J Neurophysiol 94: 4068–4081, 2005.
First published August 24, 2005; doi:10.1152/jn.00130.2005.

Selectivity of Inferior Temporal Neurons for Realistic Pictures Predicted by Algorithms for Image Database Navigation

Sarah Allred, Yan Liu, and Bharathi Jagadeesh
Departments of Physiology and Biophysics, University of Washington, Seattle, Washington

Submitted 3 February 2005; accepted in final form 14 August 2005

Allred, Sarah, Yan Liu, and Bharathi Jagadeesh. Selectivity of inferior temporal neurons for realistic pictures predicted by algorithms for image database navigation. *J Neurophysiol* 94: 4068–4081, 2005. First published August 24, 2005; doi:10.1152/jn.00130.2005. Primates have a remarkable ability to perceive, recognize, and discriminate among the plethora of people, places, and things that they see, and neural selectivity in the primate inferotemporal (IT) cortex is thought to underlie this ability. Here we investigated the relationship between neural response and perception by recording from IT neurons in monkeys while they viewed realistic images. We then compared the similarity of neural responses elicited by images to the quantitative similarity of the images. Image similarity was approximated using several algorithms, two of which were designed to search image databases for perceptually similar images. Some algorithms for image similarity correlated well with human perception, and these algorithms explained part of the stimulus selectivity of IT neurons. Images that elicited similar neural responses were ranked as more similar by these algorithms than images that elicited different neural responses, and images ranked as similar by the algorithms elicited similar responses from neurons. Neural selectivity was predicted more accurately when the reference images for algorithm similarity elicited either very strong or very weak responses from the neuron. The degree to which algorithms for image similarity were correlated with human perception was related to the degree to which algorithms explained the selectivity of IT neurons, providing support for the proposal that the selectivity of IT neurons is related to perceptual similarity of images.

INTRODUCTION

Neurons in the inferotemporal (IT) cortex can be selective for realistic images of people, places, and things. These images can be ranked from effective to ineffective, according to response strength evoked by each image in an individual neuron (Erickson et al. 1999; Fig. 4). On inspection, it can be difficult to identify common characteristics among effective images, although sometimes image simplification can find these features (Tanaka 1993). Experiments that correlate behavior with neural selectivity for images that vary in predefined dimensions show that IT neurons are more selective for images that primates find easier to discriminate and less selective for images that primates find more difficult to discriminate, even when physical differences between images are equivalent (Baylis and Driver 2001; Kayaert et al. 2003; Op de Beeck et al. 2001; Rollenhagen and Olson 2000; Vogels and Orban 1996; Vogels et al. 2001). In addition, visual experience and training with images can modify the neural and behavioral responses to those images (Erickson and Desimone 1999; Sakai and Miyashita 1991; Sigala and Logothetis 2002). These studies demonstrate that when images vary along a small number of dimensions, neurons in IT cortex are selective for perceptually important differences in images and tolerant to perceptually unimportant differences in images and that experience can change both the perceptual significance and tuning of IT neurons. We propose that selectivity is predominantly dependent on the perceptual similarity of the stimuli, even when images vary along many dimensions that have not been predefined: neurons will respond similarly to perceptually similar realistic images and differently to perceptually dissimilar realistic images. Dimensions relevant to perception may include visual features of the image but may also include cognitive information that monkeys have about images.

Previous experiments designed to discover underlying dimensions of IT response have usually involved stimuli that have varied along only a few predefined dimensions. Although these experimental stimuli have often been designed using a principled approach based on known coding in visual areas preceding IT cortex, such as V1 and V4 (Komatsu 1992; Schwartz et al. 1983), using stimulus sets that vary in few dimensions limits the understanding of global coding mechanisms in IT. The responses of IT neurons presumably give rise to perception, and perception may be modified and shaped by the global content of an image set or the particular context surrounding an image. Simple stimuli may elicit selectivity from IT neurons when presented in isolation, but that selectivity may not be maintained when IT neurons encounter the simple stimuli embedded in real-world stimuli (Sheinberg and Logothetis 2001). Furthermore, if images in a particular set vary only in one or two dimensions, the selectivity for an image pair in that set might be very different from that occurring if those two images were part of a wider set that varied in many different dimensions. In the former case, two images might be perceived as being very different relative to other images, whereas in the latter case, the two images might be perceived as being quite similar to each other.

To test our proposal that the dimensions of selectivity in IT are inherently perceptual even in uncontrolled image sets, we collected neural responses to stimuli that varied along a wide range of physical and cognitive dimensions, and we quantified image similarity in a way that reflects high-level perceptual processes. We then compared the selectivity of neural responses to measured similarity of images.

Quantifying similarity of realistic images is difficult to do both because the appropriate dimensions of physical space are poorly understood and because computational methods cannot

Address for reprint requests and other correspondence: B. Jagadeesh, Dept. of Physiology and Biophysics, University of Washington, Box 357330, Seattle, WA 98195 (E-mail: bjag@u.washington.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

take into account experience that may make certain images more similar to some individuals than to others. Advances have recently been made using algorithms developed for the practical purpose of searching large photographic databases (Rogowitz et al. 1998; Rubner 1999; Wang et al. 2001). These algorithms attempt to find numerical descriptions of image similarity so that images qualitatively similar to a target image can be located without resorting to textual descriptions of image content. We have used two such algorithms, Semantics-Sensitive Integrated Matching for Picture Libraries (SIMPLIcity) (Wang et al. 2001) and EMD, (Rubner 1999) to assess similarity of realistic images. Because such algorithms are static and cannot take into account human learning or experience, they are only an approximation of perceptual similarity. We also used three other measures (based on color and contrast) that are commonly used to evaluate image similarity although not in the context of searching image databases

In our experiments, we determined that SIMPLIcity and EMD judged similarity of realistic images in a way that corresponded to human perception. We were thus able to examine the relationship between IT neural response and perceptual similarity of realistic images that vary in many uncontrolled dimensions, building on the work of studies that found IT selectivity for particular features in controlled image sets. In individual cells and across the population, algorithms for perceptual similarity predicted neural responses to realistic images. Furthermore, with respect to all five image similarity measures, we found a correlation between the predictive value of algorithms for perception and neural response; that is, algorithms that correlated well with human perception were better able to predict neural selectivity than those that were not.

Parts of this work have been published previously in abstract form (SFN 2002, ACM SIGGRAPH 2004).

METHODS

Neural responses

We recorded from 199 IT neurons (222 experiments) in three adult rhesus macaques (*monkeys I, L,* and *G*). Surgeries on each animal were performed to implant a head restraint, a cylinder to allow neural recording, and a scleral search coil to monitor eye position (Fuchs and Robinson 1966). Materials for these procedures were obtained from Crist Instruments (Hagerstown, MD) or produced in-house at the University of Washington.

On each day, an *x-y* stage for positioning and an electrode holder containing a sterile guide tube and tungsten microelectrode (FHC, Bowdoinham, ME) were attached to the top of the recording cylinder. The electrode was moved using a microdrive, (David Kopf Instruments, Tujunga, CA), and after passing through a pre-amp, filter, and amplifier, signals from the electrode were sorted on-line using the Alpha-Omega (Nazareth, Israel) spike sorter or a template-matching algorithm (Signal Processing Systems). Responses of single IT neurons were collected while monkeys viewed realistic images presented under a central fixation spot (see task). Coded spikes were stored on a PC at a rate of 1,000 Hz using CORTEX, a program for neural data collection and analysis developed at the National Institutes of Health (Bethesda, MD). On-line histograms were created to judge qualitative selectivity, but for the purposes of this paper, all data analysis was performed post hoc on stored data. Eye movements were monitored and recorded (1,000 Hz) using an eye coil box (Crist, Hagerstown, MD) or an eye coil box from DNI (Newark, DE). All animal handling, care, and surgical procedures were performed in accordance with guidelines established by the National Institutes of Health and approved by the Institutional Animal Care and Use Committee at the University of Washington.

Neurons were selected using anatomical and physiological criteria. Structural MRI was used to guide placement of the recording chamber, which was centered in stereotaxic coordinates 18–22 mm anterior and 15–18 mm lateral. Neural recordings were made near the center of the chamber, near the perirhinal sulcus and the anterior middle temporal sulcus. Specifically, all data from *monkey L* were collected from coordinates 21.5 mm anterior and 16.7 mm lateral, between depths of 31.8 and 35.1 mm. All data from *monkey G* were collected from coordinates 16.5 mm anterior and 15.5 mm medial, between depths of 29.6 and 35.5 mm. In *monkey I,* the stereotaxic coordinates translate to recording sites spanning a 7-mm-diam area centered ~1 mm medial to the anterior tip of the anterior middle temporal sulcus (AMTS), including parts of area TE and perirhinal cortex. No anatomical confirmation of recording sites is available from these monkeys. To isolate neurons, we moved the electrode while monkeys performed the passive fixation task (see next section) with a set of 24 images. When the experimenter judged qualitatively that a neuron responded to at least 1 of the 24 images, she continued to record from that neuron. Post hoc analysis of selectivity for the 24 images (1-way ANOVA, $P < 0.05$) determined whether neural data would be kept for the experiment or discarded. Images were presented sequentially, and each image was presented for 5–15 trials. Responses to multiple image sets were collected when neural stability permitted.

Task

Monkeys were rewarded for completing a passive fixation task. A fixation point appeared. The image appeared 617 (or 317) ms after the monkey acquired fixation (fixation window either $\pm$ 2 or $\pm$0.75°, depending on experiment). Images were presented for two successive 300 (or 200)-ms epochs, intervened by a 300-ms epoch in which only the fixation spot was presented. Monkeys were required to maintain fixation throughout the trial and were rewarded with drops of water or juice. Failure to maintain fixation resulted in a time-out period followed by a repeat trial. Our image sets consisted of 17 predetermined groups of 24 images, and most neurons (199) were tested with only one image set. Monkeys were familiar with the image sets, and some image sets were used repeatedly for multiple neurons.

Analysis of neural data

Only responses from successfully completed trials were included in this analysis. Data were analyzed after the onset of the first presentation of each stimulus, and data from the second image presentation were discarded. All analysis in this paper was performed on the mean response across trials from the period 75–250 ms after stimulus onset. Responses were initially analyzed by calculating the mean response during 10 time windows of different durations and onsets. In all time windows, the pattern of response was nearly identical to the 75- to 250-ms time window. Peristimulus time histograms were constructed by smoothing with a Gaussian time window of 60 ms. No substantial response differences were found among monkeys, image resolutions, or image sizes.

