

Semantic Control of Feature Extraction from Natural Scenes

Peter Neri

Institute of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, United Kingdom, Laboratoire des Systèmes Perceptifs, CNRS UMR 8248, and Département d'études cognitives, Ecole normale supérieure, 75005, Paris, France

In the early stages of image analysis, visual cortex represents scenes as spatially organized maps of locally defined features (e.g., edge orientation). As image reconstruction unfolds and features are assembled into larger constructs, cortex attempts to recover semantic content for object recognition. It is conceivable that higher level representations may feed back onto early processes and retune their properties to align with the semantic structure projected by the scene; however, there is no clear evidence to either support or discard the applicability of this notion to the human visual system. Obtaining such evidence is challenging because low and higher level processes must be probed simultaneously within the same experimental paradigm. We developed a methodology that targets both levels of analysis by embedding low-level probes within natural scenes. Human observers were required to discriminate probe orientation while semantic interpretation of the scene was selectively disrupted via stimulus inversion or reversed playback. We characterized the orientation tuning properties of the perceptual process supporting probe discrimination; tuning was substantially reshaped by semantic manipulation, demonstrating that low-level feature detectors operate under partial control from higher level modules. The manner in which such control was exerted may be interpreted as a top-down predictive strategy whereby global semantic content guides and refines local image reconstruction. We exploit the novel information gained from data to develop mechanistic accounts of unexplained phenomena such as the classic face inversion effect.

Key words: inversion effect; natural statistics; noise image classification; orientation tuning; reverse correlation

Introduction

Early electrophysiological recordings from primary visual cortex have established that individual neurons can be remarkably selective for fundamental image features such as texture orientation (Hubel, 1963); subsequent studies have demonstrated that this selectivity can be altered in systematic ways by presenting stimuli outside the local region associated with measurable spiking output from the cell (Bolz and Gilbert, 1986). There are well known perceptual counterparts to this class of contextual effects (Schwartz et al., 2007), which can be demonstrated using synthetic laboratory stimuli not necessarily relevant to natural vision (Rust and Movshon, 2005).

Subsequent research shifted emphasis from the question of whether the functional properties of local feature detectors are altered by presenting a stimulus outside the overtly responsive region, to the question of how such well established alterations depend on the specific characteristics of the surrounding stimulation (Allman et al., 1985). An issue of topical interest in recent years has been to what extent it matters whether contextual stimuli resemble those encountered during natural vision (Simoncelli and Olshausen, 2001; Felsen and Dan, 2005; Geisler, 2008). This

question can be asked at two conceptually distinct levels of analysis.

At the more basic level, the issue is whether image statistics not necessarily associated with semantic content may be relevant. For example, it is known that finer detail in natural scenes typically carries less contrast than coarser detail according to a systematic relationship between detail and image power (Ruderman and Bialek, 1994; $1/f^2$). Existing evidence indicates that this characteristic may impact perceptual (Párraga et al., 2000) and neural (Simoncelli and Olshausen, 2001) processing. This class of studies does not pertain to semantic content: an image may possess naturalistic power characteristics but no recognizable objects/landscapes (Piotrowski and Campbell, 1982; Felsen et al., 2005).

At a higher level of image analysis, the issue is whether the meaning of the scene may be relevant. This issue is distinct from the one outlined in the previous paragraph: two images may both conform to various statistical properties of natural scenes (Gerhard et al., 2013), yet one and not the other may deliver content that is meaningful to a human observer (Torralba and Oliva, 2003). The question is whether image representation at this advanced level may impact the properties of the feature detectors it relies upon for image reconstruction in the first place (Rao and Ballard, 1999; Bar, 2004). The present study is concerned with this specific and important question, for which there is no available evidence from human vision.

Our results demonstrate that, contrary to widely adopted assumptions about the static nature of early visual function (Carandini et al., 2005; Morgan, 2011), higher level semantic representations affect local processing of elementary image features. In doing so, these results offer a compelling example of the

Received April 26, 2013; revised Nov. 18, 2013; accepted Nov. 24, 2013.

Author contributions: P.N. designed research; P.N. performed research; P.N. contributed unpublished reagents/analytic tools; P.N. analyzed data; P.N. wrote the paper.

This work is supported by the Royal Society of London, Medical Research Council, UK.

The author declares no competing financial interests.

Correspondence should be addressed to Peter Neri, Institute of Medical Sciences, University of Aberdeen, Forest-erhill, Aberdeen AB25 2ZD, UK. E-mail: peter.neri@abdn.ac.uk.

DOI:10.1523/JNEUROSCI.1755-13.2014

Copyright © 2014 the authors 0270-6474/14/342374-15\$15.00/0

integrated nature of human vision whereby different hierarchical levels interact in both feedforward and feedback fashion (Lamme and Roelfsema, 2000; Bullier, 2001), from the earliest and most elementary stage (feature extraction; Morgan, 2011) to the furthest and most advanced one (image interpretation; Ullman, 1996).

Materials and Methods

Natural image database

We initially obtained eight image databases from <http://cvcl.mit.edu/database.htm> (at the time of downloading they contained 335 images per database on average); the category labels assigned by the creators (Oliva and Torralba, 2001) were “coast and beach,” “open country,” “forest,” “mountain,” “highway,” “street,” “city center,” and “tall building.” We therefore started with a total of ~ 2.7 K images (resolution 256×256 pixels). Of these we selected 320 (approximately 1 of 8) using an entirely automated software procedure (no pick-and-choose human intervention). We first acquired each image as grayscale, rescaled intensity to range between 0 and 1 and applied a smooth circular window so that the outer edge of the image (5% of diameter) was tapered to background gray (Fig. 1A). We subsequently applied a Sobel filter of dimension equal to $\sim 15\%$ image size to identify the location of peak edge content. Subsequent to edge detection we applied a broad low-pass Gaussian filter (SD equal to half image size), rescaled intensity to range between 0 and 1, and set all image values above 1/2 to bright, all those below to dark; we refer to this image as the “thresholded” image. We then created an image of size equal to the Sobel filter containing an oriented sharp edge, centered it on the previously determined location of peak edge content, and matched its orientation to the local structure of the thresholded image by minimizing square error (MSE); the resulting MSE value was used as a primary index of how well that particular image was suited to the purpose of our experiments. We focused on peak edge content rather than selecting arbitrary edge locations to maximize the quality of edge insertion and therefore increase the viability of the assigned congruent/incongruent discrimination; task viability is a pressing issue when testing naive observers (as was done in this study). We analyzed all images using the procedure just described and only retained the top 40 for each database (those with smallest MSE value within their database). All images were rescaled to have the same contrast energy; when projected onto the monitor, they spanned a range between 4 and 60 cd/m^2 on a gray background of 32 cd/m^2 and occupied 12° at the adopted viewing distance of 57 cm.

Natural movie database

We initially extracted 60 sequences, each 5 min long, from a random selection of recent (post 2000) mainstream movies spanning a wide range of material (e.g., *Inside Man*, *The Assassination of Jesse James*, 3:10 to Yuma, *Batman*, *Lord of the Rings*, *Spiderman*, *Casino Royale*). At the original sampling rate of 25 Hz this resulted in 450 K images of 256×256 pixels (cropped central region). Each sequence was processed by a motion-opponent energy detector via convolution (in x - y - t) with a moving dark–bright edge in stage 1, convolution with an edge moving in the opposite direction in stage 2, and subtraction of squared stage 2 output from squared stage 1 output. This procedure was performed separately for 24 different directions around the clock and for local patches measuring $\sim 10\%$ of image size. The moving edge lasted eight frames (equivalent to 320 ms in movie time) and moved by three pixels on each frame (traversing a distance of $\sim 1^\circ$ in actual stimulus space, equivalent to a speed of $\sim 0.3^\circ/\text{s}$). Each point in time throughout the movie dataset was associated with the spatial coordinates and motion direction of the local patch that returned the largest output from the motion detector applied over the immediately following 320 ms segment. We then ranked segments according to the associated output, and iteratively retained those with largest outputs under the constraint that each newly retained segment should not fall within a 3 s window of previously retained segments (to avoid selecting segments from almost identical parts of the movie). The outcome of this automated selection process returned ~ 4000 candidate segments for probe insertion (approximately 1 every 4.5 s of foot-

age). Further selection was performed via visual inspection of each segment to ensure that inserted probes did not span scene changes or similar artifactual motion signals; the outcome of this additional selection process returned 1000 segments, which were used for the experiments. Each extracted segment lasted 880 ms (we retained 280 ms before and after the 320 ms segment analyzed by the motion detector) to allow for smooth temporal insertion of the probe (see below).

Probe design and insertion into natural scenes

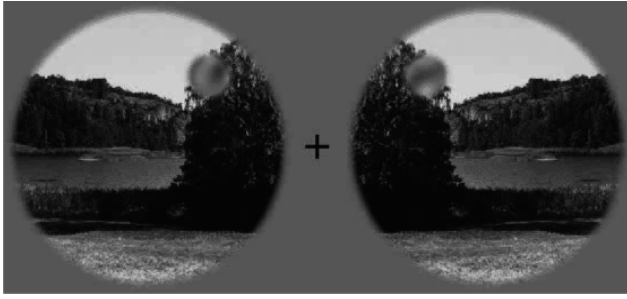
Probe parameters. Probes (indicated by magenta dashed circles in Fig. 1A, B) consisted of superimposed pseudo-Gabor wavelets at 16 different orientations uniformly spanning the 0 – π range, each taking one of four random phases ($0, \pi/2, \pi, 3/2\pi$; Fig. 1E, F). Carrier spatial frequency was fixed at ~ 1 cycle/degree. The envelope was constant over a circular region spanning $\sim 1.4^\circ$ (diameter) and smoothly decreasing to 0 around the edge following a Gaussian profile with SD of 9 arcmin. The 16 contrast values assigned to the different wavelets on trial i are denoted by vector $\mathbf{s}_i^{[q]}$ ($q = 0$ for congruent probe, $q = 1$ for incongruent probe), where the first entry into the vector corresponds to the congruent orientation. In image space, the congruent orientation was that associated with smallest MSE (see above), i.e., the orientation of the best-fit edge to the local structure of the natural scene.

Injection of orientation noise. On each trial $\mathbf{s}_i^{[q]} = \mathbf{t}^{[q]} + \mathbf{n}_i^{[q]}$; the contrast distribution across orientation consisted of a fixed (independent of trial i) target signal $\mathbf{t}^{[q]}$, summed onto a noise sample \mathbf{n}_i that varied from trial to trial. The target signal vector $\mathbf{t}^{[q]}$ consisted of 0s everywhere except the first entry when $q = 0$ (congruent probe) or the ninth entry when $q = 1$ (incongruent probe), which was assigned a value denoted by ρ (target intensity). Each entry of the noise vector \mathbf{n} followed a Gaussian distribution with mean 2.7% and SD 0.9% (contrast units) clipped to ± 3 SD. We adjusted ρ individually for each subject to target threshold performance ($d' \sim 1$; Fig. 3B) following preliminary estimation of threshold point via a two-down one-up staircase procedure; when expressed as multiple of noise mean, ρ was ~ 4 (mean across observers).

Probe insertion and sampling along vertical meridian. The probe was smoothly inserted (by using wavelet envelope to control probe/image ratio contribution to image) into the local region of the natural scene identified by the automated edge-detection procedure detailed above; see examples in Figure 1, C and D, for incongruent and congruent probes, respectively. Because the above-detailed algorithm did not select a random location for probe insertion, probe distribution across images may have been biased along some retinotopically specified coordinate, most relevant here is the vertical meridian. If, for example, there was a tendency on the part of the algorithm to preferentially insert edges within the upper part of images in our dataset as opposed to the lower part, inverting images upside down would affect the probability that probes appear on the upper versus lower part of the visual field; the effects on orientation tuning we observed for inversion (Figs. 1G, 3A) may then simply reflect differential orientation tuning for the upper versus lower visual fields. Probe distribution along the vertical meridian across our image database is plotted in Figure 1B (histogram to the right of image), where it can be seen that it is symmetric around the midpoint (fixation). We can therefore exclude the possibility that the effects observed for inversion may reflect asymmetric probing of upper versus lower visual fields.

Probe design and insertion into natural movies

Probes (Fig. 5A, B, magenta dashed circles) consisted of superimposed pseudo-Gabor moving wavelets at 16 different directions uniformly spanning the 0 – 2π range (Fig. 5C, D) each taking a random phase between 0 and 2π . Carrier spatial frequency and envelope were identical to those used in the orientation tuning experiments. The temporal envelope smoothly ramped up (following a Gaussian profile with SD of 100 ms) over 300 ms, was constant for the following 280 ms, and ramped down over the remaining 300 ms. We can use the same notation and logic adopted for orientation tuning experiments in describing the 16 contrast values assigned to the different moving wavelets. Each entry of the noise vector \mathbf{n} followed the same Gaussian distribution detailed above for oriented static probes; ρ was ~ 5 on average (units of noise mean). The



Movie 1. Orientation discrimination in natural scenes. Eleven sample trials from the orientation discrimination experiments demonstrating upside-down inversion (trials 2, 3, 6, 9, 10) and precue versus postcue (precue trials are 1, 4, 5, 8).

