ARTSCENE: A neural system for natural scene classification

Stephen Grossberg

Department of Cognitive and Neural Systems, Center for Adaptive Systems, Center of Excellence for Learning in Education, Science, and Technology, Boston University, Boston, MA, USA

Tsung-Ren Huang

Department of Cognitive and Neural Systems, Center for Adaptive Systems, Center of Excellence for Learning in Education, Science, and Technology, Boston University, Boston, MA, USA

How do humans rapidly recognize a scene? How can neural models capture this biological competence to achieve state-of-the-art scene classification? The ARTSCENE neural system classifies natural scene photographs by using multiple spatial scales to efficiently accumulate evidence for gist and texture. ARTSCENE embodies a coarse-to-fine Texture Size Ranking Principle whereby spatial attention processes multiple scales of scenic information, from global gist to local textures, to learn and recognize scenic properties. The model can incrementally learn and rapidly predict scene identity by gist information alone, and then accumulate learned evidence from scenic textures to refine this hypothesis. The model shows how texture-fitting allocations of spatial attention, called *attentional shrouds,* can facilitate scene recognition, particularly when they include a border of adjacent textures. Using grid gist plus three shroud textures on a benchmark photograph dataset, ARTSCENE discriminates 4 landscape scene categories (coast, forest, mountain, and countryside) with up to 91.85% correct on a test set, outperforms alternative models in the literature which use biologically implausible computations, and outperforms component systems that use either gist or texture information alone.

Keywords: scene classification, gist, texture, spatial attention, coarse-to-fine processing, attentional shroud, multiple-scale processing, ARTMAP

Citation: Grossberg, S., & Huang, T.-R. (2009). ARTSCENE: A neural system for natural scene classification. *Journal of Vision, 9*(4):6, 1–19, http://journalofvision.org/9/4/6/, doi:10.1167/9.4.6.

1. Introduction

Scene understanding is a hallmark of human natural vision and is a challenging goal for machine vision because a scene contains predictive information on multiple scales of processing. Computational models of scene understanding have attempted to identify scene signatures and use them for image classification. For example, Oliva and Torralba (2001) used spectral templates that correspond to global scene descriptors such as roughness, openness, and ruggedness. Fei-Fei and Perona (2005) decomposed a scene into local common luminance patches or textons. Bosch, Zisserman, and Muñoz (2006) applied the Scale-Invariant Feature Transform (SIFT: Lowe, 1999) to characterize a scene. Although successful in benchmark studies, these approaches often stress one representation over the others, either local or global, and many include computations that are non-local and implausible biologically. In contrast, Vogel, Schwaninger, Wallraven, and Bülthoff (2006) showed that human subjects did a better job in categorizing rivers/lakes and mountains when the presented images were globally blurred than locally scrambled, but conversely in categorizing coasts, forests and plains. In addition, intact images are always easier to identify than either of the manipulated ones. Such evidence indicates that neither global nor local information is more predictive than the other at all times, and that the brain makes use of scenic information from multiple scales for scene recognition.

Scene understanding from multiple-scale gist and texture categories

The ARTSCENE model assumes that global information is quickly available before more local information is acquired using attentional focusing and scanning eye movements. This assumption is consistent with several studies in global-to-local visual processing (e.g., Navon, 1977; Schyns & Oliva, 1994) and with the fact that human viewers can detect a named object in a scene within ~150 ms that is less than the average fixation time (~300 ms) (Potter, 1975). ARTSCENE furthermore proposes that global gist and local texture information are both computed using similar mechanisms, albeit at different spatial scales, and that selective attention to more local scales collects texture evidence to revise and refine a global gist prediction.

doi: 10.1167/9.4.6 Received September 27, 2007; published April 6, 2009 ISSN 1534-7362

The challenges of the model are thus to clarify what constitutes scene gist, where and what scale to attend next, how these statistical scenic measures are learned, and how to integrate gist and texture information to achieve state-of-the-art scene classification. In ARTSCENE, the gist of a scene is a learned category of its spatial layout of colors and orientations. Spatial attention is then sequentially drawn to the scene's principal textures, in order of decreasing size, which are also categorized. Scene identity is predicted via a learned mapping from multiple-scale gist and texture category activations.

ARTSCENE is one of an emerging family of Adaptive Resonance Theory, or ART, neural models that clarify how the visual system can strategically deploy attention and combine information from multiple scales to learn useful predictions about the world. ART is used because it models how the brain can rapidly and stably learn to categorize large non-stationary databases using incremental learning (Carpenter & Grossberg, 1991). Recent review articles summarize behavioral and neurobiological data that support all of the main ART predictions about how the brain does this (Grossberg, 2003b; Raizada & Grossberg, 2003).

Since gist is just one of several textures in our treatment, ARTSCENE may be viewed as a generalization of the ARTEX (Grossberg & Williamson, 1999) and dARTEX texture classifier (Bhatt, Carpenter, & Grossberg, 2007). ARTSCENE also adapts heuristics of the ARTSCAN model of view-invariant object learning (Fazl, Grossberg, & Mingolla, 2009) by incorporating multiple views of a scene that are presumed to be derived from spatial attention shifts and scanning eye movements.

In the following sections, we first describe the image and annotation dataset used to test ARTSCENE. Then, the ARTSCENE system is defined mathematically and simulation results are presented. Finally, strengths and weaknesses of the current approach are discussed, as well as possible model extensions.

2. The image and annotation dataset

2.1 The image dataset

ARTSCENE simulations ran on the natural image dataset from Oliva and Torralba (2001) that has also been used by other researchers (e.g., Bosch et al., 2006; Fei-Fei & Perona, 2005). The dataset contains 4 landscape scene categories including coast (360 images), forest (328 images), mountain (374 images), and countryside (410 images). All images are chromatic and of size 256 × 256 pixels. Figure 1 shows 8 exemplars in the dataset and illustrates the great variation within each scene category.

2.2 The annotation dataset

To study how humans parse a scene into local elements, we make use of human annotations on the same image dataset, which are available from the *LabelMe* webpage (Russell, Torralba, Murphy, & Freeman, 2005). Although this annotation scheme embodies polygon coordinates and label names of local regions, it is not an error-free dataset

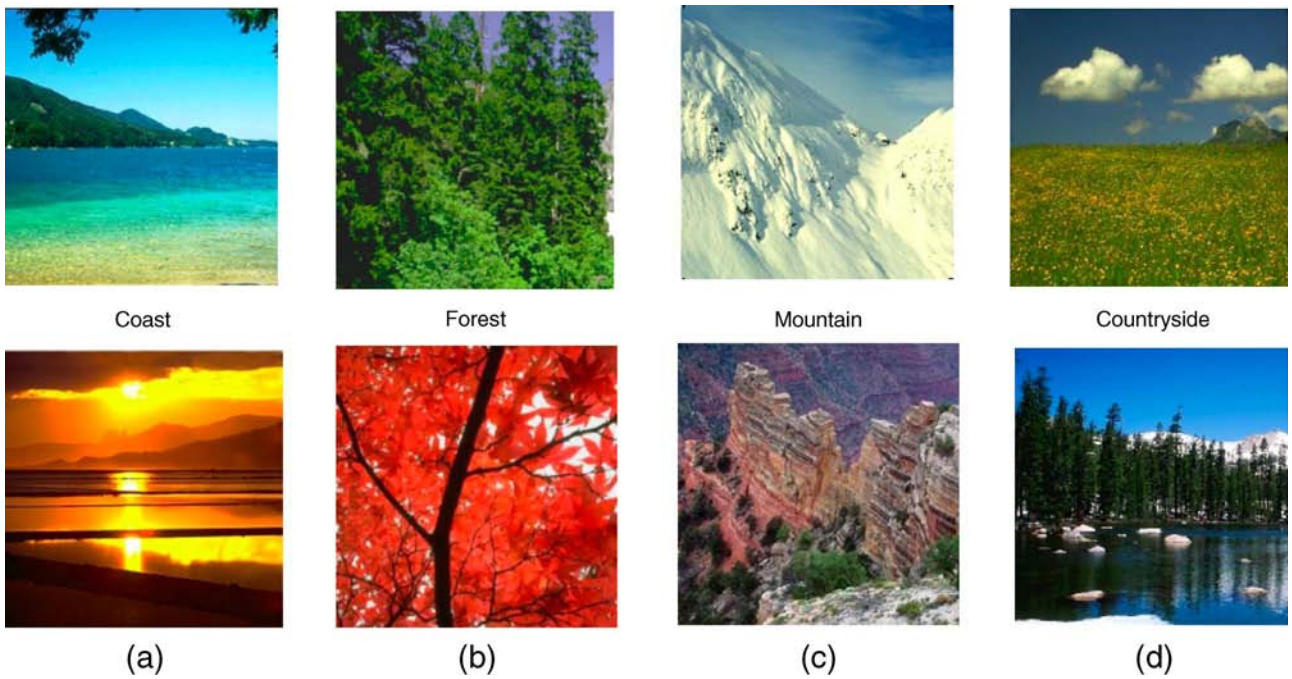


Figure 1. Example images in the dataset. Each column is an image pair in the same category to illustrate within-class variation.