Stimuli

Realistic images of people, animals, natural and manmade scenes, and objects were used. All images were cropped to be 90 $\times$ 90 pixels and were drawn from a variety of sources, including the World Wide Web, image databases, and personal photo libraries. The majority of neural data (128 of 222 experiments) were collected using five sets of 24 images that were chosen at random from a database of 4,832 images. Images were presented on a computer monitor with either 800 $\times$ 600 or 600 $\times$ 480 resolution (refresh rate: 60 Hz). Images were

displayed at either 4° of visual angle or 1.5° of visual angle, and viewing distance was ~57 cm.

Image analysis

Five different methods were used for analyzing the images, and all methods were performed on cropped images.

SIMPLICITY. The SIMPLIcity algorithm first segments images into regions to create signatures and computes similarity between signatures using an integrated region-matching scheme. The code for implementing SIMPLIcity on image databases is available from http://wang.ist.psu.edu and is described in detail in Wang et al. (2001). Here we briefly describe the process. An image is first divided into 4×4 pixel blocks, and a feature vector is calculated for each block. Each feature vector has six dimensions. Three are color dimensions, corresponding to the mean color in each of the L,U,V dimensions of LUV color space. The other three dimensions are related to spatial frequency of the 4×4 block, computed by wavelet analysis on the L component of the image. One dimension each is computed from the HL, LL, and HH bands by calculating a weighted sum of the 2×2 coefficients returned by the wavelet analysis. This particular form of wavelet analysis was chosen because the individual bands are correlated with texture perception (Wang et al. 2001).

The number of regions is then calculated using a *k*-means algorithm. The algorithm starts with one region ($k = 1$) and continues to segment the feature vectors into more regions ($k = 2, 3, \ldots n$) until certain conditions are met. Intuitively, feature vectors from blocks are grouped together as one region if the feature vectors are deemed similar enough by the algorithm.

Once images have been segmented into a variable number of regions, they are classified into different semantic categories of texture, nontexture, graph, and photograph. The SIMPLIcity algorithm considers these semantic categories to be mutually exclusive and cannot calculate the similarity between images included in different semantic categories. In this paper, all data described are for responses to images classified as photographs, so we focus here on the elements of SIMPLIcity devoted to photograph images. Examples of images and their segmented versions are illustrated in the center and left panels of Fig. 1*A* (*right panel* to be discussed later). The segmented version represents the number of different features. Each displayed color represents the average color of that particular feature; spatial frequency components are not represented in this illustration.

After sorting into semantic classes, the similarity (SIMPLIcity value) between two images is computed by comparing regions using integrated region matching (IRM). Different images have different numbers of regions; in IRM, a region in one image can be matched to more than one region in another image. The SIMPLIcity value between images is the weighted sum of the similarity values between regions. The procedure for calculating similarity is described in detail in Wang et al. (2001). Briefly, a significance matrix is computed between all possible region pairs, where significance is related to the relative importance (size) of each region in an image. Then a distance between regions is calculated by summing the differences between each color and spatial frequency dimension. The similarity between regions is assigned by multiplying the distance between each region pair by the significance of that region pair, and the SIMPLIcity value is the weighted sum of all similarity values. Smaller SIMPLIcity values indicate more similar images. Figure 1*B* shows examples of SIMPLIcity values between image pairs. Note that SIMPLIcity values are not additive.

In our database, the mean SIMPLIcity value is 42.35 across all 21.3 million values in our database, which compares to a mean value of 44.30 across 5.3 million values in the COREL database using 100 target images as reported in Wang et al. (2001).

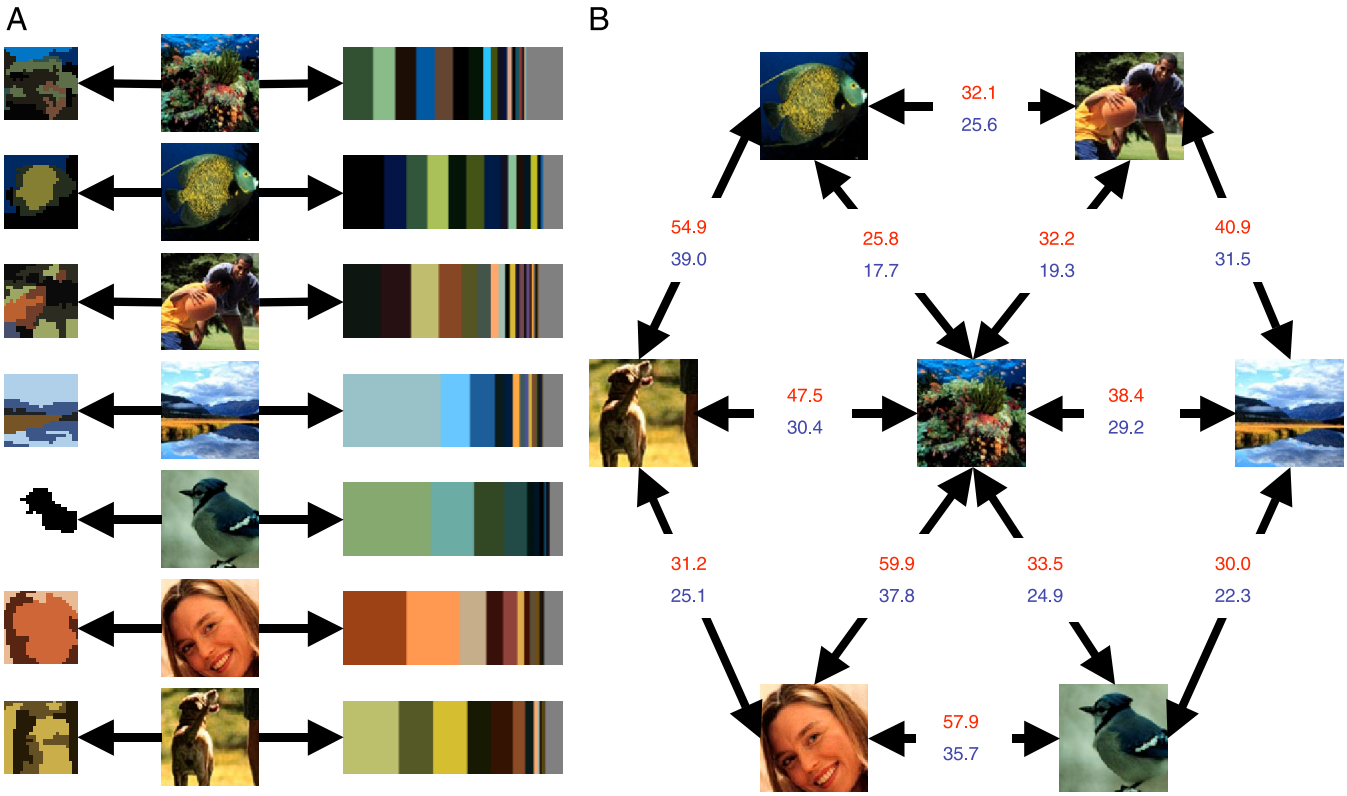

FIG. 1. Illustration of algorithm mechanisms. Images obtained from Comstock: www.comstock.com and the authors' personal picture libraries. *A*: segmentation process. *Middle*: 7 images from 1 image set. *Left*: segmented version of the image using SIMPLIcity. *Right*: color histograms of images using EMD. *B*: example algorithm values between images are shown. Red numbers indicate SIMPLIcity values between image pairs, and blue numbers indicate EMD values between image pairs. Smaller values indicate more similar images.

*Image Analysis: Earth Mover Distance (EMD).*The EMD (Rubner 1999) first analyzes images to create signatures and then computes similarity between signatures by computing the cost to transform one signature into another signature. The code for computing similarity between signatures is available as C code from http://www.cs.duke.edu/~tomasi/software/emd.htm, and the code for creating signatures was created in-house using MATLAB. We modified the Rubner code to run in MATLAB. Details of the EMD are found in Rubner (1999) and are briefly outlined here.

Each image is first transformed into CIE Lab space, and color signatures are then produced that consist of a variable number of bins corresponding to the color content of the image. Each feature within a signature consists of a weight corresponding to the fraction of total pixels contained in that feature, and three values corresponding to the mean L, a, and b components contained within the feature. Examples of images and their EMD signatures are illustrated in Fig. 1*A, right* and *middle*. Each color represents the average color of one feature; the width of the bin represents the weight of that feature in the signature. The EMD between two images is the least-possible work involved in transforming one signature into another signature. Smaller EMD values indicate more similar image pairs. The mean EMD within our database is 32.98 across all 21.3 million values in our database. Like SIMPLIcity values, EMD values are not additive.

IMAGE ANALYSIS: RMS CONTRAST. The RMS contrast is the SD of luminance values and was calculated using the following equation for contrast (Bex and Makous 2002), where L is pixel luminance in CIE Lab space, and N is the total number of pixels

$$C_{\mathrm{RMS}} = \left[\frac{\Sigma L_{(x,y)}^{\hat{}2} - \frac{(\Sigma L_{(x,y)})^{\hat{}2}}{N}}{N} \right]^{-1/2}$$

The contrast similarity between two images is the normalized difference between the two RMS contrast values

$$\text{Contrast Similarity} = \frac{|(\mathrm{RMS}_{\mathrm{image1}} - \mathrm{RMS}_{\mathrm{image2}})|}{(\mathrm{RMS}_{\mathrm{image1}} + \mathrm{RMS}_{\mathrm{image2}})}$$

IMAGE ANALYSIS: CORRESPONDING PIXEL IMAGE (CPI) SIMILARITY AND MEAN COLOR SIMILARITY. Each image is transformed into CIE Lab space. There are several ways to determine corresponding pixel image similarity. We used the equation described in Op de Beeck et al. (2001) for luminance similarity and expanded it to two color dimensions. The CPI similarity is calculated by taking the square root of the sum of the squares of individual pixel differences. The CPI similarity between two 90 × 90 pixel images, i and j, in the a and b color dimensions of CIE Lab space is as follows

CPI Similarity

$$= \sqrt{\Sigma_{x=1:90}\Sigma_{y=1:90}(a_i(x,y) - a_j(x,y))^2 + \Sigma_{x=1:90}\Sigma_{y=1:90}(b_i(x,y) - b_j(x,y))^2}$$

Mean color similarity is calculated by taking the mean color for each image in a and b dimensions, and then finding the Euclidean distance between the two mean colors. The mean color similarity between two N-pixel images, i and j, in the a and b dimensions, is as follows

Mean Color Similarity

$$= \sqrt{\left(\frac{\Sigma_{x=1:90}\,\Sigma_{y=1:90}\,a_i(x,y)}{N} - \frac{\Sigma_{x=1:90}\,\Sigma_{y=1:90}\,a_j(x,y)}{N} \right)^2 + \left(\frac{\Sigma_{x=1:90}\,\Sigma_{y=1:90}\,b_i(x,y)}{N} - \frac{\Sigma_{x=1:90}\,\Sigma_{y=1:90}\,b_j(x,y)}{N} \right)^2}$$

IMAGE ANALYSIS: ELIMINATING IMAGES. In our database, 4610/4832 (95%) of images were classified as photographs by the SIMPLIcity algorithm. SIMPLIcity cannot calculate similarity between images of different semantic classes. For the purposes of this paper, we therefore eliminated those images that were not classified as photographs by SIMPLIcity. This reduced our overall database to 4,610 pictures. This elimination was done post hoc, and meant discarding 6 of our 200 target images (Fig. 2) as well as 13 images that appeared in our experimental image sets. In the case where neural data were collected to an image that was discarded, the neural data for that particular image were also discarded, and analysis continued without that data point.