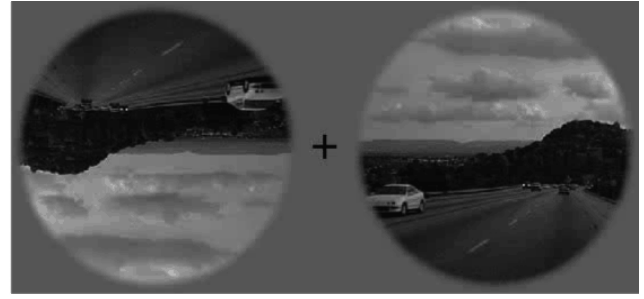
probe was smoothly inserted (by using wavelet spatial and temporal envelopes to control probe/image ratio contribution to the movie) into the local region of the natural movie identified by the automated motion-energy edge-detection procedure; see examples in Figure 5, *A* and *B*, for incongruent and congruent probes, respectively.

Stimulus presentation of static scenes and orientation discrimination task

Stimulus presentation and inversion. The overall stimulus consisted of two simultaneously presented images (duration 300 ms except for 2 observers [indicated by circle and square symbols in Figure 3] at 200 ms), one to the left and one to the right of fixation (Fig. 1*A, B*; see supplemental Movie 1). On every trial we randomly selected an image from the database and created both congruent and incongruent stimuli from the same image, but using independent noise samples for the two (randomly generated on every trial; Fig. 1*E, F*). We then presented the incongruent on the left and the congruent on the right (each centered at 7.3° from fixation; Fig. 1*A, B*), or vice versa (randomly selected on every trial). Whichever was presented to the right was mirror imaged around vertical, so that the probes were symmetrically placed with respect to fixation. On “inverted” trials, both images were flipped symmetrically around the horizontal meridian (upside down).

Spatial cueing. On “precue” trials the main stimulus just described was preceded by a spatial cue (duration 100 ms) consisting of two Gaussian blobs (matched to probe size) that colocalized with the two probes; the interval between cue and main stimulus was uniformly distributed between 150 and 300 ms. On “postcue” trials the same cue was presented but it followed the main stimulus (after the same interval detailed for precue). Observers were clearly instructed to maintain fixation at all times and stimulus duration was below typical eye-movement rate (1/3 Hz), thus minimizing any potential role of eye movements. If eye movements were occurring at all, eye movement strategy would be greatly affected by the cueing manipulation: on precue trials, observers would be tempted to foveate the two probes sequentially at the positions indicated by the cues, while no such strategy would be prompted by the scene-first-cue-second sequence of postcue trials. Had sharpening of the orientation tuning function been due to differential deployment of eye movements, this effect would be largest for the precue/postcue comparison; contrary to this prediction, there was no effect of cueing on orientation tuning (Fig. 3*A*, black symbols) excluding a potential role for eye movements in reshaping the orientation tuning function.

Response acquisition and feedback. Observers were required to select the incongruent stimulus by pressing one of two buttons to indicate either left or right of fixation. They were explicitly informed that both congruent and incongruent stimuli were presented on every trial, so by selecting the side where they thought the incongruent stimulus had appeared they also implicitly selected the side where they perceived the congruent stimulus. The task “select the incongruent stimulus” is therefore entirely equivalent to the task “select the congruent stimulus”: there is no logical distinction between the two and one can be easily converted into the other by switching response keys. Our instructions and analysis were structured around the option select the incongruent stimulus, because in pilot experiments (Neri, 2011b) using similar tasks observers



Movie 2. Upright versus inverted discrimination. Eleven sample trials from the upright versus inverted discrimination experiments used to define the unambiguous/ambiguous orientation split in Figure 3*F*.

found it more natural to look out for features that do not fit and “stick out” (i.e., are incongruent). When instructed to select the congruent stimulus, they all reported that they had perceptually converted this task into one of looking for the incongruent stimulus. In other words, no matter how task instructions are worded, observers look out for the incongruent stimulus. Their response was followed by trial-by-trial feedback (correct/incorrect) and initiated the next trial after a random delay uniformly distributed between 200 and 400 ms. Feedback was introduced for three reasons: (1) to prompt alertness on the part of observers and discourage prolonged phases of disengagement from the task, two pertinent issues when collecting large numbers of trials as in the present experiments; (2) to minimize response bias, which we have found to be more pronounced in the absence of feedback, possibly due to observer disengagement and associated choice of the same response button as a fast strategy to ignore/skip trials; and (3) to push observers into their optimal performance regime, so that interpretation of sensitivity (d') measurements would not be confounded by extraneous factors such as lack of motivation. It is conceivable that observers experienced perceptual learning, particularly in the presence of trial-by-trial feedback. We examined this possibility by comparing performance between first and second halves of data collection across each participant and found no statistically significant difference ($p = 0.64$ for upright, 0.94 for inverted). We further split data collection into 10 epochs; there was no evident trend of either increase or decrease, as quantified by computing correlation coefficients between performance and epoch number across participants with no statistically significant trend for these to be either positive or negative ($p = 0.54$ for upright, $p = 0.74$ for inverted). If learning took place, its effect was too small/inconsistent to measure given the power of our dataset. At the end of each block (100 trials) observers were provided with a summary of their overall performance (percentage of correct responses on the last block as well as across all blocks) and the total number of trials collected to that point. We tested eight naive observers (four females and four males) with different levels of experience in performing psychophysical tasks; they were paid 7 GBP/h for data collection. All conditions (upright/inverted, precue/postcue) were mixed within the same block. We collected ~11 K trials on average per observer.

Upright/inverted discrimination. Stimulus parameters for the upright versus inverted experiments (Fig. 3*F*) were identical to those detailed above, except: 1) both upright and inverted images of the same unmanipulated scene (without inserted probes) were presented on each trial but on opposite sides of fixation (see supplemental Movie 2), and observers were asked to select the upright image and (2) no feedback was provided to avoid the possibility that observers may not report the perceived orientation of the scenes, but rather the association between specific image details and the response re-enforced by feedback. We tested all eight observers that participated in the main experiment except one (Fig. 3*F*, downward-pointing triangle) who was no longer available at the time of performing these additional experiments. We collected ~2250 trials on average per observer. Because the amount of data collected by any given observer was insufficient to generate accurate estimates of upright/inverted discriminability (values on y -axis in Fig. 3*F*) across the 320 images we used in the experiments (the amount of trials allocated to each

image by each observer was only ~ 7 on average), we performed the database split into ambiguous and unambiguous images by relying on the aggregate (across observers) curve shown in Figure 3F (average number of allocated trials was ~ 50 /image).

Stimulus presentation of moving pictures and direction discrimination task

We presented congruent and incongruent stimuli (for a randomly selected movie from the database) in temporal succession (random order) on every trial, separated by a 500 ms gap. Whichever was presented second was mirror imaged around vertical to avoid repetition of identical features. We initially piloted a peripheral configuration similar to that adopted in the orientation discrimination experiments, where the two movies were presented on opposite sides of fixation; this configuration made the task impossible to perform. Having opted for the two-interval foveal configuration, spatial cueing became inapplicable/meaningless (due both to the relatively long stimulus duration and to the differential impact/meaning of the cue for first and second intervals), so no spatial cueing manipulation was adopted in these experiments. On inverted trials, both stimuli were flipped symmetrically around the horizontal meridian (upside down); on “reversed” trials, the frame order of both stimuli was reversed immediately before display. Reversed clips alter the acceleration profile (accelerations become decelerations and vice versa), but the algorithm that performed probe insertion prioritized insertion points with uniform velocity and the directional signal within the probe drifted at constant speed (see above), making it unlikely that acceleration per se would play an important role in these experiments (except for its potential impact on higher level representation). Observers were required to select the incongruent stimulus; feedback, block structure, and payment were identical to the orientation discrimination experiments. We initially tested the same eight naive observers that participated in the orientation discrimination experiments; four of them were unable to perform the direction discrimination task above chance in the absence of noise (which was the primary criterion for inclusion) and were therefore excluded from participating in the study (symbols refer to same individuals in Figs. 3, 6). All three conditions (upright, inverted, and reversed) were mixed within the same block. We collected 9.6 ± 2 K trials per observer.

Derivation of tuning functions

Orientation tuning. Each noise sample can be denoted by $n_i^{[q,z]}$: the sample added to congruent ($q = 0$) or incongruent ($q = 1$) probe on trial i to which the observer responded correctly ($z = 1$) or incorrectly ($z = 0$). The corresponding orientation tuning function \mathbf{p} was derived via application of the standard formula for combining averages from the four stimulus-response classes into a perceptual filter (Ahumada, 2002): $\mathbf{p} = \langle \mathbf{n}^{[1,1]} \rangle + \langle \mathbf{n}^{[0,0]} \rangle - \langle \mathbf{n}^{[1,0]} \rangle - \langle \mathbf{n}^{[0,1]} \rangle$ where $\langle \rangle$ is average across trials of the indexed type. Examples are shown in Figure 1G. The analysis in Figure 3, D–E, was performed on folded orientation tuning functions to compensate for the loss of measurement signal-to-noise ratio (SNR) associated with halving the dataset: we adopted the assumption that tuning functions are symmetric around their midpoint (i.e., observers are assumed to show no bias either clockwise or anticlockwise of incongruent/congruent) and averaged values on opposite sides of the midpoint (0 on x -axis in Fig. 1G). The results in Figure 3, D–E, remain qualitatively similar (albeit noisier) without folding.

Directional tuning. Directional tuning functions (Fig. 5F) were computed using the same combination rule, but were also subjected to subsequent smoothing and symmetry assumptions due to poorer SNR than the orientation tuning functions. We smoothed \mathbf{p} using a simple moving average of immediate neighboring values (box-shaped pulse of three values); we then folded it using the symmetry assumption above (no clockwise/anticlockwise bias), and further folded it under the assumption that tuning at incongruent and congruent directions is mirror symmetric (identical peaks of opposite sign). The latter symmetry assumption was clearly violated by the orientation tuning data (Fig. 1G), but we found no indication of asymmetry between incongruent and congruent directions for directional tuning [consistent with balanced motion opponency (Heeger et al., 1999) and previous measurements of similar descriptors (Neri and Levi, 2009)].

Quantitative characterization of tuning functions

Spectral centroid. The primary index of tuning sharpness adopted in this study is the spectral centroid from previous investigations (Neri, 2009, 2011a); this metric does not involve fitting and is therefore relatively stable when applied to noisy measurements. For each orientation tuning function \mathbf{p} we derived (via discrete Fourier transform; DFT) the associated power spectrum \mathbf{P} (normalized to unit sum) defined along the vector \mathbf{x} of orientation frequencies. We then computed the spectral centroid $\langle \mathbf{P}, \mathbf{x} \rangle$ where $\langle \cdot, \cdot \rangle$ is the inner product. This index may be affected by noisiness of the measured orientation tuning function: a smooth orientation tuning function would return power in the lower frequency range (small sharpness values), while a tuning function with sharp intensity transitions would return power in the higher frequency range. Measurement noise may introduce such sharp amplitude transitions, and may therefore mediate the inversion effect in Figure 3A. To check for this possibility we computed SNR as $\text{SNR} = \log(\text{RMS}/\text{RMS}^*)$, which evaluates the root-mean-square (RMS) of the tuning function \mathbf{p} against the expected RMS^* for a tuning function consisting of noise alone, i.e., originating

from a decoupled input–output process. $\text{RMS}^* = \sqrt{\frac{2N}{N^{[1]}N^{[0]}}} \sigma_n$ where

N is the total number of collected trials, $N^{[1]}$ and $N^{[0]}$ are the number of correct and incorrect trials, respectively, and σ_n is the SD of the external noise source (Neri, 2013). For a decoupled process (output response is not a function of input stimuli) the expected value of SNR is 0; furthermore, this metric scales inversely with measurement noise: the noisier the measurement, the lower the associated SNR. If measurement noise underlies the effect of inversion, we expect SNR to be lower for inverted versus upright tuning functions. Contrary to this prediction SNR was on average (across observers) twice as large for inverted functions (0.45 for inverted, 0.24 for upright). We can therefore exclude the possibility that the effect of inversion exposed by the sharpness metric was a byproduct of measurement noise; instead, it was driven by genuine structure of orientation selectivity as also corroborated by the tuning estimates returned by the fitting procedure detailed below.

