Visualizing fMRI BOLD responses to diverse naturalistic scenes using retinotopic projection

Karl Zipser

To view brain activity in register with visual stimuli, a technique here referred to as “retinotopic projection,” which translates functional measurements into retinotopic space, is employed. Retinotopic projection is here first applied to a previously acquired fMRI dataset in which a large set of grayscale photos of real scenes were presented to three subjects. A simple model of local contrast integration accounts for much of the data in early visual areas (V1 and V2). However, consistent discrepancies were discovered: Human faces tend to evoke stronger responses relative to other scene elements than predicted by the model, whereas periodic patterns evoke weaker responses than predicted by the model. Next, in new fMRI experiments, three subjects directed attention toward various elements of naturalistic scenes (Vermeer paintings). Retinotopic projection applied to these data showed that attending to an object increased activation in cortex corresponding to the location of that object. Together the results suggest that even during passive viewing, the visual system differentially processes natural scenes in a manner consistent with deployment of visual attention to salient elements.

Introduction

For studying the visual system, fMRI has an advantage over other techniques such as electrode recording or optical imaging because it permits study of multiple fragmented retinotopic maps simultaneously. This opens the possibility of exploring how complex stimuli, such as images of real scenes, are processed across these retinotopic areas. Human beings can process complex scenes seemingly without effort (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016; Greene & Fei-Fei, 2014; Thorpe et al., 1996). An fMRI dataset (Kay, Naselaris, Prenger, & Gallant, 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009) of fixating human subjects viewing a large number of photographs of real scenes provides the opportunity for exploration of this process. Given an intuitively direct way to visualize the data, the brain responses to each stimulus photo have the potential to yield valuable insights into visual processing.

In the current study fMRI signals in anatomical space were projected into a two-dimensional retinotopic space directly comparable to the visual stimulus—a simple visualization of fMRI data from retinotopic visual areas referred to as “inverse retinotopy” (Thirion et al., 2006) or “reconstructed retinotopy” (Kok & de Lange, 2014). The idea of projecting neural activity into retinotopic space originates with Creutzfeldt and Nothdurft (1978) and has been used by others (Amano, Wandell, & Dumoulin, 2009; Kok & de Lange, 2014; Lamme, Zipser, & Spekreijse, 1998; Lee, Mumford, Romero, & Lamme, 1998; Miyawaki et al., 2008; Naselaris et al., 2009; Papanikolaou et al., 2014; Winawer, Horiguchi, Sayres, Amano, & Wandell, 2010). The current study describes the first systematic use of a variant of the above procedures we term “retinotopic projection” to study the responses of early visual areas to a large set of naturalistic scenes.

A prerequisite for retinotopic projection is to measure the visual-field selectivity of each voxel. Once voxel receptive fields are characterized, retinotopic projection is a straightforward process for visualizing voxel activity by using the receptive fields to project voxel responses into the visual field. When applied to the dataset of Kay et al. (2008) and Naselaris et al. (2009), retinotopic projection allows for viewing the retinotopic activation of a particular cortical visual area and comparison of activation of separate retinotopic maps within an individual, as well as comparison or pooling of results across different subjects, all in the same two-dimensional space as the stimulus photographs. The implementation of retinotopic projection used here was developed in parallel with and independently from other versions cited above and serves as a useful validation of the method.

The diverse stimulus photos of real scenes used in the studies of Kay et al. (2008) and Naselaris et al. (2009) are, by their nature, not easily parametrized. A comprehensive assessment of the dataset thus requires
looking at brain responses to each stimulus photo and asking, what can be learned about visual processing? Viewing brain responses in the form of retinotopic projection images (RP-images) makes this task tractable. In interpreting these data-based RP-images, it was found helpful for comparison to employ retinotopic projection to visualize the output of a simple computational model of visual processing (i.e., to yield model-based RP-images). This process of subjective analysis led to unexpected observations which were tested quantitatively, yielding a more complete view of the dataset than achieved by the original two studies. Inspired by these findings, we then conducted new fMRI scanning experiments of subjects instructed to attend to different regions of naturalistic scenes while fixating. Vermeer paintings were chosen because they were appealing to the subjects and contained variations of similar scenes. The RP-images from these experiments suggest that visual attention can explain some of the results observed in the passive fixation experiment of Kay et al. (2008) and Naselaris et al. (2009).

**Methods**

**Passive fixation stimulus set and its presentation**

Passive fixation data were collected previously from three subjects (Subjects 1, 2, and 3) as described by Kay et al. (2008), Kay, Naselaris, and Gallant (2011) and Naselaris et al. (2009). The stimuli and a version of these data are available at http://crcns.org/data-sets/vc/vim-1; the version of the dataset used in the current study (provided by Kendrick Kay) has more refined preprocessing than the online version. The full stimulus set consisted of 1870 grayscale digital photographs of real scenes, each bounded by a circular blurred aperture (e.g., Figure 1A). Each stimulus photo was 500 × 500 pixels, which subtended 20° of visual angle on the display screen. A central white fixation spot (4 × 4 pixels) remained on throughout the 10-min scanning runs. The task was to fixate continuously. Between separate photos the screen was a mean-level gray for 3 s. A given
photo was presented for 1 s, flashing on and off at 5 Hz against the gray background. There were two stimulus subsets: In one subset of 1,750 photos, each photo was presented twice; in the other subset of 120 photos, each photo was presented 13 times. Photos were presented in pseudorandom sequence. Photos were from Corel Stock Photo Libraries (http://www.corel.com/en/clipart-and-photos/), copyright-free images from the Berkeley Segmentation Dataset (Martin, Fowlkes, Tal, & Malik, 2001); http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/, and the authors’ personal collections; in the current paper images from the Berkeley Segmentation Dataset and one photo by Kendrick Kay (used with permission) are presented.

BOLD signals to voxel responses

The fMRI scanning of BOLD signals by Kay et al. (2008) was done with 4T scanning of the occipital lobe with a surface coil, with a sampling rate of 1 Hz. These time course data were preprocessed (slice time correction and motion correction) and then analyzed using a GLM. The estimated beta weights for each voxel were z-scored to yield a single estimated response value for each voxel with respect to each stimulus photo.