HUMAN SIMILARITY JUDGMENTS. We collected similarity judgments from four naïve human subjects with 5 of the 17 image sets used in neural recordings. These five image sets were used in 128 of the 222 experiments. A Microsoft PowerPoint file was created with six square slides on a Macintosh G3 laptop. The first slide contained 10 practice images that were not members of the experimental image sets, and images were placed in random locations on the square slide. Each subsequent slide held the 24 images from one image set that were in a randomly ordered rectangular grid that was different for each subject. Subjects were asked to move the images so that similar images were next to each other, and dissimilar images were far from each other. Subjects were also instructed to use the entire slide area, and image overlap was explicitly permitted. No definitions of similarity were given. After subjects completed their sorting of the images and saved the file, we recorded the location (in pixels) of the images on the slide. The distance (in pixels) between two images was used as the measure of image similarity with small distances indicating similar images and large distances indicating dissimilar images. Although allowing only two dimensions for sorting clearly constrains the relative distances possible between image pairs, this method does correlate with the perceptual similarity of images measured with other methods (Mojsilovic et al. 2002).

RESULTS

We recorded from 199 IT neurons (222 experiments, as some neurons were recorded with >1 image set) in three monkeys while they performed a passive fixation task with realistic images (see METHODS). We first demonstrate that the SIMPLIcity algorithm for navigating image databases returns image similarity values that are correlated with human perception. Second, we examine the relationship between neural response to images and SIMPLIcity-assessed similarity of images. Third, we show that other quantitative measures of image similarity predict neural selectivity to the extent that they also predict human judgments of image similarity.

Algorithm-judged similarity and human perceptual similarity

To test the degree to which algorithm-assessed similarity provides accurate assessments of human perceptual similarity, we added a category of 194 target images in our greater database of 4,610 images, and tested algorithm performance at retrieving within-category images. Each of the 194 target images contained the face of a particular child. Figure 2*A* shows an example of using one of the 194 target images as a query, and shown are the 11 images most similar by the SIMPLIcity algorithm (smallest SIMPLIcity values). All of the 11 images returned are target images. The chance frequency of return is 193/4,609 (0.042). Using the same image as a target, *B* shows the results of a random query of the database. No target images are returned in this random query. We quantified

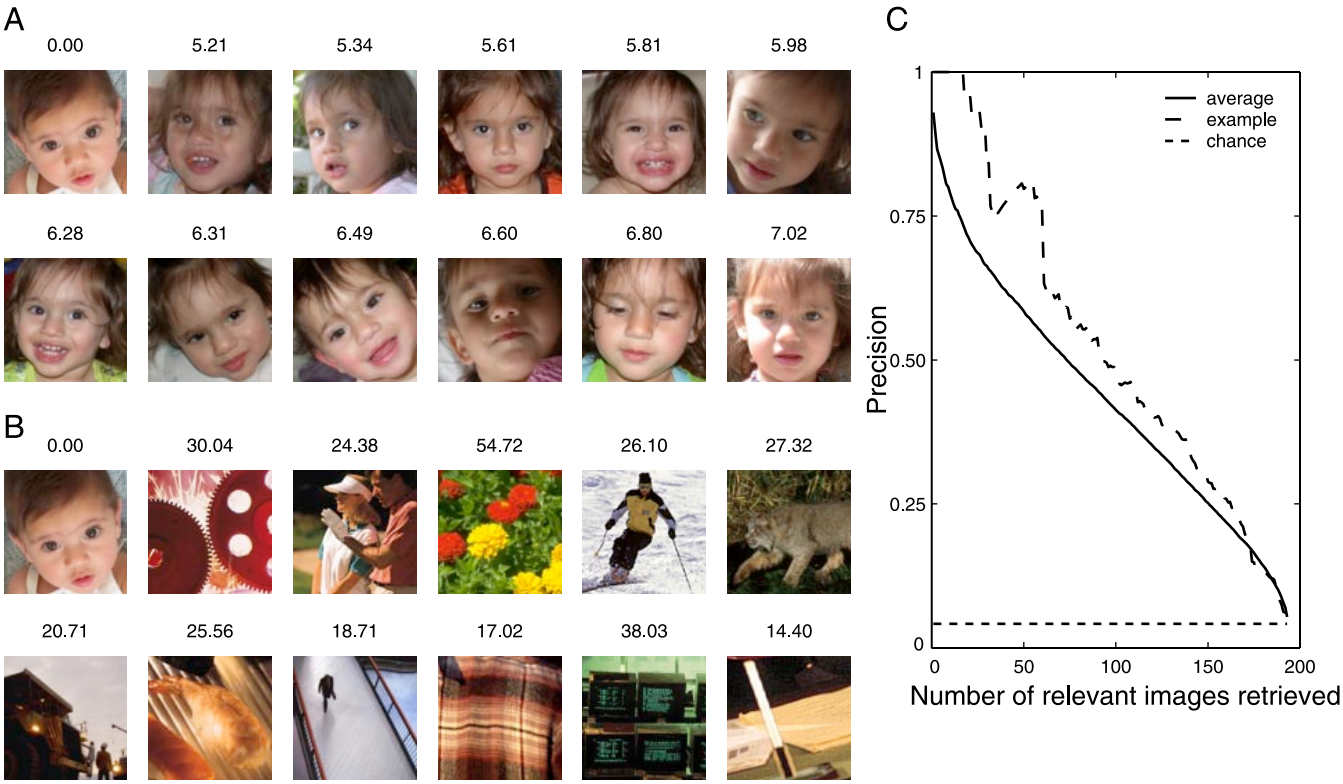

FIG. 2. Using SIMPLIcity to query our database. Images in *B* obtained from Comstock: www.comstock.com and the authors' personal picture libraries. *A*: using the *top left image* as a reference image, we queried our database using the SIMPLIcity algorithm. Above each image is the SIMPLIcity value between that image and the reference image. Shown are the 11 images most similar (smallest SIMPLIcity values) to the reference image. Our database contains 4,610 images, 193 of which contain the target. *B*: using the same reference image shown in the *top left corner,* results of pulling 11 images randomly from the database. *C*: precision of algorithm as a function of number of images recalled. —, average using each of the 193 targets as reference image. Long dashes, the precision using the example target image in *A* and *B* as the reference image. Short dashes, chance precision.

the results using a standard precision-recall plot (Rodden et al. 2000). If the algorithm performed perfectly, it would retrieve all of the 193 target images before recalling any other images in our database. For each target image, we can calculate the precision with which the algorithm recalls other target images. For the example shown in Fig. 2*C*, the precision is 1.00 until 18 target images have been retrieved; at this point, an irrelevant image is retrieved before the next target, so the precision drops to 18/19 (0.95). The results, averaged across each of the 194 target images, are shown quantitatively in Fig. 2*C*. If the algorithm performed at chance, we would expect a flat precision line of 0.042, with about 1 of 24 images recalled containing a target image. The SIMPLIcity algorithm performs quite well, with an average precision of 0.43 $\pm$ 0.010 (SE) across all recall numbers and all target images.

In addition to the standard precision-recall graph, we also computed a direct correlation between algorithm-judged similarity and human judgments of similarity on the subset of images that comprised the majority of the neural data collected. Briefly (see METHODS for details), four naïve subjects were shown sets of images (corresponding to individual sets seen by the monkeys) and asked to sort them on a computer monitor by perceptual similarity. They were given no instructions on how to sort images other than to place similar images close to each other and different images apart from each other and to use the entire space provided by the computer monitor. We then calculated how far apart (in pixels) two images were and used that pixel distance as a measure of how different the images are. An example of one subject's sorting is shown in Fig. 3*A*. This subject reported that the images of animals were similar to each other (demonstrated by the cluster of animals in the *top right*), and judged the image of the mountain on the bottom left to be quite different from those animals. In Fig. 3*B*, we plot pixel distance between images as a function of SIMPLIcity between images. When image pairs were similar by SIMPLIcity, they were perceptually similar to this subject (images placed close together on slide, r value = 0.45, $P < 0.00001$). This result was found across all image sets (Table 1). Because pixel distance and SIMPLIcity values are both measures of how different two images are, these correlations demonstrate that algorithm-judged similarity is correlated with human judgments of perceptual similarity.