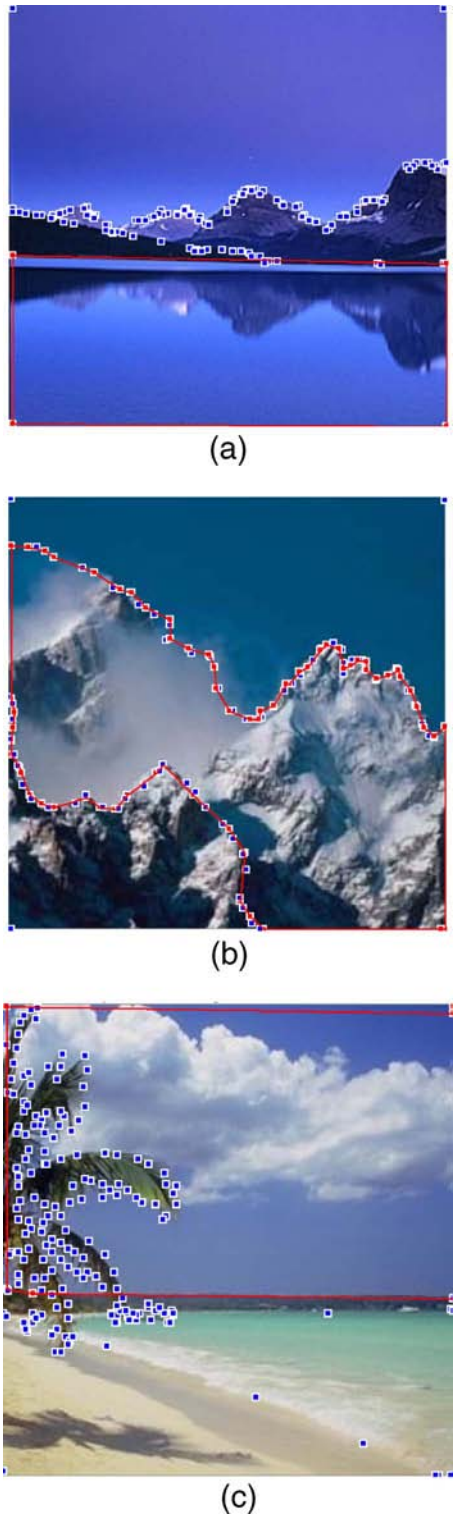


Figure 2. Red curves and blue spots circle labeled regions in human annotations. (a) Water–sky–mountain confusion due to reflection. (b) Cloud–rock confusion due to ill-defined texture boundary. (c) Leaf–cloud–sky confusion due to careless labeling.

for texture classification. The major issue is the poor segmentation. A related problem is that the label names are ambiguous if taken locally without a context. For example, a label 'water' can include a sky and mountains due to reflection (Figure 2a), and a label 'rock' can be confounded with clouds due to occlusion (Figure 2b). In addition, people tend to avoid tedious labeling in the cases of abundant occlusions or clutter (Figure 2c).

3. The ARTSCENE system

3.1 Overview

ARTSCENE consists of gist and texture subsystems (Figure 3). For gist, a 304-dimensional feature vector is constructed for each image (G in Equation 17 below), incorporating properties of orientations (O in Equation 15) and colors (C in Equation 16). A Default ARTMAP 2 classifier (Amis & Carpenter, 2007) learns recognition categories and an association between the gist category and its scene label. For texture, ARTSCENE identifies the largest labeled area (i.e., *first principal texture*) for each image and represents it by a 22-dimensional texture feature vector (T^{δ} in Equation 27). Again, Default ARTMAP 2 learns a recognition category and an association between the category and its scene label. The same procedure is applied to the second and third largest labeled regions in each image. The output of the texture system is the average of three scenic prediction vectors mapped from categories of principal textures (Equation 28). The system output is the most active scene class in the average of both gist and texture prediction vectors (Equation 29).

Complementary computing of boundaries and surfaces

As in the FACADE model (Grossberg, 1990, 1994), ARTSCENE computes both oriented boundary and unoriented surface color information. Boundaries and surfaces are defined in parallel processing streams by computationally *complementary* properties (Grossberg, 2000, 2003a, 2003b). In the brain, the boundary stream passes from LGN through V1 interblobs and V2 pale stripes to V4. The surface stream passes from LGN through V1 blobs and V2 thin stripes to V4. Boundaries pool signals from multiple achromatic and chromatic channels in order to define the strongest boundary signal possible. Pooling of signals from multiple detectors, including detectors with opposite contrast polarity, enables the boundaries of objects lying in front of textured backgrounds to be computed.

Due to the pooling of opposite contrast polarities, boundaries cannot represent visible brightness and color properties. Boundaries are predicted to be amodal, or

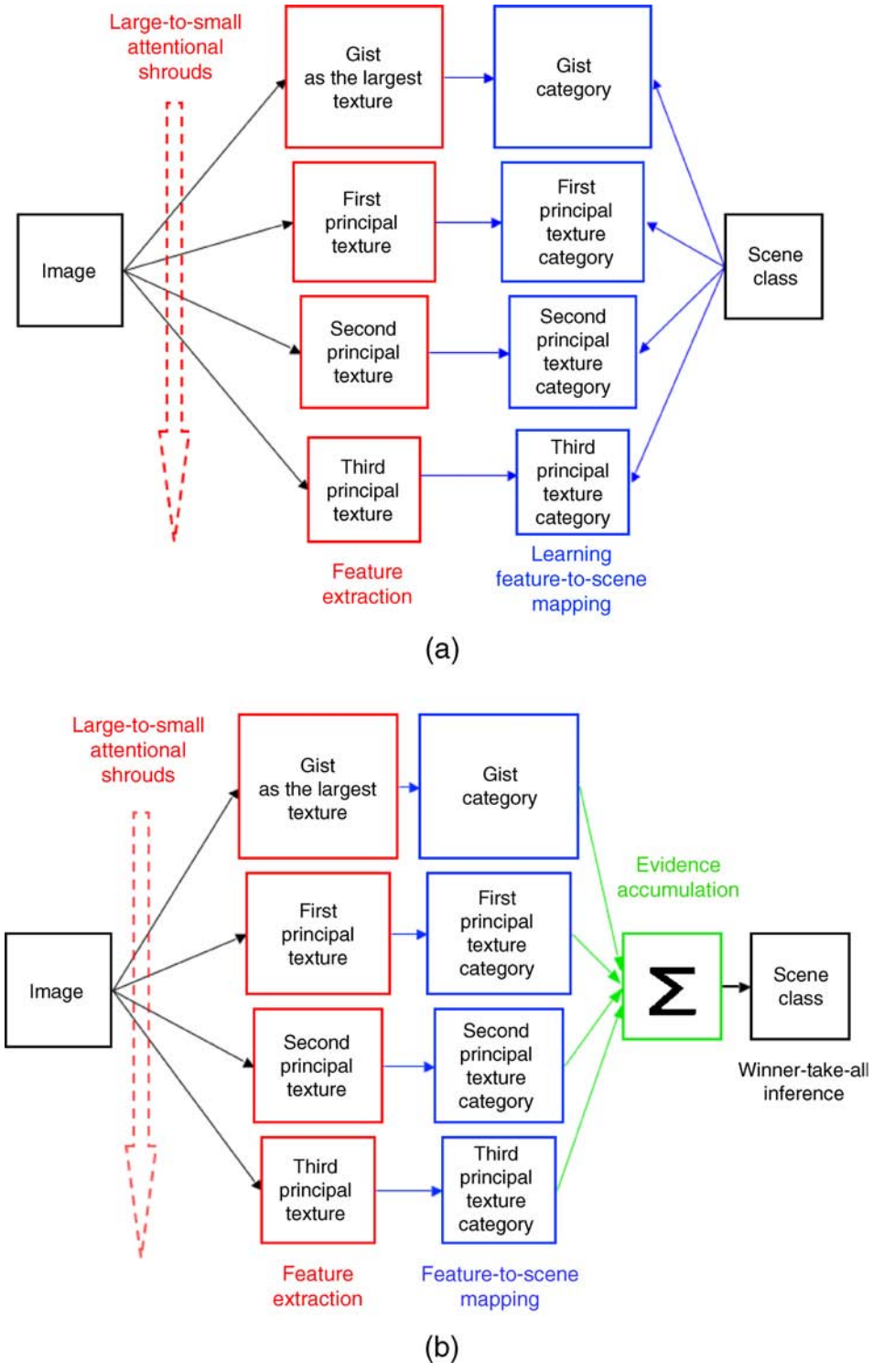


Figure 3. (a) ARTSCENE training mode. (b) ARTSCENE testing mode.

invisible, within the boundary processing stream. Visible percepts are processed within the surface stream, with the help of signals from the boundary stream. The surface stream segregates achromatic and chromatic signals in separate channels, and thereby defines the visible surface qualities that we see. The ARTSCENE model applies these properties to the problem of scene understanding.

3.2 Oriented boundary filtering

In ARTSCENE, multiple-scale oriented filtering is used to compute both gist and texture boundary properties. Such an early filtering computation can be achieved using a variety of possible kernels: Gabor functions (Daugman, 1980; Gabor, 1946), differences of normalized offset

oriented filters (Grossberg & Mingolla, 1985a, 1985b), log-Gabor functions (Field, 1987), differences of offset Gaussians (Grossberg & Todorović, 1988; Young, 1987), differences of offset differences of Gaussians (Parker & Hawken, 1988), Gaussian derivatives (Young, 1986), and a steerable pyramid (Simoncelli & Freeman, 1995). In ARTSCENE, oriented filtering is carried out by ON-cell and OFF-cell activities sampled through normalized differences of offset oriented Gaussians (Bhatt et al., 2007):

Stage 1: Color-to-gray image transformation

In the brain, boundaries pool signals from multiple color channels (Grossberg, 1994) to compute the strongest boundary possible at each position. In ARTSCENE, the values of three RGB channels are averaged:

$$I_{pq} = \frac{1}{3}\left(I^R_{pq} + I^G_{pq} + I^B_{pq}\right), \tag{1}$$

where p and q are pixel indices and I^R_{pq}, I^G_{pq}, I^B_{pq} are, respectively, the image intensities of red, green and blue channels.

Stage 2: Contrast normalization

This stage corresponds to early neural processing in the retina and lateral geniculate nucleus (LGN) that generates contrast signals using multiple-scales of antagonistic ON-cells (ON-center OFF-surround) and OFF-cells (OFF-center ON-surround) (Bhatt et al., 2007; Grossberg & Hong, 2006; Hubel & Wiesel, 1961). An on-center, I_{ij}, off-surround, $-S^g_{ijpq}I_{pq}$, shunting network normalizes local luminance for contrast enhancement:

$$\frac{d}{dt}x^g_{ij} = -x^g_{ij} + \left(1 - x^g_{ij}\right)I_{ij} - \left(1 + x^g_{ij}\right)\sum_{(p,q)} S^g_{ijpq}I_{pq}, \tag{2}$$

where x^g_{ij} is the normalized activity of the cell at position (i, j) with scale $g = 1, \ldots, 4$, the surround kernel S^g_{ijpq} is Gaussian:

$$S^g_{ijpq} = \frac{1}{2\pi\sigma^2_{sg}}\exp\left[-\frac{(i-p)^2 + (j-q)^2}{2\sigma^2_{sg}}\right], \tag{3}$$

and scale parameters $(\sigma_{s1}, \sigma_{s2}, \sigma_{s3}, \sigma_{s2}) = (1, 4, 8, 12)$. The LGN ON-cell and OFF-cell output signals are

$$X^{g+}_{ij} = [x^g_{ij}]^+, \tag{4}$$

and

$$X^{g-}_{ij} = [-x^g_{ij}]^+, \tag{5}$$

where the signal function $[x]^+ = \max(0, x)$ denotes half-rectification.

