

2007-11

ARTSCENE: A Neural System for Natural Scene Classification

<https://hdl.handle.net/2144/1955>

"Downloaded from OpenBU. Boston University's institutional repository."

ARTSCENE: A Neural System for Natural Scene Classification

Stephen Grossberg and Tsung-Ren Huang*

Department of Cognitive and Neural Systems
Center for Adaptive Systems
Center for Excellence in Learning, Science and Technology
Boston University
677 Beacon Street
Boston, MA 02215
Phone: 617-353-7858
Fax: 617-353-7755

Corresponding Author: Stephen Grossberg, steve@bu.edu,

Submitted November, 2007
CAS/CNS Technical Report 2007-017

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

* Authorship in alphabetical order. Both SG and TH are partially supported in part by the National Science Foundation (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624).

Copyright © 2007

Permission to copy without fee all or part of this material is granted provided that: 1. The copies are not made or distributed for direct commercial advantage; 2. the report title, author, document number, and release date appear, and notice is given that copying is by permission of the BOSTON UNIVERSITY CENTER FOR ADAPTIVE SYSTEMS AND DEPARTMENT OF COGNITIVE AND NEURAL SYSTEMS. To copy otherwise, or to republish, requires a fee and / or special permission.

Abstract

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, ranging from global gist to local properties of textures. The model can incrementally learn and predict scene identity by gist information alone and can improve performance through selective attention to scenic textures of progressively smaller size. ARTSCENE discriminates 4 landscape scene categories (coast, forest, mountain and countryside) with up to 91.58% 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. Model simulations also show that adjacent textures form higher-order features that are also informative for scene recognition.

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 & Torralba (2001) used spectral templates that correspond to global scene descriptors such as roughness, openness, and ruggedness. Fei-Fei & Perona (2005) decomposed a scene into local common luminance patches or textons. Bosch, Zisserman, & Muñoz (2006) applied the Scale-Invariant Feature Transform (SIFT) 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, & Bühlhoff (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.

The ARTSCENE model assumes that global information is quickly available before local information is acquired using attention 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 ~150ms that is less than the average fixation time (~300ms) (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.

The challenges of the model are thus to clarify what constitutes scene gist, where and what scale to look at next, 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 drawn to the scene's principal textures 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 generate useful predictions about the world. Since gist is just one of several textures in our treatment, ARTSCENE may be viewed as a generalization of the dARTEX texture classifier (Bhatt, Carpenter, & Grossberg, 2007). ARTSCENE also adapts heuristics of the ARTSCAN model of invariant object learning (Fazl, Grossberg, & Mingolla, 2007) 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 system is defined mathematically and simulation results are presented. Finally, the 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 & Torralba (2001) that has also been used by other researchers (e.g., Fei-Fei & Perona, 2005; Bosch et al., 2006). 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 256x256 pixels. Figure 1 shows 8 exemplars in the dataset and illustrates the great variation within each scene category.

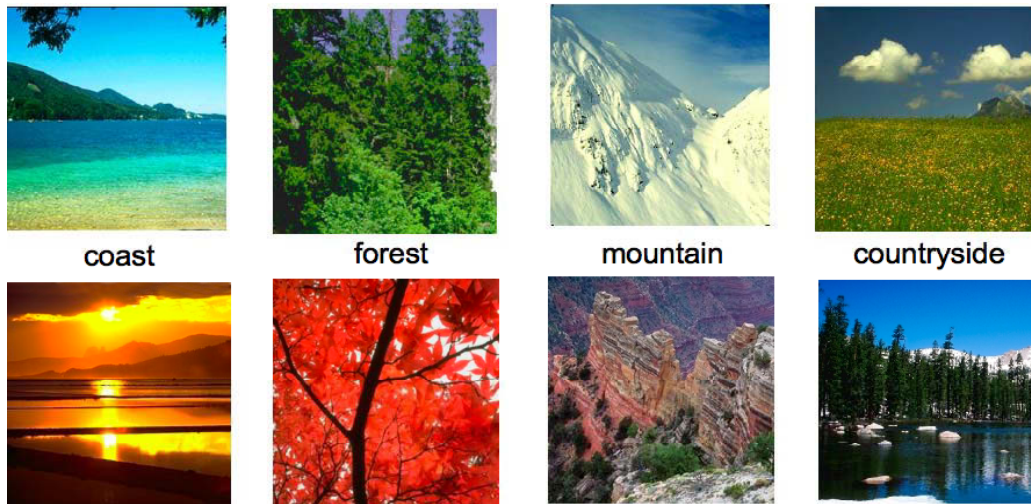


Figure 1: Example images in the dataset. Each column is an image pair in the same category to illustrate within-class variation.