Algorithm-judged similarity and neural selectivity

While the monkeys performed a passive fixation task with sets of 24 images, we recorded the responses of isolated IT neurons. The responses of one neuron to a set of 24 images are shown in Fig. 4. These peristimulus time histograms were sorted post hoc in order of decreasing neural response, and magnitude of response was calculated as the average spike rate in the window 75–250 ms after stimulus onset. This neuron was selective; it responded significantly differently to stimuli that were presented, (as analyzed post hoc by a 1-way ANOVA, $P < 0.00001$), with a minimum response of 21 spikes/s and a maximum response of 68 spikes/s. The images

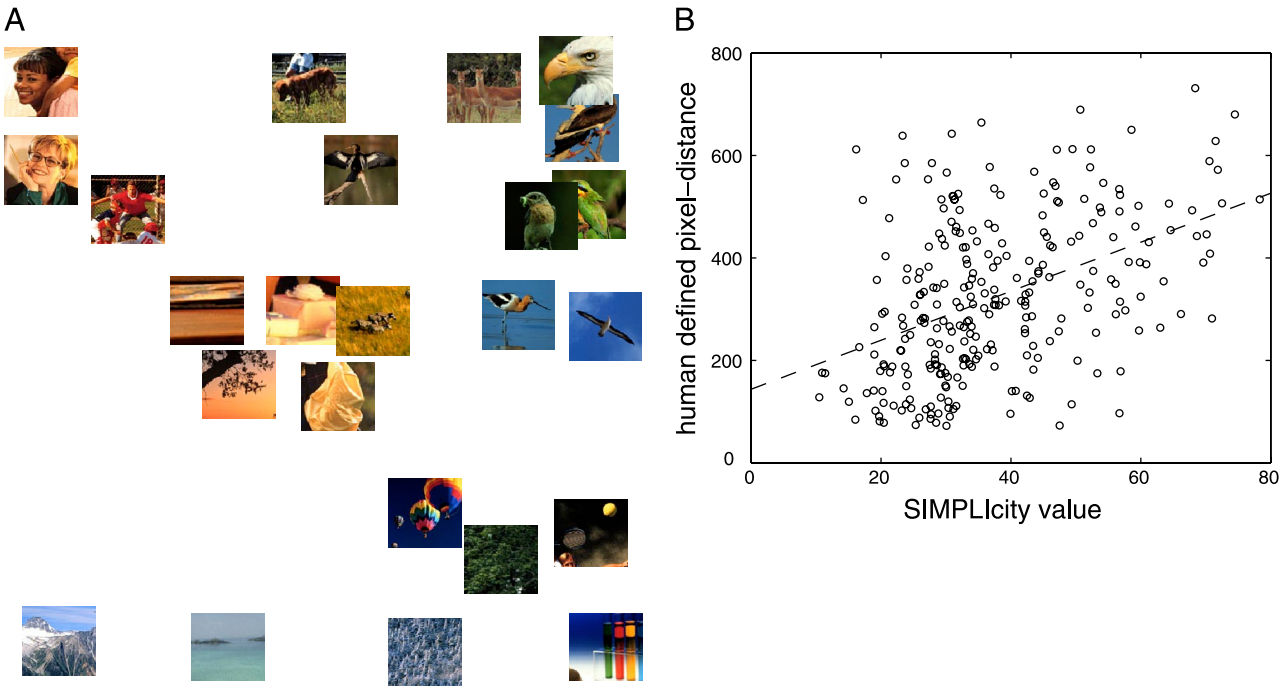

FIG. 3. Comparison of image similarity assessed by both human subjects and the SIMPLIcity algorithm. Images in *A* obtained from Comstock: www.comstock.com and the authors' personal picture libraries. *A*: example of 1 subject's sorting of a set of 24 images. *B*: pixel-distance between each possible image pair vs. SIMPLIcity value between each image pair. Each point represents 1 image pair in this set of 24 images. - - -, the least-squares regression of pixel distance vs. SIMPLIcity value (r value = 0.45, $P < 0.0001$).

we used spanned a large range of IT response space; the average minimum response was 10 spikes/s (ranging from 0 to 84 spikes/s), and the average maximum response was 43 spikes/s (ranging from 3 to 177 spikes/s).

If the underlying parameters of response space in IT are related to perceptual similarity, then images that elicit similar responses should be judged more similar by the algorithms than images that elicit dissimilar responses. We computed algorithm similarity between the "best" image, the "next-best" image, and the "worst" image, defined as the images that elicited the strongest response, the next strongest response, and the weakest response from the neuron, respectively (68, 61, and 22 spikes/s). In Fig. 4, the best, next-best, and worst images are indicated by the text above them. For the example cell, the best and next-best images were found to be much more similar by the SIMPLIcity algorithm than were the best and worst images (Fig. 5*A*, 25.8 vs. 38.6). This means that two images judged as similar by the neuron (they elicited similarly strong responses) were also judged as similar by the algorithm; likewise, two images judged as dissimilar by the neuron (they elicited the maximum response difference) were judged as dissimilar by the algorithm.

TABLE 1. *Correlation between human-judge similarity and simplicity*

Image List	SIMPLIcity	Mean Inter-Subject Correlation
List 1	0.34***	0.75 ± 0.016
List 2	0.33***	0.78 ± 0.014
List 3	0.15*	0.73 ± 0.016
List 4	0.45***	0.78 ± 0.013
List 5	0.20**	0.74 ± 0.017

Four naïve human subjects were asked to sort five sets of 24 images on a computer screen. Subjects were asked to place similar images close to each other, to place dissimilar images far from each other, and to use the entire monitor space. Subjects were not instructed on the definition of similarity. We calculated the distance in pixels between the centers of each possible image pair (276 possible pairs). Distances were averaged between subjects and then correlated with algorithm values between the same image pairs. The *middle column* represents the r value for the correlation between the average subject distance and SIMPLIcity. The *right column* contains the average inter-subject correlation for each image list, which was calculated by taking the similarity judgments for each subject and correlating them with the similarity judgments of every other subject (yielding 6 r values for each image list). *$P < 0.05$, ** $P < 0.01$, *** $P < 0.00001$.

This result was replicated across the population. In Fig. 5*B*, we plotted algorithm-judged similarity between best and next-best images (mean: 36.66) against algorithm-judged similarity between best and worst images (mean: 45.10). The majority of points fell above the diagonal, indicating that with reference to the best image, the pair that elicited the most similar responses from the neuron were more similar to each other than the pair that elicited the most different responses from the neuron. On average, best and next-best images were significantly more similar than the best and worst images (mean difference: 8.44, $P < 0.00001$, paired t-test). The distribution of neural responses that we encountered to our individual sets of images was not a normal distribution. To assure that we calculated an unbiased P value, we included a shuffle control ($n = 500$). The shuffled mean (−0.0057) was not different from zero, and the maximum of the shuffled differences (5.41) was less than the observed shuffled value, showing empirically that our result was significant ($P < 0.002$).

Comparing responses elicited by best and next-best images is somewhat arbitrary because we have no reason to assume that the range of responses elicited by our images encompasses the possible response range of a neuron. To test whether the correlation between algorithm-assessed similarity and neural response similarity continued through other response ranges,

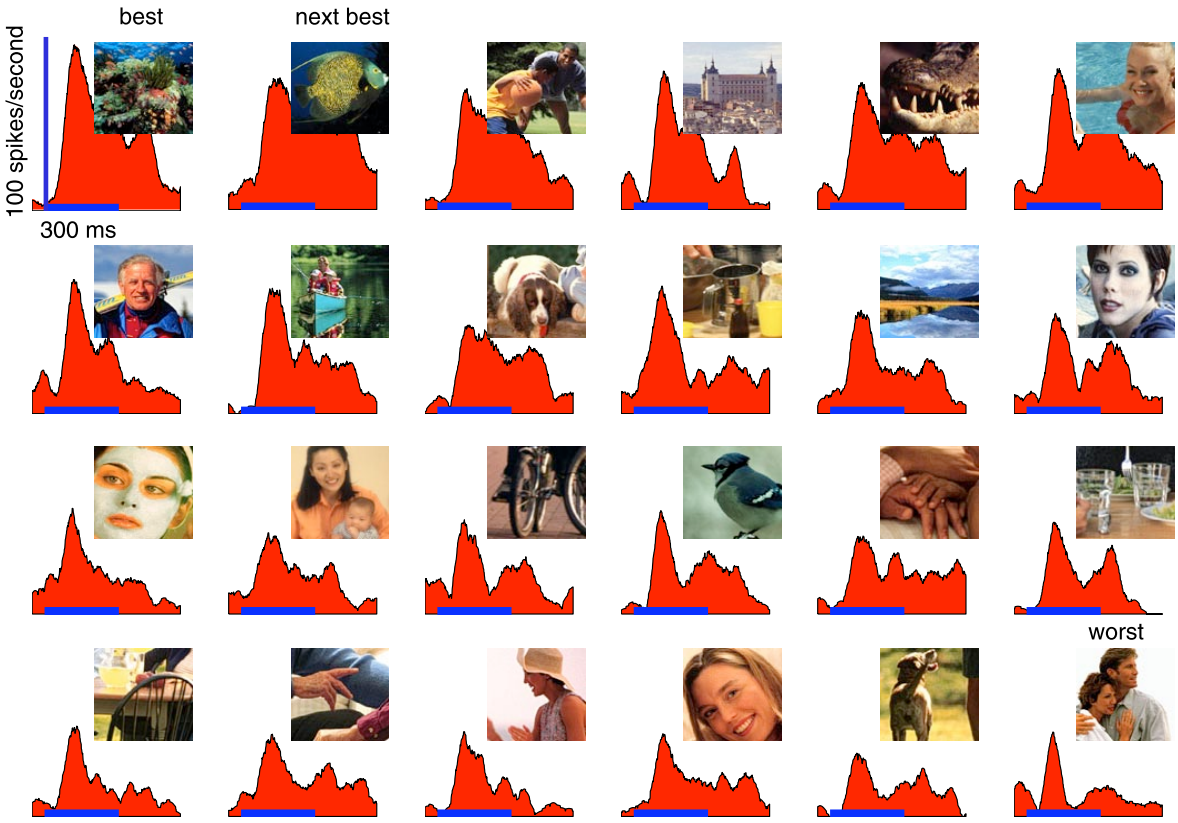

FIG. 4. Peristimulus time histogram of responses of 1 neuron to a set of 24 images. Images obtained from Comstock: www.comstock.com and the authors' personal picture libraries. Histograms were sorted post hoc in order of decreasing neural response, where response is defined as the average spike rate in the window 75–250 ms after stimulus onset. Above each histogram is the image that elicited that neural response. Horizontal blue bars, the 300 ms of stimulus presentation. Vertical blue bar, 100 spikes/s. This neuron was selective (1-way ANOVA, $P < 0.00001$). The text above images indicates those that elicited the strongest, next strongest, and weakest responses from the neuron.

we calculated similarity between two other groups of images. We defined effective images as those images that elicited responses not significantly different from the strongest response of the neuron, and ineffective images as those that elicited responses not significantly different from the weakest response of the neuron (*t*-test, $P < 0.05$). For the example neuron illustrated in Fig. 4, the first 7 stimuli were calculated to be effective, and the last 10 stimuli were calculated to be ineffective. The definition of effective and ineffective depends on the responses of the individual cell being analyzed. We calculated algorithm-judged similarity within the effective group by comparing all possible pairs of effective images, and we calculated algorithm-judged similarity between effective and ineffective groups by comparing each effective image to each ineffective image. For the example cell (Fig. 6*A*), the effective images are more similar (mean: 34.65) than the effective and ineffective images (mean: 37.51). Intuitively, this means that the group of images that elicited statistically indistinguishable responses from the neuron (effective images) were more similar to each other than they were to the images that elicited responses that were highly discriminable from the best responses from the neuron (ineffective images).

