note, while the underlying machinery is still a mystery, we do know that the ventral object vision pathway "has the capacity to generate distinct representations for a virtually unlimited variety of individual faces and objects" [2]. Even more, humans have the ability to recognize objects despite drastic changes in lighting or viewpoint. Simultaneously, humans can make fine distinctions in order to differentiate between objects in the same category. Thus, our visual system has the capacity to either "neglect or amplify input images" depending on the context [6]. Surprisingly though, almost nothing is known about how the higher level human visual system accomplishes any of these feats. For these reasons there is great interest in understanding the functional organization of this impressive visual system.

There are effectively two approaches for studying the functional architecture for object recognition in the ventral temporal cortex. The first, principally associated with the field of neuroscience, is a single-cell approach, in which researchers record impulses from individual neurons as they respond to a variety of stimuli. These studies have shown that neurons are selectively tuned for specific image features, and occasionally, appear to be tuned for entire objects. Single-cell research is particularly effective because it provides low-level results about feature-coding across small regions of the cortex. Crucially, these studies have produced evidence that specific "columns"

of cells responding to similar features or objects tend to cluster together.

The other approach, used by cognitive scientists, focuses on brain activation at a much larger scale. fMRI, or functional magnetic resonance imaging, measures the localized hemodynamic response, or change in blood oxygenation, in the brain. Crucially, this localized change in blood oxygenation is correlated to neural activity. Thus, researchers can use fMRI to gain broader insight into how vast populations of neurons respond in different areas of the brain in response to stimuli. Unlike the single-cell approach, neuroimaging allows researchers to explore large scale patterns of activation. When analyzing fMRI data, a three dimensional map of the brain is reconstructed, and activation is measured across voxels (volumetric pixels). A one to three mm voxel (typical for fMRI studies) in VT cortex contains on the order of 100,000 to 1,000,000 neurons. Thus, in brain imaging, there is significant averaging of neural activity. So, while fMRI allows researchers to identify entire category-specific brain activation, much of the fine-grained neural information is lost. Ultimately, since each technique has its own advantages, they are both essential to the study of the VT cortex.

Recent research applying both approaches for studying the ventral temporal cortex have begun to reveal clues about the functional architecture of high-level visual cortex. At the cellular level, Keiji Tanaka has investigated the role of neuronal "columns" in higher level vision, specifically the inferotemporal cortex. Tanaka showed that within a column, neighboring cells code for similar features, and that multiple columns representing different but related features may overlap with one another creating larger-scale units [6]. His research provides evidence for columns of cells in the visual object pathway which can be understood as modules coding for specific object features. These results are consistent with the architecture of early visual processing systems

in V1, where edge orientation, color, and retinotopic location are all organized in spatial maps or neuronal columns [2].

While the results of this single-cell study emphasize the spatial organization of VT cortex , recent neuroimaging research has begun to suggest that representations for objects in VT cortex are distributed and overlapping. These studies reveal a unique pattern of activation across the brain for distinct object categories. Still, O'toole *et al.* [5] note that these results are supportive of a "feature-based" model of objects. In this model, objects in different categories could share neural codes if they had some lower-level features in common. These shared feature codes would be consistent with distributed and overlapping responses, because a region coding for a specific feature would potentially activate in response to a variety of objects which contain some variant of a distinct feature. In support of this theory, Grill-Spector *et al.* [1] use high-resolution fMRI to show category-selective regions in finer detail, and obtain results that are consistent with the distributed theory of representation. In particular, they demonstrate that the fusiform face area of VT cortex (FFA), is not solely selective for faces (Fig. 1). Rather, there are multiple subregions in FFA selective to a variety of object categories, with the largest and most subregions devoted to face selectivity.

Importantly, evidence for distributed object coding in VT cortex is not inconsistent with the localized spatial organization of features discovered by Tanaka [6]. Imaging by Haxby has demonstrated that, despite a distributed brain response, at a finer level there is still a spatially organized functional architecture within separable subregions [2]. Within a larger brain region, each subregion still responds maximally to a specific category (Fig. 1). Significantly, no studies to date have attempted to thoroughly explore the neural codes for object features within subregions of category-

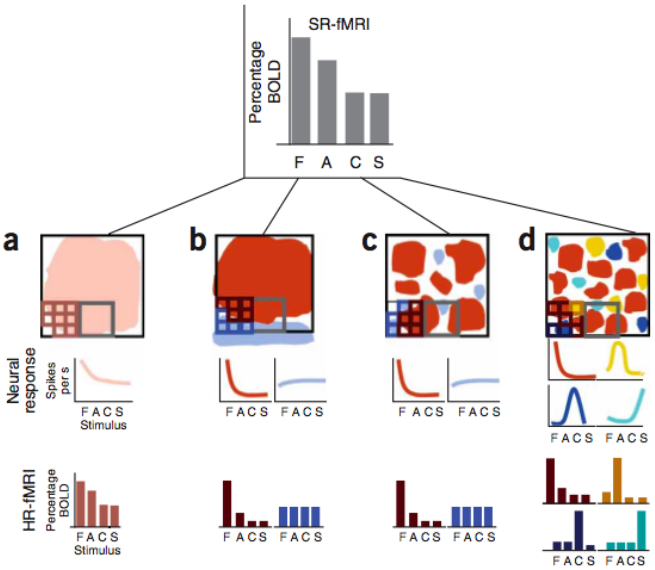


Figure 1: In this figure, from Grill-Spector *et al.* high-resolution fMRI data reveals a variety of category-specific subregions in FFA. In particular, there are subregions which activate in response to faces, animals, cars and abstract sculptures. Importantly, no research has been done investigating the underlying neural codes for these specific subregions. [1]

specific brain patterns. The aim of our research is to explore the "peak" subregions of category-specific responses in an attempt to uncover intermediate-level feature codes for objects. To this end, we plan to use fMRI to explore these peak responses for a given object category. Our hope is that we will discover an underlying neural code for a specific feature in that object.