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 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).

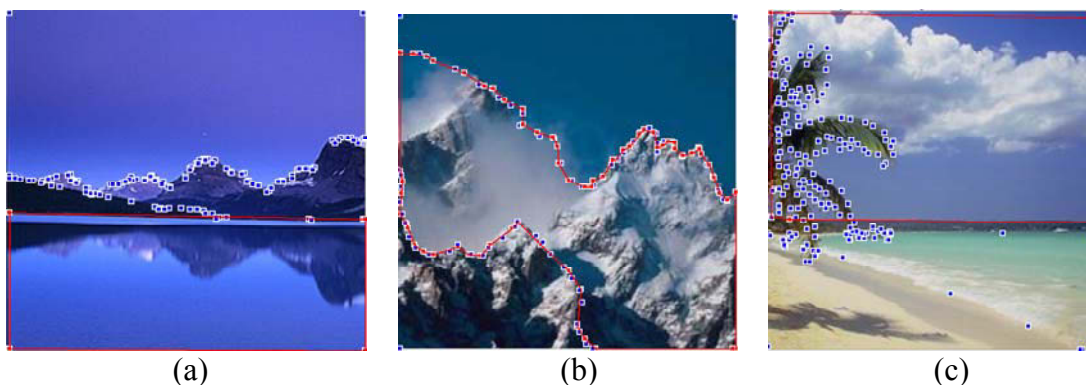
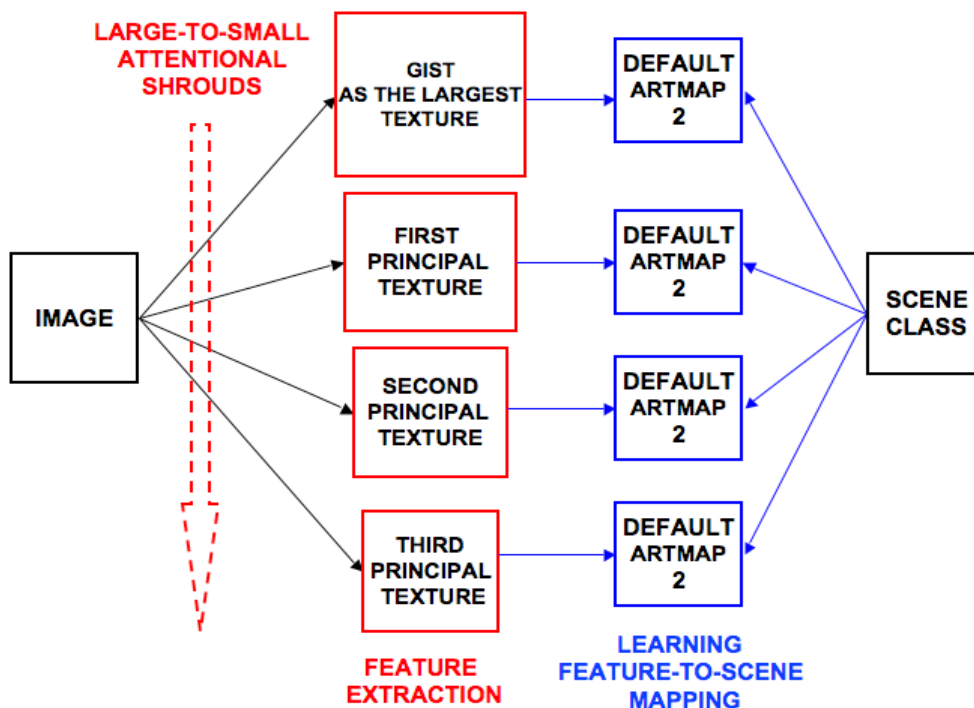