Raised Gaussian fit of orientation tuning. Based on our earlier characterization of similar descriptors (Paltoglou and Neri, 2012) and on visual inspection of the data (Fig. 1G), orientation tuning functions were fitted using a raised Gaussian profile $\Phi(\sigma)^\beta + \delta$ where $\Phi(\sigma)$ is a Gaussian function of SD σ centered at the incongruent orientation, β is a positive exponent, δ is a baseline shift, and the final fitting function was constrained to take as maximum value the peak value in the data (this last constraint was introduced to avoid an additional free parameter accommodating overall scaling across the y -axis). All three parameters (σ , β , δ) were optimized to minimize mean square error between data and fit (see smooth traces in Fig. 1G for examples). Tuning sharpness is controlled by both σ and β ; these were combined via log-ratio (β/σ) to obtain a composite sharpness index (Paltoglou and Neri, 2012; plotted in inset to Fig. 3A). The above log-ratio combination rule ensures that both parameters contribute equally to the sharpness estimate regardless of the units in which they are expressed. Virtually identical results for the composite sharpness index were obtained when the baseline shift parameter (δ) was omitted from fitting (i.e., set to 0).

Metrics/fits for directional tuning functions. Smooth traces in Figure 5F were solely for visualization purposes and consisted of a sinusoid with fixed 0 phase, i.e., constrained to take a value of 0 halfway between congruent and incongruent orientations (the data were also effectively constrained to do so by symmetry averaging as detailed above), but adjustable frequency and amplitude. The latter two parameters were optimized to minimize mean square error. The metrics plotted in Figure 6 do not rely on the above-detailed fitting. The metric plotted on the x -axis in Figure 6A is normalized RMS difference: we subtracted either inverted or reversed tuning function from the corresponding upright tuning function, we computed RMS of the resulting difference, we estimated the expected value of this quantity due to measurement noise via the standard error associated with the RMS difference between the upright tuning function and its own bootstrap sample replicas, and we divided RMS by this error. The metric plotted on the y -axis in Figure 6B is the standard Pearson product-moment correlation coefficient between either in-

verted or reversed tuning function and the corresponding upright tuning function.

Computational model for orientation discrimination

The model operates directly on $s_i^{[q]}$ (orientation energy distribution within probe q on trial i), not on the actual stimulus image (we used $\rho = 2/3$ to target threshold performance; this value is smaller than the human value because the model, differently from human observers, does not suffer from internal noise). It is therefore defined in orientation space and does not account for idiosyncratic features of individual images. The front-end stage involves convolution (*taking into account the axial nature of orientation) of the input stimulus (Fig. 4D) with a bank of orientation-selective units (Fig. 4E) defined by orientation tuning function γ (Gaussian profile with SD of 10 degrees): $\mathbf{o}_i^{[q]} = s_i^{[q]} * \gamma$. We then applied a static nonlinearity Ψ to simulate neuronal conversion to firing rate (Heeger et al., 1996): $t_i^{[q]} = \Psi(\mathbf{o}_i^{[q]})$. Ψ was defined by a Gaussian cumulative distribution function with mean and SD equal to the mean and $1/16 \times$ the SD of \mathbf{o} , the vector listing all values of the convolution output across orientation (dimension of convolution), trials (i), and stimulus identity (q). This parameterization was chosen to ensure that the operating range of Ψ would effectively span the relevant output of the front-end layer. It should be noted that this nonlinear stage is necessary to simulate the lack of performance change for the model; in the absence of this nonlinearity, the residual model predicts different d' values for upright and inverted conditions. The thresholded vector was integrated using a weighting function $\mathbf{w} = 1 - \mathbf{w}_{fill}$ where 1 corresponds to the flat prior (Fig. 4F) and \mathbf{w}_{fill} defines the predicted orientation range from the top-down filling-in process (Fig. 4G). \mathbf{w}_{fill} was constrained to have equal RMS for upright and inverted conditions, but different boxcar widths around the congruent orientation; the corresponding shapes for upright (black) and inverted (red) are shown in Figure 4G. The final scalar output produced by the model in response to each stimulus (decision variable) can therefore be written as $r_i^{[q]} = \langle \Psi(s_i^{[q]} * \gamma), \mathbf{w} \rangle$ (we omit indexing into upright and inverted to simplify notation). The model responded correctly (i.e., selected the incongruent stimulus) on trial i when $r_i^{[1]} - r_i^{[0]}$ (difference between response to incongruent and congruent stimulus) was >0 , incorrectly otherwise.

Computational model for direction discrimination

The input stimulus (defined across direction of motion with $\rho = 2/3$) was initially convolved with a bank of direction-selective units defined by tuning function γ (Gaussian profile with SD of 20 degrees): $\mathbf{o}_i^{[q]} = s_i^{[q]} * \gamma$. We then applied a weighting function $\mathbf{w} = \mathbf{w}_{bottom\ up} + \mathbf{w}_{top\ down}$. Both bottom-up and top-down direction-selective priors were defined by the interaction between a directionally selective “motion” filter \mathbf{m} , consisting of one cycle of a sinusoidal modulation peaking at the incongruent direction, and a direction-insensitive “form” filter \mathbf{f} (Fig. 7A), modeled around the top-down filling-in process developed in the previous section. More specifically, \mathbf{f} weighted positively the orientation range aligned with the local region occupied by the probe (the “congruent” orientation in the previous model); because both congruent and incongruent motion directions correspond to the same congruent orientation for the moving edge, \mathbf{f} weighted the two in the same way (Fig. 7A, double-headed arrow); in other words \mathbf{f} was insensitive to motion direction, consistent with its being a form-only module. On upright and reversed trials, its shape was two cycles of a triangular function ranging between 0 (at directions orthogonal to congruent/incongruent axis) and 1 (at congruent/incongruent directions). On inverted trials, the region of maximum weighting (1) was extended by ± 45 degrees, effectively implementing similar broadening of predicted orientation range to that adopted for the orientation discrimination model (see black vs red traces in Figs. 7A, 4G). The bottom-up direction-selective weighting function was $\mathbf{w}_{bottom\ up} = \mathbf{m} \circ \mathbf{f}$ where \circ is the Frobenius (element-by-element) product; it only changed under inversion insofar as \mathbf{f} did so as detailed above, but remained otherwise unchanged. The top-down direction-selective weighting function, which implemented the action of a higher level motion module (Fig. 7D), was $\mathbf{w}_{top\ down} = \mathbf{m} \circ (1 - \mathbf{f})$, i.e., the form signal delivered by \mathbf{f} facilitated the bottom-up directional signal and concurrently inhibited the top-down signal (Fig. 7, red plus and minus symbols). Under inversion, the

form signal not only became less precise (broader \mathbf{f}) but also less intense; this effect of inversion was implemented by increasing b from a value of 3 for upright/reversed to 10 for inverted (more weight to the top-down motion module). Under movie reversal (Fig. 7C), the top-down motion module sends a weighting signal that is opposite in direction to the bottom-up module (Fig. 7F); this effect of reversal was implemented by shifting the sinusoid in \mathbf{m} by π (or equivalently inverting its sign) in the expression for $\mathbf{w}_{top\ down}$. The final scalar output produced by the model was $r_i^{[q]} = \langle s_i^{[q]} * \gamma, \mathbf{w} \rangle$ and the same rules detailed above for the orientation processing model were adopted for the motion processing model to generate a binary psychophysical response (correct/incorrect).

Coarse metrics of scene structure

To understand whether orientation ambiguity of individual scenes as assessed by human observers (Fig. 3F) was driven by relatively coarse image features, we extracted six scalar measures of overall image structure. The “vertical asymmetry” index (Fig. 3F, orange trace) was computed from image \mathbf{A} as $\log \left[\frac{\text{RMS}(\mathbf{A}_{\text{odd}})}{\text{RMS}(\mathbf{A}_{\text{even}})} \right]$ where $\mathbf{A}_{\text{odd}} = \mathbf{A} - \mathbf{A}^*$, $\mathbf{A}_{\text{even}} = \mathbf{A} + \mathbf{A}^*$ and \mathbf{A}^* is \mathbf{A} flipped upside down. The “edge richness” index (Fig. 3F, cyan trace) was obtained by applying a Sobel filter of dimension equal to $\sim 15\%$ image size and taking the mean of the filter output across the image. We plot these two indices in Figure 3 because they returned the best correlation with the image sorting returned by human observers (correlations of 0.2 ($p < 0.001$) and -0.24 ($p < 10^{-4}$), respectively). We also derived four more measures, none of which showed substantial correlation with the human classification of orientation ambiguity. One involved taking the max (as opposed to the mean) of the Sobel filter output across the image (correlation of -0.05 , $p = 0.34$). Two more measures coarsely assessed the shape of the image 2D power spectrum: one was the log-ratio of cardinal (vertical + horizontal) orientation energy to off-cardinal energy, the other one was the log-ratio of vertical to horizontal energy (correlations of ~ 0 ($p = 0.97$) and -0.17 ($p < 0.01$), respectively). The last measure we attempted was the MSE value for probe insertion (correlation of 0.14, $p < 0.01$).

Results

Observers saw two mirror images of a natural scene on opposite sides of fixation (Fig. 1A,B). The “congruent” scene (Fig. 1B) contained an oriented probe that was aligned with the local structure of the image, while the probe in the “incongruent” scene (Fig. 1A) was orthogonal to it. Observers were required to select the incongruent stimulus. This task can only be performed by integrating probe with context: if the probe is removed, observers see two identical mirror images and cannot perform above chance; similarly, if the context (i.e., natural scene) is removed, observers see two oriented gratings with no indication of which one is incongruent.

We injected orientation noise into each probe by assigning random contrast to a set of oriented gratings (Fig. 1E,F) and by adding those gratings to the probe; as a result of this stimulus manipulation, signal orientation within the resulting probes (Fig. 1C,D) was difficult to extract and observers’ responses were largely driven by fluctuations in the injected noise, allowing us to deploy established reverse correlation techniques (Ahumada, 2002; Murray, 2011) for the purpose of retrieving the orientation tuning function used by observers in performing the assigned task (Paltoglou and Neri, 2012). An example is shown by the black curve in Figure 1G: this curve peaks at the incongruent orientation (orange vertical dashed line) indicating that, as expected, observers classified orientation energy within this region as target for selection.

Orientation tuning is altered by inverting the scene upside down

We then inverted both images upside down. This manipulation is known to interfere with the semantic interpretation of the scene

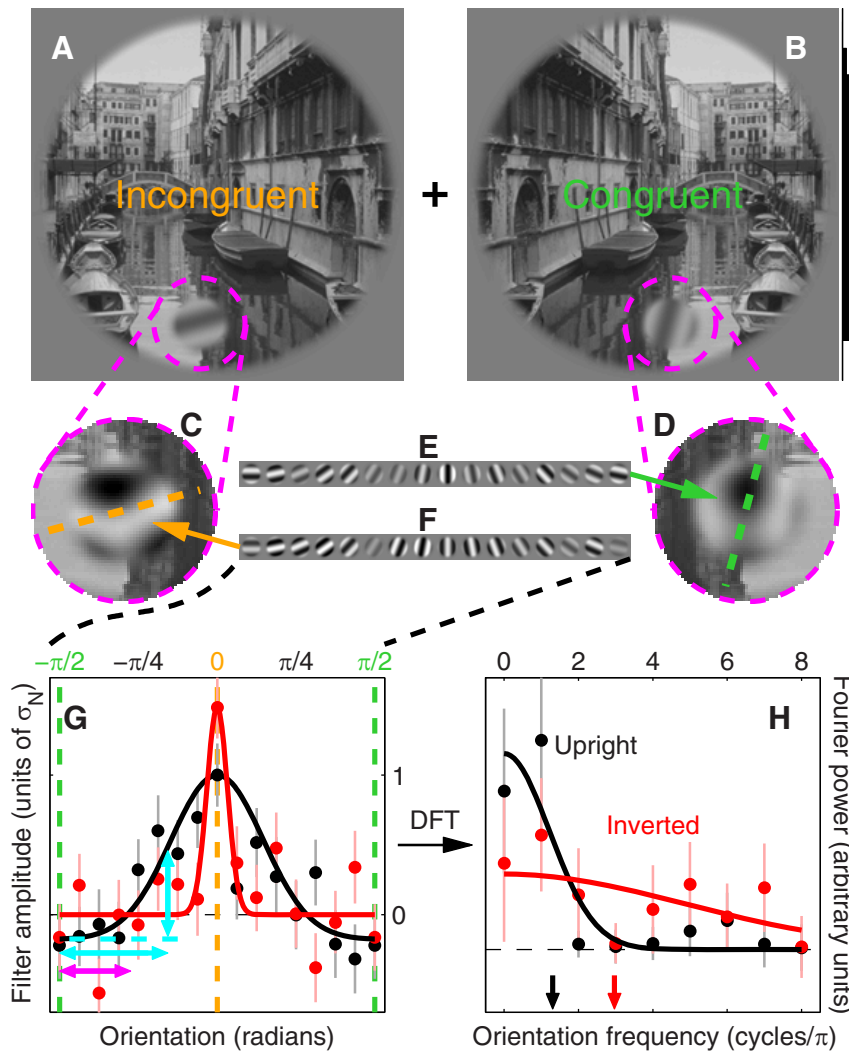


Figure 1. Orientation tuning is altered by image inversion. Congruent and incongruent images were generated by grafting an oriented probe (indicated by magenta dashed circle) that was either orthogonal to (A) or aligned with (B) the local orientation structure defined by the natural scene (histogram to the right of B shows probe location distribution along the vertical meridian across image database). Orientation noise (E, F) was added to both probes thus degrading their orientation signal (C, D). The injected orientation noise was reverse correlated with the responses generated by observers (Ahumada, 2002; Murray, 2011) to yield orientation tuning functions (G) for performing the task of identifying the incongruent stimulus (filter amplitude is expressed in units of noise SD σ_N). Black traces refer to trials on which the two scenes were in upright configuration (as shown in A and B), red traces to inverted (flipped upside down) configuration. Tuning sharpness of individual traces was estimated by the centroid (H, arrows) of the associated spectrum (Bracewell, 1965; Neri, 2011a; traces in H). Fits (smooth traces) rely on a three-parameter raised Gaussian profile (Paltoglou and Neri, 2012) in G (see Materials and Methods), and on a Gaussian profile (constrained to peak at DC on the frequency axis) with optimized amplitude and SD in H. G, H, Show aggregate data across observers (~87 K trials). Error bars indicate ± 1 SEM.

