Behavioral/Systems/Cognitive

Multiple Object Response Normalization in Monkey Inferotemporal Cortex

Davide Zoccolan,[1,2,3]* David D. Cox,[1,2]* and James J. DiCarlo[1,2]

[1]McGovern Institute for Brain Research, [2]Department of Brain and Cognitive Sciences, and [3]Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

The highest stages of the visual ventral pathway are commonly assumed to provide robust representation of object identity by disregarding confounding factors such as object position, size, illumination, and the presence of other objects (clutter). However, whereas neuronal responses in monkey inferotemporal cortex (IT) can show robust tolerance to position and size changes, previous work shows that responses to preferred objects are usually reduced by the presence of nonpreferred objects. More broadly, we do not yet understand multiple object representation in IT. In this study, we systematically examined IT responses to pairs and triplets of objects in three passively viewing monkeys across a broad range of object effectiveness. We found that, at least under these limited clutter conditions, a large fraction of the response of each IT neuron to multiple objects is reliably predicted as the average of its responses to the constituent objects in isolation. That is, multiple object responses depend primarily on the relative effectiveness of the constituent objects, regardless of object identity. This average effect becomes virtually perfect when populations of IT neurons are pooled. Furthermore, the average effect cannot simply be explained by attentional shifts but behaves as a primarily feedforward response property. Together, our observations are most consistent with mechanistic models in which IT neuronal outputs are normalized by summed synaptic drive into IT or spiking activity within IT and suggest that normalization mechanisms previously revealed at earlier visual areas are operating throughout the ventral visual stream.

Key words: **inferotemporal cortex; monkey; object recognition; multiple objects; normalization; clutter tolerance**

Introduction

Visual object recognition in cluttered scenes is extremely difficult for artificial vision systems yet is accomplished effortlessly by the brain. In primates, it is believed that object identity is extracted through processing along the ventral visual stream and is represented in patterns of neuronal activity in the highest stages of that stream: the anterior inferotemporal cortex (IT). Electrophysiological studies show that IT neurons can be selective for complex objects while also being tolerant to some transformations (object position, scale, and pose) (for review, see Logothetis and Sheinberg, 1996; Tanaka, 1996). In this context, some have suggested that IT neurons should ideally be tolerant to visual clutter, i.e., their response to an effective object should be essentially unaffected by the presence of other, less-effective objects (Rousselet et al., 2003, 2004).

However, this idealized notion of IT is inconsistent with available data showing that IT responses are altered in cluttered scenes (Sheinberg and Logothetis, 2001; Rolls et al., 2003), and responses to object pairs are typically weaker than responses to isolated, preferred objects (Sato, 1989; Miller et al., 1993; Rolls and Tovee, 1995; Chelazzi et al., 1998; Missal et al., 1999). Moreover, although some contemporary models of the ventral stream use MAX operations (Riesenhuber and Poggio, 1999a), even these models do not predict complete clutter tolerance in IT (Riesenhuber and Poggio, 1999b).

At present, we lack a systematic understanding of IT clutter tolerance, even in simple clutter conditions (e.g., two objects). Because recognition performance shows remarkable clutter tolerance even for brief presentation conditions (e.g., ~100 ms) without explicit attentional instruction (Potter, 1976; Intraub, 1980; Rubin and Turano, 1992), this suggests that top-down attention is not strictly required for robust recognition in clutter but that powerful, primarily feedforward processing mechanisms are also at work. Thus, although previous studies have investigated how attention modulates processing of targets in the presence of distracters (Moran and Desimone, 1985; Maunsell, 1995; Connor et al., 1997; Chelazzi et al., 1998), here we seek to understand the "core" feedforward processing of multiple objects.

Although some progress on understanding such processing has been made in area V4 (Reynolds et al., 1999; Gawne and

Received May 20, 2005; revised July 22, 2005; accepted July 23, 2005.

This work was supported by the National Institute of Mental Health–National Institutes of Health (NIH) Grant P20-MH66239, the National Eye Institute–NIH Grant R01-EY014970, and The Pew Charitable Trusts (University of California, San Francisco 2893sc). D.Z. was supported by a Postdoctoral Fellowship of The International Human Frontier Science Program Organization, and D.D.C. was supported by a National Defense Science and Engineering Graduate Fellowship. We thank C. Hung, M. Kouh, H. Op De Beeck, T. Poggio, and M. Riesenhuber for valuable help and discussion, B. Heisele and the Max Plank Institute for Biological Cybernetics for help in generating the face stimuli, and J. Deutsch and J. P. Mayo for technical support. We also thank K. O. Johnson for inspiration and unwavering support.

*D.Z. and D.D.C. contributed equally to this work

Correspondence should be addressed to James DiCarlo, McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: dicarlo@mit.edu.

DOI:10.1523/JNEUROSCI.2058-05.2005

 0270-6474/05/258150-15$15.00/0

Martin, 2002; Reynolds and Desimone, 2003), IT remains poorly understood. In particular, although several IT studies have touched on the issue of IT responses in clutter (Miller et al., 1993; Rolls and Tovee, 1995; Chelazzi et al., 1998; Missal et al., 1999; Sheinberg and Logothetis, 2001; Rolls et al., 2003), no study has systematically tested the relationship between responses to object pairs and responses to constituent objects (but see Missal et al., 1999), and there has been no attempt to understand how IT responses to multiple objects depend on object shape similarity.

In this study, we systematically examined IT neuronal responses to brief presentations of two or three objects in three passively viewing monkeys. Our results show that, at least under these conditions, IT neuronal responses to multiple objects are well predicted by the average of their responses to the constituent objects. This finding suggests that divisive normalization mechanisms analogous to those proposed to explain response rescaling in early visual stages (Heeger, 1992; Desimone and Duncan, 1995; Heeger et al., 1996; Carandini et al., 1997; Reynolds et al., 1999) and area MT (Recanzone et al., 1997; Britten and Heuer, 1999) may be at work in IT.

Materials and Methods

Animals and surgery

Experiments were performed on three male rhesus monkeys (*Macaca mulatta*) weighing ~8, 9.5, and 10 kg. Before behavioral training, aseptic surgery was performed to attach a head post to the skull of each monkey and to implant a scleral search coil in the right eye of monkeys 1 and 2. After 2–5 months of behavioral training (below), a second surgery was performed to place a recording chamber (18 mm diameter) to reach the anterior half of the left temporal lobe (chamber Horsley-Clark center, 15 mm anterior). All animal procedures were performed in accord with National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

Eye position monitoring

Horizontal and vertical eye positions were monitored using the scleral search coil (monkeys 1 and 2) or a 250 Hz camera-based system (monkey 3; EyeLink II; SR Research, Osgode, Ontario, Canada). Each channel was digitally sampled at 1 kHz. Methods for detecting saccades and calibrating retinal locations with monitor locations are described in detail previously (DiCarlo and Maunsell, 2000).

Visual stimuli

Stimuli were presented on a video monitor (43.2 × 30.5 cm; 75 Hz frame rate; 1920 × 1200 pixels) positioned at 81 cm from the monkeys so that the display subtended approximately ±15° (horizontally) and ±10° (vertically) of visual angle. Different visual objects were used in each experiment (see Fig. 1 and below).

Experiment 1. Monkeys 1 and 2 were tested with three simple, solid geometric forms (see Fig. 1*A*, left, a star, a cross, and a triangle), presented at full luminance (57 cd/m^2) on a gray background (27 cd/m^2). Each object was 2° in size (diameter of a bounding circle).

Experiment 2. Monkey 3 was tested using objects drawn from three object sets with parametrically controllable shape similarity within each set (see Fig. 1*A*, right). To ensure generality of results, three different spaces of morphed shapes were generated: (1) a car space; (2) a face space; and (3) a NURBS space (nonuniform rational B-spline generated two-dimensional silhouettes). Each space was generated from a set of 15 initial shapes: (1) 15 three-dimensional models of car brand prototypes; (2) 14 three-dimensional models of human heads plus their average; and (3) 15 randomly generated NURBS (44 free parameters, see below). For each space, one of these initial shapes was chosen as "center" of the space, and 14 sets of morphed shapes were built as blends (see below) of the center shape and each of the other 14 prototype shapes, thus resulting in 14 morph lines per space (see examples in Fig. 1*A*, right). In each of the three object spaces, the distance (d) between the center shape and each of the 14 prototype shapes was defined to have value 1. As shown for the three exemplar morph lines of Figure 1*A*, morphed shapes were generated not only between the center and each of the 14 prototypes (see Fig. 1*A*, right, the five middle shapes in each row) but also by extrapolating beyond the initial prototypes (see Fig. 1*A*, right, first and last shapes in each row), thus resulting in shape distances $d > 1$ and $d < 0$.

Slightly different morphing methods were used to generate the objects in each of the three shape spaces. Cars were built using an algorithm (Shelton, 2000) that found corresponding points in each pair of three-dimensional car prototypes and represented each car prototype as a vector of point coordinates. Faces were built using a face-morphing algorithm (Blanz and Vetter, 1999), in which point correspondences between pairs of face prototypes were established based on the three-dimensional structure of the head models. Car and face morphs were then created as linear combinations of the correspondence points and rendered as gray-scale two-dimensional images (with fixed viewpoint, illumination, and size; see Fig. 1*A*, right, first and second row). The center shape of the face space was the average face (see Fig. 1*A*, right, second stimulus in the second row). NURBS objects were filled shapes defined by closed third-order NURBS curves with 22 equally weighted control vertices (Rogers, 2000). NURBS morphs were generated using weighted averages of control vertices of pairs of prototypes, and all NURBS curves were filled at full luminance (72 cd/m^2) (see Fig. 1*A*, right, third row). All objects were presented at 2° in size (bounding circle diameter) on a gray background (12 cd/m^2).

Behavioral task and training

All three monkeys were trained to fixate a central point (0.2 × 0.2°) for several seconds while a series of visual stimuli were presented in rapid succession (rapid, passive viewing paradigm). In particular, stimulus conditions were presented in a random sequence in which each stimulus condition was on for 100 ms, followed by 100 ms of a gray screen (no stimulus), followed by another stimulus conditions for 100 ms, etc. (see Fig. 1*E*). That is, stimulus conditions were presented at a rate of five per second. At this presentation rate, IT neurons show robust object selectivity (Keysers et al., 2001), and this rate is consistent with that produced spontaneously by free-viewing monkey (DiCarlo and Maunsell, 2000). Single, pair, and triplet object conditions were pseudorandomly interleaved (see schematic in Fig. 1*E*). The screen background was always kept at a constant gray. The total number of stimulus conditions presented on each fixation trial ranged from 3 to 20, and the monkey was rewarded for maintaining fixation throughout the trial (±0.5° fixation window in monkeys 1 and 2; ±1.5° fixation window in monkey 3). Failures to maintain fixation throughout the trial resulted in the trial being aborted, and all stimulus conditions in that trial were re-presented.

The data presented in the current study were all acquired during this rapid, passive viewing paradigm. However, all three monkeys are also involved in ongoing studies that require behavioral training with the stimuli used in this study. Monkeys 1 and 2 were trained to perform an object identification task with single geometrical shapes presented either at the center of gaze, 2° above, or 2° below fixation. Monkeys were required to saccade to a different, fixed peripheral target for each object. Monkey 3 was trained to perform a sequential object recognition task that required the detection of a fixed target shape (the center object in each object set) embedded in a temporal sequence of shapes drawn from the same object set (blocked trials).

Recording and data collection

For each recording, a guide tube (26 gauge) was used to reach IT using a dorsal to ventral approach. Recordings were made using glass-coated platinum/iridium electrodes (0.5–1.5 MΩ at 1 kHz), and spikes from individual neurons were amplified, filtered, and isolated using conventional equipment. The superior temporal sulcus (STS) and the ventral surface were identified by comparing gray and white matter transitions and the depth of the skull base with structural magnetic resonance images from the same monkeys. Penetrations were made over a ~10 × 10 area of the ventral STS and ventral surface (Horsley-Clark coordinates: anteroposterior, 10–20 mm; mediolateral, 14–24 mm) of the left hemisphere of each animal. All recordings were lateral of the anterior middle temporal sulcus. Thus, the recorded regions included anterior and central inferotemporal cortex (Felleman and Van Essen, 1991). In all three animals, the

penetrations were concentrated near the center of this region, in which form-selective neurons were more reliably found. The animals cycled through behavioral blocks as the electrode was advanced into IT. Responses from every isolated neuron were assessed with an audio monitor and online histograms, and data were collected according to specific criteria for experiments 1 and 2.