Stage 3: Contrast-sensitive oriented filtering

The third stage models oriented simple cells in primary visual cortical area V1 (Hubel & Wiesel, 1959) that are bottom-up activated by LGN ON and OFF activities (Hubel & Wiesel, 1962) sampled through spatially elongated and offset Gaussian kernels (see Figure 4). Each receptive field of simple cells consists of ON- and OFF-subregions. The ON-subregions receive excitatory ON LGN signals and inhibitory OFF LGN signals, while the OFF-subregions have the converse relation to the LGN channels (Hirsch, Alonso, Reid, & Martinez, 1998; Reid & Alonso, 1995). In particular, model V1 simple cell activity y^g_{ijk} at position (i, j), orientation k, and scale g obeys the shunting equation:

$$\frac{d}{dt}y^g_{ijk} = -\alpha y^g_{ijk} + \left(1 - y^g_{ijk}\right)\sum_{(p,q)}\left(X^{g+}_{pq}G^{g+}_{pqijk} + X^{g-}_{pq}G^{g-}_{pqijk}\right) - \left(1 + y^g_{ijk}\right)\sum_{(p,q)}\left(X^{g+}_{pq}G^{g-}_{pqijk} + X^{g-}_{pq}G^{g+}_{pqijk}\right), \tag{6}$$

where passive decay rate $\alpha = 1$. In the excitatory term of Equation 6, LGN ON-cell activities X^{g+}_{pq} are sampled by an oriented spatially elongated and offset Gaussian kernel G^{g+}_{pqijk}. LGN OFF-cell activities X^{g-}_{pq} are sampled by a similar kernel G^{g-}_{pqijk}. The centers of kernels G^{g+}_{pqijk} and G^{g-}_{pqijk} are offset in mutually opposite directions from each simple cell's centroid along an axis perpendicular to the simple cell's direction of elongated sampling. In the inhibitory term of Equation 6, the same kernels sample an LGN channel complementary to the one in the excitatory term. The net activity of simple cells is thus a measure of image feature contrast in its preferred orientation.

The oriented, elongated, and spatially offset kernels G^{g+}_{pqijk} and G^{g-}_{pqijk} in Equation 6 are:

$$G^{g+}_{pqijk} = \frac{1}{2\pi\sigma_{hg}\sigma_{vg}} \exp\left(-\frac{1}{2}\left\{\left[\frac{(p-i+m_k)\cos\left(\frac{\pi k}{4}\right) - (q-j+n_k)\sin\left(\frac{\pi k}{4}\right)}{\sigma_{hg}}\right]^2 + \left[\frac{(p-i+m_k)\sin\left(\frac{\pi k}{4}\right) + (q-j+n_k)\cos\left(\frac{\pi k}{4}\right)}{\sigma_{vg}}\right]^2\right\}\right), \tag{7}$$

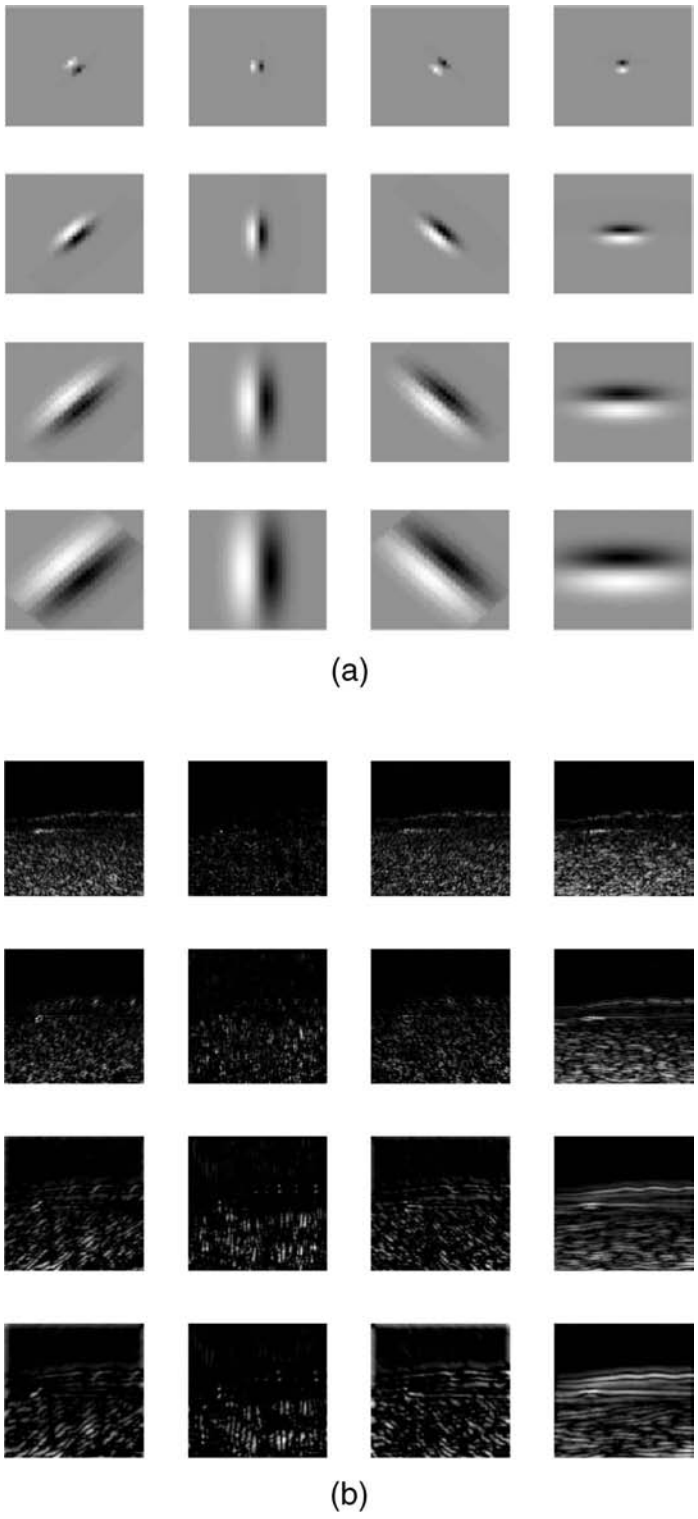


Figure 4. (a) Odd-symmetric filters used to model V1 simple neurons. (b) Corresponding filter responses to the coast image in Figure 5.

and

$$G^{g-}_{pqijk} = \frac{1}{2\pi\sigma_{hg}\sigma_{vg}} \exp\left(-\frac{1}{2}\left\{\left[\frac{(p-i-m_k)\cos\left(\frac{\pi k}{4}\right)-(q-j-n_k)\sin\left(\frac{\pi k}{4}\right)}{\sigma_{hg}}\right]^2 + \left[\frac{(p-i-m_k)\sin\left(\frac{\pi k}{4}\right)+(q-j-n_k)\cos\left(\frac{\pi k}{4}\right)}{\sigma_{vg}}\right]^2\right\}\right), \tag{8}$$

with offset vector $(m_k, n_k) = (\sin\frac{\pi k}{4}, \cos\frac{\pi k}{4})$ short-axis variance $(\sigma_{v1}, \sigma_{v2}, \sigma_{v3}, \sigma_{v4}) = (1/4, 1, 2, 3)$, and long-axis variance $(\sigma_{h1}, \sigma_{h2}, \sigma_{h3}, \sigma_{h4}) = (3/4, 3, 6, 9)$.

The outputs from model simple cells of opposite contrast polarity are half-rectified activities

$$Y^{g+}_{ijk} = [y^{g}_{ijk}]^+, \tag{9}$$

and

$$Y^{g-}_{ijk} = [-y^{g-}_{ijk}]^+, \tag{10}$$

where $[y]^+ = \max(y, 0)$.

Stage 4: Contrast-insensitive oriented filtering

This stage models oriented, contrast-polarity insensitive, and phase-insensitive V1 complex cells by pooling outputs from oriented simple cells of opposite contrast polarities (Hubel & Wiesel, 1959, 1962):

$$z^{g}_{ijk} = Y^{g+}_{ijk} + Y^{g-}_{ijk}. \tag{11}$$

Thus, model complex cells respond to oriented energy of either polarity.

Stage 5: Orientation competition at the same position

Contrast between orientations at the same pixel position is enhanced by a shunting on-center off-surround network across orientation at each position:

$$\frac{d}{dt}Z^{g}_{ijk} = -Z^{g}_{ijk} + \left(1 - Z^{g}_{ijk}\right)\sum_{\ell} z^{g}_{ij\ell} g^{+}_{\ell k} - \left(1 + Z^{g}_{ijk}\right)\sum_{\ell} z^{g}_{ij\ell} g^{-}_{\ell k}, \tag{12}$$

where the on-center kernel $g^{+}_{\ell k}$ and off-surround kernel $g^{-}_{\ell k}$ are 1D Gaussians:

$$g^{+}_{\ell k} = \frac{1}{\sigma^{+}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{\ell - k}{\sigma^{+}}\right)^{2} \right\}, \quad (13)$$

and

$$g^{-}_{\ell k} = \frac{1}{\sigma^{-}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{\ell - k}{\sigma^{-}}\right)^{2} \right\}, \quad (14)$$

with $\sigma^{+} = 0.5$ and $\sigma^{-} = 1$.

3.3 Surface color statistics

Color-pooled boundary information such as edge orientation computed in Oriented boundary filtering section is not the only feature informative for scene identification. Surface properties such as luminance or color are also useful cues for scene classification (e.g. Bosch et al., 2006; Szummer & Picard, 1998; Vailaya, Figueiredo, Jain, & Zhang, 2001; Vogel & Schiele, 2007). As noted above, the FACADE model (Grossberg, 1990, 1994) explains that boundary and surface properties are complementary (Grossberg, 2000) and interact to generate representations of brightness, color, depth, texture, and form. Consistent with this view, Oliva and Schyns (2000) conducted psychophysical experiments and confirmed that subjects bring color into play when it is a diagnostic scene attribute. They showed that reaction time (RT) decreases for normally colored displays and increases for abnormally colored ones when compared to the luminance-only condition in the (canyon, forest, coastline, desert) scene classification task. Color is thus part of the ARTSCENE feature vectors.

In the computer vision literature, color histogram is a well-established feature for both object and scene recognition (e.g. Swain & Ballard, 1991; Szummer & Picard, 1998; Vailaya, Jain, & Zhang, 1998; Vogel & Schiele, 2007). It is often constructed from the Red/Green/Blue (RGB) color space, the Hue/Saturation/Value (HSV) color space (Smith, 1978), the Ohta color space (Ohta, Kanade, & Sakai, 1980), or the Hue/Saturation/Intensity (HSI) color space (Keim & Kriegel, 1995). Other alternative color representations include color coherence vectors that further divide each bin in a color histogram into coherent and non-coherent pixels, based on whether or not a pixel is part of a large similarly colored region (Pass, Zabih, & Miller, 1997), color correlograms that compute the spatial correlation of pairs of colors as a function of the distance between pixels (Huang, Kumar, Mitra, Zhu, & Zabih, 1997), and color concentric circles of Scale-Invariant Feature Transform (Bosch et al., 2006).