One crucial question is how to go about discovering what object feature, if any, the neurons in a specific subregion are responding to. As Tanaka notes, one problem with single-celled studies is that "the variety of object features existing in the world is too great to test its entire range for a single cell". Various researchers have applied different procedures to address this problem. For example, Tanaka's approach [6] was to use an empirical feature reduction method. He took the image of the most effective stimulus and simplified it systematically to determine which feature in the image was

necessary for maximal activation. Others [7] have used genetic algorithms to "evolve" features which maximally stimulate a cell. Despite these approaches, single-celled studies still require that the researchers must first present many arbitrarily chosen objects to find an initial effective stimulus.

In this respect, by using fMRI to study object representation, we have a distinct advantage. Rather than poking around in the dark for effective stimuli, we can choose any object category, find the responsive brain regions for that category, and then use an image search algorithm to determine which specific features maximally activate a particular voxel. Because we know that some aspect of an object shown recruits a distinct brain region, the problem for us is to determine whether or not neurons in the subregion are activating in response to an individual feature or whether they respond to the presence of the entire object. Based on the previous research demonstrating evidence for feature-specific columns and spatially organized functional architecture, we expect to find some intermediate-level object representation or feature coding within active subregions.

Not surprisingly, there are several computational problems associated with using neural imaging as a tool for understanding the functional architecture of the visual cortex. Crucially, in order to discover feature-specific neural coding in the brain we need more control over our experiments than fMRI researchers have typically had. In general, fMRI studies require a fixed paradigm, in which the stimulus set is determined before the experiment begins. Those stimuli are delivered to the subject during the course of a scanner session and the subsequent data is analyzed offline. This presents a major hurdle for our research, because there are countless features in high dimensional image space to which a voxel could be responding. For our purposes, there is no practical way to choose stimuli a priori and analyze data offline after the

trial, and still efficiently identify brain response to specific image features.

Fortunately, LaConte *et al.* [3] have demonstrated the effectiveness of a new paradigm which gives researchers more control over their imaging experiments. In particular, they use a real-time fMRI system, in which the data is analyzed online. Thus, "rather than using a fixed paradigm, experiments can adaptively evolve" depending on the brain response to different stimuli. For example, recent studies have demonstrated how real-time fMRI paradigm has opened the door for a variety new work including biofeedback research and brain state classification experiments. Real-time fMRI experiments are appropriate for our purposes because it will allow us to adaptively control the stimulus in response to brain activation. Most importantly, using this paradigm, we can efficiently search image space for maximally activating features by adapting our "questions" in response to previous "answers".

While real-time fMRI creates new freedom in experimental design, it also provides several unique challenges. Specifically, real-time fMRI introduces several temporal limitations. For one, since all analysis is done online during the course of the experiment, computations must be especially fast. However, there are other more limiting factors. As LaConte [3] and others have noted, the hemodynamic delay associated with fMRI imposes problematic restrictions. As mentioned earlier, fMRI measures the hemodynamic response of particular brain regions. Unfortunately, after stimulus presentation, the hemodynamic response function (HRF) takes nearly six seconds to reach its peak, and a full fifteen seconds to return to baseline levels. In other words, measuring the voxel response for a given stimulus is very slow in a real-time system. In traditional fMRI experiments, on the other hand, new stimuli can be presented before the HRF from the previous stimulus returns to baseline. This is because when analyzing the data offline, researchers can pull apart overlapping HRFs by deconvolv-

ing the pooled data. Crucially, because we are using a real-time system and doing the analysis online, we have limited ability to deconvolve overlapping HRFs. Accordingly, we must wait a full fifteen seconds for the hemodynamic response to complete before presenting the next stimulus.

This real-time limitation is one of the central challenges in our research. Since we presume a subject can only remain in the scanner for one to two hours, we have a hard limit on the number of stimuli that we can present in the course of one run. For this reason, we frame our image search as a "20 questions" problem – we must discover a voxel activating image feature in a high dimensional image space using a relatively small number of stimuli. Here, we define a "question" to be a particular stimulus, and consider voxel activation the "response". After asking a "question", we do not get the complete response until the HRF function settles down around fifteen seconds later.

The ultimate goal of our research is to find evidence of neural coding for intermediate level features in VT cortex. Based on previous research, we expect that there is indeed a feature-based functional architecture in high level object recognition pathways of the brain. To facilitate this search for object feature codes, we plan to use a real-time fMRI system to adaptively control our stimulus presentation. Using an intelligent search algorithm, we hope to find and quickly identify the image features that activate a particular voxel. However, since it is time consuming to analyze voxel responses in real-time, we are limited in the number of stimuli we can present during the course of one scanner session. This problem requires a particularly efficient search algorithm. In this paper, I develop a novel algorithm for efficiently searching through image space by asking the most informative "questions" possible. Before using this algorithm in a functioning real-time fMRI system, we demonstrate its utility

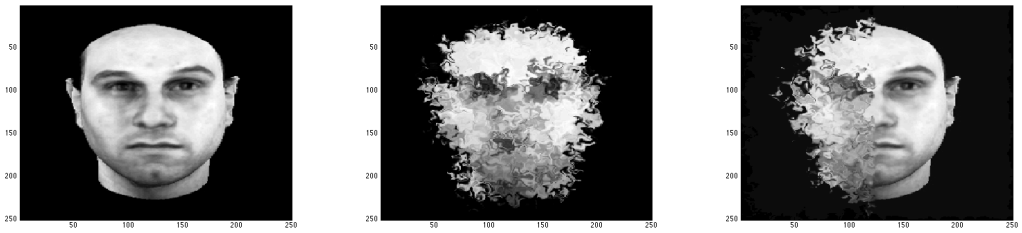
in simulation.