Experiment 1. As the electrode was advanced into IT, monkeys 1 and 2 performed the object identification task described above. Neurons that responded to any of the geometric objects at any of the three positions were further probed while the animal passively viewed the same objects (described above) (see Fig. 1*E*). Neurons that responded with a mean firing rate significantly higher than background rate to any shape at any position (t test, $p < 0.05$) were studied further. The main experimental conditions included the following: (1) each of the three shapes presented in isolation in each of three positions (Fig. 1*B*, left; 3 shapes × 3 positions = 9 stimulus conditions); (2) pairs of objects in all possible arrangements that did not include object duplicates (see Fig. 1*C*, left; 18 stimulus conditions); and (3) triplets of objects in all possible arrangements that did not include object duplicates (see Fig. 1*D*, left; 6 conditions). Object size (2°) and positions (fixation, 2° above fixation, and 2° below fixation) were chosen before data collection so that the objects did not touch or overlap but that objects were close enough to likely activate IT neurons in one or more positions. In this experiment, no attempt was made to optimize the objects or positions for the neuron under study. Instead, the exact same 33 stimulus conditions were tested for each neuron. These conditions were pseudorandomly interleaved and presented using the rapid, passive viewing paradigm described above. All neurons in which these conditions were tested were considered in Results if 10–30 presentations of each condition were completed during the time that the neuron was isolated.

Experiment 2. As the electrode was advanced into IT, monkey 3 was either engaged in the rapid, passive viewing paradigm or engaged in a recognition task similar to the behavioral task described above (except that the target object was a red triangle). To try to optimize the objects for each collected neuron in this experiment, each isolated neuron was tested with a sequence of screening procedures that always included at least 10 repetitions of each stimulus condition (pseudorandomly interleaved). During the first screening, 15 objects from each morphed space (a total of 45 objects) were presented at the center of gaze. These 15 objects were the center shape (see above) plus one stimulus randomly sampled (at a distance of 0.5 or 1.0 from the center object) from each of the 14 morph lines. Neurons that responded to one of these stimuli with a mean firing rate significantly higher than background rate (t test, $p < 0.005$) were further tested using objects within the space to which the most effective stimulus belonged (all tested during the rapid, passive viewing paradigm described above). In particular, the center object and four objects sampled (at distances $d = 0.25, 0.5, 0.75$, and 1 from the center object) from each of the 14 morph lines were presented in isolation at the center of gaze. A neuron was considered to be selective if the mean firing rates elicited by the set of five objects belonging to at least one of the morph lines were significantly different (ANOVA, $p < 0.05$). If so, the object along this morph line that was most effective in driving the cell was taken to be the "preferred object" of the neuron, and more tests of object selectivity were done using objects drawn from this morph line.

The main experimental conditions in experiment 2 included the following two primary conditions. (1) The first included 8–12 isolated objects from the most selective morph line (morphing step distance d_{step} ranging from 0.1 to 0.5). For most neurons, this set of objects included shapes generated by moving beyond the limits of the initial morph line, as well as one randomly chosen object from one of the two other object sets. Each object was presented at each of two, fixed positions (1.25° above the center of gaze and 1.25° below the center of gaze; see Fig. 1*B*, right). Thus, a total of 16–24 isolated object conditions were tested for each neuron. As in experiment 1, object size (2°) and positions were chosen and fixed before data collection so that the objects did not touch or overlap but that objects were close enough to likely activate IT neurons in one or more positions. However, unlike experiment 1, the tested range of objects was both parameterized (morph line) and chosen to obtain maximal selectivity from each neuron. (2) Pairs of objects were presented to all neurons to systematically test the ability of each neuron to tolerate the presence of a second object given the presence of a preferred object. In particular, the preferred object of the neuron (resulting from the previous screening at fovea) was presented at one position in combination with each of the objects tested in isolation (see above), including the preferred object itself (8–12 conditions) (see Fig. 1*C*, right). This was also done with the preferred object in the other position (see Fig. 1*C*, right). In summary, a total of 16–24 isolated object conditions and 16–24 paired object conditions were tested for each neuron. Fifteen to 30 repetitions of each stimulus condition were recorded for each neuron (pseudorandomly interleaved) using the rapid, passive viewing paradigm described above.

Analysis

Only neuronal responses collected during correctly completed behavioral trials were included in the analysis. The background firing rate of each neuron was estimated as the mean rate of firing over all trials in a 100 ms duration window that directly preceded the onset of the first stimulus in each trial. For all of the data recorded from the three monkeys, we quantified the response of each neuron to each of the stimulus conditions as the mean firing rate in a 100 ms window that began 100 ms after stimulus onset. The statistical tests used to assess neuronal responsiveness and selectivity to the different stimulus conditions are explained in Results, as well as the criteria to include subsets of recorded neurons in each analysis. In the following, details about some of the analysis performed in Results are provided.

Goodness-of-fit analysis (see Fig. 4). To assess, for each neuron, how much of the variance of the responses to objects pairs could be accounted for by considering responses to the constituent objects presented in isolation, a goodness-of-fit (GOF) index was computed. The GOF index provides an unbiased estimate of the percentage of true data variance explained by a given model, by removing the fraction of data variance that is merely attributable to noise (i.e., the trial-by-trial variability of the neuronal response). The GOF index calculation is based on well known mathematical relationships that are at the base of ANOVA statistics. Following the convention used by Rice (1995), let us assume we recorded J neuronal responses to each of I different stimulus pairs (I and J are, respectively, the number of groups and trials in ANOVA statistics). Let σ^2_{expl} be the true (or "explainable") variance of the mean recorded responses to the stimulus pairs. Let σ^2_{noise} be the variance of the noise that contaminates neuronal responses. Let SS_B and SS_W be the sum of squares, respectively, between groups and within groups of the ANOVA statistics for the recorded responses. The following relationship holds for the expectation of SS_B: $E[SS_B] = J(I-1)\,\sigma^2_{expl} + (I-1)\,\sigma^2_{noise}$ (Rice, 1995). Because the noise variance can be estimated as $\sigma^2_{noise} = SS_W/[I(J-1)]$, the explainable variance can be estimated as: $\sigma^2_{expl} = SS_B/[J(I-1)] - SS_W/[IJ(J-1)]$.

Given a model providing a prediction for the mean response to each object pair, the deviations from the model predictions can be computed for each trial J and each group (stimulus pair) I, so as to obtain trial-by-trial residual responses to each stimulus pair. Again, the variance of the mean residual responses to the stimulus pairs is composed of two terms: the noise variance σ^2_{noise} and the variance σ^2_{res} of the true deviations from the tested model. Therefore, σ^2_{res} can be estimated by the same equation that gives σ^2_{expl}, but with SS_B and SS_W obtained for the ANOVA statistics of the residual responses.

Once σ^2_{res} and σ^2_{expl} are estimated from the data, the GOF index can be computed as follows: $GOF = 100\,(1 - \sigma^2_{res}/\sigma^2_{expl})$. We verified that this method provides an unbiased estimate of the percentage of explainable variance explained by a model by running simulations in which data points were generated according to a linear model contaminated by different amounts of noise.

The SE of the GOF index was estimated by bootstrap resampling. For each of the I stimulus pair conditions, J responses were resampled with replacement 200 times from the J responses obtained during recordings. The GOF index was computed for each of these redrawing of the response matrix, and the SD of the resulting 200 bootstrapped GOF indexes was taken to be the SE of the GOF (Efron and Tibshirani, 1998).

Selectivity and monotonicity criteria for the tuning curves included in the population averages (see Fig. 5). Neurons recorded in experiment 2 were

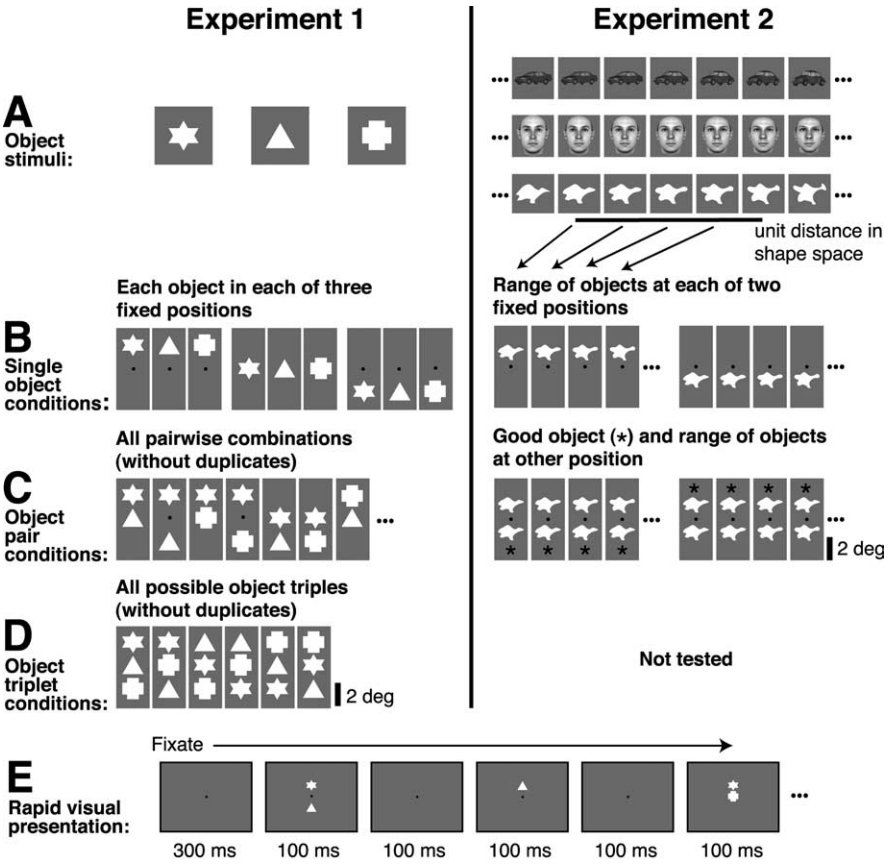

Figure 1. Stimulus conditions during recordings. ***A***, Left, The three geometrical shapes used in experiment 1. Right, An example morph line (i.e., set of parametric shapes) from each of the three shape spaces (i.e., cars, faces, and 2-dimensional silhouettes) used in experiment 2. The horizontal line indicates the unit shape distance within a morph line. This is the distance between the object prototypes used to generate the morph line (i.e., the 2nd and 6th stimulus in each row). ***B***, Single object conditions. Left, All nine single object arrangements of experiment 1 (3 shapes in each of 3 visual field locations; at center of gaze and 2° above and below center of gaze). Right, Single objects sampled from the most selective morph line (in the example, 2-dimensional silhouettes) were presented in two visual field locations: 1.25° above center of gaze (top) and 1.25° below center of gaze (bottom) in experiment 2. ***C***, Object pair conditions. Left, A subset of the 18 object pairs used in experiment 1 (3 objects in 2 of 3 positions without duplicate objects). Right, Examples of object pairs used in experiment 2. In each pair, the preferred object (indicated by the asterisk) of each neuron is presented in either the top or bottom position and is paired to a second object drawn from a range of shapes along the morph line containing the preferred object. ***D***, Object triplet conditions. Left, All six object triplet arrangements used in experiment 1 (the 3 objects in the 3 positions without duplicate objects). Right, No object triplets were tested in experiment 2. ***E***, Rapid visual presentation. Each panel is a schematic of the visual display (not to scale). The monkey was required to hold fixation on a central point while stimulus conditions were randomly interleaved and presented at a rate of five per second (see Materials and Methods).

tested with parametric objects sampled from morphed object spaces (see above). Therefore, tuning curves of neuronal responses to objects along continuous, parameterized changes in object shape (i.e., along a morph line) were obtained. The range of shape distances spanned by each morph line during the probing phase of the recordings in the parafoveal positions varied from neuron to neuron. However, each morph line spanned at least a unit shape distance (Fig. 1*A*, right, horizontal line) and included the preferred stimulus of the neuron obtained from the screening phase of the recordings, whose shape distance was defined as $d = 0$ (see above). To get a meaningful population average of the neuronal tuning properties, the following criteria were used to include each tuning curve in the average curve shown in Figure 5. (1) Responses across all tested single object conditions (i.e., both top and bottom positions) were highly selective (ANOVA, $p < 0.001$). (2) The tuning curve in the tested position was significantly selective in a shape range spanning the unit distance [i.e., in $d \in (0, 1)$; ANOVA, $p < 0.05$]. (3) The tuning curve was approximately monotonic in $d \in [0, 1]$, with peak at or near the preferred stimulus (i.e., at $d \leq 0.25$).