while leaving all other image properties unaffected (Yin, 1969; Valentine, 1988). For example, the upside-down image allocates contrast to detail in a manner identical to the upright image; it also contains the same set of contours/objects/edges, and probes sampled the vertical meridian symmetrically across the image dataset (Fig. 1B, histogram on the right; see also Materials and Methods). For this reason, inversion represents a powerful tool for selectively targeting the higher level representation of the image while sidestepping its low-level characteristics (Walther et al., 2009). This phenomenon is convincingly demonstrated by the Thatcher illusion (Thompson, 1980), where gross distortions of eye and mouth regions go perceptually unregistered when the face is upside down (Fig. 2A) but are readily available in the upright configuration (Fig. 2B). Similar effects can be demonstrated for

natural scenes such as those used here (Fig. 2C,D; Kelley et al., 2003; Rieger et al., 2008; Walther et al., 2009). It is especially important that the inverted manipulation has no effect on the amount of useful information delivered by the stimulus: from the viewpoint of a machine that strives to optimize performance in the assigned probe discrimination task (Geisler, 2011), upright and inverted stimuli are identical. Any difference we measure must therefore be attributed to idiosyncrasies of how the human brain operates (Valentine, 1988; Maurer et al., 2002).

Under inversion, we observed substantial changes in the associated orientation tuning function (Fig. 1G, red curve) whereby the peak at the incongruent orientation becomes sharper (Fig. 1G, compare black and red peaks aligned with orange vertical dashed line). To confirm this result using statistical tests that probed data structure at the level of individual observers, we applied to each curve a scalar metric borrowed from existing literature (Neri, 2009, 2011a) that quantified tuning sharpness via the associated spectral centroid (Bracewell, 1965) (Fig. 1H). This index of tuning sharpness was larger for the inverted as opposed to the upright configuration (Fig. 1H, compare red and black arrows), consistent with qualitative inspection of Figure 1G.

Figure 3A plots sharpness across observers and confirms the result that tuning functions from inverted scenes (y -axis) are sharper around the incongruent orientation than their counterparts from upright scenes: all red data points (one point per observer) fall above the diagonal line of equality ($p < 0.01$; all p values reported in this study are generated by paired two-tailed Wilcoxon signed rank tests, except for p values associated with correlation coefficients, which refer to a t statistic). On average across observers, the sharpness index increased by 48% with inversion (this effect does not reflect measurement noise, see Materials and Meth-

ods). Red data points in Figure 3B plot sensitivity (d') for performing the task; there was no measurable difference between inverted and upright conditions with respect to this metric (red data points scatter around equality line): observers performed the assigned task equally well in upright and inverted conditions; however, the properties of the mechanisms they used to deliver such performance differed. Both features are accounted for by the same model, which we discuss later in the article (Fig. 4).

We confirmed the above-detailed effect of image inversion on tuning via a different metric of sharpness derived from a raised-Gaussian fit to the orientation tuning function (see Materials and Methods), which we have extensively validated in a previous large-scale study of similar descriptors (Paltoglou and Neri, 2012); examples of such fits are shown by smooth lines in Figure

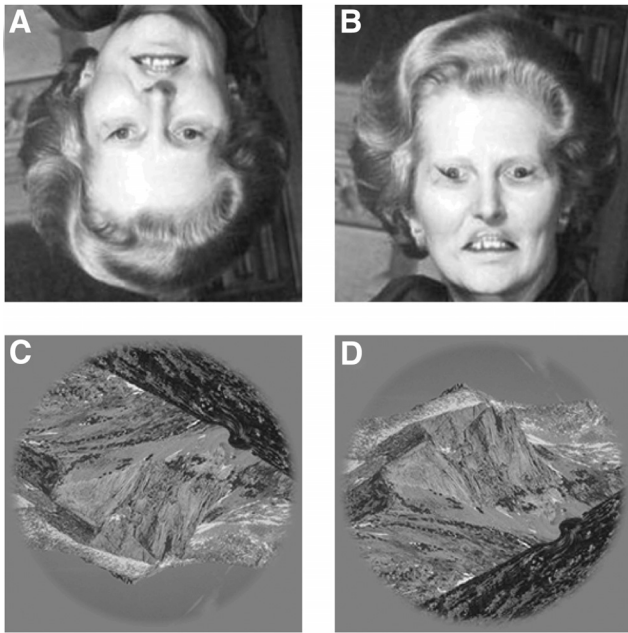


Figure 2. The face/scene inversion effect. Except for being flipped upside down, the image in **A** is identical to the image in **B** (turn this page upside down to check); it was obtained by inverting the orientation of eyes and mouth and smoothly inserting them into the original face (Thompson, 1980). The gross distortions introduced by this manipulation are clearly visible in **B**, but are perceptually inaccessible in **A**. A similar (albeit weaker) effect is demonstrated in **C** and **D** for a natural scene from our database. A local distortion of orientation has been introduced at the location of probe insertion in the experiments. The content of the image in **C** is not clearly interpretable, possibly encompassing a satellite picture; the scene in **D** is immediately recognized as depicting a mountain landscape, and the whirling pattern along the dark ridge in the bottom-right quadrant is easily interpreted as the product of image distortion.

1G. Tuning sharpness derived from fits to individual observer data is plotted in the inset to Figure 3A; this metric was significantly larger on inverted trials (red data points fall above equality line at $p < 0.04$), confirming the results obtained from spectral centroid estimates. Further indication that the two metrics captured the same quantity (i.e., tuning sharpness) comes from the strong correlation (coefficient of 0.68 at $p < 0.005$) between spectral estimates and those returned by fitting. When estimated via fitting the inversion effect returned a twofold increase in sharpness and was therefore larger than indicated by the spectral centroid measurements; however, we choose to rely on the latter metric in further analysis because it does not suffer from the many pitfalls associated with fitting algorithms (Seber and Wild, 2003; see Materials and Methods).

An intuitive interpretation of the potential significance associated with the tuning changes reported in Figure 1G may be gained by considering those functions as descriptors of the probability that the observer may report a given energy profile across orientation as being incongruent (Murray, 2011). A relatively small image distortion of approximately ± 40 degrees away from the aligned (i.e., congruent) undistorted edges (Fig. 1G, magenta double-headed arrow) will not elicit much differential response from either upright or inverted filters: the associated response will be similar to that obtained in the presence of an aligned congruent feature, prompting observers to classify it as congruent. On the other hand, a more pronounced image distortion that introduces energy beyond ± 60 degrees away from the original image structure (horizontal cyan double-headed arrow) would generate substantial response on the part of the upright filter (Fig. 1G; vertical cyan double-headed arrow); such a distortion would

lead observers to report it as incongruent, yet it would go unnoticed by the inverted filter. For the latter to prompt an incongruent classification on the part of observers, the distortion would need to be nearly orthogonal to image structure (i.e., fully incongruent). In other words, the retuning effect reported in Figure 1G is consistent with the notion that observers apply a tighter margin of tolerance around image structure to retain/exclude distortions as being congruent/incongruent; this margin is less stringent in the case of inverted images. The above notion, expressed here in qualitative terms, is implemented via computational modeling in later sections (Figs. 4, 8); it is intended as an intuitively useful (albeit inaccurate) tool for conceptualizing the main properties of the perceptual representation, and should not be confounded with response bias or confidence [we adopted two-alternative forced-choice protocols throughout (Green and Swets, 1966), making these issues largely inapplicable/irrelevant].

Effect of spatial attention is orthogonal to inversion

For clarity of exposition we have so far omitted an important detail of stimulus design: two spatial cues appeared at probe locations either before or after the stimulus (precue vs postcue configurations). Observers were thus afforded the opportunity to deploy attention to the probes on precue trials, but not on postcue trials. The question of what role (if any) is played by attention in processing natural scenes is regarded as critically relevant (Biederman, 1972; Li et al., 2002; Rousset et al., 2002), prompting us to adopt the cueing manipulation in this study.

Consistent with existing electrophysiological and psychophysical measurements using both simple (McAdams and Maunsell, 1999; Paltoglou and Neri, 2012) and complex laboratory stimuli (Biederman, 1972; Rolls et al., 2003), spatial cueing had no significant effect on tuning sharpness (black symbols in Fig. 3A scatter around equality line at $p = 0.65$) as also confirmed by the fitted metric (Fig. 3A, black symbols in inset; $p = 0.74$); however, it affected performance in the direction of increased sensitivity on precue trials (black symbols fall below diagonal equality line in Fig. 3B). The latter result demonstrates that observers did exploit the cues (which they may have potentially ignored altogether), while the former result demonstrates that the change in tuning associated with inversion (Fig. 3A, red symbols) is specific to this manipulation.

Figure 3C combines data from Figure 3, A and B, to emphasize the orthogonal effects of inversion and spatial cueing on tuning and sensitivity (x - and y -axes, respectively): inversion causes a change in tuning but no change in sensitivity (red data points fall to the right of vertical dashed line and scatter around horizontal dashed line), while the complementary pattern is associated with spatial cueing (black data points). The observed orthogonality is consistent with an earlier exploratory study from our laboratory (Neri, 2011b), despite using substantially different stimuli/measurements and probing less relevant properties of the sensory process (the earlier study reported small (marginally significant) effects with tangential relevance to sensory tuning (no characterization of feature tuning was afforded), failing to enable meaningful specification of informative computational models). Incidentally, the pattern in Figure 3C excludes any potential role for eye movements in driving the tuning changes (see Materials and Methods).

Differential analysis based on image content

Our database of natural scenes spanned a large range of image content, from mountain landscapes to skyscrapers (see Materials and Methods). For some images the effect of inversion is percep-