2 Methods

We eventually hope to explore evidence for image features across a variety of object categories. However, initially we plan to focus on faces. Accordingly, our stimuli are derived from a set of nearly two hundred unique registered faces from the Tarrlab face database. Here, each face is a black and white image, 250 pixels square. Using this face data set, we generate new stimuli by applying a distortion algorithm developed for this task. We use this distortion algorithm to selectively warp different regions of the face, therefore effectively reducing the number of possible activating features present in the image. Ultimately, we hope to discover maximally activating features by showing that when a particular image feature is distorted, the voxel response drops. By selectively warping different regions, we can look for the smallest image region that maintains a high brain response.

In this research, we do not run our experiment using a true real-time fMRI paradigm on human subjects. Rather, we create an "artificial voxel" which serves as a placeholder for our real-time system. In a real-time fMRI study, there are many complications and unknown factors. By testing our algorithms in simulation, we can ensure their functionality in a controlled environment, before pressing on to more challenging tasks. Thus, our aim is simply to demonstrate the utility of our image search algorithm, which we will eventually employ for the actual experiment.

In our simulation, for a particular warped image, we can determine the "brain response" using our artificial voxel. First, we develop a paradigm that warps the image systematically, independent of voxel response to demonstrate the utility of our warping algorithm. Later, we develop a more efficient well-defined algorithm. In



(a) Original image (b) Warped imaged (c) Half warped

Figure 2: Illustration of the warping algorithm. This algorithm effectively distorts image features, while preserving smoothness.

this paradigm, our image search algorithm generates new stimuli according to the previous voxel response. Since we design our artificial voxel, we know which feature our search algorithm should find. Thus, we can test the algorithm's effectiveness by studying whether or not we can consistently locate the correct feature region.

Warping Algorithm – For our warping function, we sought an algorithm that would distort image regions enough that previously present image features were unrecognizable. Additionally, we required an algorithm that could selectively distort specific areas while maintaining smoothness between distorted and undistorted regions. To meet these requirements, we distort the image mesh grid, and then use an interpolation algorithm to fill in the missing pixels. More specifically, we randomly add gaussian "bumps" to the mesh grid in the regions we seek to warp. This effectively amounts to repeatedly choosing random groups of pixels and translating them to a new location within a small neighborhood of their original location. Since we can selectively choose where we add the distortions to the mesh grid, we can correspondingly choose which image regions to warp. Moreover, as Figure 2 shows, since we are using interpolation to generate the new image stimulus, smoothness is maintained at the boundary between warped and unwarped regions.

Image Search – For the purposes of this study, we propose a viable image search algorithm, and test its effectiveness in simulation. To this end, we developed an "artificial voxel" so that we could test whether our image search algorithm would consistently discover the appropriate features. To accomplish this, we simply crop out an image feature from the "average face", and use that as our test feature. Now, for a particular image stimulus, we correlate this test feature over all image regions, and take the response to be the maximum correlation. This max correlation is our "voxel response".

Before running the image search algorithm, we generate prior probability distributions over "voxel responses" for warped and unwarped faces. To produce these two distributions, we calculate the max correlation of our template with both completely warped and unwarped versions of the two hundred faces in our database. We chose to fit the data to gamma distributions because in both cases the data was unimodal, but also skewed (see fig. 3). Since the gamma distribution is only defined for positive values, and our "voxel response" takes on values between negative one and one, we actually use "one minus correlation" as our statistic. Thus, our gamma distributions are defined on the interval from zero to two.

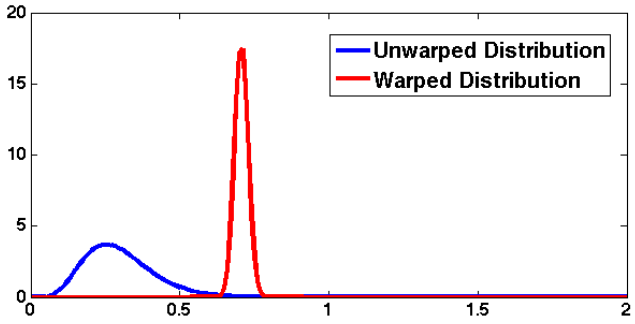


Figure 3: Our prior distributions, for a particular test feature, after the data was fit to gamma distributions. Because our test statistic was one minus the correlation, the domain is from zero to two.

Given these prior distributions, we selectively apply our warping algorithm to generate stimuli. By presenting the stimulus, we are asking a "question" about where the feature occurs. For example, if we warp half of the face and present that to our artificial voxel, we can use our prior model to determine whether it is more likely that the activating feature occurs in the warped part of the face or the unwarped part. By systematically changing which regions are warped, we can calculate a likely bounding box for the image feature.

To demonstrate the feasibility of an image search which employs our warping algorithm for stimulus generation, we initially used a methodical, but inefficient, paradigm. At multiple levels, we warp the image on one side of a horizontal or vertical line. Again, for a particular stimulus, we want to determine whether or not the activating feature occurs in the warped portion of the image, or the unwarped portion. Importantly, we use the likelihood ratio test to decide where the feature resides in the stimulus. According to the Neyman-Pearson Lemma, the likelihood ratio, $L(\frac{\theta_u}{\theta_w}) < k$, is the uniformly most powerful test for whether our data comes from the distribution governed by the parameters θ_u or θ_w. Here we take θ_u to be the parameters for the distribution of correlations on the unwarped images and θ_w the distribution over warped images. In other words, we simply take the ratio of the probability that the response came from the unwarped distribution to the probability that the response came from the warped distribution, and study its behavior as we warp different portions of the face. First we warp to the left of vertical lines, moving the line from left to right. We follow the same process warping right to left, then top to bottom, and finally bottom to top. When this ratio is large, we can say with high confidence that the feature resides in the unwarped portion of the face. Conversely, when the ratio is very small, we can presume that the feature resides in the warped

part of the face. Under these deterministic conditions, the likelihood ratio maintains the same maximum value until part of the feature region is warped. Thus, as soon as the likelihood ratio drops, we can presume that our feature is partly warped. This provides a simple and convenient way to choose our feature region. Thus, to identify the feature, we crop the tightest bounding box where the likelihood ratio is still maximized.