Simulated neuronal responses for the average and complete clutter invariance models (see Fig. 6). One goal of this study was to understand whether neuronal responses to pairs of objects could be more reliably modeled as (1) the average of the responses to the constituent objects presented in isolation (average model), or (2) the maximum of the responses to the constituent objects presented in isolation [complete clutter invariance (CCI) model]. To understand how well measures of explained variance or transformations of the data were suitable for comparing these two models, we simulated neuronal responses to object pairs that followed either the average model or the CCI model (see Fig. 6*B*). The response of each model neuron to single objects was assumed to have some tuning across a hypothetical continuous shape dimension (a Gaussian tuning was assumed, but any arbitrary tuning function could be used). Then, the response R_{AB} to each pair of stimuli A and B sampled from the same shape dimension was modeled as (1) the average of individual responses, i.e., $R_{AB} = (R_A + R_B)/2$, or (2) the maximum of individual responses, i.e., $R_{AB} = \mathrm{MAX}(R_A, R_B)$. Random fluctuations (zero mean) were added to the responses of the model neurons to the pairs to simulate more realistic neuronal responses.

Pair response distributions under alternative hypotheses (see Fig. 9). In Results, we directly compare the distribution of the observed responses to object pairs with the distributions predicted by two alternative hypotheses: (1) the normalization hypothesis and (2) the attention hypothesis. For each included object pair condition (see Results), we obtained the distributions of the spike counts resulting from each presentation of the pair and its constituent objects (10–30 presentations per pair, see above; standard 100 ms analysis window, see above). Then, we computed the distributions predicted by the two alternative hypotheses, i.e., (1) by combining the observed spike count distributions of the constituent objects (attention hypothesis) or (2) by sampling 200–600 responses (spike counts) from a Poisson distribution with average count equal to the average response to the pair (normalization hypothesis). An example of these distributions for one pair condition is shown in Figure 9*B*. For each neuron and each tested object pair condition (see Results), we normalized each of the observed and simulated pair response distributions by the average response to that pair. Finally, we combined the distributions obtained across all neurons and all pair conditions to obtain three normalized population distributions: (1) predicted by the attention hypothesis, (2) predicted by the normalization hypothesis, and (3) observed. These distributions were compared as described in Results.

Results

Complete recordings using our battery of visual conditions were obtained from 104 well isolated single IT neurons of three monkeys (35 from monkey 1; 33 from monkey 2; and 36 from monkey 3). During recordings, all neurons were tested with both single and multiple objects using rapid visual presentation according to one of the two experimental paradigms (Fig. 1) (see Materials and Methods). Each recorded neuron was tested for responsiveness to single objects, and neurons that responded significantly to at least

one of the presented single objects (relative to background rate) were included in the analyses described through the paper (t test on each single object condition, $p < 0.05$; 79 of 104 neurons; 29 of 35 cells in monkey 1, 19 of 33 neurons in monkey 2, and 31 of 36 neurons in monkey 3). This weak inclusion criterion without correction for multiple tests was done to minimize sampling bias in that all IT neurons with even weak responsivity were considered.

Responses to object pairs and triplets

In experiment 1, the same three objects (Fig. 1*A*, left) were presented in each of three fixed positions to each neuron (center of gaze and 2° above and below the center of gaze). Using those same objects and retinal positions, all pairwise and triplewise combinations were also tested (see Materials and Methods) (Fig. 1*B–D*, left). That is, a total of 33 stimulus conditions were tested for all isolated neurons (9 single object conditions, 18 object pair conditions, and 6 triple object conditions). Figure 2, *A* and *B*, shows the response of a typical IT neuron to some of these conditions. For this neuron, the single object that produced the strongest response was the cross located at the center of the gaze (Fig. 2*A*, middle panel). When the cross was flanked by a nonpreferred object located in one of the eccentric positions (2° above or below fixation) (Fig. 2*B*, first and third panel), the response to the resulting object pair was intermediate between the responses to the individual constituent shapes. Similar intermediate responses were observed when the cross was flanked by two nonpreferred objects (Fig. 2*B*, last panel).

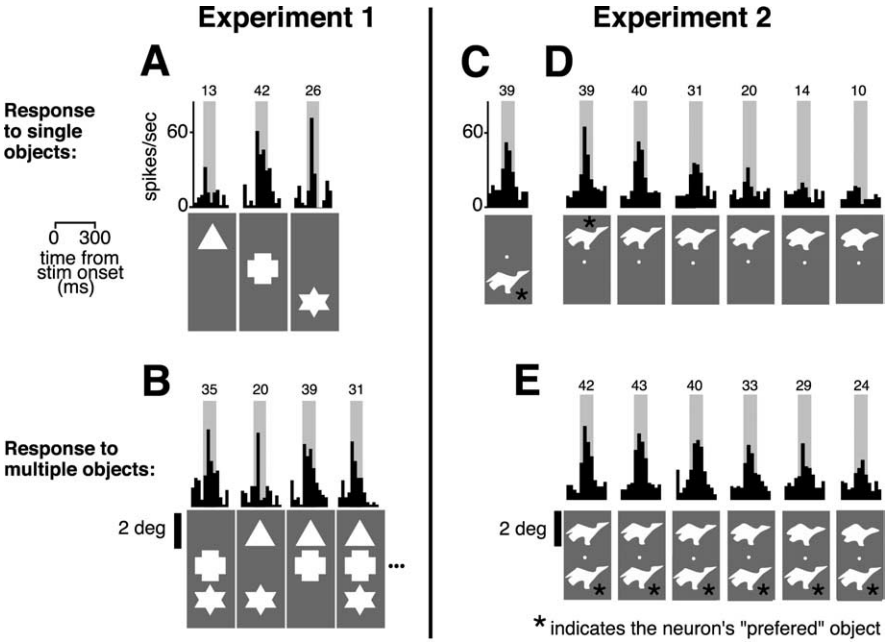

Figure 2. Examples of IT neuronal responses to single and multiple objects. ***A***, The black histograms are the average firing rates (computed in time bins of 25 ms) of a neuron recorded in experiment 1, after presentation of some of the single object conditions (stimuli are shown below the histograms). Objects were presented at time 0. The average neuronal response to each object was computed between 100 and 200 ms (gray patch), and its value (spikes per second) is reported above each histogram. ***B***, Examples of responses of the same neuron to object pairs and triplets. ***C***, Response of a neuron recorded in experiment 2 to its preferred object presented in the bottom position. ***D***, Responses of the same neuron to a range of objects sampled from the morph line containing the preferred object and presented in the top location. The response of the neuron decays as the second object is made more dissimilar to the preferred object (indicated by the asterisk). ***E***, Responses of the same neuron to stimulus pairs composed by the preferred object (asterisk; bottom position) and the range of shapes shown previously in ***D*** (top position). In both ***B*** and ***E***, responses to the object pairs are intermediate between responses to the constituent objects of the pairs.

Intermediate responses to multiple objects (relative to the responses to single objects) were also obtained in experiment 2 using sets of objects with parametrically defined shape similarity (see examples in Fig. 1*A*, right) that were presented in isolation or in pairs at two fixed retinal positions (see Materials and Methods) (Fig. 1*B*,*C*, right). Like experiment 1, the same retinal positions were tested for all neurons, but, in contrast to experiment 1, the presented objects were optimized for each neuron. Specifically, neurons were screened to have selectivity within at least one of three object spaces, a large range of objects was tested in each object space, and the set of objects (morph line) that yielded the most reliable selectivity was studied in detail (see Materials and Methods) (Fig. 1*A*, right). Figure 2*C–E* show a typical activation pattern of an IT neuron recorded in experiment 2. The response of this neuron to individual objects was significantly selective (ANOVA, $p < 0.01$) across a set of 11 objects sampled at consecutive distances along one of the morph lines of the NURBS space (responses to 6 of the 11 stimuli are shown in Fig. 2*D*). The selectivity pattern was unchanged and significant in all three tested locations [center of the gaze, 1.25° above the center of gaze (top) and 1.25° below the center of gaze (bottom); data not shown]. The neuron responded maximally to objects at one extreme of the shape space (the preferred shape) (Fig. 2, *C* and first histogram in *D*), whereas the response to the other extreme was not significantly higher than background (Fig. 2*D*, last histogram) (t test, $p > 0.05$). Responses between these two extremes of object shape showed an approximately monotonic decrease from maximal response as the object was made more dissimilar to the preferred shape (Fig. 2*D*). The response of the neuron to pairs of objects was tested by presenting the preferred object (bottom position) together with a nonpreferred object (top position) sampled across the whole morph line. The resulting activation pattern is shown in Figure 2*E*. For each stimulus pair, the response of the neuron was intermediate between its responses to the individual constituent objects of the pair (Fig. 2, compare *D*, *E*). A nearly identical response pattern was obtained when the identity of the object in the bottom position was varied while the preferred object was presented in the top position (data not shown).

To determine whether a systematic relationship existed between responses to individual objects and multiple objects, we first plotted the response to each object pair against the sum of the responses to the constituent objects of the pair. Figure 3, *A* and *B* (first panel), shows the resulting scatter plots for the two neurons just described, respectively, in the left and right side of Figure 2. As expected from previous studies, responses to object pairs were smaller than the simple sum of individual responses (i.e., well below the diagonal dashed lines in Fig. 3). However, the responses to each object pair condition (18 conditions and 22 conditions in these two cases) did not fall haphazardly on the scatter plot but clustered along a line of slope 0.5 (solid line). That is, the response of these neurons to pairs of objects was in good agreement with the average of the responses to the constituent objects presented in isolation. Most of the neurons recorded in the three monkeys showed a very similar response pattern: object pair re-

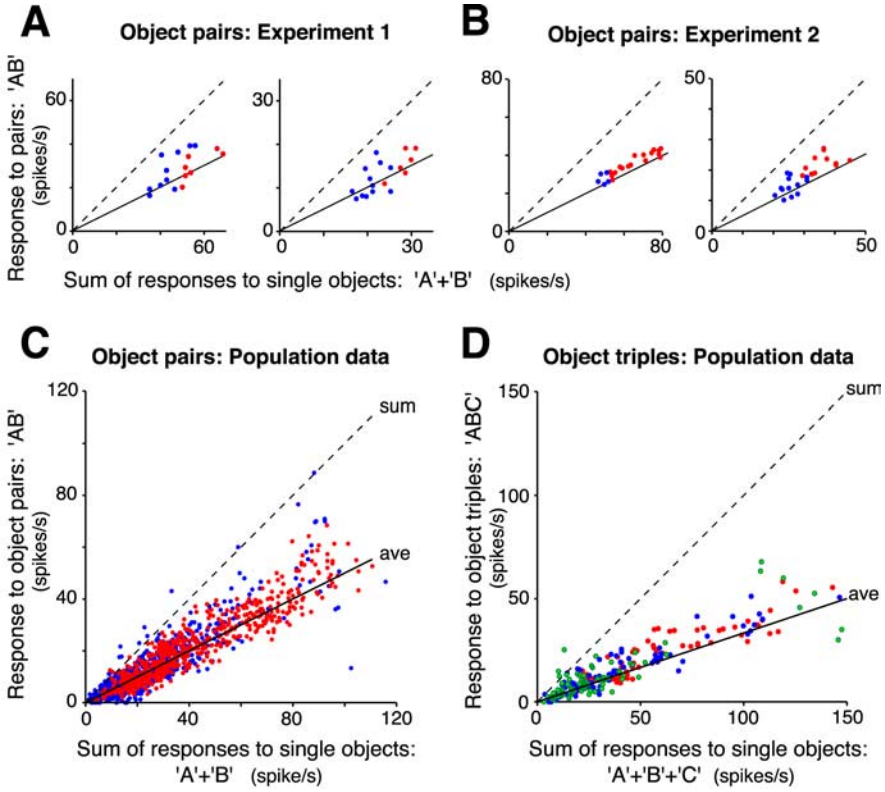

Figure 3. Responses to multiple objects as a function of the sum of responses to single objects. In each scatter plot, responses to object pairs (***A–C***) or object triplets (***D***) are plotted against the sum of the responses to the constituent objects presented alone. The dashed and solid straight lines indicate, respectively, the sum and the average of the responses to single objects. The slope of the solid line is one-half in ***A–C*** and one-third in ***D***. ***A***, Example data from two individual neurons recorded in experiment 1. Data in the left panel are from the same neuron shown in Figure 2, *A* and *B*. Red and blue dots refer to pairs in which, respectively, both or only one of the objects in the pair produced a response significantly higher than background rate (*t* test; $p < 0.05$). ***B***, Examples of scatter plots for two individual neurons recorded in experiment 2. Data in the left panel are from the same neuron shown in Figure 2*C–E*. Color code as in ***A***. ***C***, Scatter plot including responses to object pairs for the whole population of 79 responsive neurons recorded in the three monkeys. Color code as in ***A***. ***D***, Scatter plot including responses to object triplets for the whole population of 48 responsive neurons recorded in experiment 1. Red, blue, and green dots refer to triplets in which, respectively, three, two, or only one of the constituent stimuli evoked a response significantly higher than background rate. In both the individual examples and the population data, responses to multiple objects are normalized in that they were approximately the average of the responses to the constituent objects presented alone.