Local contrast images

A simple nonlinear transformation of the stimulus photos, to make them more relevant to voxel responses, is suggested by decades of electrophysiology which have shown that visual neurons are sensitive to edges and spatial contrast, rather than to absolute light level at a given point (De Valois, Albrecht, & Thorell, 1985; Hubel & Wiesel, 1968; Schiller, Finlay, & Volman, 1976). The filter-transformation of the grayscale images begins by taking a given image (e.g., Figure 1A) and convolving it with a Gaussian kernel of SD of 6 pixels to create a blurred image (Figure 1B). After the blurred image is subtracted pixel-wise from the original image, the resulting image is one in which low spatial frequencies have been removed (Figure 1C). Next, a nonlinearity is introduced by taking the absolute value of this high spatial-frequency image to yield what we refer to as a local contrast image (Figure 1D). By taking the absolute value to produce the local contrast image, we throw away information about sign of contrast—that is, which regions of the scene are light or dark. While the filtering process does not change the visual-field position of the pixels, it does change the meaning of the pixel intensity values: The individual pixel values in the resulting local contrast images are greater (brighter) where the original photos contain edges or textures (such as the stripes of the tiger) and smaller (darker) in more blank regions of the photos. That is to say, the images express local contrast (Caselles, Coll, & Morel, 1999; Frazor & Geisler, 2006). This filtering process was applied to all stimulus photos for analysis.

Correlation between local contrast and voxel response

The correlation between a given pixel’s local contrast values across photos and a given voxel’s responses to these photos is given by the standard equation for the Pearson correlation coefficient. This correlation between pixel and voxel is made separately for each pixel. The ensemble of these correlations, mapped to the corresponding pixel coordinates, is a correlation RF image. An example for one V1 voxel is shown in Figure 1E. It is noteworthy how the voxel receptive field is revealed by such a simple method—correlation to local contrast. Previous studies have employed more complex, model-based approaches to estimate voxel receptive field size and position (Dumoulin & Wandell, 2008; Kay et al., 2008; Papanikolaou et al., 2014). What we show here is that, given a suitable dataset, detailed images of a receptive field are arrived at with a direct, model-free approach. The technique is similar to reverse correlation used in electrophysiology to measure single unit receptive fields (Jones & Palmer, 1987; Ringach, 2002; Theunissen et al., 2001). Other approaches to mapping receptive fields without an a priori model were used by Lee, Papanikolaou, Logothetis, Smirnakis, & Keliris (2013) and Greene, Dumoulin, Harvey, and Ress (2014).

The specificity of the V1 voxel examined in Figure 1 can be verified by inspecting which of the stimulus photographs evoked the highest and lowest voxel responses among a subset of 120 photos not used to map the receptive field. The top row of Figure 2 shows the five of these photos which gave the voxel’s strongest responses: In each photo, texture or edges fall within the receptive field. The bottom row shows the five of these photos which produced the voxel’s weakest responses. In each of these photos, a blank region falls within the receptive field. In contrast to this spatial selectivity, the voxel has no obvious selectivity for the content of the photos. Without knowledge of the voxel’s receptive field size and position, it would be difficult to discern why this voxel responded well to the photos in the top row and poorly to the photos in the bottom row.

Identifying voxels with localized receptive fields

Many voxels have single-peaked receptive fields such as the voxel in Figure 1. Some have double receptive
fields, while others have less clear receptive fields, or no obvious receptive field at all. A data-driven process for identifying voxels with clear, single-peaked receptive fields was devised, a process which yields results similar to the time-consuming task of manually selecting voxels with clear receptive fields. As a first step in this classification process, we thresholded the correlation RF image at 2/3 of the peak correlation value. In Figure 3, this isolates the pixels indicated in color in the middle column. Thresholding the correlation image leaves some pixels isolated from the main receptive field cluster. Starting from the pixel with the peak correlation value for each receptive field, above-threshold pixels that are contiguous with this pixel are indicated in yellow, whereas those above-threshold pixels not contiguous with the peak pixel are indicated in red. We then calculate a “contiguity proportion,” the ratio of the number of yellow pixels to the total number of colored pixels for a given voxel. For a clearly localized receptive field (e.g., Figure 3A2), the value of this ratio is close or equal to 1.0 (i.e., most or all colored pixels are yellow). If the pixel correlation of the receptive field is lower, more pixels are red and the contiguity proportion is lower). For voxels with multiple, localized receptive fields (e.g., Figure 3B2), or no obvious receptive field, the contiguity proportion is 0.5 or less. Across the population of V1 voxels, contiguity proportion exhibits a bimodal distribution (results for Subject 1 are shown in Figure 3C; other subjects showed similar results). We found that by choosing voxel receptive fields with peak correlations of 0.1 or greater, and contiguity proportions of 0.7 or greater, we could reliably select voxels with clear-looking correlation RFs for areas V1 and V2 in all three subjects. Across subjects, an average of 44% of V1 voxels were thus selected, whereas in V2, an average of 30% of voxels were selected. Examples for Subject 1 are shown in Figures 4 and 5. These dense receptive field coverages of the stimulus space will make it possible to project brain activation into the visual field. The variation of receptive field size with eccentricity will have a direct bearing on the level of detail which can be resolved at different eccentricities.

For those voxels with selected receptive fields, we further refined the representation as follows: First, subtract 2/3rds of the peak correlation value from the
correlation RF image and set pixels with negative values to zero—i.e., thresholding. Second, determine which nonzero pixels are contiguous with the pixels containing the peak (the pixels labeled yellow in the Figure 3A2), and set the remainder to zero. Third, set the sum of all pixel values to one. The resulting normalized RF image for a selected voxel is shown in Figure 3A3. This is the representation we use for retinotopic projection.