Not surprisingly this procedure works reasonably well (see results) and demonstrates the effectiveness of our warping algorithm for stimulus generation. Nevertheless, under noisy conditions, we have no rigorous process by which to chose a bounding box. This is because in the presence of noise, the likelihood ratio will fluctuate some, and thus we cannot determine the precise boundaries of our feature region. Even more, because we warp at all levels systematically, the process is relatively inefficient. For these reasons, we develop a more well defined likelihood model, which builds up a probability distribution over all possible rectangles containing the activating feature.

For simplicity, in this algorithm, we construct our probability distribution over a discrete set of equivalent sized rectangles . Specifically, each candidate rectangle is the same size as our artificial voxel test feature. At each step in this iterative approach, we update our probability distribution based on our "voxel" response and the probability distribution from the previous time step. After several iterations, the distribution will become more peaked around certain rectangles. Ideally, these highly probably rectangles will contain our test feature. Since the real-time experiment will limit the number of stimuli we can present, for this study we initially only run the algorithm for ten iterations. After we run it for a fixed number of iterations, we can simply look at the maximum likelihood rectangle or study the distribution more broadly, to gain a sense of where the feature might occur.

In this model, we designed our algorithm to choose the most efficient "questions" possible at each time step. In contrast to our previous model, where we warp at multiple levels in each direction regardless of the "voxel response", here we generate a new stimulus which we expect to give us as much information as possible. This is crucial for our research, because there is a strict time penalty for asking a "question". Consequently, we want our algorithm to learn as much as possible about the rectangle probability distribution in as few questions as possible. Here, we return to our twenty questions analogy. While we are already limiting the number of "questions" we are allowed to ask, we have an additional complication because the responses we get from fMRI data will be inherently noisy. Thus, we say that we are playing "twenty questions with a liar", since our information won't always be accurate. Interestingly, "twenty questions with a liar" is known to be an NP-hard problem. In other words, there is no fast algorithm which will consistently tell us the best question to ask at each time step. Accordingly, we developed an algorithm that uses a greedy heuristic, where we choose the question that will yield the most expected information at a specific time step.

In order to choose the most informative stimulus at each step in the algorithm, we must consider the entropy of the probability distribution. Information entropy measures the uncertainty associated with a random variable. The entropy of a uniform distribution is very high, since each outcome has equal likelihood. Conversely, the entropy of a very peaked distribution is small, since we have a good guess as to what the outcome might be. For a probability distribution P(x), its entropy is defined as $\sum_x P(x)log(P(x))$. In our experiment, initially, we have no knowledge about where a voxel activating feature might reside, so the entropy is high. Accordingly, our algorithm chooses the question which will minimize the average entropy of the

probability distribution at the next time step. Thus, in a well-defined sense, we can present the stimulus which will yield the largest expected information gain.

Again, we use our prior probability distribution over warped and unwarped faces to determine the conditional probability of the voxel response to a particular "question". Each question is a face which is warped on one side of a vertical or horizontal line and unwarped on the other side. Unlike our previous attempt, in this model, we define the conditional probability distribution to be a linear combination of our two prior distributions. In particular, if a candidate rectangle intersects a warping line, then we take the conditional probability for that rectangle to be $\alpha P_u + (1-\alpha)P_w$, where P_u and P_w are the unwarped and warped distributions respectively and α is the fraction of the rectangle which is unwarped. Since α is between 0 and 1, $\alpha P_u + (1-\alpha)P_w$, is still a probability distribution. This approach is more appropriate, because if that rectangle contained our feature, only part of it would be visible, and thus we would expect a diminished response. Thus, we expect that a response that lies somewhere in between responses governed by the complete warped and unwarped distributions.

The Likelihood Model – We assume that for a given voxel response y, there is a rectangle, x, to which that voxel is responding. Since presumably we start with no information about possible activating features, we take the probability distribution over all rectangles to be uniform. We start with the following definitions:

- $P_0(x) \sim$ uniform
- $P_t(x)$: The probability, after t questions, of the activating feature residing in rectangle x
- $P_q(y|x)$: The conditional probability under question q, of y given x. Under our model this is a mixture of our two initial distributions ($\alpha P_u + (1-\alpha)P_w$,). More precisely, the stimulus, q, (a warping line), and rectangle, x, determine what

fraction, α, of the rectangle is warped.

For a given a question, the joint distribution on x and y is simply:

$$P_q(y|x)P_t(x)$$

Thus, by Bayes' theorem the posterior probability of rectangle, x, given the voxel response to a question is:

$$P_{q,t}(x|y) = \frac{P_q(y|x)P_t(x)}{\sum_x P_q(y|x)P_t(x)}$$

For a given voxel response and a rectangle the entropy of the posterior is defined as

$$H(P_{q,t}(x|y)) = -\sum_x P_{q,t}(x|y)log(P_{q,t}(x|y))$$

Now, we want to compute the average or expected value under y of the entropy. Since $P_{q,t}(y) = \sum_x P_q(y|x)P_t(x)$, the formula reduces to

$$E_y[H(p_{q,t}(x|y))] = -\sum_y \sum_x P_q(y|x)P_t(x)log(P_{q,t}(x|y))$$

Using this formula, we choose a question q^*, which minimizes this expected entropy. This "question" is the stimulus which will yield the highest expected information gain. For the voxel response y^* , we define the new rectangle probability distribution at time t+1 to be