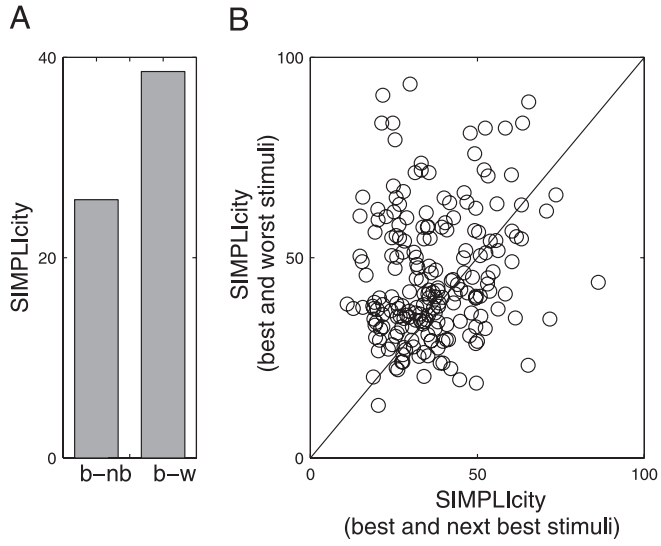

FIG. 5. Comparison of SIMPLIcity between best and next best stimuli and best and worst stimuli. *A*: example cell. The b-nb bar represents SIMPLIcity between best and next best stimuli (25.8), and the b-w bar represents SIMPLIcity between best and worst stimuli (38.6). *B*: population. Each point represents 1 experiment ($n = 204$). Solid line has unity slope. For points above the line, SIMPLIcity between best and worst (mean: 45.1) is larger than SIMPLIcity between best and next best (mean: 36.7).

The results for the population are shown in Fig. 6*B*, with the majority of points above the diagonal (mean within effective group = 38.69 ± 0.64; mean between effective and ineffective groups = 41.72 ± 0.55). This shows that with reference to groups of images rather than single images, algorithm-judged similarity is correlated with neural response similarity. Subtracting within-group similarity from between group similarity yields consistently positive values (mean difference = 3.03; $P < 0.00001$, paired *t*-test). Because the effective and ineffec-

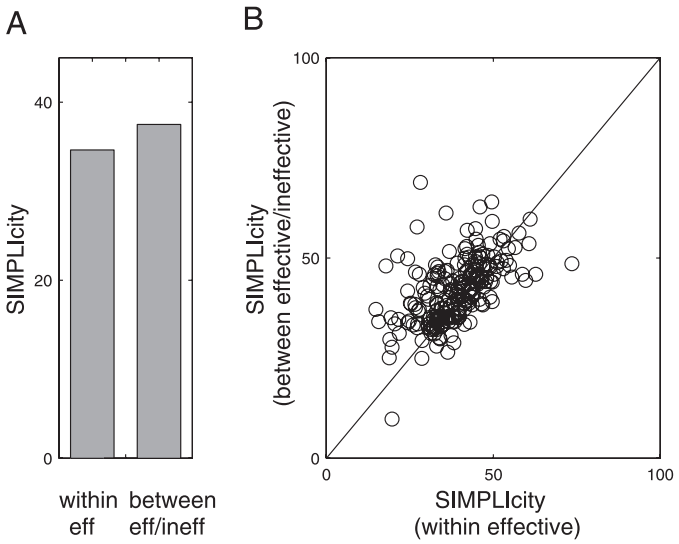

FIG. 6. Comparison of SIMPLIcity within effective group and SIMPLIcity between effective and ineffective groups. *A*: example cell. The within eff bar represents SIMPLIcity within effective group (34.7) and the between eff/ineff bar represents SIMPLIcity between effective and ineffective groups (37.5). *B*: population. Each point represents one experiment ($n = 220$). Solid line has unity slope. For points above the line, SIMPLIcity between effective and ineffective groups (mean: 41.7) is larger than SIMPLIcity within effective group (mean: 38.7).

tive groups were defined by the neuron, rather than predefined by the image set, they often contain different numbers of images (median number of effective images = 5; median number of ineffective images = 8). The number of comparisons is also different because images within the effective group are being compared only with each other (median number of comparisons = 10), while each ineffective image is compared with every effective image (median number of comparisons = 40). The possibly greater variability in the second group of values compared with the first group might skew the average similarity values aside from any inherent relationship between neural response and algorithm-judged similarity. We examined significance by keeping the number of effective and ineffective stimuli the same for each neuron; we then calculated similarity within the effective group and between effective and ineffective groups using 500 different random assignations of stimuli to groups. The asymmetry of stimulus number does have an effect; with random assignment, the effective images are actually less similar than the effective and ineffective images (mean difference: −0.35). The maximum difference found in the shuffle control (1.64) was less than the observed difference (3.03), demonstrating the significance of our result ($P < 0.002$).

For the previous set of analyses, we allowed the neural response to determine the pairs of images to be compared by the algorithm. In our second set of analyses, we test whether images defined as similar by the algorithms evoked similar responses from neurons.

We again defined the best stimulus as the image that evoked the strongest response from the neuron. However, instead of ranking the images by decreasing neural response, as we did in Fig. 4, we sorted the images by decreasing algorithm similarity to the best image. We then calculated the correlation between neural response to an image, and the algorithm judged similarity between that image and the best image. As algorithm-judged similarity to the best stimulus decreased (SIMPLIcity values increased), the neural response also decreased. This is shown in Fig. 7*A*, where the response of the example neuron in Fig. 4 is plotted as a function of decreasing similarity to the

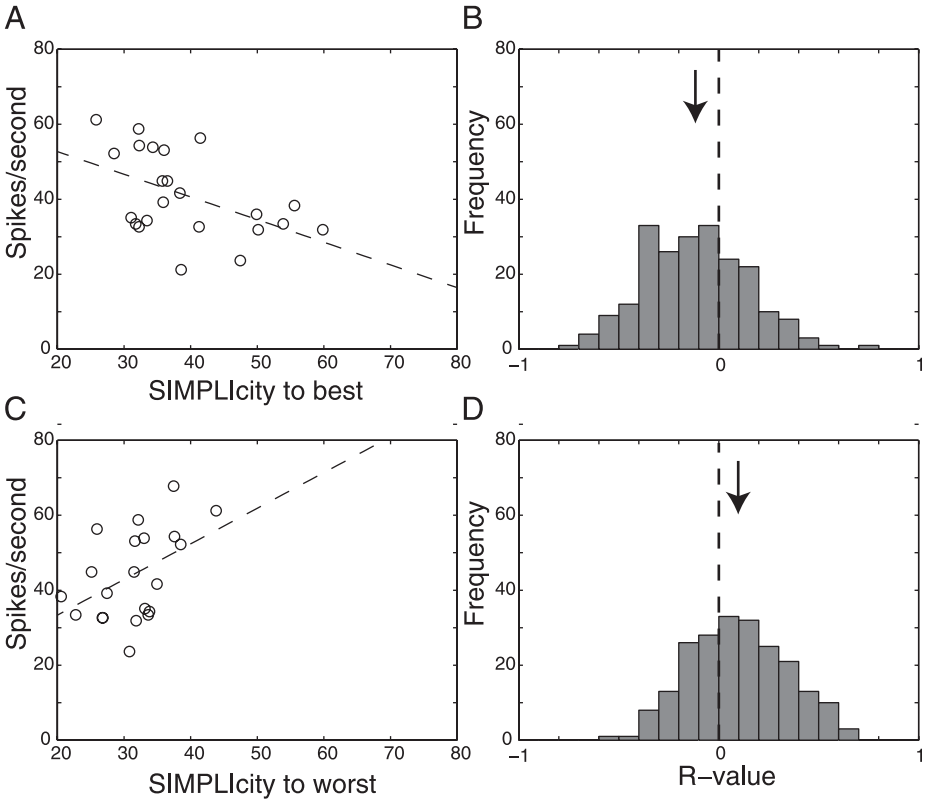

FIG. 7. Relationship between SIMPLIcity to best (or worst) image and neural response. *A*: example neural response to an image as a function of SIMPLIcity between that image and the best image (r value, −0.49, $P < 0.05$). - - -, least squares regression of neural response vs. SIMPLIcity. *C*: example neural response to an image as a function of SIMPLIcity between that image and the worst image (r value = 0.53, $P < 0.01$). - - -, least squares regression of neural response vs. SIMPLIcity. *B* and *D*: population. Distribution of r values for each experiment (*B*, $n = 217$ *D*, $n = 214$). Each r value is the correlation between neural response to an image and SIMPLIcity between that image and the best image (*B*) or the worst image (*D*). - - -, 0 correlation; —, the population means (*top*: mean r value = −0.12 ± 0.02, $P < 0.00001$; *bottom*: mean r value = 0.10 ± 0.02, $P < 0.000001$).

best image. As SIMPLIcity value increases, neural response decreases ($r = -0.49$, $P < 0.05$).

We also calculated the algorithm-judged similarity between each image and the worst image. The algorithm values in this case are all different from in the preceding analysis because the reference image is different, though the neural responses are the same. In this case, as the image became less similar to the worst image (SIMPLIcity values increased), the neural response increased (Fig. 7*C*, $r = 0.53$ $P < 0.01$) because we moved away from an image that caused a poor response.

Our population (Fig. 7, *B* and *D*) confirms the correlation between neural response and algorithm-judged similarity: when images were similar to the best image, neural response was high, and when images were dissimilar to the best image, neural response was low (mean r value $= -0.12 \pm 0.02$, $P < 0.00001$). Conversely, when images were similar to the worst image, neural response was low, and when images were dissimilar to the worst images, neural response was high (mean r value $= 0.10 \pm 0.02$, $P < 0.00001$). In calculating these correlations between neural response and SIMPLIcity, the reference image is the best image in one case and the worst image in the second case. We removed the best-best and worst-worst SIMPLIcity values because SIMPLIcity between an image and itself is 0 by definition and would artificially inflate our correlations. This left 22 distinct SIMPLIcity values for each correlation and 1 overlapping value because SIMPLIcity between best-worst will be present in both lists.