Retinotopic projection

The basis of retinotopic projection is to pixel-wise add the RF images of a given set of voxels, with each voxel’s RF image weighted by the response of that voxel to the stimulus. This weighted sum operation serves to project voxel responses to appropriate parts of the visual field. Three issues of scaling need to be taken into consideration to make this sum of images interpretable. First, voxel responses should be in directly comparable units. Z-scoring voxel responses across a large stimulus set accomplishes this purpose here. Second, because the receptive fields vary in size, we adjust each receptive field image to have the sum of its pixel values equal 1.0, after first applying a threshold to remove the influence of correlation measures outside the central receptive field area (e.g., Figure 3A3). Third, the sum of the response-weighted normalized RF images is divided, pixel-wise, by the sum of the normalized RF images unweighted by voxel responses. This serves to factor out uneven distributions of receptive fields across the visual field. The formation of a retinotopic projection image (RP-image) is thus described by the following equation:

$$P_i = \left( \sum_{j=1}^{N} z_j(i)R_j \right) \odot \left( \sum_{j=1}^{N} R_j \right)$$  \hspace{1cm} (1)

where matrix $P_i$ is the RP-image for photo $i$, $N$ is the number of voxels, $z_j(i)$ is the response of voxel $j$ to photo $i$, and matrix $R_j$ is the normalized RF image of voxel $j$. $\odot$ indicates element-wise (i.e., pixel-wise) division. In summary, retinotopic projection utilizes the receptive field images to map voxel responses to specific pixels in the RP-image; at each pixel of the RP-image, the intensity is computed from the weighted average of

Figure 3. Identifying voxels with single-peaked receptive fields. Some voxel correlation RF images exhibit clear localization (A); these voxels are useful for retinotopic projection imaging. We need to distinguish these from other voxels which lack clear receptive fields, or have multiple receptive field peaks (B). Applying a threshold allows for isolation of peaks in these receptive fields (A2 and B2). Above-threshold pixels are distinguished according to whether they are contiguous with the pixel having the highest correlation (yellow) or are not contiguous (red). The proportion of colored pixels which are yellow for a given voxel yields a metric we term contiguity proportion. (A3) Representation which isolates the yellow pixels from A2 to form a normalized RF image used for retinotopic projection imaging. See Methods for details. (C) Plot of the peak correlation value from each voxel’s correlation RF image against the RF contiguity proportion measured from the correlation RF image. Shown are all V1 voxels for Subject 1. Two separate clusters of voxels are readily apparent. The letters A and B on the graph indicate the plot locations of voxels (A) and (B) described above. Voxels in the lower left cluster lack recognizable receptive fields. The red lines indicate selection criteria for choosing voxels useful for retinotopic projection. Areas V1 and V2 of each subject show similar distributions. The same selection criteria were used for all datasets.
voxel responses, where the weights come from the normalized RF images.

The RP-image for any given photo is always made with receptive fields measured independently from that photo. To accomplish this, complete sets of correlation RFs were measured in batches. Each batch left out a different set of 120 stimulus photos. This means that for any given photo, a full set of receptive fields could be measured independently of responses to that photo.

Some pixels in a RP-image could be dependent on a single voxel—when only one normalized RF image is nonzero at that pixel. When this occurs, this pixel will tend to dominate the intensity range in the RP-image because of the wider range in variance of a single voxel.
compared to a weighted sum of voxels. Because this reduces the effectiveness of the RP-images as a visualization tool, in this study we require each pixel to have contributions from a minimum of five voxels whose normalized RF images are nonzero at the given pixel.

Because our goal was to study how brain activation varies across any given scene, pixel values in each RP-image were $z$-scored. Thus, when comparing across RP-images, we are comparing relative activations within RP-images, not absolute brain responses across stimulus photos.

Local contrast integration model

It is helpful to have some reference for comparison when attempting to interpret RP-images based on fMRI data. For this purpose we use a simple calculation based on the local contrast images and the normalized RF images (e.g., Figure 3A3). For a given voxel $j$ the modeled response $m_j(i)$ to a given stimulus photo $i$ is found by element-wise multiplying that photo’s local contrast image $T_i$ with the voxel’s normalized RF image $R_j$ and then summing the values for all pixels:
By making this calculation for each local contrast image for each voxel, we generate a complete set of computational outputs that model voxel behavior. Applying a compressive nonlinearity to the modeled responses improves the resemblance between model and data RP-images. We did this by scaling the model responses to a zero-to-one range, then taking the log. These values are then $z$ scored across the stimulus set to generate values directly comparable to voxel responses, with $z$ score values above and below ±3 limited to these values respectively. $Z$ scored model responses can be pushed through the retinotopic projection process to compare to RP-images based on actual voxel response data. (Specifically, the $z(i)$s in Equation 1 can be based on a computational model, whereas the original data-based $R_j$s are unchanged.) Local contrast integration output is compared to real subject data, beginning with Figure 6 in the Results section.

The basis of Kay et al. (2008) and Naselaris et al.’s (2009) model of visual responses in early visual areas, a Gabor-wavelet pyramid, was more complex than the local contrast integration model we use here. However, the lack of orientation specificity and the broad spatial frequency specificity of the voxels observed in these studies renders the models essentially equivalent.

### Segmenting photos for quantitative analysis

In order to make quantitative analyses of RP-image results, a naïve subject was given the task of producing segmentation masks for stimulus photos. The subject was instructed to (a) select photos with clear examples of extended repetitive textures, then produce a mask (using Adobe Photoshop) which
covered these regions in each photo; and (b) select photos which depicted human heads and torsos, and similarly produce masks covering these regions. Once these masks were generated, they were automatically modified as follows: (a) Texture masks were slightly shrunk so that they did not include parts of the photo immediately at the edge between texture and non-texture regions; and (b) Torso masks were reduced in area so that they equaled the area of the head masks by automatically eliminating pixels farthest from the center of mass of the head mask for each stimulus photo.

We used these segmentation masks as follows: A given mask was used to select pixels in RP-images corresponding to the selected area (for example, the part of the RP-image corresponding to repetitive texture in a particular photo). Then we found the average value of these selected pixels in the RP-image. Applying this process for each segmentation mask allows the comparison of responses between regions of RP-images corresponding to different parts of the stimulus photos.