In ARTSCENE, the mean RGB values, rather than color histograms, are computed to represent both gist and texture surface properties. These 3-dimensional color quantities are a direct generalization of the 1-dimensional brightness value crucial for texture classification in ARTEX (Grossberg & Williamson, 1999). Although this color-coding scheme is much more compressed than color histograms, simulations using both representations yielded comparable results (histograms therefore not shown) in model behavior and final classification performance.

3.4 Gist feature vector

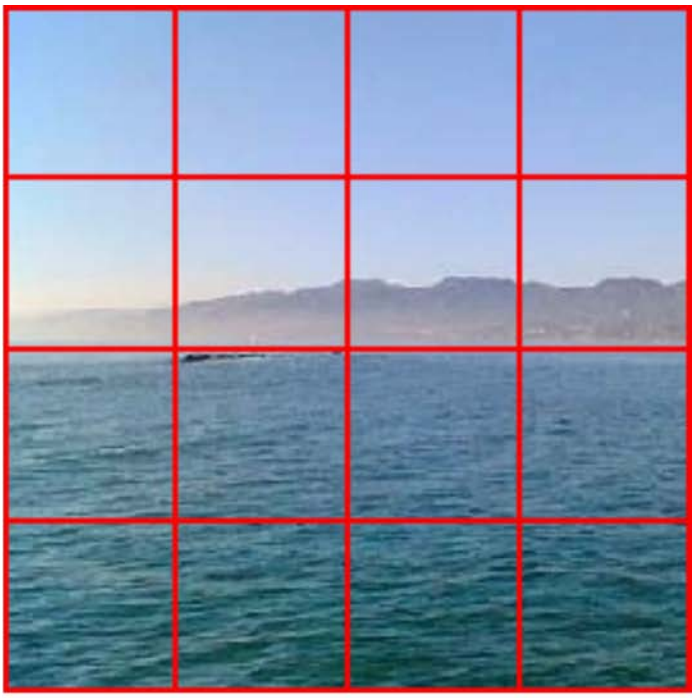
In ARTSCENE, the gist of a scene is defined and recognized as a global texture category. Previous studies have proposed a variety of perceptual dimensions to compute scene gist, such as mean depth, openness, expansion, degree of navigability, level of camouflage, degree of movement, and temperature (Greene & Oliva, 2006; Oliva & Torralba, 2001). In ARTSCENE, gist is computed using more basic properties that underlie general visual perception and categorization. Specifically, we propose that the brain learns and predicts regular scenic patterns via mechanisms of texture categorization operating at large scales for early rapid scene identification; cf. Renninger and Malik (2004). One such example would be the dominant orientation; that is, the visually most compelling orientation perceived by humans from natural textures (Picard & Gorkani, 1994) or city/suburb scenes (Gorkani & Picard, 1994). For example, the dominant energy is often horizontal in a coast scene due to the horizon and waves, vertical in a forest scene because of tall trees, and diagonal in a mountain scene due to ridges.

ARTSCENE assumes that boundary, surface, and spatial information is combined to represent scene gist. To compute this composite of boundary, surface, and spatial information, boundary and surface information is filtered within 16 evenly spaced local surface patches π (see Figure 5). Each of the 16 patches is characterized by the average values of four orientation contrasts at four different scales, yielding a 16-dimensional orientation vector, in addition to the average values of three RGB channels, yielding a 3-dimensional color vector. Mathematically, the components of the 16-dimensional orientation vector and 3-dimensional color feature vector of a surface patch π are:

$$O^{\pi g}_{k} = \frac{1}{|\pi|} \sum_{(i,j)\in\pi} Z^{g}_{ijk}, \quad (15)$$

and

$$C^{\pi\omega} = \frac{1}{|\pi|} \sum_{(p,q)\in\pi} I^{\omega}_{pq}, \quad (16)$$



(a)

(b)

Figure 5. (a) The 4 × 4 partition used for the grid gist representation. (b) Annotated regions in the LabelMe database used for texture representation. Compared with the three largest regions—'sea water', 'sky', and 'mountain', labels named 'houses occluded' and 'quay' are relatively obscure.

where the four orientations k = 1, 2, 3, 4; the scales g =1, 2, 3, 4; the colors $\omega = R, G, B$; and $|\pi|$ specifies the number of pixels in region π. The final gist feature vector G is a concatenation of 19 normalized $O_k^{\pi g}$ and $C^{\pi\omega}$ values across all 16 surface patches π:

$$G = \left(\frac{O_k^{\pi g}}{\sum_{\ell=1\ldots4} O_\ell^{\pi g}}, \frac{C^{\pi\omega}}{\sum_{v=\{R,G,B\}} C^{\pi v}} : k = 1,2,3,4; \quad \omega = R,G,B; \pi = 1,\ldots,16 \right). \tag{17}$$

In all, the gist vector G has 304 dimensions: a 19-dimensional orientation-and-color feature vector in each of 16 surface patches: (19 × 16 = 304).

We also tested another gist representation in which the only sub-area π was the whole image. In this case, the 19-dimensional feature vector G was a global average of different orientations and colors. To distinguish these two gist implementations in the later discussion, we call *grid gist* the representation with spatial partition, and *frame gist* the one without. The performance difference between the grid gist and frame gist representations can be attributed to the availability of spatially formatted boundary and surface features in one but not the other.

3.5 Attended texture feature vectors: Texture size ranking principle

A texture is a nearly homogeneous surface exhibiting certain statistical regularities, such as a clear sky, a piece of grass, or a body of rippled water. A texture itself can be a strong indicator of scene identity. For instance, a big white patch of rocks is very likely part of a snowy mountain. Other textures, such as the sky, are shared across several scene categories and not very predictive. A challenge for an efficient scene classifier is to discover and learn scene-specific texture categories.

Attentionally modulated learning of principle textures

We have found that principal textures, defined and ordered by their relative size in the visual field, are informative regions for landscape scene identification. We call this coarse-to-fine strategy the *texture size ranking principle*. In particular, it is sufficient to combine texture information from the three largest annotated regions in an image to achieve good scene recognition. These texture measures highly correlate with scene identity, and the correlation strength is proportional to the texture size (see Section 4, Table 1). Moreover, on average, three principal textures together constitute 92.7% of the total area of a

	Pure textures Mean ± STD Min-Med-Max	Shroud textures Mean ± STD Min-Med-Max	Box textures Mean ± STD Min-Med-Max
1st texture → Scene	71.39 ± 2.32% 66.30–71.47–77.17%	74.60 ± 1.97% 68.48–74.46–78.80%	74.80 ± 2.18% 69.02–75.00–79.62%
2nd texture → Scene	62.32 ± 2.45% 56.79–62.23–67.93%	66.31 ± 2.64% 59.51–66.30–72.01%	67.49 ± 2.40% 61.41–67.39–73.64%
3rd texture → Scene	55.20 ± 2.57% 50.00–55.16–62.77%	59.80 ± 2.10% 55.16–60.05–64.95%	61.80 ± 2.29% 56.79–61.68–66.30%
1st + 2nd textures → Scene	78.88 ± 1.94% 75.27–78.80–84.24%	81.77 ± 1.80% 77.17–81.52–86.41%	81.33 ± 2.09% 76.90–81.25–86.41%
1st + 2nd + 3rd textures → Scene	81.08 ± 1.73% 76.90–81.25–84.78%	83.05 ± 1.78% 79.08–82.88–86.96%	83.14 ± 1.57% 79.62–83.15–86.96%

Table 1. Predictive power of principal textures in different representations. Pure textures refer to principal textures originally annotated in the LabelMe database. Shroud textures are computed by extending the boundaries of pure textures to incorporate texture interfaces. Box textures are derived from the minimum bounding boxes of pure textures (see Figure 6 and Section 3.5 for details). Abbreviations: Std = standard deviation; Min = minimum; Med = median; Max = maximum.

landscape image in the dataset that we studied, and appear much more salient than small objects and textures, as illustrated in Figure 5. Attention shifts thus have a 92.7% likelihood of falling within these regions during free viewing.

We thus hypothesize that humans deploy spatial attention and make eye movements onto these principle textures to refine scene identification, and that the search order tends to be from large textures to small textures for most efficient evidence accumulation. This Texture Size Ranking Principle is implemented algorithmically in ARTSCENE by sequentially processing one principal texture in an image at a time during learning, recognition, and evidence accumulation. For example, in Figure 5, ARTSCENE scans in sequence through the regions of "sea water," "sky," and "mountain" in the order of their relative size.

In ARTSCENE, spatial attention defines an information window that masks out information outside the window. Bhatt et al. (2007) and Fazl et al. (2009) have shown how an attentional spotlight (e.g. Eriksen & Yeh, 1985; LaBerge, 1995; Moran & Desimone, 1985; Posner & Petersen, 1990) can spread into a form-fitting shroud of spatial attention (Tyler & Kontsevich, 1995) that selects an entire textured region, while down-regulating other scenic regions. We assume that such a shroud selects the texture-specific filtered quantities that comprise a texture feature vector. Several types of experiments, including studies of change blindness (Rensink, 2000; Simons & Levin, 1997) and scene perception (Schyns & Oliva, 1994), have indicated that not all scenic information is available in a glimpse, and scene gist mainly delivers coarsely coded information (Bar, 2004). In the same spirit, we construct the scene gist in Equation 17 by averaging orientations and colors over the regions π, which implicitly leaves out the fine-scale image content.

Bounding box attentional windows

We compare results using region-fitting attentional shrouds with a simpler attentional window, δ, that is defined to be the minimum bounding box of a principal texture (box textures in Figure 6). The 2D spatial extents of δ range from $\min(x_k)$ to $\max(x_k)$ in the x direction and from $\min(y_k)$ to $\max(y_k)$ in the y direction, where (x_k, y_k) are polygon vertices of the chosen texture in the LabelMe database (see Figure 5b and Section 2.2).