sponses were normalized in that they were approximately the average of the constituent objects responses. Figure 3 also shows data from two additional example neurons.

Because previous work in other visual areas showed that response normalization for multiple stimuli does not always hold when one stimulus is poorly effective (e.g., low contrast) (Britten and Heuer, 1999; Heuer and Britten, 2002), we specifically considered object pair conditions in which each of the two constituent objects drove the neuron significantly above background when presented alone (Fig. 3, red dots) and conditions in which only one of the two objects did (Fig. 3, blue dots). This did not reveal any obvious difference between such conditions in that both sets of points cluster along the same average line (Fig. 3). Moreover, the fact that the blue points are well below the diagonal shows that objects that have no significant effect on IT neuronal responses when presented alone can strongly impact responses to more preferred objects.

As a first look at our entire population of IT neurons in the three monkeys, we pooled the data from all 79 responsive neurons in a scatter plot using the same axes shown for the example neurons (Fig. 3*C*). Like the individual examples, responses to pairs of objects were highly correlated with the sum of responses to the constituent objects ($r = 0.92$), and the slope of the best linear fit to the data was 0.55. This value is very close to the 0.5 slope expected if the responses to object pairs were the average of the responses to individual objects (solid line, referred to as the average model). Like the single neuron examples (Fig. 3*A*,*B*), this relationship was independent of the effectiveness of the less optimal object of each pair (red and blue dots are as in Fig. 3*A*,*B*). Neurons recorded in experiment 1 were also tested with triplets of simultaneously presented objects (Figs. 1*D*, 2*B*). Figure 3*D* shows that responses to triplets were also highly correlated with the sum of the responses to the constituent objects of the triplets ($r = 0.91$), and the slope of the best linear fit to the data was 0.37. This value is very close to 0.33, i.e., the slope expected if responses to the object triplets were the average of the responses to individual objects (solid line).

To remove variance in the responses to the pairs (Fig. 3*C*) and triplets (Fig. 3*D*) attributable to differences in the range of firing rates over the population of neurons, these same analyses were repeated after normalizing by the response of each neuron to its most effective stimulus. Normalized responses to pairs and triplets of objects were still well correlated with the sum of normalized responses to the constituent objects ($r = 0.58$ and $r = 0.43$, respectively, for pairs and triplets), and the slope of the best linear fit to the data were very close to the slope predicted by the average model (i.e., slope of 0.44 and 0.27, respectively, for pairs and triplets).

Assessment of goodness-of-fit of the average model for individual IT neurons

Because these previous analyses suggested that a simple average model might explain a great deal of the IT response to multiple objects, we sought to assess, for each recorded neuron, how well responses to object pairs could be accounted for by the average model. To do that, we determined the GOF of the average model for each neuron (see Materials and Methods). The advantages of the GOF measure are that it provides an unbiased estimate of the percentage of data variance not attributable to noise (explainable variance) that is explained by the model, and it follows directly from well established statistical methods (see Materials and Methods). In this case, the data variance to explain for each neuron is the variance of the mean response across all of the tested object pair conditions. However, the advantages conferred by quantitative fit measures (like the GOF) come at the price of requiring sufficient neuronal response variance to support a reliable measure. To concentrate on cases for which reliable GOF estimates could be obtained without biasing our dataset for or against the model under scrutiny (the average model), we focused on neurons that showed the most reliable response modulation across the single object conditions: that is, the most selective neurons (ANOVA, $p < 0.001$). These neurons were 34 of the original 79 responsive neurons (15 of 48 for experiment 1 and 19

of 31 for experiment 2). For each of these neurons, the response R_{AB} to a pair of simultaneously presented stimuli A and B was modeled as a linear function of the sum of the responses R_A and R_B to the constituent stimuli presented in isolation, i.e., $R_{AB} = p + m\,(R_A + R_B)$. Two linear models were tested: (1) the average model, with $p = 0$ and $m = 0.5$ fixed for each neuron (i.e., the average of the individual responses); and (2) the best linear fit to the data (with intercept p and slope m being free parameters of a least squares fit for each neuron). For both models, the GOF and its bootstrap SE were computed (for details, see Materials and Methods). The median GOF across the 34 tested neurons was 63 and 67% for the average and best linear model, respectively (median SE of 16%). Figure 4 shows the distribution of the GOF values obtained for the two models. Figure 4*B* (last panel) also shows the distribution of the slopes m obtained by the best linear fits to the data. The median of this distribution was 0.45, which is very close to the 0.5 slope expected for the average model. The distribution of GOF values was not significantly different in experiment 1 and experiment 2 (Kolmogorov–Smirnov tests, $p > 0.05$). Overall, these analyses showed that responses to object pairs are very well predicted by the average of the responses to the constituent objects of each pair. Indeed, the median GOF value (63%) corresponds to a correlation coefficient of ~0.8 (similar to the correlation coefficient of the data plotted in Fig. 3*B*, right panel). Because the average model was nearly as good as the general linear model (above; only 4% difference in explained variance) but required no free parameters, all other figures and analyses in this manuscript refer only to the average model.

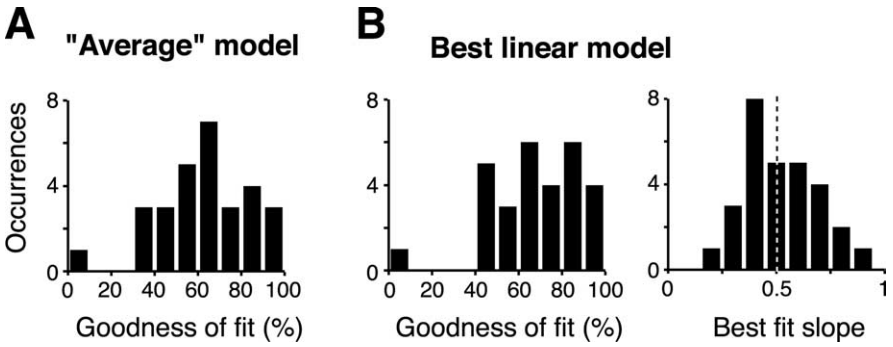

Figure 4. GOF of the responses to pairs. ***A***, GOF distribution for the average model (see Results). The GOF was computed for each of 34 highly selective neurons (see Result). Twenty-nine of those 34 neurons with GOF bootstrap SE < 40% are shown in the plot. ***B***, GOF distribution for the best linear fit to the data (left) and distribution of the slopes resulting from that fit (right). Same neuronal population as in ***A***. Both models explain a very large fraction of the response variance, and the best linear model yields a slope distribution centered around 0.5.

Although the very high level of fit for the average model at the level of single IT neurons was only confirmed for neurons in which sufficient data were available to obtain a reliable GOF measure (i.e., neuron with good selectivity for our test objects), this set is a priori unbiased with respect to the average model. Moreover, this does not imply that less selective neurons do or do not follow an average model but only that, for such neurons, the data did not allow this very quantitative assessment of model fit. Likewise, reliable GOF could not be obtained for the responses to object triplets because time allowed only six triplet configurations to be tested in experiment 1 (Fig. 1*D*), and, as a result, the amount of explainable variance in the triplet responses was, on average, only ~10% of the explainable variance in the pair responses. In summary, although our data indicate that the average model also holds for weakly selective neurons (Fig. 3*C*) and triple object conditions at the level of the IT population (Fig. 3*D*), in these cases, the data do not have sufficient power to reliably assess the average model or any other model at the level of individual neurons.

Responses to single objects and pairs of objects across continuous shape dimensions

One advantage of experiment 2 is that it allowed us to closely examine the response of each neuron to pairs of objects over a continuous shape space with very similar objects (Figs. 1*A*, right, 2*C–E*). That is, we were able to find neurons that were sensitive to one of these continuous shape dimensions and measure their responses along that parametric shape dimension. This, in turn, allowed us to place the preferred object of each neuron in the receptive field (RF) and then determine the effect of adding a second object of decreasing effectiveness (when presented alone). The methods used to generate the parametric objects and to optimize the examined shape dimension for each neuron are described in detail in Materials and Methods. These methods allowed us to build a tuning curve across a very selective shape dimension optimized for each of the recorded neurons (Fig. 5). The origin of each such tuning curve plot was set to be the preferred shape of the neuron (Fig. 2, *C* and first plot in *D*) obtained during initial screening at the center of gaze. Examples of such tuning curves are shown as red lines (open circles) in Figure 5, *A* and *B*, and thin black lines in the inset of Figure 5*C*. Besides obtaining tuning curves for single objects, we also presented object pair conditions containing both the preferred object of the neuron (defined as shape distance $d = 0$) and a second object drawn from along the tuned shape dimension (and, in some cases, drawn from another shape space; see Materials and Methods) (Fig. 2*E*).

Figure 5*A* shows the data obtained from the neuron already described in Figures 2*C–E* and 3*B* (left). The red line (open circles) shows the response of the neuron to 11 different objects sampled at increasing distances from the preferred object ($d = 0$) along the most selective shape dimension of the neuron (all presented at 1.25° above the center of gaze) (Fig. 2*D*). The black line shows the response of the neuron to 11 object pair conditions in which the preferred object of the neuron (1.25° below fixation) was presented along with a second object (whose identity is indicated by the abscissa) at 1.25° above fixation (Fig. 2*E*). The addition of the second object clearly causes the response of the neuron to drop below the response to the preferred object presented alone (i.e., the black line falls below the horizontal dashed line). In fact, the response to each object pair (black line) is always in between the response to each of the constituents of the pair, i.e., between the dashed line and the red line (open circles). At a more quantitative level, the green line shows the average of the responses to the constituent objects in each pair, i.e., the average of the dashed line and the red line (open circles). The green and black lines are almost exactly superimposed, indicating that the responses to object pairs are well predicted by the average model, regardless of the similarity of the two objects.