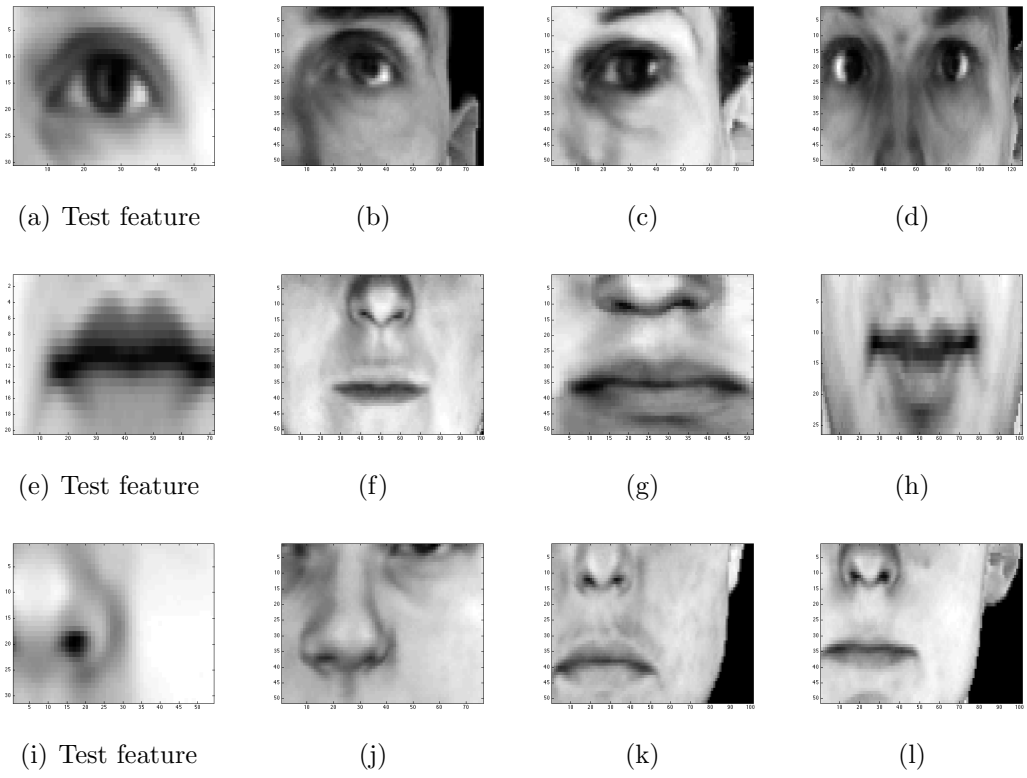
$$P_{t+1}(x) = P_{q^*,t}(x|y^*)$$

. In this way, we have a rigorous procedure for determining the likelihood that a particular rectangular is our feature-containing region. In our simulation, we consider only rectangles of equivalent size. Nevertheless, this model would still be appropriate if we were considering a richer class of rectangle sizes.

3 Results

Initially, to demonstrate the effectiveness of the warping algorithm, we systematically warped different areas of the face, and used the likelihood ratio test to determine whether it was more likely the feature occurred in the warped part or the unwarped part. After asking twenty eight questions, (seven for each of the cardinal directions), we used the likelihood ratios to crop out a candidate region for the feature location. While this cropped region was much larger than the original feature size, it often contained the test feature. Empirically these results demonstrate the effective use of the warping algorithm for stimulus generation. Figure 4 shows examples where our algorithm correctly identified a region containing the test feature. Impressively, in all trials, the cropped region contained the feature that our artificial voxel was responding to.

However, for these trials, the bounding box that we identified tended to be very large. Because of this, while the region contained our test feature, it also contained other features. For example, for the "mouth voxel", the candidate regions contained mouths, but some also contained noses. Conversely, the identified regions for the "nose voxel" all contained fragments of a nose, but some also contained mouths. Interestingly, perhaps because of the symmetry of the face, in one of our "eye" runs, our algorithm identified a region containing both eyes, rather than just one. Importantly, if we hope to find evidence for intermediate level features in VT cortex, the candidate regions must tightly bound a particular feature so that we can say with confidence what that region is coding for. This data suggested that we needed a more rigorous procedure for identifying a candidate region. Additionally, this algorithm is quite inefficient because it systematically warps across multiple levels, rather than intelligently warping to generate a stimulus that will yield more information. As noted,



(a) Test feature (b) (c) (d)

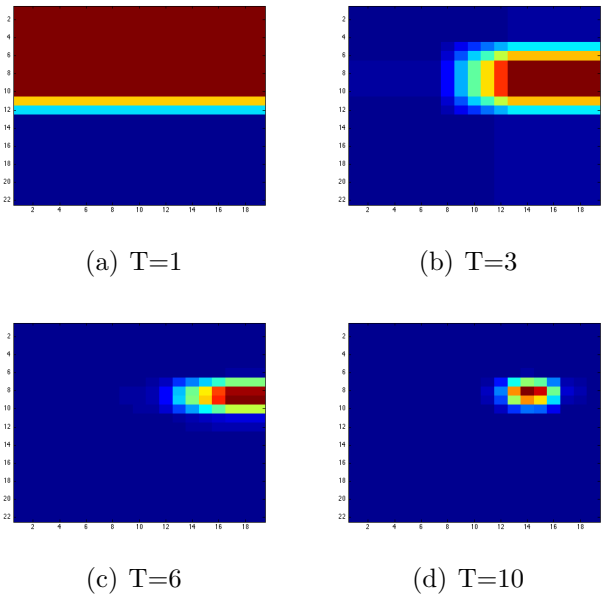
(e) Test feature (f) (g) (h)

(i) Test feature (j) (k) (l)

Figure 4: Features discovered using the likelihood ratio test over warped images. The first column contains the test features (the "artificial voxel"). Each row contains three candidate regions for that test feature. While the region generally contains the test feature, the region does not always tightly bound the right feature.

inefficiency is a major drawback because our real-time system imposes several time constraints. Lastly, under noisy conditions, our procedure for choosing the candidate region would fail. Our algorithm must be robust to noise, since we expect a very noisy signal from the MRI images. Thus, while the initial results seemed promising, they also demonstrated the need for a more rigorous model.