**fMRI data collection with attentional tasks**

A new attention dataset was collected with three subjects (Subject 2 from the passive fixation experiments of Kay et al. (2008) and Naselaris et al. (2009) and new Subjects 4 and 5). Full or half-brain scanning was done with a 32 or 16 channel Siemens 3T scanner at the Henry H. Wheeler, Jr. Brain Imaging Center at UC Berkeley. A multiband protocol (Moeller et al., 2010) with acceleration factor 6 was employed, allowing for 2 mm isotropic voxels and a TR of 900 ms (see Supplementary Materials for more details on scanning protocols).

In the passive fixation experiment, many natural images were used as stimuli; these images turned out to be well suited to mapping the voxel receptive fields. In the attentional study described here, only a small number of stimuli were used. This meant that the stimuli themselves would not be diverse enough to map the receptive fields. Therefore, voxel RFs were mapped using a variation on the standard rotating wedge/expanding ring paradigms (Dumoulin & Wandell, 2008; Kay, Winawer, Mezer, & Wandell, 2013; Kriegeskorte et al., 2008) which used the entire monitor screen. These synthetic stimuli were employed because their strong contrast variations have previously been shown to allow for relatively rapid measurement of retinotopic maps.

The correlation RF mapping technique is based on local contrast (e.g., Figure 1D), but there is no requirement that local contrast comes from natural images. If we consider a single pixel in the display during the time course of the moving wedge and ring stimuli, the pixel will either be part of a high local contrast region (wedge or ring) or a zero local contrast region (the gray background). Thus, a pixel’s local contrast can be approximated by alternating from high to low (one to zero) over time. To get a prediction of what voxel responses would be if driven by a single pixel’s local contrast, the pixel’s one/zero local contrast time course was convolved with a hemodynamic response function (Kay et al., 2013). Correlating each pixel’s local contrast-predicted hemodynamic response time course with the BOLD activation of a given voxel gives a correlation value at each point in the image, the ensemble of which yields a correlation image directly comparable to the correlation RF images for the first part of the study. An important distinction, however, comes from the fact that the wedge and ring stimuli have strong and stereotypical intraintrope contrast correlations. For example, if a pixel is in a high local contrast region during the ring part of the sequence, it is strongly correlated with all pixels of similar eccentricity. This is distinct from the average spatial correlations of local contrast in the natural image stimulus set in the passive fixation study; these spatial correlations have a spread substantially smaller than the smallest receptive fields mapped in the passive fixation data.

Despite this difference between the stimuli, the RF correlation images measured from the wedge-ring fMRI data displayed clear receptive field peaks consistent with the results in the first part of this study on the passive fixation data. However, these peaks ride on top of a background correlation pattern reflecting the preferred ring and wedge patterns for each voxel. Therefore, it was necessary to change the threshold used in going from correlation RF images to normalized RF images used in retinotopic projection. It was found that changing this threshold from 2/3 (~0.67) to 31/32 (~0.97) produced RFs for Subject 2 that corresponded well in eccentricity size progression to the RFs mapped from the same subject in the passive fixation study. Furthermore, the peak correlation value threshold for identifying voxels with single-peaked receptive fields (see Figure 3C) was shifted from 0.1 to 0.2 with the wedge-ring mapping dataset. Thus, the correlational RF method used with natural images was successfully applied, with parametric adjustments measured from the common Subject 2, to mapping receptive fields using synthetic wedge-ring RF mapping stimuli. It is clear, however, that for future correlation RF mapping experiments, synthetic stimuli should be designed to reduce stereotypic image correlations, (e.g., by deploying random distributions of high and low local contrast).

During the attention experiments, subjects fixated a stream of small letters (< 0.5° width) appearing one
after the other in the center of the display during 5-min runs. The display subtended 23° of visual angle in the width dimension. Images, four Vermeer paintings, were individually presented for 4 s, with a 2 s blank between images, during which the stream of letters at fixation continued. Eye tracking was used for Subject 4 who was not a dedicated vision science researcher. The fMRI time course data were preprocessed (slice time correction and motion correction) and then analyzed using a GLM. The estimated beta weights for each voxel were z scored to yield a single estimated response value for each voxel with respect to each stimulus photo/attentional task combination. Because the total number of stimuli was smaller than in the data set of Kay et al. (2008) and Naselaris et al. (2009), analysis was applied to the naturalistic images and the receptive field mapping stimuli as a complete set, in order to insure a wide range of responses for each voxel.

Different attentional tasks were done in blocks, either of 30 s, or for a full 5-min run. The core tasks were read letters, attend face, attend vase, and attend ground. For Subjects 2 and 5, there was one location in the ground of each image to attend to, whereas for Subject 4 there were two locations tested in separate runs, one location being near the vase and the other near the face. Subjects 4 and 5 also had additional tasks in separate runs which will be described elsewhere. As there were only a small number of painting stimuli, subjects were easily acquainted with each. The different elements that would be attended to were clearly described to the subjects before scanning began. Therefore, there was no need for spatial precueing of attention target location. Rather, the subject knew for a particular run what element of each painting to attend to, which in practice was unambiguous and easy for each subject. The read letters task was intended to prevent the subject from covertly attending to other parts of the image. For Subjects 2 and 5, paintings were shown in 5-min runs during which the subject had only one attentional task (e.g., attend to the face in each picture). In this paradigm, paintings appeared in a pseudorandom order. Subject 4 also saw 5-min runs, but the task switched every 30 s, typically from an attend task (e.g., attend face) to the read letters task. In this experiment, the color of the fixation letters indicated which task should be performed. The order of tasks was intentionally not randomized in order to prevent the subject from being confused about what to do at any given moment. In this paradigm, paintings appeared repeatedly in the same order (two other paintings, landscapes, were also used, so the actual length of the image stream was six images). There was no indication of any difference in fMRI results between the two paradigms.