When best and worst images were used as reference images, correlations between neural response and SIMPLIcity were significant. We also used every other intermediate image as a reference image to test the extent of the correlation between neural response similarity and algorithm-judged similarity. To test whether algorithm similarity to images that evoked intermediate responses from the neuron also predicted response differences, we repeated the calculations in Fig. 7, *A* and *C*, using each image (not just best and worst) as a reference image. In this case, rather than using the neural response evoked by each image, we used the difference in response evoked by the reference image and each other image. Each point in Fig. 8 represents the average correlation of response difference and algorithm-judged similarity for all the experiments where SIMPLIcity could be defined for the relevant image pairs. The mean r values to best image and to worst image in Fig. 7, *B* and *D*, are seen in Fig. 8 as the data points for rank 1 and rank 24, respectively. Again, because our values were not normally distributed, we performed a shuffle control to verify that the curve is not a result of the possible neural responses to each reference image. The experimental values lie significantly above the shuffled values when comparing to images of rank 1–5, and rank 20–24, ($P < 0.05$, Bonferroni corrected) but not when comparing to images of rank 6–19. Clearly the correlations between algorithm-judged similarity and neural response differences are the strongest when images evoked either very strong or very weak responses from the neuron. Finally, we computed for each cell the correlation between image similarity of all possible image pairs and neural response difference for all possible image pairs (276). The distribution of correlations demonstrates that although still significant (mean r value, 0.042 ± 0.0086, $P < 0.00001$, data not shown), the mean correlation was less than that observed in Fig. 7, when only response differences to best and worst images were used. This is the result of averaging in the much lower correlations when using intermediate stimuli as reference images as seen in Fig. 8.

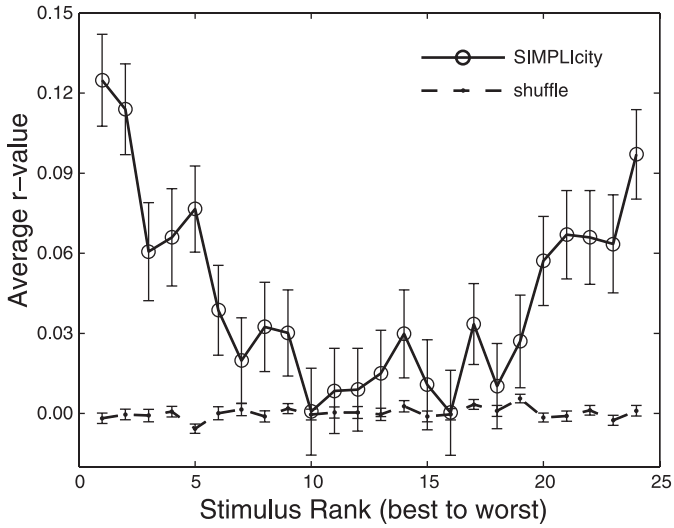

FIG. 8. Relationship of response differences between pairs of images and SIMPLIcity values between pairs of images. ○, average correlations across all experiments between neural response differences and SIMPLIcity values for each stimulus rank. - - - shuffle average where responses were randomly assigned to images. - - -, represents mean chance correlations. Error bars are across different random assignations ($n = 100$).

The previous analyses have all relied on the average spike counts in the time period 75–250 ms after stimulus onset. To explore the time course of the relationship between SIMPLIcity and neural response across our population, we calculated the average peristimulus time histogram (Fig. 9). Because we were interested in pursuing the time course of the relationship, we confined this analysis to the 134 neurons where SIMPLIcity was a good predictor of neural selectivity (points above the line, Fig. 6*B*). For each neuron, we used the best stimulus as a reference stimulus and ranked the 23 remaining stimuli by SIMPLIcity to the best stimulus. We divided the ranked stimuli into three groups, consisting of ranks 2–9, ranks 10–17, and ranks 18–24. Lower numbered ranks indicate images more similar to the reference image. For each cell, we averaged the spike trains from all trials within each group of rankings. We then averaged across neurons for our population histogram. Figure 9*A* shows that groups of images that are similar to the best image elicit strong responses from the neuron for the entire duration of stimulus presentation. The difference in response between the ranked groups of images is present from the initial peak of maximum response throughout the period of stimulus presentation.

The time course of the correlation between algorithm-judged similarity and neural selectivity is seen clearly in Fig. 9*B*. Here we calculated the correlations from Fig. 8*A* as a function of time after stimulus onset. In Fig. 7, we calculated the average firing rate for each image in the time window from 75 to 250 ms after image onset and then found a correlation between neural response to an image and SIMPLIcity between each image and the best (or worst) image. In Fig. 9*B*, we calculated the instantaneous firing rate for each image at every millisecond. We then calculated the correlation between SIMPLIcity to best (or worst) image and the instantaneous firing rate for each neuron and averaged across the population. The correlations

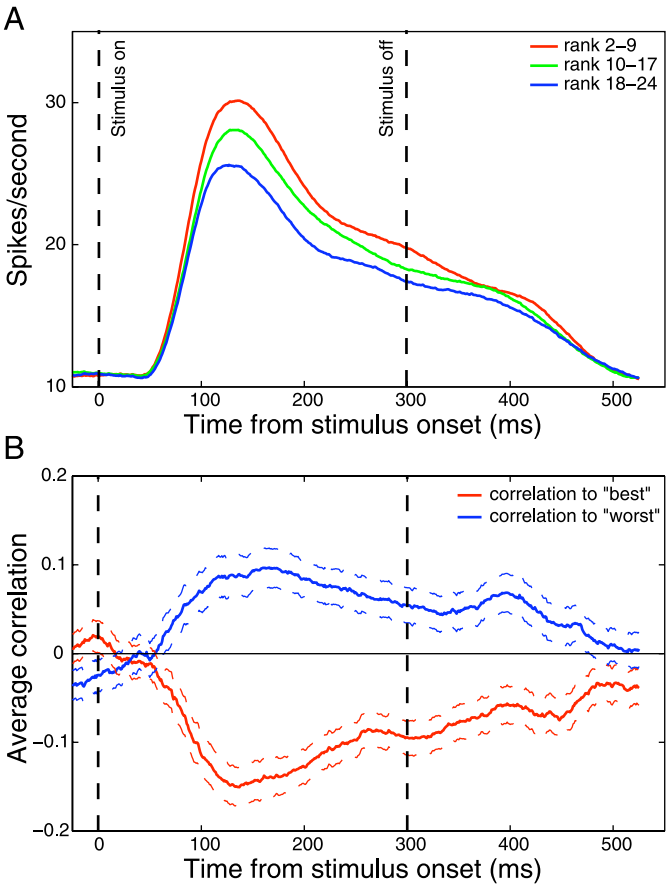

FIG. 9. Time course of results. *A*: average peristimulus time histogram grouped by SIMPLIcity rank to best stimulus. Histograms are time-locked to stimulus onset, and the stimulus was turned off at 300 ms, leaving only a fixation spot. For each neuron, we ranked the stimuli in order of SIMPLIcity value to best stimulus (not shown) and then used that ranking to divide stimuli into 3 groups. Red line, stimuli of rank 2–9; green line, stimuli of rank 10–17; and blue line, stimuli of rank 18–24. Histograms were created for each neuron, then averaged across neurons to obtain the plots shown here. Dashed lines. stimulus onset and offset. *B*: average time course of the correlation between SIMPLIcity to best stimulus and neural response (red line), and SIMPLIcity to worst stimulus and neural response (blue line). For each neuron, we calculated the average response to each stimulus as a function of time. Then for each point in time (1-ms steps), we calculated the correlation described in Fig. 8. Dashed lines, error bars across neurons. The correlation was significantly different from 0 ($P < 0.05$, 2-tailed *t*-test) for the period 69–477 ms after stimulus onset (red line) and for the period 79–431 ms after stimulus onset (blue line). $n = 134$.

began to emerge when the cell began to respond [$P < 0.05$ from 69 to 477 ms after stimulus onset, (to best); $P < 0.05$ from 79 to 431 ms after stimulus onset (to worst)], and the correlations reached their maximum values about the same time that neural response peaked [time of peak *r* value: 138 ms (to best), 167 ms (to worst); average time of peak neural response: 136 ms].

Although all the image sets tested were familiar to the monkeys (see METHODS), we also examined the effect of stimulus repetition on the correlation between algorithm-judged similarity and neural selectivity. For each data set, we separated neural responses for each stimulus into those collected on the first half of trials and those collected on the second half of trials. We repeated the correlation between algorithm-similarity to best and neural response (described in Fig. 7) for each group separately. We found no difference between the first half and second half data for the correlation to best (1st half, mean *r* value = −0.095, ± 0.018; 2nd half, mean *r* value = −0.11, ±0.17; *P* value = 0.56, unpaired *t*-test), and in neither group were the correlations significantly different from the correlations in the data set as a whole [whole set, mean *r* value = −0.12, ± 0.018; *P* = 0.38 (1st half); *P* = 0.75 (2nd half)]. Likewise, there was no difference between the first half and second half data for the correlation to worst, (1st half, mean *r* value = 0.094, ± 0.017; 2nd half, mean *r* value = 0.072, ± 0.17; *P* value = 0.37, unpaired *t*-test) nor were these correlations different from the data set as a whole [mean r value = 0.097, ± 0.017; *P* = 0.88 (1st half); *P* = 0.29 (2nd half)].

To further investigate the effects of stimulus familiarity on neural response and its correlation with algorithm judgments of image similarity, we compared data collected in repeated experiments with the same image set. Each image set was used in 1–47 experiments. For each monkey, we divided experiments into two groups; the first group consisted of the first half of experiments collected with each image set, and the second group consisted of the second half of experiments collected with each image set. We then calculated the correlation to best (and worst) for each of the two groups, and we found no effect of stimulus familiarity (to best: mean *r* value 1st half = −0.10, mean *r* value 2nd half = −0.10, *P* = 0.97; to worst: mean *r* value 1st half = 0.087, mean *r* value 2nd half = −0.11, *P* = 0.45).