Figure 5*B* shows neuronal tuning curves of another neuron (same cell analyzed in Fig. 3*B*, right) along its most selective shape dimension. Like the neuron described above, the responses to

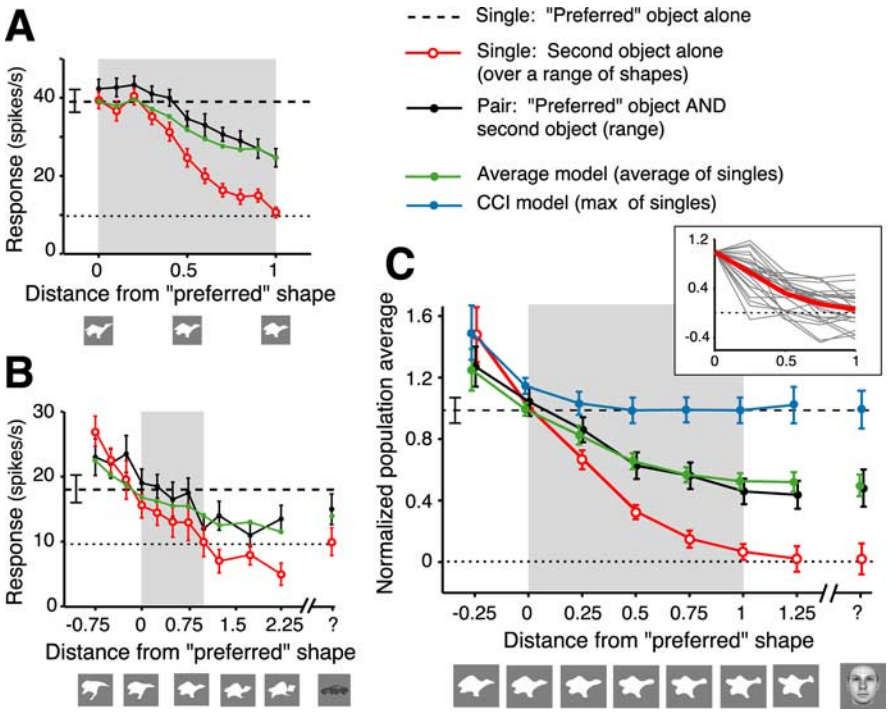

Figure 5. Response normalization for object pairs along continuous shape dimensions. ***A***, ***B***, Individual examples of tuning curves obtained for two neurons recorded in experiment 2 (***A***, same neuron as in Figs. 2*C–E* and 3*B*, left; ***B***, same neuron as in Fig. 3*B*, right). The abscissa is the shape distance (i.e., shape dissimilarity) within the tested morph line (shapes corresponding to some of the tested distances are shown below each shape axis). The origin of the shape axis is the preferred shape of the neuron obtained from the recording screening procedure (see Materials and Methods). The gray patch shows the region of shape space initially tested to obtain the preferred shape (unit shape distance; see Materials and Methods). The horizontal dotted line indicates the background rate of the neuron. Morphed shapes were sampled within either the unit distance (***A***) or a larger shape range (***B***) that included a stimulus drawn from a different shape space (data points at the far right in ***B***). For both neurons, responses to the object pairs (black line) are very close to the average (green line) of the responses to the constituent objects of the pairs presented in isolation, i.e., to the average of the horizontal dashed line and the red line (open circles). Error bars are SE of the mean firing rate. ***C***, Population average of 26 tuning curves obtained from the 15 most selective neurons recorded in experiment 2 for single and pair object conditions (see Results). These tuning curves were background subtracted, aligned to the preferred object (0 on the abscissa), and normalized by the response to the preferred object. The inset shows these 26 normalized tuning curves for responses to single objects (thin gray lines) and their average (thick red line). The red line (open circles) in the main panel shows the population average of the responses to single objects and included single object conditions outside the unit shape distance (gray patch). The horizontal dashed line shows the population average of the responses to the preferred object of each neuron (i.e., the shape at value 0 on the abscissa). The black line shows the population average of the responses to object pairs containing both the preferred object and another object sampled along the abscissa. The green line shows the average model prediction (population normalized average of average model curves as in ***A*** and ***B***). The cyan line shows the prediction of the CCI model (see Results). Error bars are SE of the population averages. Although different morph lines were tested for different neurons, example shapes are shown below the abscissa from a representative morph line. The dotted line is the background rate. Only 11 of 15 neurons (for a total of 18 of 26 responses) were tested outside the unit distance (gray patch) and contribute to the points outside this range. This plot is nearly identical when constructed from conditions in which the preferred object was either in the best or second best RF position (top or bottom; data not shown), i.e., forcing every neuron to contribute only one tuning curve to the population average.

object pairs (black line) that include the preferred object were very close to the average of the responses to the constituent objects presented in isolation (green line, see above). In addition, this neuron was also tested with objects sampled beyond the range of the morph-line unit distance (beyond the gray patch in Fig. 5*B*) and the response to pairs continued to primarily track the average. Moreover, an object belonging to a different shape space (a car) was also tested both in isolation (last open red circle on right) and paired with the preferred shape (last black point on right). Even for this very dissimilar object drawn from a completely different set of shapes, the response to an object pair containing this object and the preferred object of the neuron was very close to the average of the response to each object presented in isolation (last green point on right).

Building tuning curves of the responses to single and paired object conditions (Fig. 5*A*,*B*) allowed us to test, for the neuronal population recorded in experiment 2, whether there were any consistent deviations from the average model that depended on the degree of shape similarity between the objects in the pairs. To obtain a population measure of the dependence of pair responses from shape similarity, we considered the 19 most selective neurons recorded in experiment 2 that were included in the GOF analysis (Fig. 4) and built tuning curves for single object responses in top and bottom positions for each of these neurons, thus obtaining a total of 38 tuning curves. The tuning curves were aligned on a single shape axis (Fig. 5*C*, abscissa) by choosing the origin to be the preferred object of each neuron obtained during the screening procedure (see Materials and Methods). This preferred object was always used as one of the two objects in each object pair tested during later recordings. To get a meaningful average neuronal tuning curve, the 38 single tuning curves were screened to be both selective and essentially monotonic in the unit shape distance range, i.e., within the gray patch of Figure 5*C* (see Materials and Methods). This resulted in a subset of 26 tuning curves recorded in 15 neurons (11 neurons contributed two tuning curves, and four neurons contributed one tuning curve). These tuning curves were then averaged after subtracting background firing rates and normalizing by the response to the preferred object ($d = 0$). These 26 normalized tuning curves are shown individually in the inset of Figure 5*C*, and the resulting population average tuning curve is shown as the red line (open circles) in Figure 5*C* and inset. By construction, this population average falls along the abscissa as the distance from the preferred object is increased ($d > 0$). Note that the response typically falls to near background firing rates (ordinate = 0; dotted line) for "distant" objects sampled both within the same shape space (e.g., $d \geq 1$) and from other shape spaces (last open red circle on right). Note also that, although the preferred object ($d = 0$) was defined during initial screening, later tests sometimes included objects sampled to the "left" of the preferred object ($d < 0$), and these tests often revealed that the response to single objects continues to increase even beyond what was taken to be the preferred object (i.e., red line, open circles, continues to rise on the left side of Fig. 5*C*).

The black line in Figure 5*C* shows the population average of the normalized tuning curves obtained for pairs of simultaneously presented objects, in which the identity of one object of the pair was fixed at $d = 0$, whereas the identity of the second object spanned the tested range of shapes. Like the individual examples (Fig. 5*A*,*B*), the population average response to object pairs (black line) was intermediate between the average response

to the fixed object of the pair (horizontal dashed line) and the average response to the single objects (red line, open circles). For each of the tested neurons, the response to the object pairs was modeled as the average of the responses to the constituent objects of each pair to obtain model prediction curves as shown in Figure 5, *A* and *B* (green lines). These curves were normalized and averaged to obtain the population average model prediction curve shown in Figure 5*C* (green line). The fact that the black line and the green line almost perfectly overlap in Figure 5*C* supports two conclusions. First, the average model holds regardless of the similarity of the shapes composing the pairs. Second, at least under the limited clutter conditions tested in the present work, the agreement of neuronal data to the average model prediction becomes virtually perfect when responses of even a small population of IT neurons are pooled (as done here). This result was unchanged when all 38 tuning curves from the 19 most selective neurons recorded in experiment 2 were included in the analysis.

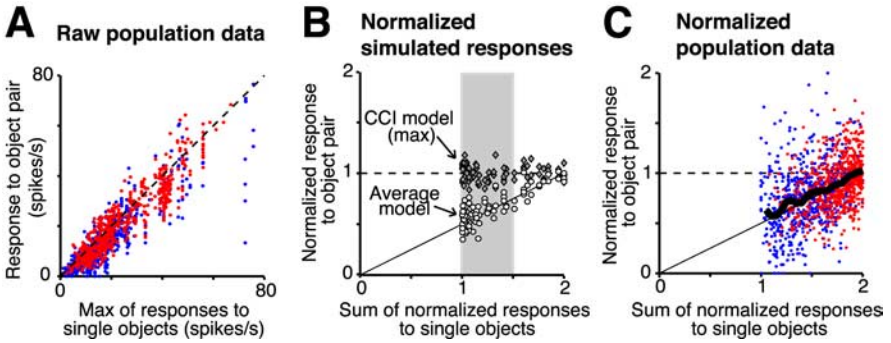

Figure 6. Comparison of the average model and the CCI model. ***A***, Responses to each object pair are plotted as a function of the maximum of the responses produced by each of the constituent objects of the pair (i.e., the CCI model prediction). Data from all 79 responsive neurons are included in the plot. Color code as in Figure 3. ***B***, Simulated normalized responses of two model neurons, one following the average rule (open circles) and the other following the CCI rule (gray diamonds; see Materials and Methods). In the first case, the neuronal response to object pairs was modeled as the average of the response to the constituent objects of the pair and, in the latter case, as the maximum of the constituent responses. Responses to each object pair and the two constituents of that pair were normalized by the maximum of the latter two. As a consequence, the sum of the normalized single object responses (in abscissa) ranges from 1 to 2, whereas the normalized responses to pairs cluster around the solid line with slope of 0.5 for the average model-simulated neuron and around the dashed line with slope of 0 for the CCI model-simulated neuron. The gray patch shows the range in which the predictions of the two models can be most easily discriminated. ***C***, Normalized responses for the whole population of 79 neurons (i.e., normalized as in ***B***). Color code as in Figure 3. The heavy black curve is the average response to object pairs as a function of the sum of individual responses. The average is computed in a running window of size 0.1 shifted in consecutive steps of 0.05.

Another model of responses to multiple objects

These findings clearly show that the responses of individual IT neurons are not unaffected by the presence of a second nonpreferred object (i.e., they are not clutter invariant), even when that second object produces no response on its own (see right side of plots in Fig. 5*A–C*). Instead, the response to an effective object is predictably reduced by the presence of a less effective "clutter" object and primarily follows an average model. However, given the relevance of this conclusion for theories of neuronal representation of multiple objects (Rousselet et al., 2003, 2004) and the difference with the some results in area V4 (Gawne and Martin, 2002), we explicitly compared the predictions of the average model with the predictions of an alternative model: the complete clutter invariance model. The CCI model predicts that the response to a pair of simultaneously presented objects is equal to the response of the most effective object of the pair, i.e., to the maximum of the responses to the individual stimuli.

The conditions used in experiment 2 are optimized to distinguish among the CCI model and the average model because the object pairs almost always include at least one condition in which both a very effective object and a noneffective object are presented together (discussed further below). Examination of the example curves in Figure 5, *A* and *B*, clearly shows that the CCI model is not correct. In particular, the addition of a second, less effective object always causes the response to decrease below that produced by the effective object presented in isolation (the black line is well below the dashed line). To examine this for this subpopulation of our data, CCI model prediction curves were built for each neuron and were normalized and averaged to obtain a predicted population average CCI curve (Fig. 5*C*, cyan line). The CCI model was consistently much poorer than the average model (green line) in predicting the population response to the stimulus pairs (black line), especially for object conditions in which a poorly effective object was part of the pair (e.g., for $d \geq 0.5$). Nevertheless, this is only a subset of our data, and we sought to fully test the predictions of the CCI model across both experiments for all of the individual IT neurons recorded in the three monkeys.