This approach relaxes the need for perfect texture segmentation and is commonly used for object recognition in real scenes (Everingham, Zisserman, Williams, & Van Gool, 2006). Our simulations show (see Section 4) that ARTSCENE classification works well with this segmentation scheme. In particular, a 16-dimensional orientation vector and 3-dimensional color feature vector for region δ are defined by:

$$O_k^{\delta g} = \frac{1}{|\delta|} \sum_{(i,j)\in\delta} Z_{ijk}^g, \tag{18}$$

and

$$C^{\delta\omega} = \frac{1}{|\delta|} \sum_{(p,q)\in\pi} I_{pq}^{\omega}, \tag{19}$$

where the four orientations $k = 1, 2, 3, 4$; the scales $g = 1, 2, 3, 4$; the colors $\omega = R, G, B$; and $|\delta|$ specifies the number of pixels in region δ.

In addition to orientation and color, we also incorporate spatial factors—notably, the region area, A^{δ}, and the region centroid $(P_x^{\delta}, P_y^{\delta})$—into the texture feature vector. Since the polygon vertex coordinates from the LabelMe database, (x_k, y_k), specify boundary pixels rather than

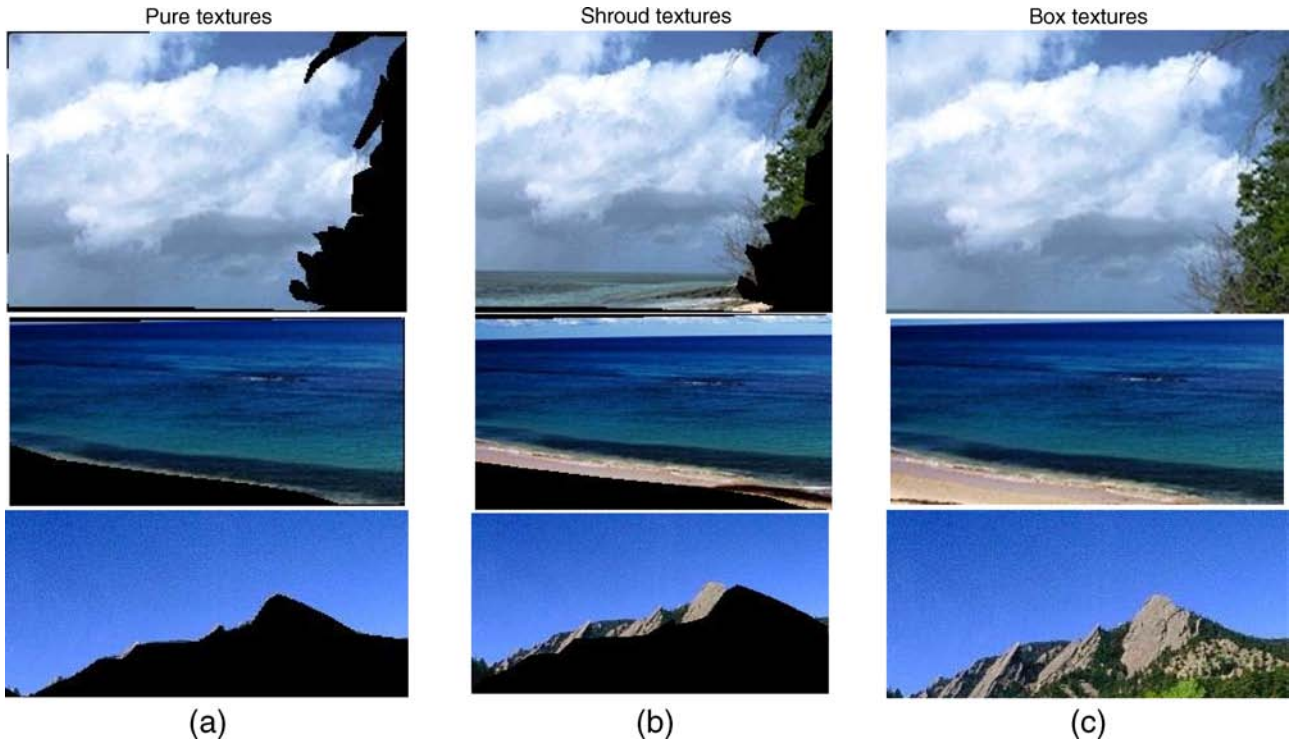


Figure 6. Examples of three different texture representations. From top to bottom are the sky in a coast, the sea in a coast, and the sky in a mountain scene. From left to right are the well-segmented pure textures, the fringy shroud textures, and the minimum bounding box textures (see Section 3.5 for details).

surface pixels of a region (see Section 2.2 and Figure 2), we derive these quantities from the following equations, instead of image moments (Pratt, 2007):

$$A^{\delta} = |\delta| = (\max_k x_k - \min_k x_k)(\max_k y_k - \min_k y_k), \quad (20)$$

$$P_x^{\delta} = (\max_k x_k + \min_k x_k)/2, \quad (21)$$

and

$$P_y^{\delta} = (\max_k y_k + \min_k y_k)/2. \quad (22)$$

The rationale here is to discriminate visually similar textures using ecological constraints in a scene. For example, a clear blue sky is hardly distinguishable from a surface of stationary water if taken out of context. In this case, the texture centroid $(P_x^{\delta}, P_y^{\delta})$ in a scene is informative because the sky often occupies the upper visual field, whereas water usually occurs in the lower field. As for the texture area A^{δ}, the same texture may occupy different portions of a scene, depending upon the scene category. For instance, the sky tends to be large in both 'coast' and 'countryside' scenes, but small in a 'forest' due to tree occlusion. We envisage these distinctions as being part of the spatial information available to the brain when studying a scene.

Surface-fitting attentional shrouds

To better understand the learning and recognition effects of attentional focusing on surfaces in space, we also implement the attentional window, δ, in the form of an annotated principal texture with or without a fringe (see pure and shroud textures in Figure 6, respectively), and contrasts the aforementioned bounding box approach (box textures in Figure 6). In other words, an *attentional shroud* is used to demarcate object surfaces. The concept of such a shroud was introduced by Tyler and Kontsevich (1995), who used it to explain how a surface map morphs to account for momentarily available depth cues during 3D vision. In our usage, a *pure texture* is a form-fitting distribution of spatial attention that completely covers the texture. A *shroud texture* is a form-fitting distribution of spatial attention that includes a fringe around the pure texture region.

Bhatt et al. (2007) and Fazl et al. (2009) further developed the shroud concept to predict how an attentional shroud can control learning of view-invariant object categories, as well as texture categories. For computa-

tional simplicity, ARTSCENE replaces the attentional shroud dynamics that are simulated in Fazl et al. (2009) with the algorithmically pre-segmented surface regions that define principal textures. ARTSCENE learns statistical measures of each such surface region, as well as the region area, A^{δ}, and the region centroid $(P_x^{\delta}, P_y^{\delta})$ of a principal texture (Press, Teukolsky, Vetterling, & Flannery, 2007):

$$A^{\delta} = |\delta| = \frac{1}{2}\left|\sum_{k=0}^{N-1}(x_k y_{k+1} - x_{k+1}y_k)\right|, \tag{23}$$

$$P_x^{\delta} = \frac{1}{6A^{\delta}}\left|\sum_{k=0}^{N-1}(x_k + x_{k+1})(x_k y_{k+1} - x_{k+1}y_k)\right|, \tag{24}$$

and

$$P_y^{\delta} = \frac{1}{6A^{\delta}}\left|\sum_{k=0}^{N-1}(y_k + y_{k+1})(x_k y_{k+1} - x_{k+1}y_k)\right|, \tag{25}$$

where $(x_N, y_N) = (x_0, y_0)$, coordinates (x_k, y_k) are polygon vertices used to define a region in the LabelMe database (see Figure 6), and N is the number of such vertices for a certain region label. In the simulations of shroud textures, we first obtain the centroid of a principal texture, $(P_x^{\delta}, P_y^{\delta})_{\mathrm{OLD}}$, via Equations 24 and 25, and extend the shroud boundary by scaling the distance between the centroid and each vertex $(x_k, y_k)_{\mathrm{OLD}}$:

$$(x_k, y_k)_{\mathrm{NEW}} = 1.3 \cdot \left[(x_k, y_k)_{\mathrm{OLD}} - (P_x^{\delta}, P_y^{\delta})_{\mathrm{OLD}}\right]. \tag{26}$$

The new region centroid and area are then computed via Equations 23–25.

It should be noted that Equations 20–22 are special cases of Equations 23–25 if (x_k, y_k) are the four corners of a region bounding box. Here we use different equations to highlight the geometric differences between these two implementations of the attentional window. The final 22-dimensional texture feature vector T^{δ} is a concatenation of normalized $C_{\omega b}^{\delta}$, O_{gk}^{δ}, P_x^{δ}, P_y^{δ}, and A^{δ} values:

$$T^{\delta} = \left(\frac{C_b^{\delta\omega}}{\sum_{\omega,b} C_b^{\delta\omega}}, \frac{O_k^{\delta g}}{\sum_{k=1..4} O_k^{\delta g}}, \frac{P_x^{\delta}}{256}, \frac{P_y^{\delta}}{256}, \frac{A^{\delta}}{256^2}\right). \tag{27}$$

If the attentional windows, δ, are bounding boxes of principal textures, we call the selected regions *box textures* to distinguish them from *pure textures* that are the exact polygons from the LabelMe database, and *shroud textures* that are the extended regions that surround pure textures.

3.6 Default ARTMAP 2 classifier

Default ARTMAP 2 (Amis & Carpenter, 2007), the latest version of the ARTMAP classifier family, was used in ARTSCENE to learn gist and texture categories $\boldsymbol{w}_j$ from feature vectors $\mathbf{f}$ (see Equations A1–A3 and A12), where $\mathbf{f} = G$ for gist features (Equation 17) and $\mathbf{f} = T^{\delta}$ for texture features (Equation 27). ARTMAP also learns the associations W_{jk} between these categories and scene labels K to compute prediction vectors ψ_k, both for gist predictions ψ_k^G and texture predictions $\psi_k^{T\delta}$ from region δ (Equations A4, A9, and A21).

ARTMAP illustrates how humans can incrementally and stably learn to categorize items in an ever-changing world by matching bottom-up inputs and top-down expectations (Carpenter & Grossberg, 1991). In Default ARTMAP 2, the only free parameter is the baseline vigilance $\bar{\rho}$, which controls how general the learned categories will be (Equations A5, A8, and A10). Low $\bar{\rho}$ causes learning of abstract and general categories, whereas high $\bar{\rho}$ enables concrete and sharp discriminations to be learned.