Measures of image similarity, human perceptual similarity, and neural selectivity

We also investigated the relationship between four other measures of image similarity and neural response: EMD, mean color, corresponding pixel image similarity (CPI), and RMS contrast. The EMD, like SIMPLIcity, was also developed for the practical purpose of searching image databases. The EMD calculation is quite different from SIMPLIcity, although EMD also uses color information (METHODS). CPI similarity is essentially calculated by summing the difference between corresponding pixels in two images. This calculation, applied to luminance rather than color, has been used in other studies comparing IT neural response and perceptual similarity (Op de Beeck et al. 2001). We chose to calculate mean color between images because color is an important element of SIMPLIcity and because early studies in IT have shown a relationship between color and neural response (Komatsu 1992). Finally, we used RMS contrast because it relies on information different from the other four calculations, but RMS contrast has been used as a measure of stimuli used in other visual cortical areas such as V1. RMS contrast has not been demonstrated to correlate with either human perception or IT response and as such serves as a useful control in comparison to the other measures of image similarity.

For each of these four measures, we calculated the correlation between human perceptual similarity and image similarity and the degree to which image similarity predicted neural selectivity. To compare human judgments of similarity and quantitative image similarity, we repeated the calculation described in Table 1 and Fig. 3*B* for each of the four new measures. To calculate the degree to which image similarity predicted neural selectivity, we repeated the correlation to best (and worst) stimulus analysis described in Fig. 7. We chose to replicate this particular analysis because it utilizes neural

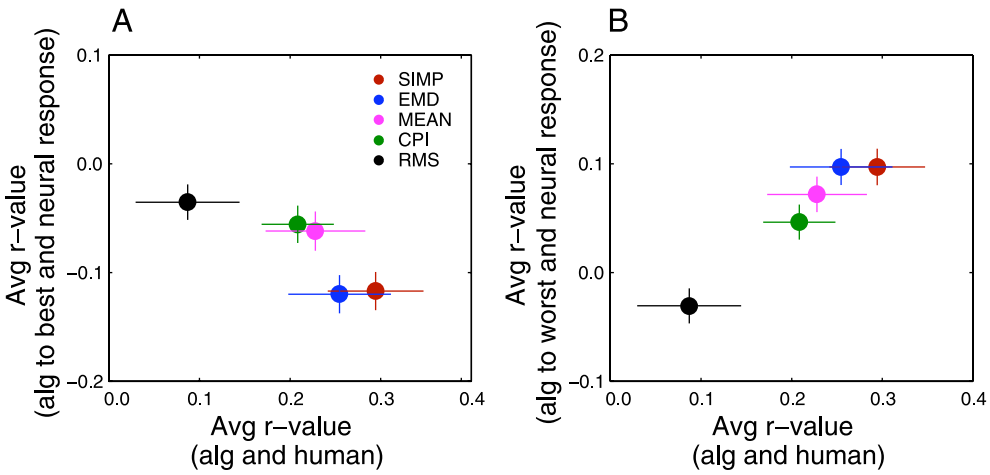

FIG. 10. Relationship between quantitative predictions of human similarity judgments and neural response. Each color represents a different quantitative measure of image similarity [red = SIMPLIcity (SIMP); blue = EMD; purple = mean color (MEAN); green = CPI; black = RMS contrast (RMS)]. Each data point represents averages across experiments (error bars are SE) In both *A* and *B*, values along the *x* axis represent the correlation between quantitative measures of image similarity and human judgments of image similarity (see METHODS and Fig. 3). $n = 5$ image sets, 4 human observers for each image set. Values along the *y* axis represent the *r* value of neural response to each image and the quantitative similarity between each image and the best image (*A*, $n = 217$) or the worst image (*B*, $n = 214$). See Fig. 7.

responses to all images. Figure 10 shows that as similarity measures did a better job of capturing human perception, their ability to predict neural response increased. The average correlations (across image sets) between algorithm and human judgments of image similarity are represented along the *x* axis, and the average correlations (across all neurons) between algorithm-judged similarity to best image (Fig. 10*A*) or worst image (Fig. 10*B*) and neural response to that image are represented along the *y* axis. Each data point shows average values for a different similarity measure. In Fig. 10*A*, stronger negative correlations between quantitative similarity to best and neural response (*y* axis) indicate that quantitative similarity better predicted neural selectivity (see Fig. 7, *A* and *B*), whereas the opposite is true for Fig. 10*B* (see Fig. 7, *C* and *D*). For both panels, the degree to which quantitative image similarity predicted human estimates of image similarity was correlated with the degree to which quantitative image similarity predicted neural selectivity (Fig. 10*A*, *r* value = −0.85; Fig. 10*B*, *r* value = 0.98).

For the human data, the SIMPLIcity algorithm ranked first in explaining human similarity judgments, followed by EMD, mean color, CPI, and last, RMS contrast, though across these five image sets, SIMPLIcity was only significantly better than CPI and RMS contrast ($P = < 0.05$, 1-tailed paired *t*-test). For the neural data, SIMPLIcity-judged similarity to best image was significantly better correlated with neural response than mean color, CPI, and RMS contrast ($P < 0.0005$ for all 3, 2-tailed paired *t*-test), though it was not significantly better correlated with neural response than is the EMD ($P = 0.70$). Likewise, SIMPLIcity-judged similarity to worst was significantly better correlated with neural response than mean color ($P < 0.05$), CPI color ($P < 0.0001$), and RMS contrast ($P < 0.000001$), though it was not significantly better than the EMD ($P = 0.996$). Mean color and CPI were, however, both significantly correlated with neural response ($P < 0.0005$ for correlation to best and worst), whereas RMS contrast was not.

We directly tested the relationship between human perception of image similarity and neural response by recalculating the correlation to best (and worst) image using human judgments of image similarity. For each neuron that responded to an image set for which we collected human data (128 neurons), we calculated the human similarity judgments between the best (or worst) image and every other image and then compared those values to the neural response for each image. As with SIMPLIcity (Figs. 7 and 10), human judgments of image similarity were significantly correlated with neural response (to best: *r* value = −0.080, $P < 0.001$; to worst: *r* value - 0.086, $P < 0.00001$; data not shown). Human judgments of image similarity predicted neural response no differently than SIMPLIcity judgments of image similarity (to best: $P = 0.20$; to worst: $P = 0.68$). This may be the result of a variety of differences between monkey and human data, such as task differences that could lead to response bias, number of data points, and inter-species differences.

DISCUSSION

We have shown that algorithms (SIMPLIcity and EMD), and elements of algorithms (mean color, CPI color) designed to navigate image databases are correlated with perceptual similarity of realistic images in human subjects. These measures of image similarity are also correlated with responses of single neurons in inferotemporal cortex. Furthermore, the degree to which measures of image similarity predict human perceptual similarity reflects the degree to which those same measures predict neural selectivity. Conversely, RMS contrast, a common method of image analysis that is not correlated with perceptual similarity of images, is also not correlated with neural response. Previous studies have shown that when images vary along carefully chosen dimensions, neural selectivity is correlated with those dimensions most relevant to perception (Op de Beeck et al. 2001; Rollenhagen and Olsen 2000; Sigala and Logothetis 2002). Using realistic pictures that have been chosen without regard to particular image subspaces, we found that neural response is correlated with algorithms that predict perceptual similarity but not with an algorithm unrelated to perceptual similarity. This study therefore provides further support for the idea that the tuning of IT neurons is related to perceptual similarity of images, even when images vary along many unpredictable dimensions.

Although SIMPLIcity utilizes color and spatial information to judge similarity between images, it is unlikely that inferotemporal cortex analyzes images in the same fashion. We chose this algorithm because it is, intuitively speaking, designed to do the same thing that the brain does. We concluded, through the results in Figs. 2 and 3 and Table 1, that SIMPLIcity judgments of similarity are correlated with human perception. We expected a correlation between quantitative measures

of image similarity and neural response in IT only to the extent that these judgments of similarity captured perceptual similarity. For the range of human perceptual similarity we could analyze, our data confirmed this expectation (Fig. 10). Although these measures of image similarity are significantly correlated with human perception, they only capture a fraction of the total variation in human similarity rankings (Table 1, Fig. 10). In addition, these measures of image similarity discard information that the brain cannot, such as relative spatial configuration. For these reasons, it is unlikely that the brain processes images in the same way that these measures of image similarity do.

Early studies in IT do demonstrate weak relationships between neural response and elements of the algorithms discussed here, such as color (Komatsu 1992) and spatial frequency (Schwartz et al. 1983). However, these results are unlikely to transfer to images that vary along more dimensions than those specifically tested in earlier studies. For example, Komatsu (1992) tested selectivity of IT neurons for circular patches of uniform color. Patches varied only in color, so any perceptual differences between them must have been based on color. Because IT neurons are putatively responsible for perception, it is not surprising that some neurons would respond selectively to the patches. However, when other information is available to contribute to perception, color per se may not determine the selectivity of neurons. When neurons respond selectively to categories of images, for example, changing to an achromatic version of the image does not dramatically affect the response (Nakamura et al. 1994; Vogels 1999). Edwards et al. (2004) found that changing from color images to achromatic versions of images decreased neural response in STS/IT but did not remove object selectivity. In addition, we found that although mean color is significantly correlated with neural response, algorithms that utilize more information are more strongly correlated with neural response (Fig. 10).

In addition to its relative failure to predict neural response, mean color also fails to predict human perceptual similarity. Studies that have compared effectiveness of algorithms for image similarity have concluded that color histograms alone are not sufficient to successfully navigate databases (Rodden et al. 2000). In addition, spatial frequency information alone (an element of SIMPLIcity) cannot explain human perceptual similarity. Changing the spatial frequencies present in a primed image does not affect the response time for identifying that image (Fiser and Biederman 2001).

More recent studies have shown that although IT neurons can show selectivity for low-level dimensions, selectivity is robust to changes in those dimensions as long as those changes are unlikely to affect global perception of the image (Bayliss and Driver 2001; Vogels 1999). However, when small changes do affect global perception, selectivity is not maintained across manipulations (Bayliss and Driver 2001; Sigala and Logothetis 2002; Vogels 1999b; Vogels et al. 2001). Because the algorithms discussed here rely on image statistics that of themselves are not particularly relevant to perception, it is unlikely that IT neurons themselves are coding those dimensions, but rather that algorithms, using divergent methodology, are converging on estimates of similarity. This is likely true because of ecology: in the real world, images that contain similar objects contain similar colors and spatial frequencies (Mojsilovic et al. 2002; Rubner et al. 1999).