Results

Retinotopic projection data from V1 and V2

The goal of retinotopic projection is to pool together the responses of thousands of voxels (with receptive fields spanning the visual field) in a systematic way to give us insight into how retinotopic areas of the brain represent a particular stimulus. Intuitively, by combining the type of specificity we see in Figure 2 with the visual field coverage we see with receptive fields in Figures 4 and 5, we should be able to generate a two-dimensional image representing how the set of voxels responds to a given photo. The process of doing this—retinotopic projection—is to pixel-wise sum the receptive field images of a given set of voxels, with each voxel’s receptive field image weighted by the response of that voxel to the stimulus (Equation 1). Appropriately scaled, this weighted sum serves to project voxel responses to appropriate parts of the visual field to create a retinotopic projection image (RP-image). Each RP-image is z scored across pixels, so that the pixel values reflect relative activation within a scene; see Methods for details. The RP-image for any given photo is always made with receptive fields measured independently of responses to that photo.

In Figure 6 (left) we show a particular stimulus photo (a roller coaster). Let us retinotopically project the responses of V1 voxels of Subject 1 during presentation of this photo (Figure 6, top center). In the resulting RP-image, yellow indicates higher ensemble voxel responses, and blue indicates lower ensemble voxel responses. White pixels indicate where receptive field coverage is fewer than five voxels (i.e., fewer than five normalized RF images are nonzero at that pixel). Here we have a visualization which assigns response values to most pixels in the visual field region corresponding to the stimulus photo. There is a clear correspondence between the RP-image and the stimulus photo itself: Areas of texture (the roller-coaster and trees) appear yellow (have larger responses) in the RP-image, whereas the untextured regions (the sky) appear blue (have lower responses) in the RP-image.

Let us also retinotopically project the responses of V2 voxels from the same subject to this photo for the same stimulus presentation in Figure 6 (bottom center). This RP-image is based on a different set of voxels, with their own receptive fields and responses to the stimulus photo. Yet the V1 and V2 RP-images are quite similar. The more “ragged” edges of the V2 RP-image result from a lack of selected V2 voxels with receptive fields in certain parts of the visual field. The high degree of similarity in V1 and V2 responses which can be observed in the RP-images would be difficult to ascertain if the voxel responses remained in anatomical
coordinates (either in situ, or in an inflated cortical representation).

How similar in general are RP-images for areas V1 and V2? We measured the correlation of the pixels between the V1 and V2 RP-images (using pixels that are represented in RP-images for both areas). For the set of 120 photos presented 13 times each, the mean correlations between V1 and V2 RP-images are 0.54, 0.48, and 0.51 for Subjects 1, 2, and 3 respectively. These results confirm that V1 and V2 yield broadly similar RP-images for a given stimulus.

Retinotopic projection of the output of a computational model

Up to now, we have used retinotopic projection to visualize the responses of voxels to a stimulus as estimated from fMRI BOLD signals. We can also use retinotopic projection to visualize the output of a computational model of voxel behavior. In this process, the same set of RF images is used to map voxel to pixels, but model output is substituted in place of the actual voxel responses. We use a simple model of voxel behavior: namely, that a voxel’s response to a stimulus is determined by integrating the amount of stimulus local contrast that falls within the voxel’s receptive field, followed by a logarithmic compression. We refer to this as a local contrast-integration model of voxel responses; see Methods for details. Pushing these model outputs through the retinotopic projection equation (Equation 1) for the set of V1 and V2 voxels, we generate model-based RP-images as shown in Figure 6 (right). The results appear qualitatively similar to the data-based RP-images. We will compare brain and model responses in more detail later.

Pooling V1 and V2

Because V1 and V2 RP-images are similar, we make visualizations with a larger number of voxels by pooling voxels from V1 and V2 and making retinotopic projections of their responses. In Figure 7 we compare the RP-images for the three subjects, for different photographs of real scenes. If we examine the RP-images in any given column—i.e., examine how a given subject’s V1/V2 responses vary with the stimulus—we see diverse patterns of activation. If, however, we compare RP-images across subjects for a given stimulus—i.e., look across a single row—we see from these examples that a given photo evokes a similar pattern of activity in each subject for the first three rows. The degree of similarity across subjects would not be easily apparent if voxel responses were viewed in anatomical space in which each subject would have a unique layout for V1 and V2. For comparison, we show (Figure 7, rightmost column) the corresponding RP-images based on output of the local contrast integration model, pooling receptive fields from all three subjects. The similarity of the RP-images based on brain responses (data RP-images) to those based on modeled responses (model RP-images) is striking. Near the fovea, these details become fairly well resolved because of the smaller voxel receptive field sizes. The examples in Figure 7 indicate qualitatively that the local contrast integration model is good at explaining the brain responses for the types of scenes shown in rows 1–3. The fourth row is an example of a highly textured scene where the model responses describe the data less well, a divergence to be explored below.

Quantitative comparison of data and model RP-images

For the set of 120 photos presented 13 times each, the mean correlations between V1-V2 pooled RP-images for data and model are 0.60, 0.59, and 0.55 for Subjects 1, 2, and 3 respectively. Averaging RP-images for Subjects 1, 2, and 3 gives a correlation coefficient of 0.69 between data and model RP-images. These results are consistent with the main qualitative impression from inspecting RP-images for the entire dataset, namely that the simple local contrast model accounts for a substantial part of voxel responses.

Divergence of data and model RP-images

The effectiveness of the local contrast-integration model, together with its sheer simplicity, makes it a useful benchmark when analyzing the data. Our goal now is to ascertain whether there are characteristic ways in which data and model RP-images diverge that could reflect properties of cortical computation beyond local contrast integration.