The entropy model we derived was an attempt to address some of the problems from our first approach. In this well-defined model, we continuously update the probability distribution over all feature-bounding rectangles. Each rectangle is defined to be the same size as the test feature, which is the strictest requirement possible.



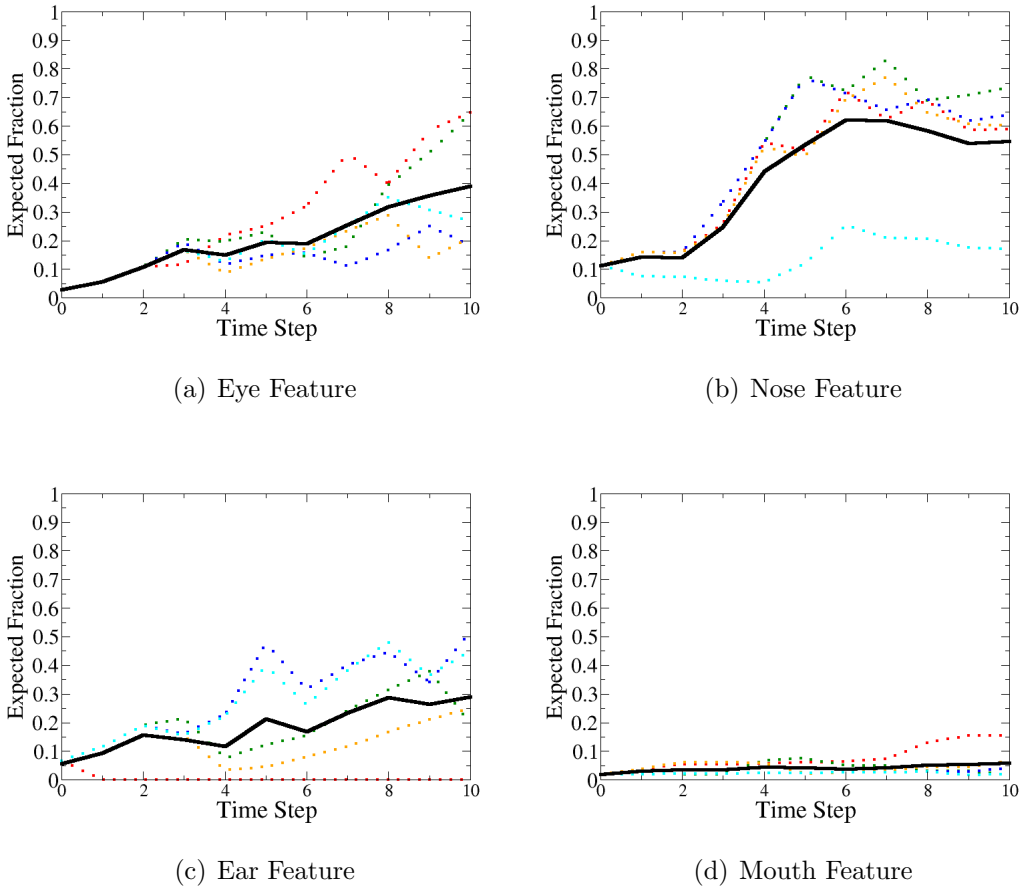
(a) T=1 (b) T=3

(c) T=6 (d) T=10

Figure 5: Rectangle Probability Distributions After T Time Steps For an Eye Feature. The first "question" determines which half of the face the feature is in. For this particular face, after ten time steps, our algorithm determines that the rectangle containing the right eye to be the most probable one. Additionally, the vast majority of rectangles have nearly zero probability.

Figure 5 shows how these probability distributions are becoming more peaked around a particular feature region as more "questions" are asked. These distributions provide empirical evidence for the success of this algorithm, but do not rigorously define their performance. Thus, as a metric for the performance of this algorithm, we computed the expected value of the visible feature fraction at each iteration. To do this, at each time step, for each rectangle, we determine the fraction of the test feature that is visible in that rectangle. To compute the average score, we take the probability of a rectangle times the fraction of feature in that rectangle, and sum this quantity over all rectangles. Since all of the faces were registered, we simply used the test feature coordinates to find the fraction of overlap between a candidate rectangle and the test feature.

Our results show, that for some test features, our entropy model does a great



(a) Eye Feature

(b) Nose Feature

(c) Ear Feature

(d) Mouth Feature

Figure 6: Expected Visible Feature Fraction vs. Time. This figure demonstrates the capacity of our image search algorithm to find the test feature in random faces. The black line is the average over all five trials whereas the dotted colored lines represent the expected feature fraction for an individual face over time. Under noiseless conditions, the algorithm generally succeeds in finding the feature region for three of the four features tested. On the mouth feature, however, in four of the five sample faces, our algorithm failed to pinpoint the correct region. These failures are most likely due to false assumptions in our model, stemming from the weakness of our "artificial voxel".

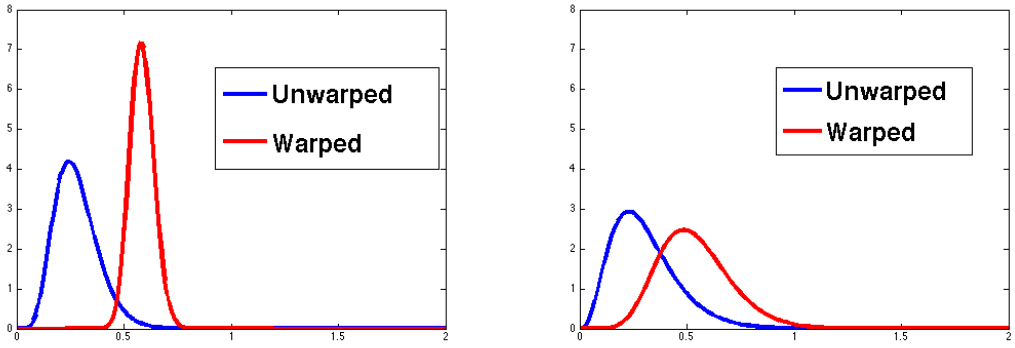
job at identifying likely feature containing rectangles. In particular, fig. 6(b) shows that the expected visible feature fraction, for a nose feature, was over fifty percent in some trials. For rectangles which are exactly the same size as the feature, this is fairly impressive. Even more, in all of these trials, we are only asking ten questions, whereas in our initial algorithm we were asking over twenty-five. If we continued to generate more stimuli, presumably the probability distribution, in this case, would

become more peaked around the correct feature. Conversely, fig. 6(d), illustrates that for the mouth feature, our algorithm failed to find the correct feature in any of the five random sample faces we tested it on. There were other isolated failures on sample faces for other features as well. For some test features, on certain faces, the probability model failed to locate the correct region.