Although Default ARTMAP 2 is trained using winner-take-all activation of category nodes, it can also generate distributed predictions of class likelihood (ψ_k in Equation A21), which enables the model to achieve hierarchical information fusion and cognitive rule discovery (Carpenter, Martens, & Ogas, 2005). In ARTSCENE, we collect such distributed predictions from Default ARTMAP 2 modules across scales for more general model averaging. Mathematically, the final prediction vector ψ_k^T from the texture system is:

$$\psi_k^T \equiv \frac{1}{3}\sum_{\delta=1}^{3} \psi_k^{T\delta}, \tag{28}$$

where k specifies the scene class and vectors $\psi_k^{T\delta}$ are the scenic predictions generated by each principal texture δ. Together with the gist prediction vector ψ_k^G, the final output of ARTSCENE is the scene class label K^* that is the most active scene node:

$$K^* = \arg\max_{k=1,\ldots,4} (\psi_k^G + \psi_k^T), \tag{29}$$

with the corresponding class label K to which K^* is associated during supervised learning trials.

4. Simulation results

To evaluate model performance and robustness on all 1472 images, we ran simulations 100 times based on different training-testing splits. For each simulation, three quarters of the images were randomly chosen for training, and the remaining quarter was used for testing. The baseline vigilance $\bar{\rho}$ was set to 0.8 for both training and testing. This value achieved the optimal validation performance in a parametric study of $\bar{\rho}$ ranging from 0 to 0.9 with a spacing of 0.1. In fact, the ARTSCENE performance was qualitatively unchanged as a function of $\bar{\rho}$. In Tables 1 and 2, model categorization performance is summarized by mean, standard deviation, median, and range of overall percentage correct over these 100 simulations.

Table 1 summarizes the predictive power of principal textures and compares the performance difference among three texture representations. Individual principal textures correlate with scene identity and thereby their classification performances are all better than chance (25%). However, such correlation declines as the texture size decreases (one-tailed pairwise t-test, $p < 0.01$). This trend is also reflected in the reduced gain when we incrementally combine smaller and smaller principal textures to make a better inference (one-tailed pairwise t-test, $p < 0.01$). Table 1 also shows that box and shroud textures, without gist, lead to comparable results and carry more scenic information than pure textures. All simulations using box and shroud textures resulted in better classification performances than ones using pure textures (one-tailed pairwise t-test, $p < 0.01$). The marginal effect presumably comes from the interface information between two adjacent textures. For example, a water texture alone may suggest coast as well as countryside. However, water and sand together form a higher-order texture—beach—that is only associated with coast. Built upon these diagnostic local regions, ARTSCENE averages the prediction vectors from three principal textures in a scene to be the output of the texture system (see Equation 28).

Table 2 summarizes how well gist predicts a scene and how much the texture information improves this prediction. Here the comparison between gist and gist-plus-texture performances quantitatively simulates how the subsequent spatial attention shifts, possibly supported by eye movements in vivo, would enable sequential scrutiny of local textures and thereby refine the hypothesis of scene identity inferred from the global gist in the first glimpse. For all simulations, the predictive power of frame gist is notably worse than grid gist in terms of classification rate because global averaging omits local statistics and under-represents an image. However, the performance boost after gist-texture integration (one-tailed pairwise t-test, $p < 0.01$) is more pronounced for frame gist than grid gist, which agrees with the notion that active vision helps to minimize expectation uncertainty, especially when gist is less sure of scene identity. Note that box and shroud textures again lead to comparable results and the performance advantage of using box or shroud textures over pure textures is less marked after gist-texture integration. It is because the information of texture interfaces is available not only through box and shroud textures but also through gist by definition (see Section 3.4). Finally, for all texture representations, the grid gist-plus-texture predictions outperform predictions from either grid gist or textures alone (one-tailed pairwise t-test, $p < 0.01$), and the gist plus three shroud textures achieved the maximal accuracy of 91.85%.

To understand where misclassification happens, we constructed Table 3 that breaks down overall performance into its component categorical performances from 100 simulations using grid gist plus three shroud textures on different testing sets. Table 3 is the confusion matrix in which each row represents a ground-truth class, each column represents a predicted class, and each cell reports the proportion of a predicted scene label on all testing images from a given scene class. Therefore, the diagonal terms are the percentages of correctly classified images, and the off-diagonal terms are the percentages of misclassified images. The first and second numbers in each table cell are

	Frame gist Mean ± Std Min-Med-Max	Grid gist Mean ± Std Min-Med-Max
Gist → Scene	77.14 ± 2.15% 70.38–77.17–82.07%	85.08 ± 1.72% 80.98–85.05–90.22%
Gist + 3 pure textures → Scene	82.13 ± 1.96% 75.82–82.07–87.50%	85.74 ± 1.64% 80.43–85.60–91.03%
Gist + 3 shroud textures → Scene	82.43 ± 1.93% 78.26–81.93–87.50%	86.60 ± 1.54% 82.88–86.41–91.85%
Gist + 3 box textures → Scene	82.54 ± 1.84% 77.99–82.34–86.96%	86.55 ± 1.64% 82.07–86.68–91.58%

Table 2. Categorization performance of gist and texture integration. Grid gist refers to the gist with 4 × 4 spatial partition and frame gist is the one without it (see Section 3.4 for details). Abbreviations: Std = standard deviation; Min = minimum; Med = median; Max = maximum.

Truth\predicted	Coast	Forest	Mountain	Countryside
Coast	79.94%	0.55%	1.02%	18.49%
	81.72%	0.55%	0.55%	17.17%
Forest	0%	87.83%	7.52%	4.65%
	0%	89.27%	5.88%	4.85%
Mountain	0.62%	1.19%	88.33%	9.86%
	0.37%	1.72%	89.25%	8.66%
Countryside	9.26%	1.86%	4.47%	84.41%
	8.86%	2.47%	2.35%	86.32%

Table 3. Confusion matrix before and after attentional shifts to three principal textures. The first and second numbers in each table cell are the prediction performances using gist and gist-plus-texture, respectively. Each row is a ground-truth category and each column is a predicted category. See Section 4 for detailed discussion.

the results from gist-alone and gist-plus-texture predictions, respectively. They separately simulate human scene recognition before and after attention to principal textures. In Table 3, the second number in a cell is greater than the first one for all diagonal cells but vice versa for most off-diagonal cells. This indicates that the gist-plus-texture predictions are generally better than gist-alone predictions, and confirms the functional benefit of attention to principal textures, notably attentional shrouds, in scene recognition.

ARTSCENE tends to misclassify 'coast' as 'countryside', 'forest' as 'mountain', 'mountain' as 'countryside', and 'countryside' as 'coast'. A post-hoc image examination reveals that the confusion between 'forest' and 'mountain' comes mostly from the co-occurrence of trees and mountains, and the confusion between 'countryside' and the other three categories is due to the loose definition of 'countryside'. We can gain better insight into these confusions from Figure 7, which shows some misclassified images in the best simulation. Significantly, these images are also ambiguous to humans and the model well captures that ambiguity.

5. Discussion and conclusions

In Table 2, the most favorable ARTSCENE representation scheme is the combination of grid gist plus three

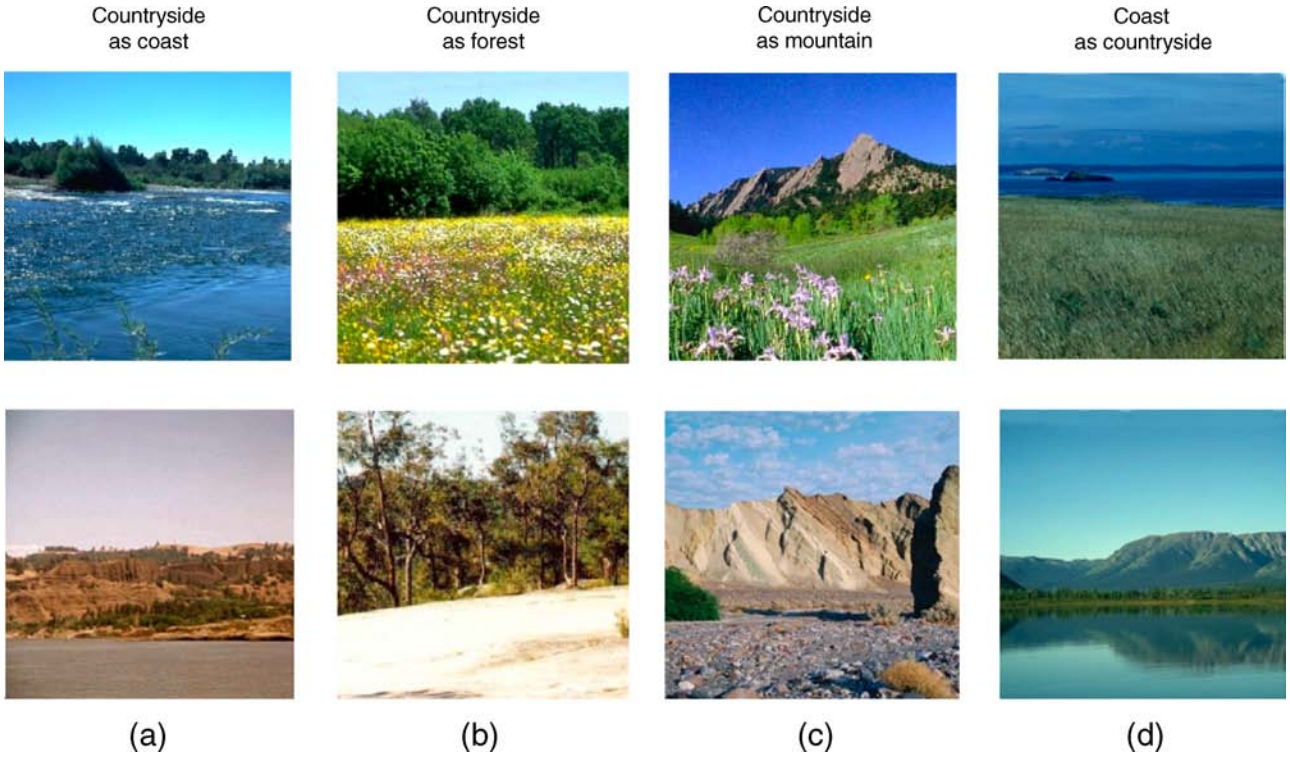


Figure 7. Misclassified testing images in the ARTSCENE best simulation. Countryside images are misclassified as coasts in the first column, as forests in the second column, and as mountains in the third column. Coast images are misclassified as countrysides in the last column. Note that these images are intrinsically ambiguous to categorize.

shroud textures. It leads to the best mean classification rate, 86.60%, as well as the best peak performance for the four landscape scene classification of 91.85%, which is around 1.5% higher than the best benchmark in the literature—namely, the peak performance of 90.28% in Bosch et al. (2006)—using the same image dataset. This represents a reduction in error of 16.15%. The performance fluctuation in 100 simulations of the same scheme is in part due to the known dependence of ARTMAP learning on input presentation order, and in part due to the different degrees of exposure to ambiguous instances during learning (see Figure 7 for example images).