In general, testing whether responses to object pairs are better predicted by the average model (or any other model that is a weighted sum of responses to individual objects) or by the CCI model is not trivial. Although these models sound very different, the predictions of the average model and of the CCI model can be nearly identical depending on the object conditions used to test the neurons. To illustrate this, Figure 6*A* shows data from our whole population of 79 responsive neurons (the same data presented in Fig. 3*C*) but now with the prediction of the CCI model on the abscissa. The data primarily fall along the diagonal, and the correlation coefficient is high ($r = 0.91$ compared with $r = 0.92$ for average model) (Fig. 3*C*). At first glance, this suggests that the CCI model can explain IT pair responses nearly as well as the average model. However, plots like that in Figure 6*A* have limited power to distinguish among the CCI model, the average model (Fig. 3*C*), and other "reasonable" models that forces responses to pairs of objects to be near the firing rate range of each individual neuron. Testing the prediction of the CCI model for each individual neuron using measures of explained variance (as done for the average model in Figs. 3*A*,*B*, 4) can also produce misleading results. For example, if, as in experiment 2, the same effective stimulus is only paired with less effective stimuli, the CCI model predicts no variance across the responses to pairs and therefore the GOF is ill defined (the computed median GOF was only 4% in this case, but this is not a fair test of the CCI model).

In light of these issues, we sought a more powerful comparison between CCI and average model for all of the responsive neurons. To do this, we transformed the data: given a pair of objects *A* and *B*, with responses R_A and R_B to the individual objects and response R_{AB} to the pair *AB*, all three responses were transformed by dividing them by the maximum of the individual responses, i.e., $\mathrm{MAX}(R_A, R_B)$. As a consequence of this transformation, for each pair of objects, the response to one object presented alone is equal to 1, the response to the other object pre-

sented alone is between 0 and 1, and the sum $R_A + R_B$ of the responses to the objects presented alone is between 1 and 2. After this transformation, the predictions of the CCI and the average model are much more distinguishable. This is shown in Figure 6*B*, in which transformed responses to pairs (R_{AB}) are plotted against the sum of transformed single responses ($R_A + R_B$) for two different simulated neurons (one following a CCI model and one following an average model; see Materials and Methods). The scatter plots in Figure 6*B* show that the average model predicts that the transformed data should fall along the straight line with slope 0.5 (open circles), whereas the CCI model predicts that the transformed data should fall along the line with slope 0 (gray diamonds).

Data from the entire population of 79 responsive neurons recorded from the three monkeys were transformed as described above and then plotted in Figure 6*C*. Although the transformation produced data that appeared very noisy (division by a noisy number), the data points were most consistent with the average model in that they were scattered around the straight line with slope 0.5 (solid line). Indeed, a running average of the transformed data were almost exactly superimposed to the slope 0.5 line (heavy black line; for details, see legend). This shows that, across the entire population, the data are much more consistent with the average model than the CCI model (for quantitative assessment of the fit of the average model, see Fig. 4). Because the average slope in Figure 6*C* remains at 0.5 across the entire range of possible abscissa values (1.0–2.0), this shows that this agreement did not depend on the effectiveness of the individual objects, confirming the conclusions obtained from the subset of neurons examined in Figure 5.

As a final comparison of the average model and the CCI model, we focused on the stimulus conditions under which the predictions of the average and CCI model are most disparate. Specifically, as suggested by a previous V4 study (Gawne and Martin, 2002), we considered only object pair conditions in which the less effective objects in isolation evoked a response that was less than half the response evoked by the more effective object. In the transformed data described above, this corresponds to data with abscissa values <1.5, so we computed the median response to pairs (R_{MED}) for each neuron across this subset of conditions [$1 < (R_A+R_B) < 1.5$] (Fig. 6*B*, gray patch). If the data are uniformly distributed across the 1–1.5 interval, then the average model predicts that the R_{MED} distribution should be centered around 0.625 (in fact, the data were not uniformly distributed over this interval, so the average model predicted an R_{MED} distribution centered around 0.68). Conversely, the CCI model predicts that the R_{MED} distribution should be centered around 1.0. The observed median R_{MED} across a population of 64 (of 79) neurons recorded in the three monkeys was 0.7 (mean of 0.7), i.e., very close to the prediction of the average model [15 neurons were excluded from this analysis because they had no points in the interval $1 < (R_A+R_B) < 1.5$]. Put another way, this shows that, on average, the response of an IT neuron to an effective object is reduced by 30% when that effective object is presented with a "less than half" effective second object. At an individual neuron level, 43 of the 64 neurons had median responses to these object pairs that were reduced by at least 20% (relative to the response to the preferred object presented alone). We also observed that ~12% of the neurons (8 of 64) had responses to these object pairs that were reduced by <5% and might thus be taken to be consistent with the CCI model. Overall, however, the vast majority of IT neuron responses to pairs of objects were far from the CCI model prediction (see Discussion).

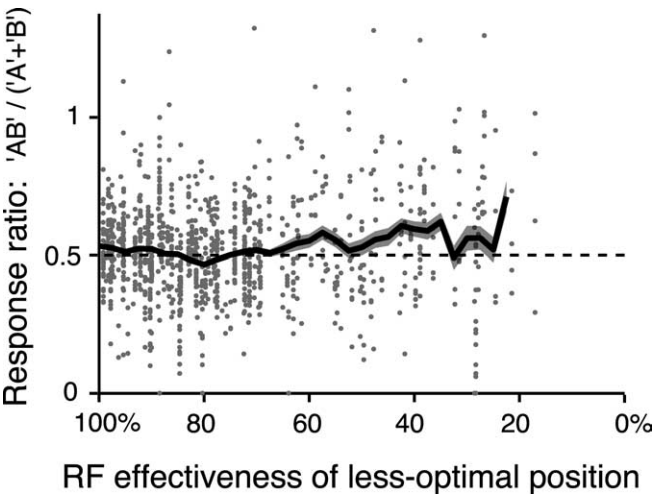

Figure 7. Agreement between responses to object pairs and prediction of the average model as a function of the RF sensitivity. The abscissa shows the RF effectiveness of the less effective position occupied by one of the two objects. The ordinate is the ratio of the responses to object pairs to the sum of responses to the constituent objects. Each gray point is one pair condition from one neuron, and all 79 responsive neurons are included. The solid black curve line is the average in a running window of size 10% shifted in consecutive steps of size 2.5%. The gray shaded region is ±1 SE of the running average. The horizontal dashed line shows the ratio predicted by the average model (0.5).

Response to objects pairs as a function of RF sensitivity

The present study was not designed to explicitly test the dependence of responses to objects as a function of their RF position in that the spatial separation of objects in the RF was not systematically varied and only two or three RF locations close to the center of gaze were tested. However, because IT neuronal RFs are not all centered at the same retinal position, have a broad range of sizes (Op De Beeck and Vogels, 2000), and can often be small relative to the separation of our objects (Op De Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003), we used these RF variations to ask whether there was any relationship between the averaging behavior described above and position in the RF. In particular, we might not expect the response to a pair of objects to be the average of the responses to the constituent objects if one of those objects was presented very far outside the RF (Missal et al., 1999), but we wondered whether we might detect some breakdown in averaging behavior when one of the objects was near the edge of the RF.

To examine this, we first defined the sensitivity of the RF at each tested position as the average response to objects that were effective in at least one position (i.e., eliciting a response significantly higher than background; *t* test, $p < 0.05$). We then examined the average model as a function of the relative effectiveness of the two RF positions. Specifically, for each tested object pair condition, we computed the response to the object pair as a fraction of the sum of the responses to the constituent objects, i.e., the ratio $R_{AB}/(R_A + R_B)$. As described above, this value tends toward 0.5 (average model) when all our data are considered together. Figure 7 shows this ratio as a function of the relative RF effectiveness of the two positions (the thick black curve is a running average). Two points can be taken from Figure 7. First, as expected based on the placement of our objects and the distribution of IT RF sizes (Op De Beeck and Vogels, 2000), for most neurons, both objects were well within the RF (relative effectiveness values are all >0%), but, for some neurons, one of the tested positions was near the edge of the RF (i.e., 20% RF effectiveness). Second, over this range of RF sensitivity conditions, we see only a very

slight trend away from averaging (and toward no effect of the second object) as we approach the edge of the RF. This trend is consistent with the reduction of response suppression produced by the less effective shape in a pair as a function of its distance from the more effective shape (Missal et al., 1999).

Dynamics of the response to object pairs

In the previous analyses, we considered IT neuronal responses in a fixed, 100 ms time interval (i.e., between 100 and 200 ms from the stimulus onset) (Fig. 2) that is constrained by IT latencies and behavioral reaction times to be most relevant for recognition tasks (Fabre-Thorpe et al., 1998; DiCarlo and Maunsell, 2000). We found that IT responses are normalized in that the mean response rate to object pairs in this time interval is well predicted by the average of the responses to the constituent objects. However, we wondered whether we could detect any deviations from this average model over this time window that might provide insight regarding the neuronal mechanisms underlying this normalization. To do this, we considered smaller time bins (25 ms width) shifted in consecutive time steps of 5 ms. For each neuron and each time bin, we computed the median ratio between responses to object pairs and the sum of responses to the constituent objects of the pair, i.e., $R_{AB}/(R_A + R_B)$, median over all object pairs tested for each neuron. As described above, this ratio tends toward 0.5 (average model) when data are considered over our standard 100–200 ms poststimulus interval. The resulting time course of the median ratio is shown for one neuron recorded in experiment 1 and another in experiment 2 (Fig. 8*A*). For comparison, the time course of the median response of each neuron to the object pairs is also shown in each panel (light gray background). Figure 8 shows that, for these two neurons, before the onset of the response (i.e., up to ~100 ms after stimulus onset), the median ratio fluctuates around 0.5. This is expected because the background rate during presentation of single objects and pairs of objects should be approximately the same. Then, at the beginning of the neuronal response (~100 ms poststimulus onset), the median ratio slightly increases above 0.5. The peak ratio then decreases, reaches a minimum (slightly below 0.5) at ~150 ms from the stimulus onset and then reaches a new peak (slightly above 0.5) at ~200 ms from the stimulus onset.

This temporal pattern suggests that, at the onset and toward the offset of the neuronal response, responses to object pairs are slightly above the average of the individual responses to the constituent stimuli, i.e., in the direction predicted by the sum of the responses to the constituent stimuli. This pattern was found for many neurons recorded in experiment 1 and for some neurons recorded in experiment 2. To examine this across the recorded neuronal population, we computed the time course of the median ratio between responses to object pairs and the sum of responses to the individual objects, median over all object pairs tested across the whole population of 48 responsive neurons of experiment 1. The resulting curve (Fig. 8*B*, solid line) showed dynamics very close to that observed in many individual neurons except that the peaks and trough were smaller. Because Figure 8 shows that the deviation from the value predicted by the average model is small (i.e., bounded between ~0.45 and ~0.55), the normalization reported throughout this paper is not highly dependent on the choice of analysis time interval (this analysis gave similar results when time bins of 50 and 75 ms were used; data not shown). However, because Figure 8 shows some consistent temporal modulation in the normalization, this provides some clues about the mechanisms that underlie response normalization in IT.