Despite these mixed results, for all features tested, the average score still increases, if only slightly. Only in a handful of individual trials did the score drop to zero. This suggests that the algorithm is at the very least, consistently identifying the general region where the feature occurs. Since the probability distribution is necessarily becoming more peaked (the entropy is decreasing), it is significant that the score is greater than zero at all. For the vast majority of rectangles, after ten iterations, the probability assigned to the the vast majority of rectangles is zero. In other words, for non-zero scores, the probability distribution must be peaked in a region very close to the actual feature-containing rectangle. While it doesn't always pinpoint the precise region, it seems to always get close.

Furthermore, there are compelling reasons for why our algorithm fails when it does. Significantly, in this simulation, the assumptions in our model often seems to be invalid. Principally, we assume that the "voxel response", comes from one of the two prior probability distributions (either warped or unwarped). However, since we are taking the maximum correlation to be our "voxel response", sometimes this is not the case. While the maximum correlation between the test feature and the image usually occurs in the feature region, there are other regions of the face that have high correlations. In other words our artificial voxel is not finely tuned.

Ideally, we want to build an artificial voxel that has a fairly peaked response around the feature region, and little to no response elsewhere. Since we are simply



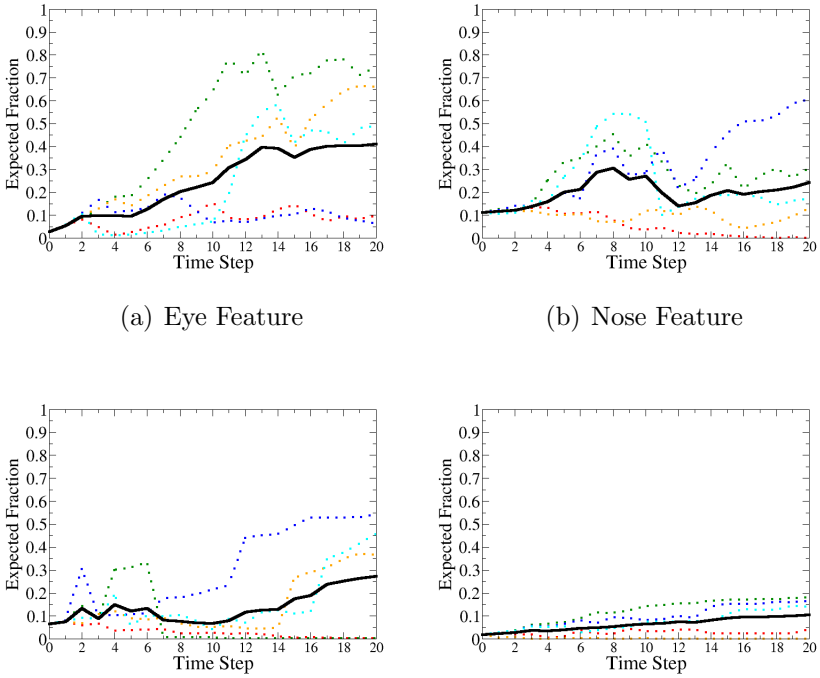
(a) No noise

(b) Noise $\sim N(0, 0.15)$

Figure 7: Warped and unwarped prior distributions under noisy conditions.

using normalized image correlation, this is not the case. Even more, since the image statistics of a natural, unwarped image are quite different from those of an image warped under our algorithm, our model often produces false positives. Specifically, when the feature is warped, the "artificial voxel" occasionally still returns a high response because it is responding to some other feature in an unwarped part of the face. In this case, our model determines that the feature resided in the unwarped part of the face, when it was in fact in the warped part of the face. Thus, a more selective "artificial voxel" would more effectively demonstrate the utility of this algorithm. However, because the computational problem of feature identification is not a trivial task, and also not the focus of this research, we chose to stick with something simple. Crucially, these failures speak more to the weakness of our "artificial voxel", rather than to the weakness of our image search algorithm. In reality, neurons tend to be finely tuned, so we expect our true voxel responses to be to be more selectively tuned as well, and thus to be less likely to get false positives. Thus, these results seem to suggest that this algorithm will succeed in a real-time system.

Despite the relatively auspicious results, this data still comes under fairly fixed



(a) Eye Feature

(b) Nose Feature

(c) Ear Feature

(d) Mouth Feature

Figure 8: Expected Visible Feature Fraction vs. Time Under Noisy Conditions. This figure demonstrates the utility of our image search algorithm under noisy conditions. For this experiment, gaussian noise, $\eta \sim N(0, 0.15)$, was added to the voxel response. The black line is the average over all five trials whereas the dotted colored lines represent the expected feature fraction for an individual face over time.

conditions. In our real-time experiment, however, we expect the true fMRI data to be much noisier. A voxel may respond principally to a specific feature, but also be responding less significantly to a variety of other features. Because a voxel may cover up to one million neurons, there is significant averaging across activity. Thus, we needed to test our paradigm under less ideal conditions. To this end, we ran our algorithm again in the presence of noise. At each step, we simply added random noise to our "voxel response". For each correlation, we add a variable drawn from a gaussian distribution with mean zero and standard deviation 0.15 ($\sim N(0, 0.15)$). As figure 7 shows, this creates more overlap between the two probability distributions,

and thus less certainty as to which distribution a response belongs to.