Our results have shown that the correlation between algorithm-judged similarity and neural response is present throughout the length of stimulus presentation (Fig. 9). Some other studies have shown that the information present in visual responses changes during the time course of response. For example, gross and fine categorization of faces may occur at different time points in the neural response (Sugase et al. 1999), while in V4, the initial transient can be the same for attended and nonattended stimuli, but the responses after the initial transient are different (Ghose and Maunsell 2002). Because static algorithms cannot take into account active processing of stimuli, we might expect correlations between algorithms and neural response to be stronger during the initial period of the response. Alternatively, some studies have suggested that color information arrives in IT later than luminance or spatial information due to the difference in processing times of the magnocellular and parvocellular pathways (Merrigan and Maunsell 1993). Because our algorithms rely heavily on color content of images, this would mean that any predictive value of SIMPLIcity in determining neural response would occur after the initial peak of IT response. In Fig. 9, we see a systematic decrease in neural response as the image similarity to the best image decreases. However, the time course of the relationship between SIMPLIcity and neural response follows the time course of IT neural response generally, remaining through the 75–500 ms after stimulus onset that neurons respond. There is no window during stimulus presentation where the correlation between SIMPLIcity and neural response is eliminated. This result supports the recent finding (Edwards et al. 2003) that color information is present in IT from the initial stages of response.

Further evidence for a relationship between selectivity of IT neurons and perceptual similarity comes from noting where the strongest correlations between neural response and algorithm-judged similarity occur. When response differences are more likely to be an accurate reflection of image difference, such as when the two neural responses compared are quite different from each other, correlations between neural response differences and algorithm similarity are relatively high. However, when response differences are less likely to accurately predict which of two images was presented, the predictive value of algorithm in determining neural response is substantially weaker. Our results show this in two different ways. First, the SIMPLIcity difference in the "best, worst" comparison (Fig. 5) is considerably higher than the SIMPLIcity difference in the "effective, ineffective" comparison (Fig. 6). Second, when a reference image elicits a response at the extreme of the neuron's tested range, SIMPLIcity between an image and the reference image correlates more strongly with neural response differences between the image pair than when the reference image elicits an intermediate response from the neuron (Fig. 8). Many images elicit intermediate responses from neurons; therefore these responses do not accurately predict which image was presented, and the demonstrated reduction in correlation between SIMPLIcity and neural response reflects this.

The relationship between algorithm-judged similarity and neural response discussed here is significant but not complete. Most importantly, no static algorithm can entirely capture perceptual judgments because those judgments are subject to change as primates experience the world or are trained to

notice particular details of images (Sigala and Logothetis 2002). Our recording sites were quite anterior and medial; posterior areas of IT are thought to be less impacted by top-down variables such as learning and might be better correlated with algorithm judgments.

In conclusion, algorithms for image similarity that reflect human perception are correlated with neural response in IT cortex. We argue that both algorithms and IT cortex are representing the perceptual similarity of images, even though the methods for achieving that representation are likely different. RMS contrast, which does not reflect perceptual similarity, also fails to predict neural response. Furthermore, neural responses that more accurately reflect which stimulus was presented are more strongly correlated with algorithm-judged image similarity. Taken together, these results suggest that even when stimuli are not chosen to vary along mathematically defined dimensions, the responses of IT neurons are related to the perceptual similarity of images.

Much more work must be done to elucidate the relationship between perceptual similarity of images and IT selectivity. While the correlations between algorithm-judged similarity and neural responses differences are significant, they still account for a relatively small part of neural selectivity. Furthermore, no static algorithm can take into account the cognitive factors that also shape IT response. Algorithms such as SIMPLIcity and EMD may prove useful tools for further investigation of the parameters of IT response. If an image is found that evokes a strong response from a neuron, for example, algorithms could be used to quickly scan a database for other images likely to elicit strong or weak responses from the neuron. Finding such images quickly would enable physiologists to better test the effects of the cognitive variables that influence perception. In addition, as algorithms improve and are better able to capture perceptual similarity of images, the mechanisms used by algorithms might yield insight into the processing performed by the brain. Finally, because algorithms do not reflect differences among individuals, they should function to provide a baseline for comparison when measuring the changes in perceptual similarity of realistic images that occur with individual experience.

ACKNOWLEDGMENTS

We thank J. Bullis for collecting human similarity judgments, J. M. Jagadeesh for assisting with image analysis, C. A. Erickson for assistance during surgeries, A. McAlister, J. Skiver Thompson, and K. M. Ahl for technical help, and M. N. Shadlen and J. I. Gold for comments on an earlier version of this manuscript.

GRANTS

This research was supported by the Royalty Research Foundation (University of Washington), the Sloan Foundation, the McKnight Foundation, the Whitehall Foundation, and National Center for Research Resources. S. R. Allred was supported by National Eye Institute Vision Research Training Grant T32 EY-07031.

REFERENCES

Allred SR, Erickson CA, and Jagadeesh B. Characteristics of pictures that evoke equivalent neural responses in macaque perirhinal cortex. *Soc Neurosci Abst* 160.8, 2002.

Allred SR, Liu Y, and Jagadeesh B. Algorithms for image database navigation and object selective neurons in the non-human primate. Symposium on Applied Perception in Graphics and Visualization. ACM SIGGRAPH, APGV04, 2004.

Baylis GC and Driver J. Shape-coding in IT cells generalizes over contrast and mirror reversal but not figure-ground reversal. *Nat Neurosci* 4: 937–942, 2001.

Bex PJ and Makous W. Spatial frequency, phase, and the contrast of natural images. *J Opt Soc Am A Opt Image Sci Vis* 19: 1096–1106, 2002.

Edwards R, Xiao D, Keysers C, Foldiak P, and Perrett D. Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *J Neurophysiol* 90: 1245–1256, 2003.

Erickson CA and Desimone R. Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J Neurosci* 19: 10404–10416, 1999.

Erickson C, Jagadeesh B, and Desimone R. Learning and memory in the inferior temporal cortex of the macaque. In: *The New Cognitive Neurosciences* (2nd ed.), edited by Gazzaniga MS. Boston, MA: MIT Press, 1999, p. 743–752.

Erickson CA, Jagadeesh B, and Desimone R. Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nat Neurosci* 3: 1143–1148, 2002.

Fiser J and Biederman I. Invariance of long-term visual priming to scale, reflection, translation, and hemisphere. *Vision Res* 41: 221–234, 2001.

Fuchs AF and Robinson DA. A method for measuring horizontal and vertical eye movement chronically in the monkey. *J Appl Physiol* 21: 1068–1070, 1966.

Ghose GM and Maunsell JH. Attentional modulation in visual cortex depends on task timing. *Nature* 419: 616–620, 2002.

Kayaert G, Biederman I, and Vogels R. Shape tuning in macaque inferior temporal cortex. *J Neurosci* 23: 3016–3027, 2003.

Kobatake E and Tanaka K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol* 71: 856–867, 1994.

Kobatake E, Wang G, and Tanaka K. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J Neurophysiol* 80: 324–330, 1998.

Komatsu H, Ideura Y, Kaji S, and Yamane S. Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *J Neurosci* 12: 408–424, 1992.

Merrigan JH and Maunsell JHR. How parallel are the primate visual pathways? *Annu Rev Neurosci* 16: 369–402, 1993.

Miyashita Y. Inferior temporal cortex: where visual perception meets memory. *Annu Rev Neurosci* 16: 245–263, 1993.

Miyashita Y, Kameyama M, Hasegawa I, and Fukushima T. Consolidation of visual associative long-term memory in the temporal cortex of primates. *Neurobiol Learn Mem* 70: 197–211, 1998.

Mojsilovic A, Gomes J, and Rogowitz B. ISee: perceptual features for image library navigations. *Proc SPIE Hum Vision Electron Imag* 2002, p. 266–277.

Nakamura K, Matsumoto K, Mikami A, and Kubota K. Visual response properties of single neurons in the temporal pole of behaving monkeys. *J Neurophysiology* 71: 1206–1221, 1994.

Op de Beeck H, Wagemans J, and Vogels R. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci* 4: 1244–1252, 2001.

Peli E. Contrast in complex images. *J Opt Soc Am A* 7: 2032–2040, 1990.

Ringach DL, Hawken MJ, and Shapley RM. Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. *J Vision* 2: 12–24, 2002.

Rodden K, Basalaj W, Sinclair D, and Wood K. *A Comparison of Measures for Visualising Image Similarity. The Challenge of Image Retrieval*, Third UK Conference on Image Retrieval, Brighton, United Kingdom. 2000.

Rogowitz BE, Frese T, Smith JR, Bournan CA, and Kalin E. Perceptual image similarity experiments. *Proceedings of SPIE 3299,* 576–590, 1998.

Rogowitz BE, Thomas F, Smith JR, Bouman CA, and Kalin E. In: *Proceedings of the SPIE,* edited by Rogowitz BE and Pappas TN. San Jose, CA: 1998, p. 3299.

Rollenhagen JE and Olson CR. Mirror-image confusion in single neurons of the macaque inferotemporal cortex. *Science* 287: 1506–1508, 2000.

Rubner Y. *Perceptual Metrics for Image Database Navigation* (Phd thesis). Stanford, CA: Stanford University, 1999.

Rubner Y, Tomasi C, and Guibas LJ. The earth mover's distance as a metric for image retrieval. *Proceedings of the 1998 Asian conference on computer vision*, Hong Kong, China.

Sakai K and Miyashita Y. Neural organization for the long-term memory of paired associates. *Nature* 354: 152–5, 1991.

Schwartz EL, Desimone R, Albright TD, and Gross CG. Shape recognition and inferior temporal neurons. *Proc Natl Acad Sci USA* 80: 5776–5778, 1983.

Sheinberg DL and Logothetis NK. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J Neurosci* 21: 1340–1350, 2001.

Sigala N and Logothetis NK. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415: 318–320, 2002.

Sugase Y, Yamane S, Ueno S, and Kawano K. Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400: 869–873, 1999.

Tanaka K. Mechanisms of visual object recognition: monkey and human studies. *Curr Opin Neurobiol* 7: 523–529, 1997.

Torralba A and Oliva A. Statistics of natural image categories. *Network* 14: 391–412, 2003.

Vogels R. Categorization of complex visual images by rhesus monkeys. II. Single-cell study. *Eur J Neurosci* 11: 1239–1255, 1999.

Vogels R, Biederman I, Bar M, and Lorincz A. Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *J Cogn Neurosci* 13: 444–453, 2001.

Vogels R and Orban GA. Coding of stimulus invariances by inferior temporal neurons. *Prog Brain Res* 112: 195–211, 1996.

Wang JZ, Li J, and Wiederhold G. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans Pattern Anal Machine Intell* 23: 947–963, 2001.

Weliky MJ, Fiser J, Hunt RH, and Wagner DN. Coding of natural scenes in primary visual cortex. *Neuron* 37: 703–718, 2003.