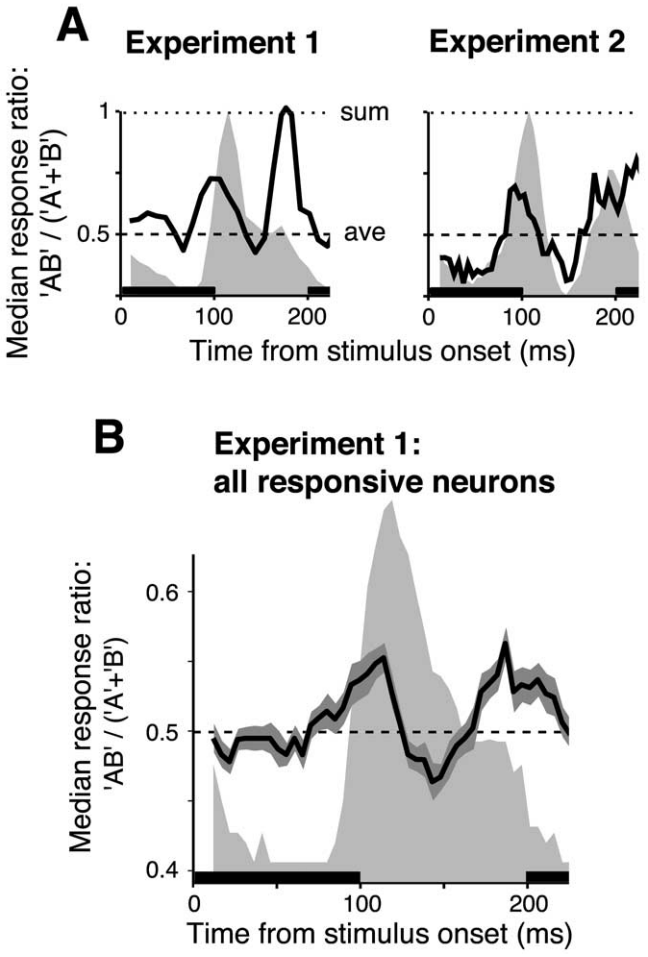

Figure 8. Dynamics of the response normalization. ***A***, Time course (solid line) of the median ratio between responses to object pairs and sum of the responses to the constituent objects of the pairs, for one individual neuron recorded in experiment 1 (left) and one neuron recorded in experiment 2 (right). Neuronal responses are computed in overlapping time windows of 25 ms shifted in time steps of 5 ms. The light gray background shows the time course of the median response to object pairs for each neuron. The heavy bars along the abscissa show timing and duration of stimulus presentation (see Fig. 1*E*), and all calculations are based on single objects or object pairs presented at time 0. The dashed line is the prediction of the average model without dynamics. The dotted line is the prediction of a sum model (i.e., a model in which the response to a pair of objects is the sum of the responses to the constituent objects). ***B***, The solid line is the median of the ratios between responses to object pairs and the sum of responses to the individual objects, median over all object pairs tested across the whole population of 48 responsive neurons of experiment 1. The shaded regions are ±1 SE of this median (the SE was computed by bootstrap resampling of the ratios). The light gray background is the running median over the responses to object pairs across the 48 neurons.

The temporal pattern shown in Figure 8 was less pronounced in experiment 2, although it was observed (Fig. 8*A*, second panel). When the 31 responsive neurons of experiment 2 were considered (as in Fig. 8*B*), the resulting curve had a time structure similar to that shown in Figure 8*B* (data not shown). However, the first peak at the time of response onset (~100 ms) was much less prominent. The absence of a clear peak may have been attributable to more frequent saturation of neuronal responses in experiment 2 because, unlike experiment 1, it involved a very effective object in all pair conditions (Fig. 2*E*).

Can attention shifts explain the average effect?

Because we did not explicitly control attention, we wondered whether the average effect described throughout this study could

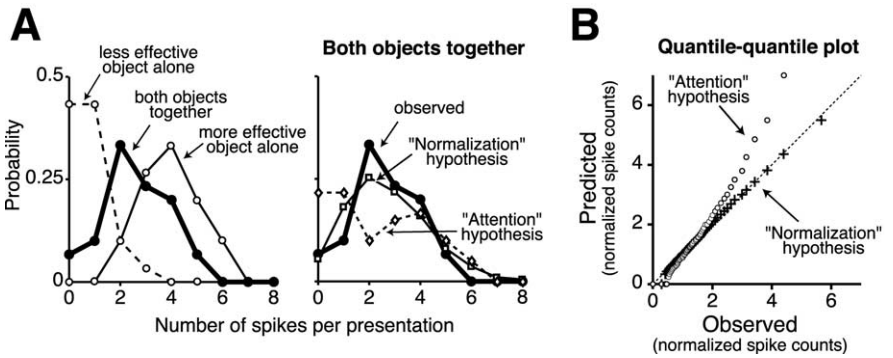

Figure 9. Spike count distributions of observed and predicted responses to object pairs. ***A***, Left, Spike count distributions obtained from one example IT neuron to repeated presentations of one example pair of objects (thick solid line) and to each of its constituent objects presented alone (thin solid and dashed lines). Right, The observed pair response distribution (thick solid line) is compared with the distributions predicted (see Materials and Methods) by the attention hypothesis (dashed line) and the normalization hypothesis (thin solid line). ***B***, Quantile–quantile plot (Rice, 1995). Each symbol plots the normalized spike count value (see Materials and Methods) at a particular quantile for the observed distribution of responses to object pairs and for the distribution predicted by the attention hypothesis (open circles) and by the normalization hypothesis (crosses). For example, the 50% quantile point shows the normalized spike count value in each distribution for which 50% of that distribution is lower than that value (quantiles 1–98% are shown in the plot). The dashed straight line (slope 1) is the expected relationship if the predicted distributions were identical to the observed one. Whereas the crosses line up very well along this line, the circles show considerable deviations from this linear relationship at both the lower and upper end of the plot, thus showing the supremacy of the normalization over the attention hypothesis in predicting the distribution of spike counts. All spike counts were taken in the standard 100 ms analysis window on individual presentations of object pairs and were normalized by the average spike count elicited by each object pair condition (see Materials and Methods).

have resulted from effects of attention. When two objects are present and a monkey is cued to attend a specific visual field location (Connor et al., 1997; Reynolds et al., 1999) or a specific target object (Moran and Desimone, 1985; Treue and Maunsell, 1996; Chelazzi et al., 1998), neuronal responses in the ventral visual stream (including IT) move toward the response elicited by the attended object, as if that object were presented alone. Thus, if one of the two objects in each of our object pairs were attended on each presentation of the pair, and the choice of the attended object was random across the 10–30 trials in which each pair was tested, the mean response over all presentations could look very much like the average of the responses to the constituent objects presented alone. Although our presentation conditions (100 ms stimulus duration) are likely far too rapid for attention shifts during a single presentation of a pair, if the animal's attention were directed toward one position for approximately half of the presentations and the other position for the rest, attention shifts might explain the average effect.

This attention hypothesis explicitly assumes that the distribution of responses across the 10–30 presentations of each object pair is drawn more or less equally from the distributions of responses to the two constituent objects (thus producing the average effect in the mean of the distribution). The alternative hypothesis (and the one we have implied throughout this paper) is that the observed average effect is a true normalization in that it is obtained for each and every stimulus pair presentation, as if a single, "average-effective" stimulus had been presented. The attention hypothesis predicts that the distribution of responses to each pair of objects (i.e., the distribution over all repetitions of that stimulus condition) should be very broad (and approaching bimodal), especially for cases in which one object is very effective and the other is noneffective. In contrast, the normalization hypothesis predicts that distribution of responses should be no different (and thus no broader) than that produced by a single, average-effective object that produces the same mean firing rate.

To illustrate the predictions of these two hypotheses, Figure 9*A* (left panel) shows the distribution of spike counts obtained from a single IT neuron to repetitions of one example pair of objects (thick solid curve) and to each of the constituent objects of the pair (thin solid and dashed curves). The thin dashed and solid curves in the right panel of Figure 9*A* show, for this object pair condition, the predictions of the attention hypothesis and the "normalization hypothesis" (see Materials and Methods). Note that, for both hypotheses, the predicted pair response distribution has the same mean (i.e., the average of the responses to the constituent objects) but different shapes. The observed pair response distribution (thick solid curve) is more consistent with the normalization hypothesis, and this was qualitatively true for all cases we examined in which the responses to the two objects of the pair were different enough so as to make the hypotheses visually distinguishable (as in this example).

To quantitatively test these two hypotheses over our entire dataset, we used two approaches to compare the observed pair response distributions with the predictions of the hypotheses. First, we measured the broadness of the distributions by calculating the Fano factor (i.e., the ratio of the variance of the average spike count and its mean) for each of the 79 responsive neurons. We included all of the pair conditions in which only one of the objects in the pair produced a response significantly higher than background when presented alone (Figs. 3, 6, blue dots) (see Materials and Methods) because these object pair conditions should give the broadest distribution of pair responses under the attention hypothesis and thus have the most power to distinguish among the two alternatives. We found that the observed average Fano factor (1.05) was as follows: (1) very close to the value of 1.0 predicted by Poisson spiking (Softky and Koch, 1993; Shadlen and Newsome, 1994; Rieke et al., 1997; Shadlen and Newsome, 1998); (2) markedly different than the value predicted by the attention hypothesis (1.24; one-tailed paired t test, $p < 0.001$) (see Materials and Methods); and (3) not significantly higher than the value predicted by the normalization hypothesis (i.e., the Fano factor obtained for single object conditions, 1.13; one-tailed unpaired t test, $p = 0.98$) but, instead, was slightly smaller (i.e., in the opposite direction from that predicted by the attention hypothesis).

As a second approach, we aimed to directly compare the entire shape of the observed pair response distribution with the distribution predicted by the two hypotheses (i.e., compare the shape of the thick solid curve in the right panel of Fig. 9*A* with the shape of the thin solid and dashed curves). Because the 15–30 trials recorded for each stimulus conditions did not typically provide enough statistical power to test this prediction at the level of individual conditions, we performed a population analysis over the same conditions described above by normalizing and combining all of the distributions (79 neurons; see Materials and Methods). Examination of quantile–quantile plots (Fig. 9*B*) showed that the observed population distribution was very similar to the distribution expected under the normalization hypothesis (no significant difference; Kolmogorov–Smirnov tests, $p = 0.17$), although, in contrast, it showed strong departure from that

predicted by the attention hypothesis (highly significant difference; Kolmogorov–Smirnov tests, $p < 10^{-15}$; experiment 1, $p < 10^{-24}$; experiment 2, $p < 10^{-8}$).

In summary, these results indicate that the observed average effect described throughout this study cannot simply be explained by attention shifts. This does not rule out all possible explanations of the average effect that involve attention or imply that attention plays no role in IT responses to multiple objects [it can play a role under some conditions (Moran and Desimone, 1985; Chelazzi et al., 1998)]. However, at least under the behavioral conditions tested here (passive fixation), our data are consistent with the hypothesis that object pairs activate IT neurons as if a single, average-effective object had been presented.

Does the average effect generalize to other stimulus presentation conditions?

In this study, we presented stimulus conditions for only 100 ms each and at a rate of five per second (see Materials and Methods) in an attempt to collect as much data as possible from each recorded neuron and to limit attentional shifts that might result from longer presentations. Although this presentation rate is arguably most physiologically relevant because it is consistent with that spontaneously produced by the eye movements of free-viewing monkeys (DiCarlo and Maunsell, 2000), we wondered whether the average result would have been found if we had used more "standard" presentation conditions. To test this, we analyzed only the responses to the first stimulus condition presented on each trial (comparable with standard conditions in that it followed a standard fixation period; see Materials and Methods). For each of the 79 responsive neurons, we computed the ratio between (first stimulus) responses to the object pairs and the sum of the (first stimulus) responses to the constituent objects (done for all available stimulus pair conditions). Across all 79 responsive neurons, the median value of this ratio was 0.51: nearly identical to the value predicted by the average model (0.5) and the value computed using all recorded responses (0.52). This shows that the observed average effect does not result from the fact that many stimuli were presented in relatively rapid succession on each trial.

Discussion

The goal of the present study was to systematically examine IT neuronal responses in limited clutter conditions (i.e., with several objects present), using two complementary experimental paradigms. In experiment 1, we tested the exact same visual object conditions across an unbiased sample of IT neurons. In experiment 2, we optimized stimulus conditions for each neuron to produce maximal selectivity across a continuous shape dimension.

Our results show that, for both experiments, a large fraction of the explainable variance in responses to object pairs was explained by the average of the responses to the constituent objects (~63%, average model) (Fig. 4*A*). The average model becomes virtually perfect when responses of even a small population of neurons are pooled (Fig. 5*C*). Thus, at least under the conditions tested here, IT responses to pairs of objects depend primarily on the effectiveness of each constituent object in driving the neuron, and it does not matter much whether that effectiveness is altered by changing the object identity (Fig. 5) or RF position (Fig. 7). Moreover, objects that are completely ineffective when presented alone powerfully reduce the neuronal responses when paired to very effective objects (at least for the conditions tested, see below). As such, most IT neurons do not have CCI, i.e., their responses are not independent of the presence of less effective objects (Figs. 5*C*, 6).

Previous studies of multiple stimuli in the ventral visual stream

Consistent with previous investigators (Sato, 1989; Miller et al., 1993; Rolls and Tovee, 1995; Missal et al., 1997, 1999; Chelazzi et al., 1998; Sheinberg and Logothetis, 2001), our study found that IT responses to very effective objects are typically reduced by the presence of less effective clutter objects. In particular, on average, IT responses to an effective stimulus were decreased by ~30% when a less than half effective stimulus was also presented, very close to the suppression reported by Rolls and Tovee (1995) and Missal et al. (1999) for similar object pairings.