Impressively, the results under noisy conditions are nearly as good, as the results from the noiseless data. In some cases, the search algorithm under noisy conditions even seems to outperform the algorithm under less noisy conditions. Still, we must run the algorithm for longer (twenty iterations in this case), thus asking more "questions". It is not surprising that more "questions" need to be asked under noisy conditions because there is more uncertainty associated with each response. For example, when playing "20 questions with a liar", you may have to ask the same question multiple times before you become confident about the correct response. Most importantly, these results suggest that our image search algorithm is fairly robust to noise.

4 Discussion

Our results demonstrate that our image search algorithm can, with relative consistency, identify regions containing a voxel activating feature. Moreover, our image search appears to be robust to noise. When our algorithm does fail to identify the correct feature, this seems to be due more to a poor design of our simulated "voxel', rather than to a weakness in the image search algorithm itself. Thus, we believe this paradigm would be effective in a true real-time fMRI system. However, some improvements need to be made before it is implemented in real-time. Importantly, the algorithm needs to be optimized for speed. We are summing over many variables when calculating the minimum expected entropy of a probability distribution, so the algorithm has the potential to become a bottleneck. Since one of the major motivations for this algorithm was to improve efficiency, it would be self-defeating if the stimulus generator itself became a bottleneck.

Importantly, our real-time analysis procedure is just one of several possible ap-

proaches for studying the role of image features in the VT cortex. For example Nestor *et al.* [4] used an approach to pre-identify candidate image fragments in the face. In particular, they determined a method for computing fragment diagnosticity. This diagnosticity was a measure of how unique a particular fragment was to faces vs. images of other natural objects. To accomplish this, they calculated the mutual information between fragment presence and image category across images with faces and other natural images. Other approaches include studying the underlying dimensional components of the images using principle components, or examining the role of edge hierarchies in object recognition [4]. Researchers could also consider using a generative approach, where images of faces are parameterized and unique features from this parameter space are artificially produced. For example, Yamane *et al.* [7] apply a generative approach using a genetic algorithm over a parameter space to produce highly-activating three dimensional surface features.

These techniques all successfully address the problem of identifying good candidate features. However, they do not suggest how to search through these candidate features to identify which ones recruit a particular brain region. Significantly, our algorithm not only identifies candidate features, but also tells us which ones are most likely to be activating a particular voxel. Our probability distribution over the space of image fragments tells us precisely which fragments are most likely to contain a voxel activating feature. Fragments with high probability are more likely to the voxel activating feature. In this respect, our algorithm simultaneously addresses the issue of identifying candidate features, and searching through the candidate features for the most highly activating feature.

Even more, in contrast to several of these approaches, we make very few underlying assumptions about the nature of object recognition in VT cortex. In particular,

we assume that a voxel is only responding to one image fragment contained within a single rectangle. However, it is quite possible that a voxel is responding to a small set of features in disjoint parts of the face. Thus, we would have to adjust our model to account for the fact that the voxel my be responding to multiple fragments. Additionally, we could adjust our algorithm to search over more complex non-rectangular image fragments. This could also include warping across non-linear boundaries.

Despite these assumptions, previous researchers typically make more assumptions about the nature of VT cortex than we do. Unlike their approaches, we do no pre-computation to identify candidate features, and also avoid the assumptions inherent in a generative approach. Rather, we let the image search itself discover key image fragments using only our prior distributions built from pre-collected fMRI data. Still, despite the relative strength of our approach, it is quite possible that our results will not definitively answer many of our questions. The success of some of these other methods suggest that using multiple approaches to study the same problem may strengthen evidence of specific featural codes in VT cortex. If we do identify subregions that respond to specific facial features, it would be beneficial to corroborate our results with other techniques.

Most importantly, while we only tested this algorithm in simulation, it is easily adaptable to the real system for use with human subjects. For example, we expect to build our prior distributions by showing subjects a large set of warped and unwarped faces, most likely during a separate scanner session. Using these prior distributions, we can run the algorithm as presented in this paper, almost without modification. However, first, we will need to create a mapping from the hemodynamic response function to a real-valued voxel response. This is nontrivial, in particular because this function can be noisy and hard to approximate parametrically. Additionally, because

of a phenomenon known as "scanner drift", the magnitude of the HRF decreases throughout the course of the experiment. While these challenges can be overcome, our solutions must be fast because the analysis occurs in real-time. Nevertheless, the image search algorithm presented in this paper represents the first major step toward the successful implementation of a real-time fMRI system for the study of VT cortex.

References

[1] Grill-Spector, K., Sayres, R., and Ress, D. (2006). High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nature Neuroscience* **9** (9) 1177–1185

[2] Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., and Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex *Science* **293** 2425–2430

[3] LaConte, S., Peltier, S., and Hu, X. (2002). Real-Time fMRI Using Brain-State Classification *Human Brain Mapping* **28** 1033-1044

[4] Nestor, A., Vettel, J., and Tarr, M. (2008). Task-specific Codes for Face Recognition: How They Shape the Neural Representation of Features for Detection and Individuation

[5] O'Toole, A., Jiang, F., Abdi, and H., Haxby, J. (2005). Partially Distributed Representation of Objects and Faces in Ventral Temporal Cortex *Journal of Cognitive Neuroscience* **17**)(4) 580–590

[6] Tanaka, K. (2003). Columns for Complex Visual Object Features in the Inferotemporal Cortex: Clustering of Cells with Similar but Slightly Different Stimulus Selectivities *Cerebral Cortex* **13** (1), 90–99.

[7] Yamane, Y., Carlson, E., Bowman, K., Wang, Z., and Connor, C. (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience* **11** (11) 1352–1360