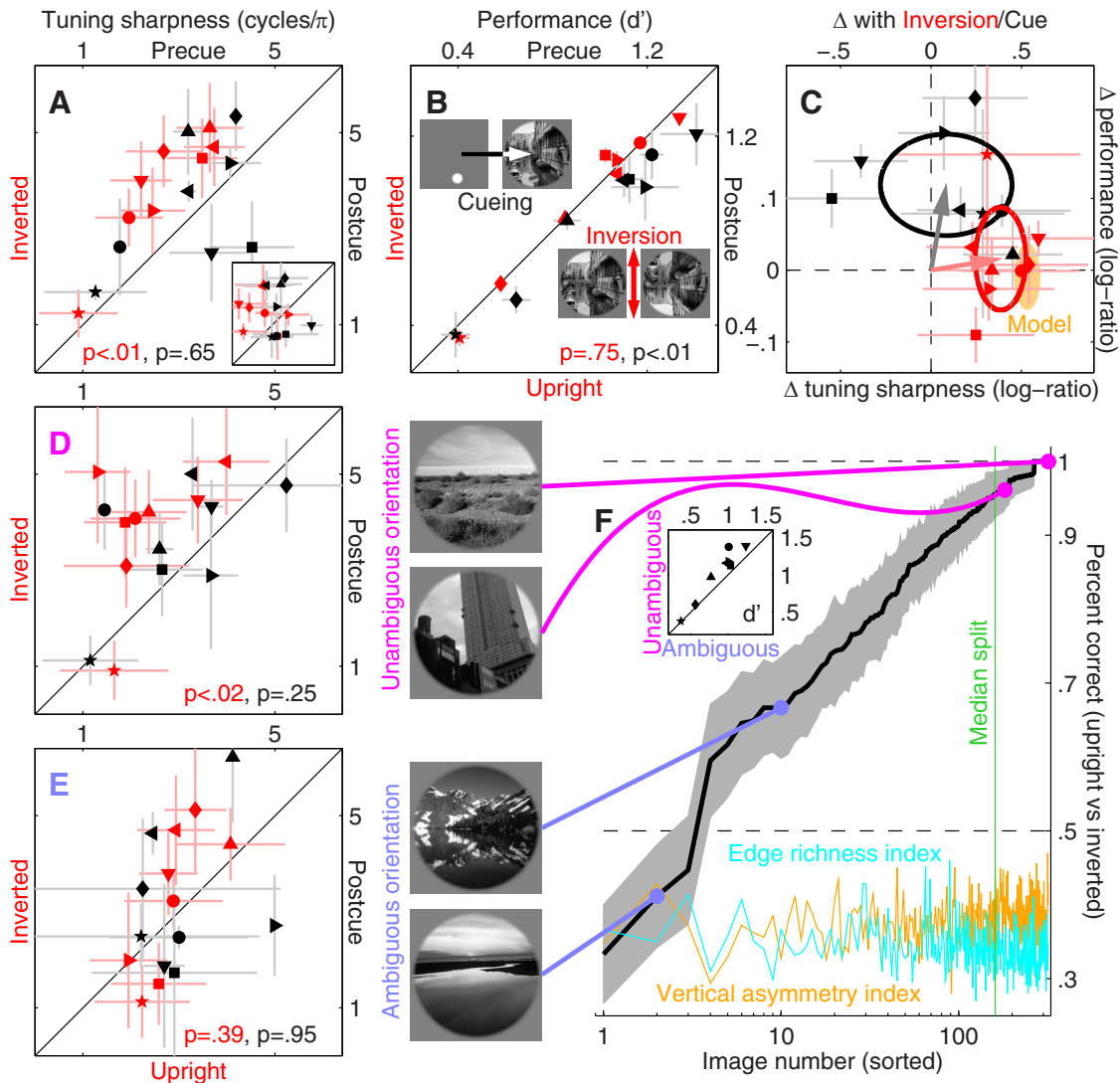


Figure 3. Tuning sharpness is affected by image inversion. Red symbols in **A** plot sharpness (Fig. 1*H*) for inverted (*y*-axis) versus upright (*x*-axis) trials, black symbols for postcue (*y*-axis) versus precue (*x*-axis) trials. **A**, Inset plots sharpness as estimated via fitting (smooth traces in Fig. 1*G*; see Materials and Methods for details). **B**, Plots sensitivity using similar conventions. **C**, Plots the log-ratio change in sensitivity (*y*-axis) versus the change in tuning sharpness (*x*-axis) for both inversion (red) and spatial cueing (black). Ovals are centered on mean values with radius matched to SD along the corresponding dimension; arrows point from origin to mean coordinates. Orange shading shows range spanned by top-down predictive model (Fig. 4, see Materials and Methods). **D**, **E**, Plot same as **A**, but after splitting the dataset into images with unambiguous versus ambiguous orientation (respectively). Images were assigned to either category based on the associated aggregate performance for discriminating upright versus inverted configuration (*y*-axis in **F**) of individual images (*x*-axis): those above the median performance value (vertical green line) were assigned to the unambiguous category, those below to the ambiguous one (see Materials and Methods for experimental details). Orange/cyan traces plot two image-based metrics (arbitrarily rescaled along *y*-axis) quantifying vertical asymmetry and edge richness of individual images (see Materials and Methods). Inset plots sensitivity for performing the probe discrimination task with unambiguous versus ambiguous scenes. **F**, Left, Examples of images from the two categories. Different symbols refer to different observers. Error bars indicate ± 1 SEM (not visible when smaller than symbol). *p* values show the result of paired two-tailed Wilcoxon signed rank tests for the comparison between values on the *y*-axis versus corresponding values on the *x*-axis.

tually evident: when the image is presented in upright configuration, it is easily recognized as upright; when in an upside-down configuration, its inversion is equally obvious. Other images, however, depict scenes that cannot be readily oriented; for these images the effect of inversion is not perceptually conspicuous. It seems reasonable to expect that the effects of inversion reported in Figure 3*A* should be limited to the former class of images, and not apply to the latter. We wished to test this prediction by restricting our dataset to either class and repeating the analysis separately.

To classify individual images as belonging to either category, in additional experiments we presented observers with both upright and inverted images of the same scene and asked them to select the upright configuration (see Materials and Methods). It

was necessary to rely on human observers to carry out this classification because there is no established algorithm/metric for assessing this type of complex image property; the metrics we attempted only correlated mildly with human classification (top two performers are plotted in Fig. 3*F*; see Materials and Methods).

Figure 3*F* plots the percentage of correct upright/inverted classifications for each image in the dataset; we split the database into equal halves below and above the median percentage correct value (respectively to the left and right of the vertical green line). Consistent with the above-detailed prediction, the effects reported in Figure 3*A* survived when the dataset was restricted to images with unambiguous orientation but disappeared when restricted to the ambiguous class (Fig. 3*D*, *E*; the inversion effect for

red data points in *D* remains significant after Bonferroni correction for multiple (2×) comparison). These results highlight not only the internal consistency of the dataset, but also the suitability of tuning sharpness as an appropriate metric for gauging the effect of image inversion.

Further evidence for the selective ability of our protocols to expose semantic effects comes from the observation that performance in the primary probe discrimination task differed for ambiguously and unambiguously classified images (points lie above the equality line in inset to Fig. 3*F* at $p < 0.01$), even though it did not differ ($p > 0.46$) between upright and inverted for either class (confirming the overall result reported in Fig. 3*B*): images that could be reliably oriented by observers, i.e., presumably those with more readily accessible semantic content, are associated with superior performance in the local orientation discrimination task. The performance metric of sensitivity (d') is therefore able to expose a robust effect of semantics across different scenes, even though as shown in Figure 3*C* it is unaffected by within-scene semantic manipulations (i.e., inversion).

Top-down predictive model of orientation tuning

To aid in data interpretation, we constructed a simple top-down predictive model (Rao and Ballard, 1999; Rauss et al., 2011; Sohoglu et al., 2012) that simulates the empirical effects reported so far. Figure 4, *A–C*, offers an intuitive view of how the model operates. We assume that the model exploits image structure around the probe to infer its expected orientation range if consistent (i.e., congruent) with surrounding context; this operation may be thought of as a filling-in process (Ramachandran and Gregory, 1991) where image structure is exploited to fill in the ambiguous region (the probe). It is also assumed that the predicted orientation range is more tightly arranged around the congruent orientation for the upright than the inverted case (Fig. 4, compare spread of light-colored lines in *A* and *C* and orientation ranges in *B*), i.e., it is assumed that the filling-in process is driven to a non-negligible extent by the higher level representation of the scene so that it is more precise in the upright configuration. This specific feature of the model is motivated by the following three concepts/findings from existing literature: (1) top-down semantic information about the “gist” of the scene may be used by the visual system to aid in local object/feature identification (Lee and Mumford, 2003; Torralba et al., 2010), possibly via a coarse-to-fine strategy associated with enhanced orientation precision (Kveraga et al., 2007); (2) single-unit recordings from primary visual cortex have demonstrated that optimally oriented gratings sharpen neuronal orientation tuning when extended outside the

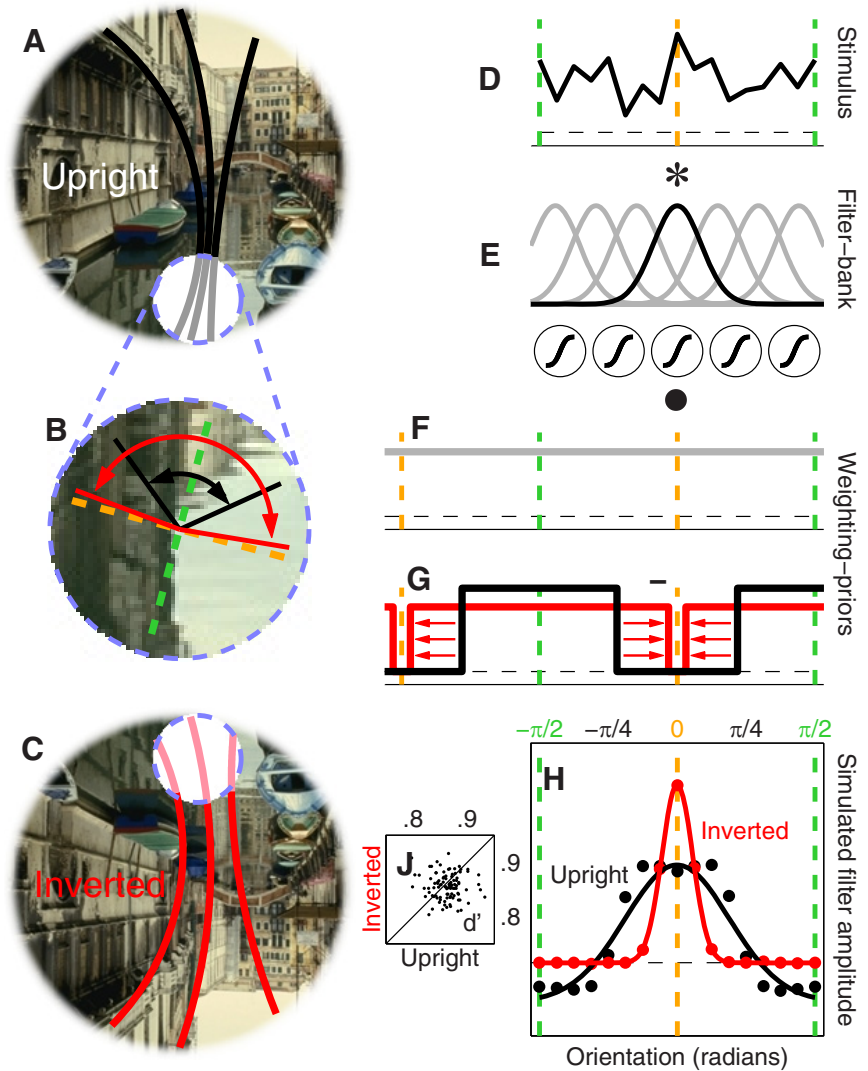


Figure 4. Top-down predictive model of orientation tuning. *A–C*, Summarize the model qualitatively. *D–G*, Show its architecture in more detail. The model exploits the edge structure of the scene to project an expected orientation range onto the probe region (indicated by dashed blue circle) similar to a filling-in process (black lines in *A* fill into gray lines). This process is less precise under inversion (*C*) so that the projected orientation range is broader for inverted (red) than upright (black) configurations (*B*). Software implementation involved filtering of the input stimulus (*D*) via a bank of oriented units (*E*); the output from this layer was subjected to a sigmoidal nonlinearity (Ⓢ symbols in *E*) and weighted by an isotropic prior (*F*) minus the predicted orientation range (*G*), thus returning the degree of incongruency associated with the stimulus. The model was challenged with the same stimuli used with human observers and on each trial selected the stimulus associated with largest incongruency as returned by the read-out rule in *F* and *G*. The associated orientation tuning functions (*H*) capture most features observed in the human data (compare with Fig. 1*G*; smooth traces were obtained via the same fitting procedure). *J*, Plots sensitivity for upright versus inverted configurations (each point refers to 1 of 100 model iterations, 10 K trials per iteration).

classical receptive field (Chen et al., 2005), indicating that contextual information is exploited by cortical circuitry to refine the orientation range spanned by congruent local edges; and (3) recent computational models successfully account for these neuronal effects via implicit encoding of statistical properties exhibited by natural scenes (Coen-Cagli et al., 2012). Those properties, in turn, are connected with semantic segmentation of the scene and object attribution of local edges (Arbelaez et al., 2012).

Under the above-detailed assumptions, software implementation of the model involves filtering the input stimulus (Fig. 4*D*) using a bank of oriented units (Fig. 4*E*); the output from the filter bank is then converted to firing rate (Heeger et al., 1996) and weighted by a read-out rule that starts out with an unoriented

isotropic assumption for the probe (Fig. 4*F*) and subtracts from it the orientation range predicted by the top-down filling-in process (Fig. 4*G*). The output of this subtraction returns the degree of incongruity of the stimulus, i.e., the amount of residual energy after the congruent prediction has been subtracted out. The model selects the stimulus with larger degree of incongruity as a target, thus simulating binary choices by the human observers (see Materials and Methods for details).

We challenged this model with the same stimuli and analysis used with human observers; the resulting simulations captured all main features of the human data (compare Fig. 4*H* with Fig. 1*G*). More specifically, the sharper peak for the inverted tuning function is a consequence of the broader predicted orientation range associated with inversion: by extending this orientation range around the congruent orientation (Fig. 4*G*) the process is effectively squeezing the tuning function around the incongruent orientation (Fig. 4*H*, orange vertical dashed line). The model also replicates the lack of inversion effect on performance (Figs. 4*J*, 3*C*, orange-shaded region), fulfilling the primary purpose of the present modeling exercise, which is to demonstrate that the measured tuning functions are compatible with the observed lack of changes in performance (i.e., the two results can be simultaneously accounted for by the same simple model). Its purpose is not to provide a detailed account of the mechanisms underlying sensory retuning; the latter effect is simply inserted into the model as given.