Extended repetitive patterns

By inspecting many RP-images, we noticed a divergence between data and model RP-images for stimulus photos with extended repetitive patterns. For example, Figure 8 shows a photo in which the most vivid features are stark shadows forming a repetitive pattern in the lower half of the image. The data and model RP-images are quite distinct for this photo—the data RP-image shows a gap over the shadows, whereas the model RP-image has a high output there.
To enable us to better visualize how the RP-image relates to the photo, we modulate the photo by the colors of the RP-image: Where the RP-image is positive (yellow), the grayscale photo is tinted toward yellow, and where the RP-image is negative (blue), the grayscale photo is tinted toward blue. A similar divergence between data and model RP-images is seen in another photo with repetitive shadows in the

| Figure 7. RP-images of four photos, pooling V1 and V2 voxels. (Column 1) Photographs of real scenes. (Columns 2–4) Data RP-images for the three subjects. The degree of similarity across subjects would not be easily apparent if voxel responses were viewed in anatomical space, in which each subject has a unique layout for V1 and V2. (Rightmost column) Model RP-images based on output of the local contrast integration model; model RP-images shown here pool receptive fields from all three subjects. As described in Methods, the receptive fields are based on subject data, but the responses used with these receptive fields in retinotopic projection can be based on model data or subject data. There is a strong similarity of the data RP-images and model RP-images for the first three photos. The photo in the bottom row is an example for which the model is a poor fit to the data. First three photos are modified from the copyright-free Berkeley Segmentation Dataset (Martin et al., 2001), https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html; bottom photo by Kendrick Kay who has made his images available for research and publication; see https://crcns.org/data-sets/vc/vim-1. |
ground plane (Figure 9, top row). Another example of a repetitive pattern that typically yielded lower than expected data RP-image activation is the coat of animals with repetitive patterns (e.g., zebras), as shown in the second and third row of Figure 9. The heads and limbs of the zebras evoked higher data RP-image activation than the center of the body, despite the high contrast stripes there. In the fourth row, the grating pattern behind the man yielded lower activation than the model predicts, whereas the major activation was directed more to the head of the figure and the foreground elements. And finally, looking back at the bottom row of Figure 7, the repetitive pattern of the fence evokes low activation from the voxels of the three subjects but high activation from the model.

Quantitative analysis

All of the examples in Figures 8 and 9 have regions of the photos with vivid and high contrast that yield patterns with lower data RP-image activation than expected from the model RP-images. A broader quantitative analysis of this observation was done as follows: A naive subject was given the task of finding photos from the stimulus set with pronounced areas of repetitive patterns, then indicating these regions with a segmentation mask (see Methods for details). We then computed the average values of data and model RP-images in these regions. Comparing them showed that the data RP-images were consistently near zero in these textured regions (recall that RP-images are averaged results from areas V1 and V2 from the three subjects. Images modified from the copyright-free Berkeley Segmentation Dataset (Martin et al., 2001), https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html. FMRI data and stimuli from Kay et al. (2008) are available at https://crcns.org/data-sets/vc/vim-1.

Human figures

Inspection of RP-images suggests that human figures are another part of scenes in which the model fits the data less well than other regions. Figure 11 shows a picture of a woman’s head in front of a river. The
subject-averaged data and model RP-images are quite distinct, as seen in the top row. Both the data and model RP-images have a minimum within the dark region of the rock. However, the data RP-image shows greater preponderance of positive activation at the woman’s head than at the other parts of the scene, whereas the model RP-image displays a more distributed activation corresponding to the diverse areas of local contrast in the stimulus image.

We observed a similar discrepancy for other stimulus photos. For example, in the top row of Figure 12, the high-contrast edges of the woman’s clothing cause the dominant model responses, whereas the face shows more activation in the data RP-image. In the second
row of Figure 12, the data RP-image activation is focused on the small area corresponding to the human being within a landscape. This region is not the dominant aspect of the model RP-image. In the third row of Figure 12, the data RP-image shows activation on the child’s head that is missing in the model output. In the fourth row, the data RP-image activation is strong on the diver, whereas in the model RP-image, the high contrast edges of the underwater vegetation are much more important components of the activation. These examples and many others suggested that there could be a consistent difference between data and model retinotopic projection when the photos contain human figures.

Quantitative analysis

A naive subject was given the task of segmenting those stimulus photos containing clear depictions of human figures into head and torso regions. This analysis brings together data for $n = 229$ stimulus photos in which the head and torso regions were at least one degree of visual angle wide. Using these manually marked regions, we calculated the average
intensity of RP-images within these head and torso parts of the photos. Recall that the mean value of each RP-image is zero because of $z$ scoring. The results are shown in Figure 13 for V1 and V2 separately for the three subjects. For each visual area in each subject, the trend is the same: Whereas head regions evoke similar levels of activation in both data and model RP-images, activation for torso regions is approximately 25% lower for the data RP-images compared to the model RP-images, the difference being significant at the $p < 0.05$ level for all three subjects for area V1, and for one subject for area V2.

**Cross-stimulus RP-images**

A cross-stimulus RP-image visualization is possible if we align face regions from different photos. We chose the subset of photos with faces in frontal or near frontal view; we then aligned the faces such that the eyes and mouth were in the same position relative to a reference face. In order to achieve the alignment, photos containing faces could be rotated clockwise or counterclockwise, scaled and mirror reflected, but not otherwise distorted, to bring the eyes and mouth into alignment. This enabled pooling together data from
208 faces in 129 photos. The data RP-images, and separately, the model RP-images, were then aligned with the same registration as their associated photos and averaged pixel-wise. The aggregate data show that the face region is the focus of data RP-image activation, whereas the model RP-images show broader activation outside the face region as well, especially the neck and upper body regions where the shirt typically begins.

Retinotopic projection of visual attention

Could the face effect (Figure 14) result from visual attention being covertly attracted to a naturally salient stimulus? In order to test this hypothesis, new experiments were conducted in which subjects not only fixated, but had various attentional tasks. The subjects (Subject 2 and new Subjects 4 and 5) were shown four different Vermeer paintings containing both face and object (vase) elements, an example of which is illustrated in Figure 15A. As a control, subjects were instructed to read small, rapidly presented letters appearing at the center of the screen while the naturalistic images were displayed. Alternatively, subjects fixated the letters but attended to a face, a vase, or a background region in each image. Attentional tasks were directed by verbal instructions given before scanning (see Methods).

Resulting V1, V2 pooled RP-images are shown in Figure 15B for Subject 2 [who was also Subject 2 in the passive fixation experiments of Kay et al. (2008) and Naselaris et al. (2009)]. Attention to face or vase elicited focused increased activation in these areas. In contrast, allocation of attention to a specific location in the featureless background of the scene resulted in spatially diffuse positive modulation within this area which spread to adjacent objects. Here attention to ground yielded higher (yellowish) activation not only in the target area (which is blue in the preceding three cases) but also diffuse activation across adjoining background and objects as well.