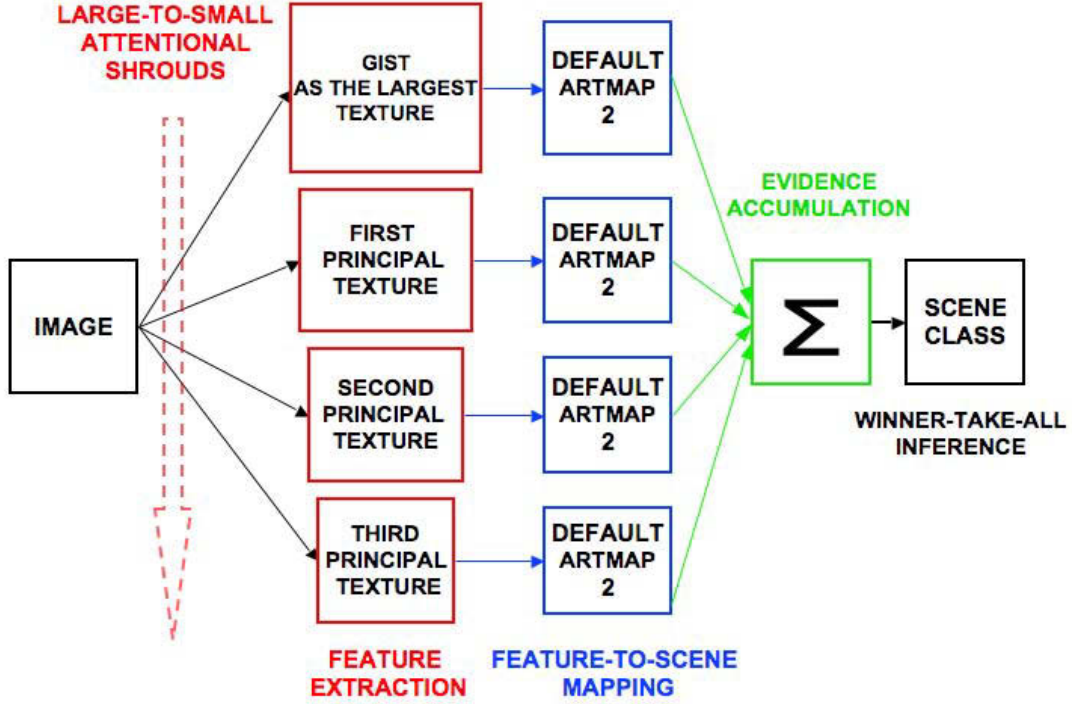
Figure 2: Red curves and blue spots circle labeled regions in human annotations. (a) water-sky-mountain confusion due to reflection; (b) cloud-rock confusion due to ill-defined texture boundary; (c) leaf-cloud-sky confusion due to careless labeling.

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 43-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)).



(a)



(b)

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

3.2 Oriented boundary filtering. As in the FACADE model, ARTSCENE computes both oriented boundary and unoriented surface color information (Grossberg, 1990; Grossberg, 1994). Multiple-scale oriented boundary filtering is used to compute both gist and textures. Oriented filtering is carried out by simplified dARTEX operations (Bhatt, Carpenter, & Grossberg, 2007):

Stage 1 : Color-to-Gray image transformation. In the brain, boundaries pool signals from multiple color channels (Grossberg, 1994). To obtain grayscale images for boundary processing, the values of three RGB channels are averaged:

$$I_{pq} = \frac{1}{3}(I_{pq}^R + I_{pq}^G + I_{pq}^B), \quad (1)$$

where p and q are pixel indices and I_{pq}^R , I_{pq}^G , I_{pq}^B 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). An on-center, I_{ij} , off-surround, $-S_{ijpq}^g I_{pq}$, shunting network normalizes local luminance for contrast enhancement:

$$\frac{d}{dt} x_{ij}^g = -x_{ij}^g + (1 - x_{ij}^g) I_{ij} - (1 + x_{ij}^g) \sum_{(p,q)} S_{ijpq}^g I_{pq}, \quad (2)$$

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

$$S_{ijpq}^g = \frac{1}{2\pi\sigma_{sg}^2} \exp\left[-\frac{(i-p)^2 + (j-q)^2}{2\sigma_{sg}^2}\right] \quad (3)$$