These results derive from efficient deployment of attention on locally computed multi-scale boundary and surface information that has proved to be necessary for explaining a wide range of other visual phenomena (Bhatt et al., 2007; Cao & Grossberg, 2005; Fazl et al., 2009; Grossberg & Hong, 2006; Grossberg, Kuhlmann, & Mingolla, 2007; Grossberg & Swaminathan, 2004; Grossberg & Yazdanbakhsh, 2005). ARTSCENE hereby further develops the idea of integrating global and local information for scene recognition, which was discussed by Vogel et al. (2006). In their computational study, scene gist was represented by spatial envelope (Oliva & Torralba), and region-based scene semantics were represented by a 10 × 10 grid of texture layout that required supervised texture learning and classification. They found a mean classification rate was 52% for simulations using gist alone and of 73.6% using both gist and region-based semantics. Their results were obtained on a different dataset consisting of five landscape scene categories (coasts, forests, mountains, plains, and rivers/lakes). Our results and their together illustrate why a state-of-the-art scene classifier needs to take into account information from multiple scales including both global gist and local textures. Our results also expand the growing number of studies showing a useful role for attentional shrouds in object, texture, and scene learning and recognition (Bhatt et al., 2007; Fazl et al., 2009).

For learning, we use ARTMAP classifiers because they are capable of fast, incremental, stable learning of recognition categories and predictions in response to non-stationary data streams, and can automatically discover the proper degree of category generalization in response to changing environmental statistics. As noted above, all the major predictions of ART since its introduction in Grossberg (1976a, 1976b) have received increasing support from psychological, neuropsychological, and neuroanatomical data over the years (Grossberg, 2003b; Raizada & Grossberg, 2003). The use of ART as a gist and texture classifier is thus compatible with a biological account of scene understanding.

Compared with scene classifiers that use either fixed gist templates or texture vocabulary, one strength of ARTSCENE is that it can adaptively update its internal category representations for all scenic predictors across scales, including multiple textures and gist, which is critical for on-line use. A human-predefined gist or texture vocabulary often demands significant human labor in search of common elements in the image dataset, as in the models of Oliva and Torralba (2001) and Vogel and Schiele (2007). Although the search can be replaced by machine learning schemes (Bosch et al., 2006; Fei-Fei & Perona, 2005), such vocabularies often require rebuilding from scratch to learn a new instance due to the use of batch learning schemes such as k-means. In these approaches, even if the vocabulary construction is replaced by incremental learning, the scene decomposition in terms of the new vocabulary and subsequent processes still need to be re-calculated for every image due to the vocabulary update. In addition, the distributed predictions in ARTMAP allow ARTSCENE to naturally perform multi-category classification and information integration across scales. In contrast to the Vogel and Schiele (2007) and Vogel et al. (2006) use of the Support Vector Machine (SVM) to carry out pairwise comparisons of scene likelihoods, ARTSCENE is free from combinatorial explosion when more scene categories are introduced into the task.

A weakness of the current implementation is the use of LabelMe polygon coordinates (see Section 2.2 and Section 3.5). However, simulations in Table 2 show that both box and fringy shroud textures yield slightly better mean performance than human segmentations (i.e., pure textures). These results suggest that perfect texture segmentation is not needed to achieve good performance on scene classification. This opens the possibility in future studies of replacing LabelMe with machine segmentations wherein principal textures along with their centroids and areas are still well defined (see Equations 21–23).

Another possible extension of the model is to include an object system to learn associations between salient learned object categories (see Fazl et al., 2009) and scene labels (see reviews in Bar, 2004). Since our model framework is essentially a mixture of experts, the system can generalize to accommodate more scenic predictors, including coherent objects. Such a generalization is now being pursued.

6. Appendix A

6.1 Default ARTMAP 2 (Amis & Carpenter, 2007)

The Default ARTMAP 2 algorithm specifies two modes of operation: Training and testing. The untrained ARTMAP network begins with a pool of uncommitted category nodes that are not bound to any class label. As learning progresses, nodes from this pool are recruited, or committed, to encode feature patterns for learned categories (see Equation A11 in the training procedure below). Thus, the population of committed category nodes grows with learning, and its size C is determined by task demands. In all simulations presented in this article, the training

procedure was repeated 3 times for the same training set to stabilize learning and consolidate feature categories. Different numbers of repetitions can be used and lead to qualitatively unchanged model behavior. For testing, a feature is compared to each learned feature category (Equation A16) and activates the category nodes in proportion to its similarity with those categories (Equations A19 and A20). The distributed output predictions are then computed by the learned mapping from feature category activations to class labels (Equation A21).

6.2 Training, with distributed next-input test

1. The M-dimensional feature vector $\mathbf{f} = (f_1, f_2, \ldots, f_M)$ represents the activities of input ON cells. It causes the corresponding OFF cells to attain the values

$$\mathbf{f}^c = 1 - \mathbf{f}. \tag{A1}$$

The total $2M$-dimensional input vector

$$F \equiv (\mathbf{f}, \mathbf{f}^c), \tag{A2}$$

is said to be *complement coded*. The L^1 norm of $\boldsymbol{F}$ is normalized at the value M.

2. Set initial values: Assign 1 to the mapping w_{ij} from feature vector F_i in the vector $\boldsymbol{F} = (F_1, F_2, \ldots, F_{2M})$ to category j for all $i = 1,\ldots,2M$ and $j = 1,\ldots,C$. Assign 0 to the mapping W_{jk} from category j to output class label k. Assign 1 to the number of committed category nodes C.

3. Select the first input vector $\boldsymbol{F}$. Associate it with the output class label K.

4. Set learned weights for the newly committed category $j = C$:

$$w_C = F, \tag{A3}$$

and

$$W_{CK} = 1. \tag{A4}$$

5. Set vigilance ρ to its baseline value $\overline{\rho} = 0.8$:

$$\rho = \overline{\rho}\,. \tag{A5}$$

6. Reset all category activities:

$$y = 0. \tag{A6}$$

7. Select the next input vector from the training set in randomized order. Associate it with the output class label K. Do this recursively until the last input of the last training epoch is presented.

8. Calculate feature-to-category matching signals T_j for committed category nodes $j = 1,\ldots,C$ using the choice-by-difference signal function (Carpenter, 1997):

$$T_j = |F \wedge w_j| + (1 - \alpha)(M - |w_j|). \tag{A7}$$

In Equation A7, $\wedge$ denotes the fuzzy intersection: $(\boldsymbol{F} \wedge \boldsymbol{w_j})_k = \min(F_k, w_{ik})$, $|.|$ denotes the L^1 norm, $(\boldsymbol{w_j})_i = w_{ij}$ is the learned weight vector for category j, and parameter $\alpha = 0.01$ specifies the preference for more local categories when more than one coded category equally matches the input feature vector.

9. Search order: Sort the committed coding nodes with $T_j > \alpha M$ in order of T_j values from max to min.

10. Search for a category J that meets the matching criterion and predicts the correct output class label K, as follows:

(a) Code: For the next sortedcategory ($j = J$) that meets the matching criterion:

$$\left(\frac{|F \wedge w_J|}{M} \geq \rho\right), \tag{A8}$$

set $y_J = 1$ and $y_k = 0$, $k \neq J$ (winner-take-all).

(b) Output class prediction:

$$\psi_k = \sum_{j=1}^{C} y_j W_{jk} = W_{Jk}. \tag{A9}$$

(c) Correction prediction: If the active code J predicts the output class label $K(\psi_K = W_{JK} = 1)$, go to Step (12) (learning).

(d) Match tracking: If the active code J fails to predict the correct output class ($\psi_K = 0$), raise the vigilance to:

$$\rho = \frac{|F \wedge w_J|}{M} + \varepsilon, \tag{A10}$$

where the match tracking parameter $\varepsilon = -0.001$. Term ε permits the system to code inconsistent cases, where two identical training set inputs are associated with different outcomes (Carpenter, Milenova, & Noeske, 1998), which is common in human annotated databases. Return to Step (10a) and continue the memory search.

11. After unsuccessfully searching the sorted list, increase C by 1 (add a committed node):

$$C = C + 1. \tag{A11}$$

Return to Step (4).

12. Learning: Update coding weights:

$$w_J^{\text{new}} = \beta(F \wedge w_J^{\text{old}}) + (1-\beta)w_J^{\text{old}}, \tag{A12}$$

where β is the learning rate and $\boldsymbol{w}_j^{\text{old}}$ is the previously learned weight vector for category j. In the present simulations, the choice $\beta = 1$ ensured fast learning.

13. Distributed next-input test: verify that the input makes the correct prediction with distributed coding:

(a) Make prediction: Generate an output class prediction K^* for the current training input $\boldsymbol{F}$ using distributed activation, as prescribed for testing (compare with Equation 29):

$$K^* = \arg\max_k \psi_k. \tag{A13}$$

(b) Correct prediction: If distributed activation predicts class label K, return to Step (5) (next input).

(c) Match tracking: If distributed activation fails to predict the correct output class label ($K^* \neq K$), raise the vigilance:

$$\rho = \frac{|F \wedge w_J|}{M} + \varepsilon. \tag{A14}$$

Return to Step (10a) (continue search).

6.3 Default ARTMAP testing (Distributed code)

1. Complement code M-dimensional test set feature vectors $\mathbf{f}$ to produce $2M$-dimensional input vectors $\boldsymbol{F} \equiv (\mathbf{f}, \mathbf{f}^{\mathbf{c}})$.

2. Select the next input vector $\boldsymbol{F}$ from the testing set in randomized order. Associate it with the output label K.

3. Reset the category activities:

$$y = 0. \tag{A15}$$

4. Calculate feature-to-category matching signals T_j for committed category nodes $j = 1,\ldots,C$:

$$T_j = |F \wedge w_j| + (1-\alpha)(M - |w_j|), \tag{A16}$$

where parameter $\alpha = 0.01$, as during training.