Although it is tempting to further specify the model to produce sensory retuning via explicit circuitry and potentially direct links to the contextual structure exhibited by the scenes in our dataset, an effort of this kind would be largely speculative: the empirical results presented here support the notion of orientation retuning, but they do not provide sufficient information to determine the exact nature of the mechanisms involved, e.g., whether actual retuning of individual units in the front-end filtering stage, differential read-out of those units between upright and inverted images, or selective modulation of gain-control circuits in the absence of unit and/or read-out retuning. The latter possibility in particular would be supported by recent experimental and theoretical work (Chen et al., 2005; Coen-Cagli et al., 2012; see also Bonds, 1989; Spratling 2010); under this view, semantic control as demonstrated here would tap onto general-purpose cortical machinery and modulate it in concomitance with nonsemantic flanker effects (e.g., via control of divisive normalization networks). Further experimental characterization will be necessary to pinpoint the relevant circuitry and guide more detailed computational efforts.

Directional tuning is altered by reversing, but not inverting, the movie

Although scenes like Figure 1*A* represent a closer approximation to natural stimuli than simpler laboratory images, they differ substantially from real-life vision in several respects, most notably the lack of movement. We wished to ask similar questions to those asked with static images but in the context of moving meaningful sequences such as those seen in films; there is no existing data that speak to this issue. For this purpose we grafted a moving probe into a film clip in either incongruent (moving in the opposite direction, Fig. 5*A*) or congruent configuration (moving along the direction defined locally by the movie, Fig. 5*B*), and asked observers to select the incongruent movie (see Materials and Methods). Similar to the orientation discrimination experiments, this task can only be performed by comparing probe and context: if either one is removed,

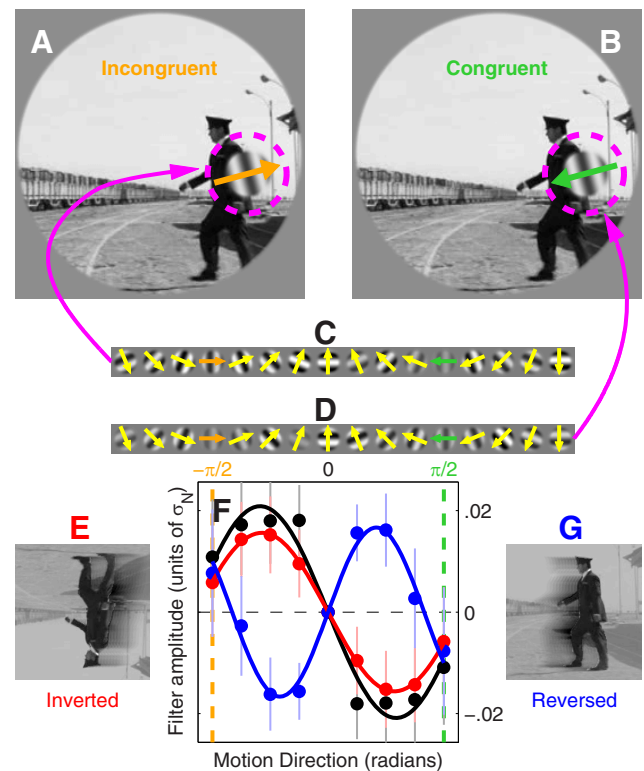


Figure 5. Directional tuning of feature detectors in natural movies. Experimental design was similar to that adopted with static pictures (Fig. 1) except probes consisted of moving gratings (arrows in *A* and *B*) embedded within movie segments in either congruent (moving along with the segment) or incongruent (moving in the direction opposite to it) configuration (*B* and *A*, respectively). Observers were asked to select the incongruent stimulus. Directional noise (*C* and *D*) was added to the probes. Directional tuning functions were derived for movie segments in their native, inverted (*E*), and reversed (*G*) configurations (black, red, and blue traces, respectively, in *F*). Error bars indicate ± 1 SEM.

congruent and incongruent stimuli become indiscriminable. Furthermore, the directional task requires observers to engage motion-selective mechanisms: if the movie is stopped, the two probes can no longer be labeled as congruent/incongruent and the task becomes unspecified.

We injected directional noise into each probe by assigning random contrast values to a set of moving gratings (Fig. 5*C,D*) and by adding those gratings to the probe. We then retrieved the directional tuning function used by observers (Fig. 5*F*, black curve); as expected it allocates positive weight to the incongruent direction (Fig. 5*F*, orange vertical dashed line) indicating that observers classified directional energy within this region as target for selection. We then applied two higher level manipulations: on some trials we inverted both clips upside down (Fig. 5*E*) and on other trials we played them backward (Fig. 5*G*). Both manipulations degrade semantic interpretation but leave other image properties unaffected (Blake and Shiffrar, 2007).

Under inversion, we observed little change in the associated directional tuning function (Fig. 5*F*, red curve). This result is perhaps unexpected given that inversion affected orientation tuning (Fig. 1*G*); later in the article we discuss an extended version of the orientation-discrimination model that accommodates this observation. When the movies were played backward, there was a substantial change in the associated directional tuning functions: no change was observed at the target (incongruent) direction (compare data points aligned with orange vertical dashed line in Fig. 5*F*), but there was a reversal of the directional

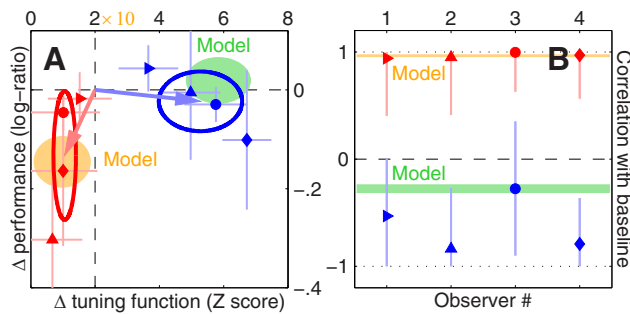


Figure 6. Inverting the movie changes performance, playing it backward changes directional tuning. Values on *y*-axis in **A** plot log-ratio sensitivity (similar to Fig. 3C) for inverted (red) or reversed (blue) d' values versus corresponding upright values; *x*-axis plots RMS of the tuning function obtained by subtracting either inverted or reversed directional tuning function from its upright counterpart (red/blue vs black traces in Fig. 5F) in units of expected deviation from measurement noise (see Materials and Methods). **B**, Plots point-by-point correlation between inverted/reversed (red/blue) and upright tuning functions across observers (*x*-axis). Different symbols refer to different observers preserving the same symbol-to-observer labeling adopted in Figure 3. Error bars indicate ± 1 SEM. Shaded orange and green areas show ranges (mean \pm SD across simulations) spanned by the model outlined in Figure 7; *x* scaling in **A** has been stretched by one order of magnitude ($\times 10$) for model to accommodate the inevitably higher precision afforded by the simulations (model has no intrinsic noise).

tuning function away from the incongruent direction (see region near $\pm \pi/4$).

To confirm this result across observers, we applied a scalar metric to inverted and reversed curves (Fig. 5F, red/blue traces) aimed at quantifying their departure from the upright curve (black trace). This metric (normalized RMS difference between two tuning functions) is plotted on the *x*-axis in Figure 6A; it is substantially different from measurement noise when it is greater than the value indicated by the vertical dashed line (see Materials and Methods). As expected from qualitative inspection of Figure 5F, all observers presented a substantial change in the tuning function associated with playing the movies backward, but there was no such effect under inversion (blue data points in Fig. 6A fall to the right of the vertical dashed line, red data points fall to the left).

Directional discrimination is reduced by inverting, but not reversing, the movie

There was no measurable change in sensitivity associated with the reversed configuration (blue data points in Fig. 6A scatter around the horizontal dashed line), i.e., observers were equally good at performing the probe discrimination task when movies were played backward as opposed to forward; however, there was a substantial drop in sensitivity under inversion (red data points in Fig. 6A fall below the horizontal dashed line). Disruption of semantic information may therefore impact both sensory tuning and sensitivity, depending on the probed perceptual attribute and on the applied semantic manipulation.

On two separate instances across our dataset and in two substantially different contexts, the metrics of sensitivity and sensory tuning exposed different manipulations in an orthogonal manner: inversion/reversal caused a change in orientation/direction tuning but not sensitivity, while cueing/inversion caused a change in sensitivity but not orientation/direction tuning (Figs. 3C, 6A). This result has important methodological consequences in that it indicates that the two metrics considered above are complementary tools (Nagai et al., 2008; Dobres and Seitz, 2010; Neri, 2011b) and should therefore be used in conjunction with each other if the experimenter is to obtain a fuller picture of the

underlying sensory process. It is potentially relevant in this context that some studies of natural scene discrimination have reported measurable inversion effects with relation to reaction times, without any concomitant change in performance (Rieger et al., 2008).

Tuning function for directional discrimination changes shape, not amplitude

The metric plotted on the *x*-axis in Figure 6A provides an indication of how much overall change occurred between the baseline upright tuning function and the tuning function under inverted/reversed configuration, but it provides no information about the specific way in which the tuning function changed. For example, a change in overall amplitude (rescaling along *y*-axis) would result in large values for this metric, even though there may be no associated change in shape. To capture specific shape changes we considered the point-by-point correlation between tuning functions (this metric is complementary to the metric plotted in Fig. 6A because it does not explicitly carry information about overall amplitude changes).

Correlation values were all positive for the inverted manipulation (red data points in Fig. 6B fall above horizontal dashed line), confirming that there was little change in shape under inversion. In contrast, the values associated with the reversed manipulation were all negative (blue data points in Fig. 6B fall below horizontal dashed line), demonstrating that the way in which the tuning function changed shape specifically involved a reversal of its tuning profile, consistent with qualitative inspection of Figure 5F.

Top-down predictive model of directional tuning

To aid in the interpretation of the directional data, we extended the model used to simulate orientation tuning functions (Fig. 4). The orientation tuning model was used to construct the form-specific higher level module within the new model (Fig. 7A); this module exploits image structure to infer the expected orientation range within the probe (Fig. 7A, red vs black dashed lines) and controls the competition between bottom-up and top-down motion-selective signals (Fig. 7B, D, respectively). More specifically, the higher level form module defines a nondirectional axis (Fig. 7A, double-headed arrow) orthogonal to the expected orientation structure within the probe (Fig. 7A, dashed lines). For directional signals along this axis, the higher level module prioritizes bottom-up input (Fig. 7B, + sign) and silences the top-down signal (Fig. 7B, - sign). This facilitatory/inhibitory connectivity was motivated by the notion that input directional signals consistent with locally expected edge structure would be more reliable than those away from it; for the latter, the form module gradually allows the higher level motion module to “take over” and deploy its own feedback to the local motion sensor within the probe (see Materials and Methods).