Retinotopic projections from Subject 4 are shown in Figure 15C for the same stimulus. The attend face and attend vase results are similar to those for Subject 2. For this subject, attention to ground was directed to different target locations, either to the ground area directly above the vase or to the left of the face, in separate scans. In each case, activation spread out across the adjacent objects and ground.

When RP-image mean activation was measured within the vase regions, the values for the attend vase
condition were significantly greater than for the read letters condition ($p < 0.05$ or better for all subjects and all paintings, Welch’s $t$ test). Likewise, in the target ground areas, attend ground yielded significantly greater mean values than read letters condition values at $p < 0.05$ or better for all subjects and all paintings (Welch’s $t$ test). However, for faces the results were less consistent. For the face region, attend face condition...
values were significantly greater than read letters condition values at $p < 0.05$ or better for three out of four paintings for Subject 2, for two out of four paintings for Subject 4, and for one out of four paintings for Subject 5. Figure 15D shows data for Subject 5 in which the attend vase and attend ground cases show characteristic modulations: For attend vase, activation is centered on the vase, and for attend ground, the blue area to the left of the figure in the previous three conditions is replaced by yellow, which spills along the side of the figure and up the wall on the left. But the face region is not significantly modulated for attend face as compared to read letters.

Figure 15E shows the quantitative results graphically. The left graph plots average activity level in the RP-images within vase regions. The $x$ axis measures this activity level under the control read letters task. The $y$ axis shows the activity level for other attention tasks. As labeled on the plot, the blue symbols are for the cases where the task was to attend to the vase region. These symbols are all above the $x = y$ identity line, indicating that the activity in the vase region increases relative to the read letters control when attention is directed there. In contrast, the red symbols indicate activity levels in the vase region when the task was to attend to the face. These fall close to the $x = y$ identity line, meaning that in this condition the activity in the vase region of the RP-image was similar to that for the read-letters control. Thus, attention to the vase region consistently increases activation in that part of the display compared to the read letters task, but attention to face region does not change activation in the vase region, as compared to the read letters task. There is one data point for each painting, per subject and attentional condition. Error bars indicate $SEM$ across runs.

The right graph is of the same format as the previous one, except that activity in the face region of the RP-images is measured. When the face region is attended to (red symbols), activation is generally higher than the read letters condition, although attention to the face region does not in every case produce significant increase in activation relative to read letters, as described above. Attention to the vase region (blue symbols) does not increase activation in the face region above the read letters control. In fact, in a quarter of cases, face region activation is significantly ($p < 0.05$) reduced by attention to the vase.

**Discussion**

The current study describes the first systematic use of retinotopic projection to investigate the responses of retinotopic visual areas to images of real scenes. The retinotopic projection technique developed for this study has varying degrees of similarity to visualization techniques employed by other researchers for the study of brain responses to simple synthetic scenes (Amano et al., 2009; Kok & de Lange, 2014; Lamme et al., 1998; Lee et al., 1998; Miyawaki et al., 2008; Naselaris et al., 2009; Papanikolaou et al., 2014; Winawer et al., 2010). Retinotopic projection is a process that channels fMRI signals from voxels (anatomical space) to pixels (retinotopic space), meaning that brain responses can be directly compared to visual stimuli. As described by Equation 1, a RP-image depends on two factors: (a) a set of voxel receptive fields, which define the projection from voxels to pixels, and (b) the responses of these voxels to the particular stimulus.

The following observations were made in the current study:

a. With stimulus photos that juxtapose blank regions and regions of texture, V1 and V2 RP-images show clear spatial contrasts corresponding to these distinct regions (e.g., Figure 6). These data RP-images are well matched by model RP-images, in which simulated responses are used in place of real voxel response data; the modeled responses are computed by integrating the local contrast within the actual voxel receptive fields (Equation 2). The model RP-images describe the data RP-images far less well with stimulus photos that lack stark juxtapositions of blank and textured regions, (e.g., Figure 7, bottom row). Although part of the failure of the model RP-images with these more textured scenes may be regarded as a signal-to-noise issue, there is also the indication that data RP-images reveal characteristic departures from the model RP-images, as described below.

b. Extended high-contrast textured regions evoked consistently less brain activation than expected from the model (e.g., Figure 10).

c. It was observed that faces could evoke a higher relative brain activation compared to surrounding scene elements than would be expected from the local contrast-integration model, e.g., Figure 11 and Figure 12. Quantitative analysis of RP-images for many photographs showed a similar level of activation for faces in data and model RP-images; however, relative to the face regions, the accompanying torso regions evoked a significantly lower response in data RP-images as compared to model RP-images (Figure 13). This result was also seen when face-aligned RP-images were compared (Figure 14). Thus, the consistent finding is that faces evoke a relatively larger response compared to surrounding scene elements.
in data RP-images as opposed to model RP-images.

d. A face enhancement effect similar to that observed with the passive fixation dataset (Figures 11 through 14) could be induced when subjects were instructed to attend to a face (Figure 15). A similar but stronger effect could be induced when the subjects were instructed to attend to an inanimate object (for example, a vase of similar size, Figure 15). Why were the attentional effects for face, compared to the read-letters task, less strong and reliable than for vase? Does this mean that faces are a weak substrate for visual attention? This seems paradoxical, given that the observation (“c” above) of higher activation for faces than expected from the contrast-integration model. We suggest that rather than being a weak substrate for attention, attention to faces is so strong that it is present to a degree even in the read letters control task which was intended to block it. Thus, when subjects attend to the face intentionally, the measured shift in attention is modest compared to the read letters task because their attention was already drawn to the face. In contrast, the read letters task more successfully blocked attention to vase; thus, when subjects attended the vase, a larger range attentional modulation could be observed. Finally, attention directed to a background region generated a diffuse increase in response which spread to adjacent objects.