5. Define Λ as the set of indices of categories satisfying the matching criterion $T_\lambda > \alpha M$:

$$\Lambda = \{\lambda = 1,\ldots,C : T_\lambda > \alpha M\}, \tag{A17}$$

and Λ' as the set of indices of categories perfectly matching the input:

$$\begin{aligned}\Lambda' &= \{\lambda = 1,\ldots,C : T_\lambda = M\} \\ &= \{\lambda = 1,\ldots,C : w_j = F\}.\end{aligned} \tag{A18}$$

6. Increased Gradient (IG) CAM Rule: The Increased Gradient (IG) CAM rule contrast-enhances the input differences in the distributed category code (Carpenter, 1997; Carpenter et al., 1998):

(a) The point box case occurs when at least one category exactly encodes the input. The activities y_j of such categories are then uniform: If $\Lambda' \neq \varphi$ (i.e., $\boldsymbol{w}_j = \boldsymbol{F}$ for some j), set

$$y_j = \frac{1}{|\Lambda'|}, \tag{A19}$$

for each $j \in \Lambda'$.

(b) In cases other than a point box code, a distributed category activation is computed for categories satisfying the match criterion:

$$y_j = \frac{\left[\frac{1}{M - T_j}\right]^p}{\sum_{\lambda \in \Lambda}\left[\frac{1}{M - T_\lambda}\right]^p}, \tag{A20}$$

for each $j \in \Lambda$, where the power law parameter $p = 1$ determines the amount of code contrast enhancement. As p increases, the category activation increasingly resembles a winner-take-all code in that only the category with highest bottom-up signal survives.

7. Calculate distributed output class predictions:

$$\psi_k = \sum_{j=1}^{C} y_j W_{jk}. \tag{A21}$$

8. Until the last test input, return to Step (2).

7. Acknowledgments

We wish to thank the reviewers for their helpful comments. Both authors are supported in part by the National Science Foundation (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624).

Commercial relationships: none.
Corresponding author: Stephen Grossberg.
Email: steve@bu.edu.
Address: Department of Cognitive and Neural Systems, Center for Adaptive Systems, Center of Excellence for Learning in Education, Science, and Technology, Boston University, 677 Beacon Street, Boston, MA 02215, USA.

8. Author's Note

Co-first authors SG and TRH contributed equally to this work.

9. Editor's note

This paper was submitted and reviewed as part of the special issue on Perceptual organization and neural computation http://www.journalofvision.org/8/7/i/.

10. References

Amis, G., & Carpenter, G. (2007). Default ARTMAP 2. *Proceedings of the International Joint Conference on Neural Networks (IJCNN'07)* (pp. 777–782). Orlando, Florida: IEEE Press.

Bar, M. (2004). Visual objects in context. *Nature Reviews, Neuroscience, 5,* 617–629. [PubMed]

Bhatt, R., Carpenter, G. A., & Grossberg, S. (2007). Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research, 47,* 3173–3211. [PubMed]

Bosch, A., Zisserman, A., & Muñoz, X. (2006). Scene classification via pLSA. *Proceedings of the European Conference on Computer Vision, 4,* 517–530.

Cao, Y., & Grossberg, S. (2005). A laminar cortical model of stereopsis and 3D surface perception: Closure and da Vinci stereopsis. *Spatial Vision, 18,* 515–578. [PubMed]

Carpenter, G. A. (1997). Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Networks, 10,* 1473–1494. [PubMed]

Carpenter, G. A., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. Cambridge: The MIT Press.

Carpenter, G. A., Martens, S., & Ogas, O. J. (2005). Self-organizing information fusion and hierarchical knowledge discovery: A new framework using ARTMAP neural networks. *Neural Networks, 18,* 287–295.

Carpenter, G. A., Milenova, B. L., & Noeske, B. W. (1998). Distributed ARTMAP: A neural network for fast distributed supervised learning. *Neural Networks, 11,* 793–813. [PubMed]

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research, 20,* 847–856. [PubMed]

Eriksen, C. W., & Yeh, Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 583–597. [PubMed]

Everingham, M., Zisserman, A., Williams, C., & Van Gool, L. (2006). The PASCAL visual object classes challenge 2006 (VOC2006) results. Retrieved from http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/results.pdf.

Fazl, A., Grossberg, S., & Mingolla, E. (2009). View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology, 58,* 1–48. [PubMed]

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR'97) Conference, 2,* 524–531.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, Optics and Image Science, 4,* 2379–2394. [PubMed]

Gabor, D. (1946). Theory of communication. *Institution of Electrical Engineers, 93,* 429–457.

Gorkani, M. M., & Picard, R. W. (1994). Texture orientation for sorting photos at a glance. *Proceedings of the IEEE Computer Vision and Pattern Recognition, 1,* 459–464.

Greene, M. R., & Oliva, A. (2006). Natural scene categorization from conjunctions of ecological global properties. *Proceedings of the 28th Annual Conference of the Cognitive Science Society,* 291–296.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134. [PubMed]

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics, 23,* 187–202. [PubMed]

Grossberg, S. (1990). Neural FACADEs: Visual representations of static and moving Form-And-Color-and-DEpth. *Mind and Language, 5,* 411–456.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics, 55,* 48–121. [PubMed]

Grossberg, S. (2000). The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences, 4,* 233–246. [PubMed]

Grossberg, S. (2003a). Filling-in the forms: Surface and boundary interactions in visual cortex. In L. Pessoa & P. DeWeerd (Eds.), *Filling-in: From perceptual*

completion to skill learning (pp. 13–37). New York: Oxford University Press.

Grossberg, S. (2003b). How does the cerebral cortex work? Development, learning, attention, and 3-D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews, 2,* 47–76. [PubMed]

Grossberg, S., & Hong, S. (2006). A neural model of surface perception: Lightness, anchoring, and filling-in. *Spatial Vision, 19,* 263–321. [PubMed]

Grossberg, S., Kuhlmann, L., & Mingolla, E. (2007). A neural model of 3D shape-from-texture: Multiple-scale filtering, boundary grouping, and surface filling-in. *Vision Research, 47,* 634–672. [PubMed]

Grossberg, S., & Mingolla, E. (1985a). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review, 92,* 173–211. [PubMed]

Grossberg, S., & Mingolla, E. (1985b). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Perception & Psychophysics, 38,* 141–171. [PubMed]

Grossberg, S., & Swaminathan, G. (2004). A laminar cortical model for 3D perception of slanted and curved surfaces and of 2D images: Development, attention, and bistability. *Vision Research, 44,* 1147–1187. [PubMed]

Grossberg, S., & Todorović, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: A unified model of classical and recent phenomena. *Perception & Psychophysics, 43,* 241–277. [PubMed]

Grossberg, S., & Williamson, J. R. (1999). A self-organizing neural system for learning to recognize textured scenes. *Vision Research, 39,* 1385–1406. [PubMed]

Grossberg, S., & Yazdanbakhsh, A. (2005). Laminar cortical dynamics of 3D surface perception: Stratification, transparency, and neon color spreading. *Vision Research, 45,* 1725–1743. [PubMed]

Hirsch, J. A., Alonso, J. M., Reid, R. C., & Martinez, L. M. (1998). Synaptic integration in striate cortical simple cells. *Journal of Neuroscience, 18,* 9517–9528. [PubMed] [Article]

Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J., & Zabih, R. (1997). Image indexing using color correlograms. *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR'97).*

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148,* 574–591. [PubMed] [Article]

Hubel, D. H., & Wiesel, T. N. (1961). Integrative action in the cat's lateral geniculate body. *The Journal of Physiology, 155,* 385–398. [PubMed] [Article]

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160,* 106–154. [PubMed] [Article]

Keim, D. A., & Kriegel, H. P. (1995). Issues in visualizing large databases. Proc. Conf. on Visual Database Systems (VDB-3), Lausanne, Schweiz, März 1995. In *Visual database systems* (pp. 203–214). London: Chapman & Hall Ltd.

LaBerge, D. (1995). *Attentional processing: The brain's art of mindfulness*. Cambridge, MA: Harvard University Press.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *International Conference on Computer Vision, 2,* 1150–1157.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229,* 782–784. [PubMed]

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9,* 353–383.

Ohta, Y., Kanade, T., & Sakai, T. (1980). Color information for region segmentation. *Computer Graphics and Image Processing, 13,* 222–241.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41,* 176–210. [PubMed]

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175.

Parker, A. J., & Hawken, M. J. (1988). Two-dimensional spatial structure of receptive fields in monkey striate cortex. *Journal of the Optical Society of America A, Optics and Image Science, 5,* 598–605. [PubMed]

Pass, G., Zabih, R., & Miller, J. (1997). Comparing images using color coherence vectors. *Proceedings of the Fourth ACM International Conference on Multimedia,* 65–73.

Picard, R. W., & Gorkani, M. (1994). Finding perceptually dominant orientations in natural textures. *Spatial Vision, 8,* 221–253. [PubMed]

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13,* 25–42. [PubMed]

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965–966. [PubMed]

Pratt, W. K. (2007). *Digital image processing: PIKS scientific inside* (4th ed.). Hoboken, NJ: Wiley-Interscience.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). Cambridge, UK: Cambridge University Press.

Raizada, R. D., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral Cortex, 13,* 100–113. [PubMed] [Article]

Reid, R. C., & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature, 378,* 281–284. [PubMed]

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition. *Vision Research, 44,* 2301–2311. [PubMed]

Rensink, R. A. (2000). Seeing, sensing, and scrutinizing. *Vision Research, 40,* 1469–1487. [PubMed]

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2005). LabelMe: A database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science, 5,* 195–200.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. *Proceedings of International Conference on Image Processing, 3,* 444–447.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1,* 261–267.

Smith, A. R. (1978). Color gamut transform pairs. *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques,* 12–19.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision, 7,* 11–32.

Szummer, M., & Picard, R. W. (1998). Indoor–outdoor image classification. *IEEE International Workshop on Content-based Access of Image and Video Databases, 98,* 42–51.

Tyler, C. W., & Kontsevich, L. L. (1995). Mechanisms of stereoscopic processing: Stereoattention and surface perception in depth reconstruction. *Perception, 24,* 127–153. [PubMed]

Vailaya, A., Figueiredo, M. T., Jain, A. K., & Zhang, H. J. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing, 10,* 117–130. [PubMed]

Vailaya, A. J., Jain, A. K., & Zhang, H. J. (1998). On image classification: City vs. landscape. *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries,* 3–8.

Vogel, J., & Schiele, B. (2007). Semantic scene modeling and retrieval for content-based image retrieval. *International Journal of Computer Vision, 72,* 133–157.

Vogel, J., Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2006). Categorization of natural scenes: Local vs. global information. *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization, 153,* 33–40.

Young, R. A. (1986). Simulation of human retinal function with the Gaussian derivative model. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'86),* 564–569.

Young, R. A. (1987). The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spatial Vision, 2,* 273–293. [PubMed]