When the movie is played in its original configuration (Fig. 7B), the directional structure expected by the higher level motion module is consistent with the signal delivered by the stimulus: the resulting tuning function therefore matches the directional tuning function of both top-down and bottom-up modules (Fig. 7E). When the movie is played backward (Fig. 7C), however, a conflict arises between the two modules: the higher level module is still signaling that the train conductor in Figure 5A must be walking forward as expected from prior experience during natural vision, but the bottom-up signal goes in the opposite direction. The resulting tuning function is again a mixture of top-down and bottom-up signals; because they no longer match, the shape of the tuning function is distorted away from its original

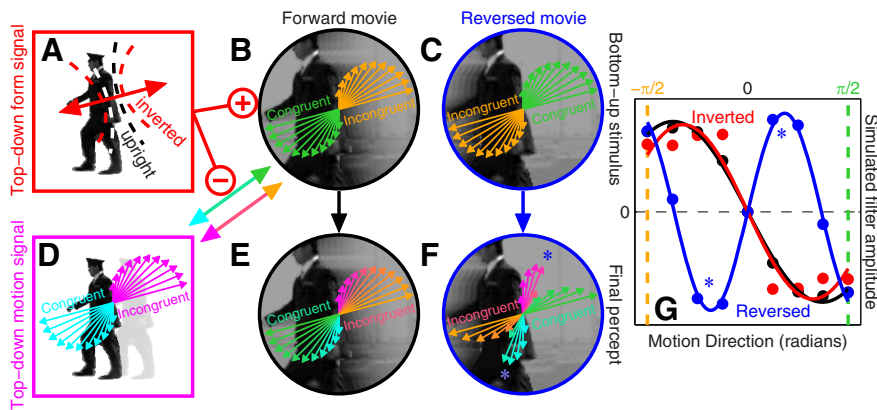


Figure 7. Top-down predictive model of directional tuning. This model incorporates the orientation tuning model (Fig. 4) as its form module (A): the projected orientation range (dashed lines in A) is less precise under inversion (red dashed lines) and is aligned with the local edge content. This module sends a facilitatory signal to the bottom-up motion module (B) and an inhibitory signal to the top-down motion module (D), thus controlling the balance between bottom-up and top-down directional processing within the system. The form module itself is not directionally selective in that its modulation is symmetric with respect to the motion directions defined by the edge (double-headed arrow in A). The relative contribution of bottom-up and top-down directional filtering to the final percept at different directions is indicated by different colors: green/orange for bottom-up, cyan/magenta for top-down. Green bottom-up units respond to the congruent direction (i.e., the one defined by the low-level properties of the stimulus) and they are coupled with the top-down cyan units, which respond to the direction congruent with the semantic content of the scene (i.e., the one defined by the higher level properties of the stimulus); orange bottom-up units are similarly coupled with magenta top-down units (coupling is represented by multicolored double-headed arrows). When the movie is played in its normal configuration (forward), bottom-up (B) and top-down (D) representations of congruent and incongruent directions agree with each other, so that the final percept (E) matches both. When the movie is played backward (reversed), the two representations are opposite to each other (C vs D) and the final percept (F) presents distortions from the veridical direction in the stimulus (C) introduced by the top-down motion module; the distortions are more pronounced away from congruent/incongruent directions because the form module (A) allows progressively more input from the top-down motion module within that region of directional space. When the model is challenged with the same stimuli and analysis used with human observers it is able to capture most features of the human data (compare G with Fig. 5F; see also shaded regions in Fig. 6). Quantitatively simulated distortions in G are linked to those diagrammed in F using asterisks. G shows average across 100 iterations (9 K trials each).

configuration (Fig. 7F). We challenged a software implementation of this model with the same stimuli and analysis used with the human observers; the model is able to capture all main features in the human data (compare Figs. 7G, 5F), including relevant effects on performance (Fig. 6A, shaded regions).

The interaction between form and motion as implemented here is reminiscent of existing models where form detectors bias the processing performed by motion detectors (Burr and Ross, 2002). In those models, motion detection is facilitated along the direction of form streaks (Geisler, 1999); similarly here, bottom-up motion processing is facilitated along the edge structure represented by the higher level form module. However, it is critically important that the distinction between low-level detection and higher level interpretation is preserved in juxtaposing the two frameworks, because the bottom-up low-level motion detector in our module is facilitated along the direction orthogonal, not parallel (Geisler, 1999), to the edge structure retrieved by the form module. The similarities between these two modeling architectures end where they begin; it is therefore more cautious to view them as largely independent.

Discussion

Mainstream models of early sensory processing do not incorporate semantic control (Heeger et al., 1996): they assume that feature detectors are relatively static in relaying stimulus-driven signals to later processing stages (Carandini et al., 2005). This state of affairs should not be interpreted to mean that current thinking is oblivious to the notion that early sensors may be shaped by feedback from higher level modules; on the contrary, several theoretical accounts of early vision have postulated the

existence of top-down instruction onto front-end feature detectors (Lamme and Roelfsema, 2000; Bullier, 2001), alongside electrophysiological evidence that the properties of neurons in early sensory cortex are affected by higher level cognitive factors (Lamme and Roelfsema, 2000; McAdams and Reid, 2005). The difficulty stems from the limitation that, when such recurrent processing is hypothesized, it cannot be incorporated into actual computational models of the perceptual process because it is simply not known how top-down modulatory effects may manifest themselves at the level of early sensory tuning for elementary image features (e.g., orientation; Lee and Mumford, 2003 for specifically relevant theoretical proposals).

The results reported in this study provide clear evidence that feature detectors for processing orientation (Fig. 1) and motion direction (Fig. 5) do not operate in isolation. Rather they are affected by natural context at its highest level of representation; the associated behavioral effects have measurable impact on both sensory tuning (Figs. 1G, 5F) and discrimination performance (Figs. 3F, inset; 6A). Most importantly, these results afford a sufficient level of specification to allow for explicit implementation of higher level effects in the form of phenomenological and computational models (see below). At the same time, our

conclusion that semantic interpretation of natural scenes controls sensory tuning of feature detectors is based on empirical observation, independent of whether the models proposed in Figures 4 and 7 are applicable or not. If we accept the widely held notion that the inverted/reversed manipulations selectively disrupt higher level representations (Valentine, 1988; Blake and Shiffrar, 2007), we must conclude that higher level representations control the shape of the tuning functions to a measurable extent.

It is important to emphasize that observers were never asked to carry out higher level object recognition: the task involved orientation discrimination of Gabor wavelets and their immediately adjacent image context, without any requirement for explicit attribution of semantic content. Although we show that this early visual process is shaped by higher level image representation (see above), our results cannot be used to draw conclusions in the opposite direction: we do not provide evidence that the observed changes in the low-level process have a measurable impact on the operation of the higher level process. The latter is manipulated via inversion/time reversal; it remains to be verified whether this experimental manipulation, artificially imposed by the experimenter, is also coupled with modifications controlled by the low-level process as characterized here.

To illustrate the general significance of our results without committing to specific modeling schemes, we refer back to the phenomenological inversion effects illustrated in Figure 2. We can compute the energy distribution introduced by local distortions across the dimension of orientation (Fig. 8E, F, blue traces)

and filter this energy using the two tuned operators derived from human data for upright and inverted configurations (Fig. 8E,F, black and red traces); the amount of distortion energy returned by these operators for the natural images in Figure 8, A–D, is indicated by red tint. Because the upright operator matches the energy distribution introduced by the distortion more closely than the inverted operator (compare match of black/red traces with blue trace in Fig. 8E,F), distortion energy is more visible in the upright images (Fig. 8, compare red tint in B, D vs A, C), which may explain why it is more readily available to our perceptual system (Thompson, 1980). The above-detailed account of the Margaret Thatcher illusion (and inversion effects more generally) is substantially different from, and more detailed than, current explanations of this phenomenon (Bartlett and Searcy, 1993). More specifically, current theories postulate that inversion disrupts the ability to integrate local features (Valentine, 1988; Maurer et al., 2002); our results indicate that, instead, inversion disrupts the ability to fine-tune local processing of those features (see also Sekuler et al., 2004; Gold et al., 2012).

Although more speculative, further interpretation of our results may be guided by the computational models and associated simulations presented in Figures 4 and 7. Both models adopt a top-down predictive architecture (Rao and Ballard, 1999; Rauss et al., 2011) whereby image semantics modulate low-level sensory tuning. These models are not intended as veritable representations of the sensory process, nor as exclusive accounts of this process. Their purpose is to offer a demonstration, among many other potential candidates, that the notion of predictive coding (Rao and Ballard, 1999; Rauss et al., 2011; Sohoglu et al., 2012) is consistent with our measurements, also in connection with recent electrophysiological (Chen et al., 2005) and computational work (Coen-Cagli et al., 2012). In line with their illustrative nature, these relatively simple models do not incorporate full-scale image interpretation in the sense of extracting explicit semantic content from image structure as attempted by some computer vision algorithms (Ullman, 1996; Torralba et al., 2010). An endeavor of the latter kind, although undoubtedly necessary for a complete understanding of the relevant issues, is beyond the scope of the current study.

The term “feature detector” as adopted here refers to a perceptual entity of conceptual validity independently of specific interpretations based on potential neural substrates (Spillmann, 2006; Morgan, 2011); similarly, we use the terms low-level and higher level in a perceptual sense as currently understood and conceptualized in the cognitive literature (Ullman, 1996). It is

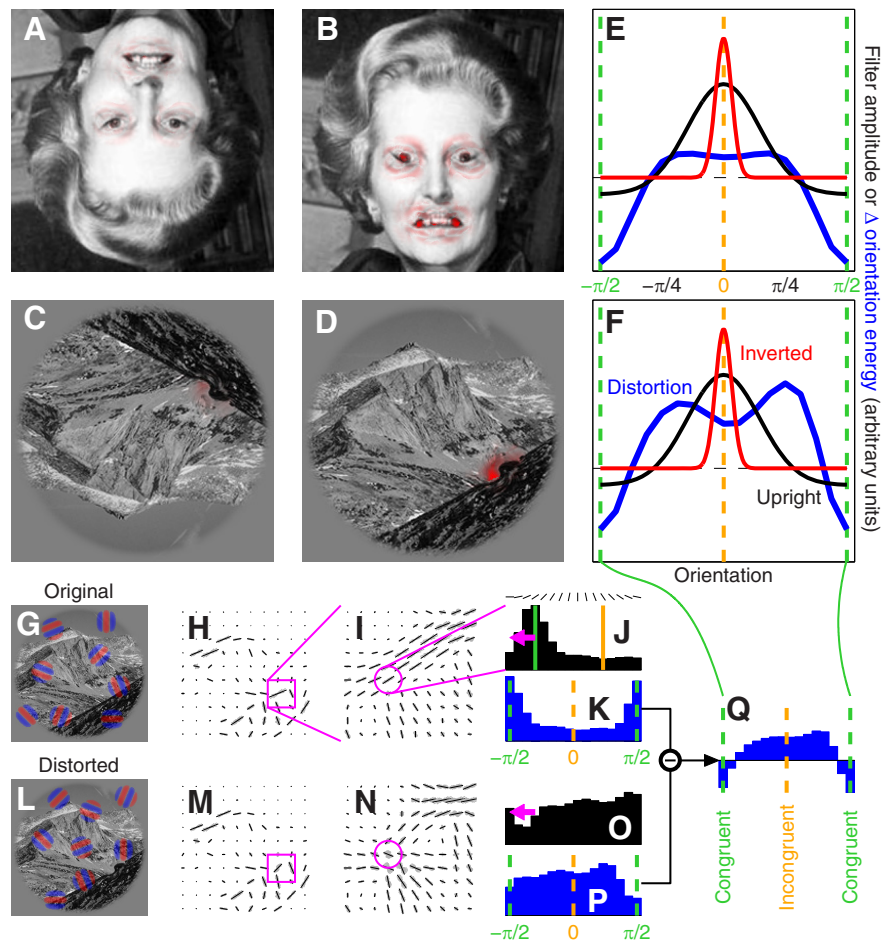


Figure 8. The face/scene inversion effect explained. **A–D**, They are the same images as shown in Figure 2, **A–D**, with superimposed output (red tint) from inverted (**A, C**) and upright (**B, D**) human orientation-selective filters (red and black traces in **E** and **F**) in response to the orientation energy introduced by local distortions in the original images (blue traces in **E** and **F**). The latter was computed by applying (via convolution) a set of quadrature-pair oriented Gabor patches (see red–blue examples in **G**) and taking the sum of their squared outputs (oriented energy), resulting in the orientation-energy map shown in **H** where orientation energy at sample locations in the image is plotted using oriented segments of length matched to the energy content for the corresponding orientation (maximum-energy orientation is plotted in black, others in gray). **I, J**, Resamples local square region of image where distortion was applied (indicated by magenta square region in **H**). **J**, Plots energy distribution for a specific location (indicated by magenta circle in **I**). The same analysis was applied to the distorted image (**L**) to obtain corresponding orientation maps (**M** and **N**) and energy distribution (**O**). The orientation associated with maximum energy content at each specific location in the original (undistorted) image (indicated by vertical green line in **J**) was labeled as congruent, and both undistorted and distorted distributions were realigned by the same amount (indicated by magenta arrows in **J** and **O**) so that the congruent orientation would correspond to $-\pi/2$ on the x -axis (and the incongruent orientation to 0); example realigned distributions are shown in **K** and **P**. We then subtracted the original distribution (**K**) from the distorted distribution (**P**) to obtain the change in energy distribution introduced by the distortion (**Q**). Blue traces in **E** and **F** show averages across entire image; orientation filters (from fits to human data in Fig. 1G) were applied not to the average but separately to each pixel in the image to estimate perceived distortion energy (red tint in **A–D**) across the scene.

nevertheless reasonable to speculate that observers used both primary sensory structures (e.g., V1) and distributed semantic neural resources (Walther et al., 2009; Huth et al., 2012; Ban et al., 2013; associative cortex) to carry out the discrimination tasks in our experiments: we specifically designed probes that matched the properties of early visual areas in their spatial extent and characteristics (Hubel, 1963; Bolz and Gilbert, 1986), so that our experiments would be directly relevant to low-level structures in the visual processing hierarchy. At the same time, probes were embedded within natural scenes and the assigned task enforced integration of both stimulus components, extending the relevance of our experiments to neural circuitry where semantic content may be represented (Rossion and Gauthier, 2002; Huth et al.,

2012; Preston et al., 2013). Studies comparing neuronal responses in early visual cortex under stimulation from artificial versus natural statistics have reported important differences (David et al., 2004; Haslinger et al., 2012) (including orientation tuning changes; Smyth et al., 2003) that may be interpreted as selective adaptations to the conditions of natural vision (Felsen and Dan, 2005).