There are hints of an average effect in previous IT studies. Miller et al. (1993) found a correlation between the amount of response suppression and the effectiveness of the RF location at which a second object was presented, implying that the amount of suppression depends on the neuronal response of the neuron to the second object presented alone. Conversely, Missal et al. (1999) did not find correlation between responses to object pairs and the sum of the responses to the constituent objects. In that study, however, because a very effective shape was always paired with a poorly effective or ineffective shape, there was likely little variance in the sum of the responses (Fig. 6*C*, abscissa), making it difficult to reliably detect the average effect. That is, the systematic relationship reported here might not have been apparent without testing a wide range of stimulus conditions.

Previous studies appear to disagree on how response suppression depends on the identity of the second, less-effective object. Miller et al. (1993) suggested that the amount of suppression did not depend on that identity, although Missal et al. (1999) found the opposite for 50% of neurons. Our results indicate that the answer depends not on object identity per se but on the activation produced by each object present alone (with important caveats; see below). For example, objects that have different identities but do not produce any response will, on average, produce the same amount of response suppression (compare with the far right points of Fig. 5*C*). Although our data do not rule out shape-dependent deviations from the average model for all individual neurons (see example in Fig. 5*B*), they suggest that any such deviations would be averaged out by pooling even a small IT population.

Our results are also in good agreement with previous investigations of earlier visual areas. In particular, Reynolds et al. (1999) found that V2 and V4 neuronal population responses to stimulus pairs tended to follow an average model when the monkey's attention was not directed to either stimulus. Thus, we speculate that a common set of stimulus interaction mechanisms may operate at each visual stage. However, mechanisms that produce averaging behavior may not be the only ones mediating stimulus interactions, because some populations of cat V1 neurons (Lampl et al., 2004) and monkey V4 neurons (Gawne and Martin, 2002) have responses that are similar to the response of the best constituent stimulus of each pair (CCI model). Nevertheless, in our IT recordings, only a small percentage of neurons (~10%) showed this behavior, even when we specifically examined the response conditions tested by Gawne and Martin in V4 (see Results).

Possible mechanisms

The average model presented here is purely descriptive and leaves open the question of underlying mechanisms (although Fig. 8 may provide clues). Reynolds et al. (1999) proposed a mechanis-

tic implementation of the "biased-competition model" (Desimone and Duncan, 1995) that can explain the weighted average responses to constituent stimuli of a pair in V2 and V4 (and possibly IT). That model assumes that each object activates separate populations of neuronal afferents, and a normalization factor proportional to the total synaptic input rescales the neuronal response. The average effect could also arise if the output of each IT neuron was normalized by the total spiking activity of a broad population of IT cells, similar to divisive normalization models proposed to explain nonlinear behavior in early visual areas (Heeger, 1992; Heeger et al., 1996; Carandini et al., 1997; Schwartz and Simoncelli, 2001; Cavanaugh et al., 2002) and in area MT (Recanzone et al., 1997; Britten and Heuer, 1999; Heuer and Britten, 2002). Finally, other biologically inspired computational models, such as those using iterated MAX and Gaussian tuning operations (Riesenhuber and Poggio, 1999a,b), can also produce average-like effects in simulated IT neurons, although they were not explicitly designed for that purpose (M. Kouh, D. Zoccolan, and T. Poggio, unpublished observations).

Generality, limitations, and implications

Although our results are very consistent across three monkeys, two experimental designs, and a wide range of object pair conditions, we have clearly not explored the entire "operating range" of the visual system. For one, neuronal responses were assessed using a relatively rapid stimulus presentation rate. However, because our presentation rate was similar to that produced by free-viewing monkeys (Motter and Belky, 1998a,b; DiCarlo and Maunsell, 2000; Sheinberg and Logothetis, 2001), gives robust object selectivity in IT (Keysers et al., 2001), and is clearly within the abilities of human recognition (Potter, 1976; Intraub, 1980; Rubin and Turano, 1992), our data provide reasonable estimates of the neuronal responses that underlie fast recognition in clutter. Furthermore, we also found the same average effect for more standard presentation conditions (first stimulus) and previous studies have reported a similar effect in areas V2 and V4 with longer presentation times (Reynolds et al., 1999).

Similarly, although our results do not speak to effects of explicit manipulation of visual attention (Treue and Maunsell, 1996; Connor et al., 1997; Chelazzi et al., 1998; Reynolds et al., 1999), our data are highly relevant to conditions without such pretrial cuing, and robust recognition in clutter is still observed under such conditions (Potter, 1976; Intraub, 1980; Rubin and Turano, 1992). Critically, we have shown that the observed average effect in IT cannot be explained by simple attentional shifts but instead behaves as if it were a primarily feedforward property of the ventral visual stream.

By design, some of the neurons contained in our dataset were not tested with their true preferred object (especially experiment 1). However, significant efforts were made in experiment 2 to achieve conditions in which objects were highly preferred. Although current techniques cannot guarantee optimal objects, most neurons in the experiment 2 dataset had very sharp tuning (Fig. 5*C*), with peak firing at the center of gaze >40 spikes/s. Thus, even for objects that are likely close to the tuning peak of each neuron, the average model closely describes responses to object pairs (Fig. 5*C*).

Although our long-term goal is to understand recognition in real-world clutter, the reduced goal of this study was to understand IT responses in limited, parameterized clutter conditions, i.e., pairs or triplets of objects near the fovea. Additional investigations will be required to understand IT response behavior (1) for larger numbers of objects, (2) over an even broader range of object identities and RF locations, and (3) for complex real-world backgrounds such as textures and scenes. For example, although the average effect clearly cannot hold for all ineffective objects (e.g., objects presented far outside the RF or with attributes that do not penetrate the visual system), these departures from the average effect may provide insight into how objects are represented in IT.

Finally, it is intriguing to ask whether the average effect has some useful role in object representation. For example, it has been shown that normalization mechanisms lead to more efficient forms of representation in early visual areas (Schwartz and Simoncelli, 2001) and can produce bell-shaped tuning curves that can support position- and scale-invariant object representation (Poggio and Bizzi, 2004). Note that the average effect does not change the preferred objects of IT neurons but rescales their tuning properties (Fig. 5), consistent with the preservation of IT selectivity profiles found in studies using natural visual scenes (Sheinberg and Logothetis, 2001; Rolls et al., 2003). Nevertheless, the average effect indicates that the presence of a second object changes the response of each neuron to its preferred object and thus, at first glance, suggests that such an effect will negatively impact recognition of the preferred object. However, although no effect of a second object may seem to be a desirable property of individual IT neurons (Rousselet et al., 2003, 2004), it may not be necessary when populations of IT neurons are considered. The impact of the average effect at the population level and the possibility that such behavior could allow simultaneous representation of multiple objects are areas of ongoing research.

References

Blanz V, Vetter T (1999) A morphable model for synthesis of 3D faces. In: 1999 Symposium on Interactive 3D Graphics—Proceedings of SIGGRAPH'99, pp 187–194. New York: ACM.

Britten KH, Heuer HW (1999) Spatial summation in the receptive fields of MT neurons. J Neurosci 19:5074–5084.

Carandini M, Heeger DJ, Movshon JA (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. J Neurosci 17:8621–8644.

Cavanaugh JR, Bair W, Movshon JA (2002) Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. J Neurophysiol 88:2530–2546.

Chelazzi L, Duncan J, Miller EK, Desimone R (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. J Neurophysiol 80:2918–2940.

Connor CE, Preddie DC, Gallant JL, Van Essen DC (1997) Spatial attention effects in macaque area V4. J Neurosci 17:3201–3214.

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222.

DiCarlo JJ, Maunsell JHR (2000) Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci 3:814–821.

DiCarlo JJ, Maunsell JHR (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. J Neurophysiol 89:3264–3278.

Efron B, Tibshirani RJ (1998) An introduction to the bootstrap. Boca Raton, FL: Chapman and Hall/CRC.

Fabre-Thorpe M, Richard G, Thorpe SJ (1998) Rapid categorization of natural images by rhesus monkeys. NeuroReport 9:303–308.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Gawne TJ, Martin JM (2002) Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. J Neurophysiol 88:1128–1135.

Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vis Neurosci 9:181–197.

Heeger DJ, Simoncelli EP, Movshon JA (1996) Computational models of cortical visual processing. Proc Natl Acad Sci USA 93:623–627.

Heuer HW, Britten KH (2002) Contrast dependence of response normalization in area MT of the rhesus macaque. J Neurophysiol 88:3398–3408.

Intraub H (1980) Presentation rate and the representation of briefly glimpsed pictures in memory. J Exp Psychol [Hum Learn] 6:1–12.
Keysers C, Xiao DK, Foldiak P, Perrett DI (2001) The speed of sight. J Cogn Neurosci 13:90–101.
Lampl I, Ferster D, Poggio T, Riesenhuber M (2004) Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J Neurophysiol 92:2704–2713.
Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annu Rev Neurosci 19:577–621.
Maunsell JHR (1995) The brain's visual world: representation of visual targets in cerebral cortex. Science 270:764–769.
Miller EK, Gochin PM, Gross CG (1993) Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. Brain Res 616:25–29.
Missal M, Vogels R, Orban GA (1997) Responses of macaque inferior temporal neurons to overlapping shapes. Cereb Cortex 7:758–767.
Missal M, Vogels R, Li CY, Orban GA (1999) Shape interactions in macaque inferior temporal neurons. J Neurophysiol 82:131–142.
Moran J, Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. Science 229:782–784.
Motter BC, Belky EJ (1998a) The zone of focal attention during active visual search. Vision Res 38:1007–1022.
Motter BC, Belky EJ (1998b) The guidance of eye movements during active visual search. Vision Res 38:1805–1815.
Op De Beeck H, Vogels R (2000) Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426:505–518.
Poggio T, Bizzi E (2004) Generalization in vision and motor control. Nature 431:768–774.
Potter MC (1976) Short-term conceptual memory for pictures. J Exp Psychol [Hum Learn] 2:509–522.
Recanzone GH, Wurtz RH, Schwarz U (1997) Responses of MT and MST neurons to one and two moving objects in the receptive field. J Neurophysiol 78:2904–2915.
Reynolds JH, Desimone R (2003) Interacting roles of attention and visual salience in V4. Neuron 37:853–863.
Reynolds JH, Chelazzi L, Desimone R (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. J Neurosci 19:1736–1753.
Rice JA (1995) Mathematical statistics and data analysis. Belmont, CA: Duxbury.
Rieke F, Warland D, Ruyter van Steveninck RR, Bialek W (1997) Spikes: exploring the neural code. Cambridge, MA: MIT.
Riesenhuber M, Poggio T (1999a) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.
Riesenhuber M, Poggio T (1999b) Are cortical models really bound by the "binding problem"? Neuron 24:87–93, 111–125.
Rogers DF (2000) An Introduction to NURBS with historical perspective. San Francisco: Kaufmann.
Rolls ET, Tovee MJ (1995) The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. Exp Brain Res 103:409–420.
Rolls ET, Aggelopoulos NC, Zheng F (2003) The receptive fields of inferior temporal cortex neurons in natural scenes. J Neurosci 23:339–348.
Rousselet GA, Thorpe SJ, Fabre-Thorpe M (2003) Taking the MAX from neuronal responses. Trends Cogn Sci 7:99–102.
Rousselet GA, Thorpe SJ, Fabre-Thorpe M (2004) How parallel is visual processing in the ventral pathway? Trends Cogn Sci 8:363–370.
Rubin GS, Turano K (1992) Reading without saccadic eye movements. Vision Res 32:895–902.
Sato T (1989) Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. Exp Brain Res 77:23–30.
Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4:819–825.
Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. Curr Opin Neurobiol 4:569–579.
Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. J Neurosci 18:3870–3896.
Sheinberg DL, Logothetis NK (2001) Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci 21:1340–1350.
Shelton C (2000) Morphable surface models. Int J Comp Vis 38:75–91.
Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13:334–350.
Tanaka K (1996) Inferotemporal cortex and object vision. Annu Rev Neurosci 19:109–139.
Treue S, Maunsell JHR (1996) Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382:539–541.