The above observations suggest the presence of contextual or top-down influences in early visual cortex. The existence of such influences have long been observed with electrode recordings in monkey early visual cortex, phenomena described as figure-ground modulation or surround suppression, as well as independent attentional effects (citations below). What these studies had in common was the use of simplistic synthetic stimuli. A key question is, in what way do the observations made with these stimuli extend to complex, naturalistic scenes such as those studied in the current work? We shall address this by discussing figure-ground, surround suppression, and attention in turn.

Figure-ground

Previous microelectrode and optical imaging studies in monkey suggested a role for areas V1 and V2 in figure-ground segmentation (Gilad, Pesoa, Ayzenshtat, & Slovin, 2014; Gilad & Slovin, 2015; Lamme, 1995; Zhou, Friedman, & Von Der Heydt, 2000; Zipser, Lamme, & Schiller, 1996), a result supported by fMRI scanning in human (Scholte, Jolij, Fahrenfort, & Lamme, 2008; Strother, Lavell, & Vilis, 2012). In these studies it was observed that brain responses corresponding to parts of the visual field containing a “figure” in simple synthetic scenes were consistently greater than for the surrounding “ground” regions. This phenomenon is not merely attentional modulation (Poort et al., 2012). Do these observations with simple synthetic figures translate in a clear way to the complex photos of real scenes we consider here?

From the retinotopic projection analysis we saw that in the case of human figures large enough that distinct parts could be well resolved in the RP-images, human figures were not uniformly elevated in response as would be expected from the previous figure-ground studies in early visual cortex. Rather, faces tend to evoke a higher relative level of brain activation compared to surrounding regions, as discussed above. Thus, from retinotopic projection, there was no evidence of segmentation of human (or animal) figures as a uniform whole. Furthermore, extended high-contrast textured regions evoked consistently less brain activation than expected from the model (Figure 10), even when these regions could be considered part of a “figure,” such as in the bodies of zebras (Figure 9). Thus, although previous studies of figure-ground processing in early visual areas indicate the presence of complex, context-specific processing, the current retinotopic projection imaging analysis does not support a simple translation of the figure-ground interpretation to real scenes. Synthetic figures may be uniformly elevated in activation, but complete figures of animals or people in the photos of complex real scenes studied here are not modulated in this way in our observations. The implication of this is that studies with simple synthetic displays should be paired with more realistic scenes, in order to test how observations of phenomena such as figure-ground modulation transfer from the one to the other.

Surround suppression

Surround suppression (Bair, Cavanaugh, & Movshon, 2003; Shushruth et al., 2012; Smith, Bair, & Movshon, 2006) may be related to (or identical with) the hypothesized figure-ground mechanisms discussed above. The effect observed here, of extended textured regions giving much lower activation than expected from the model, could be well described by a surround suppression phenomenon with a simple like-inhibits-like mechanism. As attractive as such an explanation may be, it seems unlikely that the brain would allow for major perturbations across scene representations without a more sophisticated system of control than the postulated surround suppression mechanisms.
Attention

Much research has focused on the spatial window and topology of visual attention (Brefczynski-Lewis, Datta, Lewis, & DeYoe, 2009; Datta & DeYoe, 2009; Dosher, Liu, Blair, & Lu, 2004; Gandhi, Heeger, & Boynton, 1999; Li, Lu, Tjan, Dosher, & Chu, 2008; Müller, Mollenhauer, Rösler, & Kleinschmidt, 2005; Puckett & DeYoe, 2015; Silver, Ress, & Heeger, 2007; Simola, Stenbacka, & Vanni, 2009). A priori, it is not clear how these findings will transfer to complex naturalistic scenes. Indeed, even in relatively simple synthetic displays, scene interpretation has interesting consequences for how visual attention is expressed neurally (Qiu, Sugihara, & von der Heydt, 2007). The focus of the fMRI attention experiment conducted here was to explore the nature of the interaction of attention and naturalistic scenes and how this could relate to the passive viewing results.

Attention to face (or to vase) elicited greater activation at the specific region in RP-images than in the control condition. This is consistent with the attention studies cited above. But in fact it was not obvious that these would be the results within the context of a complex scene. For example, attention to the face could have resulted in elevated activation for the entire figure. Attention to the vase could have elevated the supporting surface. Attention directed to the background area did in fact give unexpected results—not a focused attentional spotlight, but a spreading modulation encompassing adjacent objects. The attention results for faces resemble the face enhancement results observed in the passive viewing dataset (Figures 11, 12, and 14). This similarity suggests that covert attention to faces by the fixating subjects could explain these results.

Scene perception

Given the facility with which human subjects perceive naturalistic scenes as a whole (Greene et al., 2016; Greene & Fei-Fei, 2014; Thorpe et al., 1996), it seems likely that the convergence of diverse and powerfully modulatory signals onto the early retinotopic cortical representation is critical to forming these percepts. By allowing us to make observations of cortical representations of complete scenes, retinotopic projection of fMRI signals, utilized with rich and carefully selected scene stimuli, should be able to offer many more insights into this process.

Keywords: face perception, scene perception, fMRI, visual attention

Acknowledgments

The author thanks David Zipser and Kendrick Kay for discussion of the current work, and also thanks Kendrick Kay for help with some of the fMRI scanning. Ben A. Inglis provided essential help in designing the scanning protocols at the UC Berkeley Brain Imaging Center, and Richard L. Redfern provided critical technical support. Two anonymous reviewers contributed valuable feedback on preparation of this manuscript. This work was supported by grants from NSF (IIS-1111765) and DARPA’s Cortical Processor program awarded to Bruno Olshausen.

Commercial relationships: none.
Corresponding author: Karl Zipser.
Email: karlzipser@berkeley.edu.
Address: Redwood Center for Theoretical Neuroscience, University of California Berkeley, CA, USA.

References


Datta, R., & DeYoe, E. A. (2009). I know where you are secretly attending! The topography of human visual attention revealed with fMRI. Vision Research, 49(10), 1037–1044.


Müller, N. G., Mollenhauer, M., Rösler, A., &

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron, 63*(6), 902–915.