Previous psychophysical research has demonstrated the perceptual relevance of specific properties of natural scenes, such as amplitude spectrum (Párraga et al., 2000) or edge distribution (Bex et al., 2009); however, no prior study has addressed the potential impact of semantic content on low-level feature extraction. For example, previous research has shown that edge density affects image discrimination for natural scenes (Bex et al., 2009). This result is relevant to the first level of analysis we considered in the introduction (i.e., role of natural statistics without meaningful content), but is distinct from the semantic level targeted by the present study: edge structure around the probe is unaffected by image inversion in our experiments, as are all other image characteristics addressed by previous research (Webster and Miyahara, 1997; Párraga et al., 2000; Bex et al., 2009). By demonstrating that semantic representations shape perceptual analysis of elementary image features, our results indicate that the sensory process must be viewed as a multidirectional system where modules at both ends of the spectrum, from higher to lower level representations, interact with and inform each other in a cohesive manner within a highly integrated architecture (Lamme and Roelfsema, 2000; Bullier, 2001; Lee and Mumford, 2003).

References

- Ahumada AJ Jr (2002) Classification image weights and internal noise level estimation. *J Vis* 2(1):121–131. [CrossRef Medline](#)
- Allman J, Miezin F, McGuinness E (1985) Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev Neurosci* 8:407–430. [CrossRef Medline](#)
- Arbelaez P, Hariharan B, Gu C, Gupta S, Bourdev L, Malik J (2012) Semantic segmentation using regions and parts. *CVPR, IEEE Conference*, June 2012.
- Ban H, Yamamoto H, Hanakawa T, Urayama S, Aso T, Fukuyama H, Ejima Y (2013) Topographic representation of an occluded object and the effects of spatiotemporal context in human early visual areas. *J Neurosci* 33:16992–17007. [CrossRef Medline](#)
- Bar M (2004) Visual objects in context. *Nat Rev Neurosci* 5:617–629. [CrossRef Medline](#)
- Bartlett JC, Searcy J (1993) Inversion and configuration of faces. *Cogn Psychol* 25:281–316. [CrossRef Medline](#)
- Bex PJ, Solomon SG, Dakin SC (2009) Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure. *J Vis* 9(10):1–19. [CrossRef Medline](#)
- Biederman I (1972) Perceiving real-world scenes. *Science* 177:77–80. [CrossRef Medline](#)
- Blake R, Shiffrar M (2007) Perception of human motion. *Annu Rev Psychol* 58:47–73. [CrossRef Medline](#)
- Bolz J, Gilbert CD (1986) Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature* 320:362–365. [CrossRef Medline](#)
- Bonds AB (1989) Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Vis Neurosci* 2:41–55. [CrossRef Medline](#)
- Bracewell RN (1965) *The Fourier transform and its applications*. New York: McGraw-Hill.
- Bullier J (2001) Feedback connections and conscious vision. *Trends Cogn Sci* 5:369–370. [CrossRef Medline](#)
- Burr DC, Ross J (2002) Direct evidence that “speedlines” influence motion mechanisms. *J Neurosci* 22:8661–8664. [Medline](#)
- Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC (2005) Do we know what the early visual system does? *J Neurosci* 25:10577–10597. [CrossRef Medline](#)
- Chen G, Dan Y, Li CY (2005) Stimulation of non-classical receptive field enhances orientation selectivity in the cat. *J Physiol* 564:233–243. [CrossRef Medline](#)
- Coen-Cagli R, Dayan P, Schwartz O (2012) Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Comput Biol* 8:e1002405. [CrossRef Medline](#)
- David SV, Vinje WE, Gallant JL (2004) Natural stimulus statistics alter the receptive field structure of v1 neurons. *J Neurosci* 24:6991–7006. [CrossRef Medline](#)
- Dobres J, Seitz AR (2010) Perceptual learning of oriented gratings as revealed by classification images. *J Vis* 10(13):8. [CrossRef Medline](#)
- Felsen G, Dan Y (2005) A natural approach to studying vision. *Nat Neurosci* 8:1643–1646. [CrossRef Medline](#)
- Felsen G, Touryan J, Han F, Dan Y (2005) Cortical sensitivity to visual features in natural scenes. *PLoS Biol* 3:e342. [CrossRef Medline](#)
- Geisler WS (1999) Motion streaks provide a spatial code for motion direction. *Nature* 400:65–69. [CrossRef Medline](#)
- Geisler WS (2008) Visual perception and the statistical properties of natural scenes. *Annu Rev Psychol* 59:167–192. [CrossRef Medline](#)
- Geisler WS (2011) Contributions of ideal observer theory to vision research. *Vision Res* 51:771–781. [CrossRef Medline](#)
- Gerhard HE, Wichmann FA, Bethge M (2013) How sensitive is the human visual system to the local statistics of natural images? *PLoS Comput Biol* 9:e1002873. [CrossRef Medline](#)
- Gold JM, Mundy PJ, Tjan BS (2012) The perception of a face is no more than the sum of its parts. *Psychol Sci* 23:427–434. [CrossRef Medline](#)
- Green DM, Swets JA (1966) *Signal detection theory and psychophysics*. New York: Wiley.
- Haslinger R, Pipa G, Lima B, Singer W, Brown EN, Neuenschwander S (2012) Context matters: the illusive simplicity of macaque V1 receptive fields. *PLoS One* 7:e39699. [CrossRef Medline](#)
- Heeger DJ, Simoncelli EP, Movshon JA (1996) Computational models of cortical visual processing. *Proc Natl Acad Sci U S A* 93:623–627. [CrossRef Medline](#)
- Heeger DJ, Boynton GM, Demb JB, Seidemann E, Newsome WT (1999) Motion opponency in visual cortex. *J Neurosci* 19:7162–7174. [Medline](#)
- Hubel DH (1963) The visual cortex of the brain. *Sci Am* 209:54–62. [CrossRef Medline](#)
- Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76:1210–1224. [CrossRef Medline](#)
- Kelley TA, Chun MM, Chua KP (2003) Effects of scene inversion on change detection of targets matched for visual salience. *J Vis* 3(1):1–5. [CrossRef Medline](#)
- Kveraga K, Boshyan J, Bar M (2007) Magnocellular projections as the trigger of top-down facilitation in recognition. *J Neurosci* 27:13232–13240. [CrossRef Medline](#)
- Lamme VA, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci* 23:571–579. [CrossRef Medline](#)
- Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis* 20:1434–1448. [CrossRef Medline](#)
- Li FF, VanRullen R, Koch C, Perona P (2002) Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci U S A* 99:9596–9601. [CrossRef Medline](#)
- Maurer D, Grand RL, Mondloch CJ (2002) The many faces of configural processing. *Trends Cogn Sci* 6:255–260. [CrossRef Medline](#)
- McAdams CJ, Maunsell JH (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J Neurosci* 19:431–441. [Medline](#)
- McAdams CJ, Reid RC (2005) Attention modulates the responses of simple cells in monkey primary visual cortex. *J Neurosci* 25:11023–11033. [CrossRef Medline](#)
- Morgan MJ (2011) Features and the “primal sketch.” *Vision Res* 51:738–753. [CrossRef Medline](#)
- Murray RF (2011) Classification images: a review. *J Vis* 11(5): pii:2. [CrossRef Medline](#)
- Nagai M, Bennett PJ, Sekuler AB (2008) Exploration of vertical bias in per-

- ceptual completion of illusory contours: threshold measures and response classification. *J Vis* 8(7):25.1–17. [CrossRef Medline](#)
- Neri P (2009) Nonlinear characterization of a simple process in human vision. *J Vis* 9(12):1.1–29. [CrossRef Medline](#)
- Neri P (2011a) Coarse to fine dynamics of monocular and binocular processing in human pattern vision. *Proc Natl Acad Sci U S A* 108:10726–10731. [CrossRef Medline](#)
- Neri P (2011b) Global properties of natural scenes shape local properties of human edge detectors. *Front Psychol* 2:172. [Medline](#)
- Neri P (2013) The statistical distribution of noisy transmission in human sensors. *J Neural Eng* 10:016014. [CrossRef Medline](#)
- Neri P, Levi D (2009) Surround motion silences signals from same-direction motion. *J Neurophysiol* 102:2594–2602. [CrossRef Medline](#)
- Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comp Vis* 42:145–175. [CrossRef](#)
- Paltoglou AE, Neri P (2012) Attentional control of sensory tuning in human visual perception. *J Neurophysiol* 107:1260–1274. [CrossRef Medline](#)
- Párraga CA, Troscianko T, Tolhurst DJ (2000) The human visual system is optimised for processing the spatial information in natural visual images. *Curr Biol* 10:35–38. [CrossRef Medline](#)
- Piotrowski LN, Campbell FW (1982) A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception* 11:337–346. [CrossRef Medline](#)
- Preston TJ, Guo F, Das K, Giesbrecht B, Eckstein MP (2013) Neural representations of contextual guidance in visual search of real-world scenes. *J Neurosci* 33:7846–7855. [CrossRef Medline](#)
- Ramachandran VS, Gregory RL (1991) Perceptual filling in of artificially induced scotomas in human vision. *Nature* 350:699–702. [CrossRef Medline](#)
- Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2:79–87. [CrossRef Medline](#)
- Rauss K, Schwartz S, Pourtois G (2011) Top-down effects on early visual processing in humans: a predictive coding framework. *Neurosci Biobehav Rev* 35:1237–1253. [CrossRef Medline](#)
- Rieger JW, Köchy N, Schalk F, Grüschow M, Heinze HJ (2008) Speed limits: orientation and semantic context interactions constrain natural scene discrimination dynamics. *J Exp Psychol Hum Percept Perform* 34:56–76. [CrossRef Medline](#)
- Rolls ET, Aggelopoulos NC, Zheng F (2003) The receptive fields of inferior temporal cortex neurons in natural scenes. *J Neurosci* 23:339–348. [Medline](#)
- Rossion B, Gauthier I (2002) How does the brain process upright and inverted faces? *Behav Cogn Neurosci Rev* 1:63–75. [CrossRef Medline](#)
- Rousselet GA, Fabre-Thorpe M, Thorpe SJ (2002) Parallel processing in high-level categorization of natural images. *Nat Neurosci* 5:629–630. [Medline](#)
- Ruderman DL, Bialek W (1994) Statistics of natural images: scaling in the woods. *Phys Rev Lett* 73:814–817. [CrossRef Medline](#)
- Rust NC, Movshon JA (2005) In praise of artifice. *Nat Neurosci* 8:1647–1650. [CrossRef Medline](#)
- Schwartz O, Hsu A, Dayan P (2007) Space and time in visual context. *Nat Rev Neurosci* 8:522–535. [CrossRef Medline](#)
- Seber GAF, Wild CJ (2003) *Nonlinear regression*. New York: Wiley.
- Sekuler AB, Gaspar CM, Gold JM, Bennett PJ (2004) Inversion leads to quantitative, not qualitative, changes in face processing. *Curr Biol* 14:391–396. [CrossRef Medline](#)
- Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. *Annu Rev Neurosci* 24:1193–1216. [CrossRef Medline](#)
- Smyth D, Willmore B, Baker GE, Thompson ID, Tolhurst DJ (2003) The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *J Neurosci* 23:4746–4759. [Medline](#)
- Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. *J Neurosci* 32:8443–8453. [CrossRef Medline](#)
- Spillmann L (2006) From perceptive fields to Gestalt. *Prog Brain Res* 155:67–92. [Medline](#)
- Spratling MW (2010) Predictive coding as a model of response properties in cortical area V1. *J Neurosci* 30:3531–3543. [CrossRef Medline](#)
- Thompson P (1980) Margaret Thatcher: a new illusion. *Perception* 9:483–484. [CrossRef Medline](#)
- Torralba A, Oliva A (2003) Statistics of natural image categories. *Network* 14:391–412. [CrossRef Medline](#)
- Torralba A, Murphy KP, Freeman WT (2010) Using the forest to see the trees: exploiting context for visual object detection and localization. *Commun ACM* 53:107–114. [CrossRef](#)
- Ullman S (1996) *High-level vision*. Cambridge, MA: MIT.
- Valentine T (1988) Upside-down faces: a review of the effect of inversion upon face recognition. *Br J Psychol* 79:471–491. [CrossRef Medline](#)
- Walther DB, Caddigan E, Fei-Fei L, Beck DM (2009) Natural scene categories revealed in distributed patterns of activity in the human brain. *J Neurosci* 29:10573–10581. [CrossRef Medline](#)
- Webster MA, Miyahara E (1997) Contrast adaptation and the spatial structure of natural images. *J Opt Soc Am A Opt Image Sci Vis* 14:2355–2366. [CrossRef Medline](#)
- Yin RK (1969) Looking at upside-down faces. *J Exp Psychol* 81:141–145. [CrossRef](#)