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

$$X_{ij}^{g+} = [x_{ij}^g]^+ \quad (4)$$

and

$$X_{ij}^{g-} = [-x_{ij}^g]^+, \quad (5)$$

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 that are bottom-up activated by LGN ON and OFF activities sampled through spatially elongated and offset positive semi-definite elongated Gaussian kernels (see Figure 4). In particular, V1 simple cell activity y_{ijk}^g at position (i, j) , orientation k , and scale g obeys the shunting equation:

$$\frac{d}{dt} y_{ijk}^g = -\alpha y_{ijk}^g + (1 - y_{ijk}^g) \sum_{(p,q)} (X_{pq}^{g+} G_{pqijk}^{g+} + X_{pq}^{g-} G_{pqijk}^{g-}) - (1 + y_{ijk}^g) \sum_{(p,q)} (X_{pq}^{g+} G_{pqijk}^{g-} + X_{pq}^{g-} G_{pqijk}^{g+}) \quad (6)$$

where passive decay rate $\alpha = 1$. In the excitatory term of Equation (6), LGN ON activities X_{pq}^{g+} are sampled by an oriented spatially elongated and offset Gaussian kernel G_{pqijk}^{g+} . LGN OFF channel activities X_{pq}^{g-} are sampled by a similar kernel G_{pqijk}^{g-} . The centers of kernels G_{pqijk}^{g+} and G_{pqijk}^{g-} 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_{pqijk}^{g+} and G_{pqijk}^{g-} in Equation (6) are:

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

and

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

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 $(\sigma_{h1}, \sigma_{h2}, \sigma_{h3}, \sigma_{h4}) = (3/4, 3, 6, 9)$.

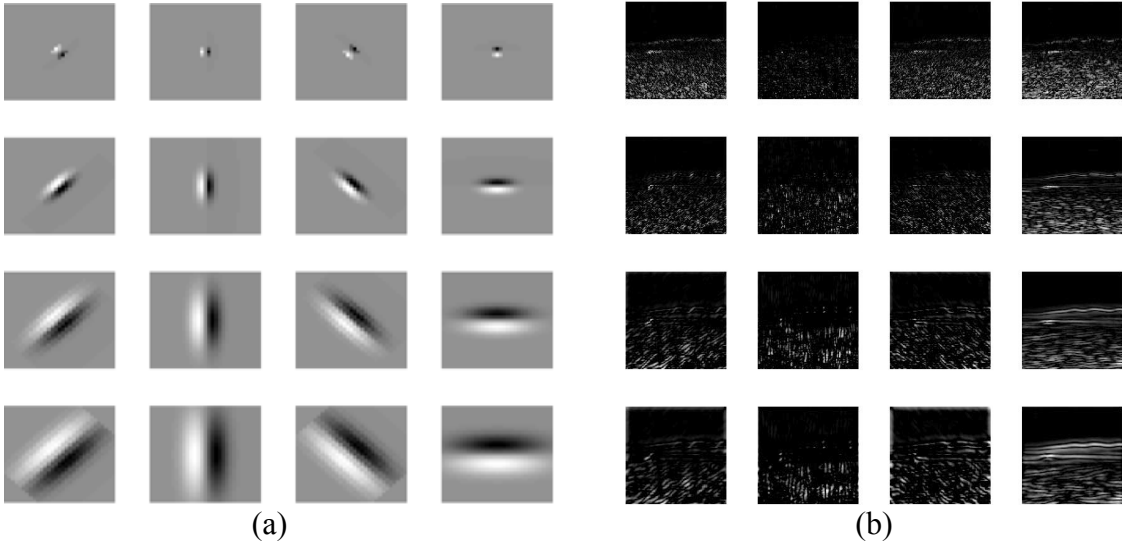


Figure 4: (a) Odd-symmetric filters used to model V1 simple neurons; (b) filter responses to the coast image in Figure 5.

The outputs from model simple cells of opposite contrast polarity are

$$Y_{ijk}^{g^+} = [y_{ijk}^g]^+ \quad (9)$$

and

$$Y_{ijk}^{g^-} = [-y_{ijk}^g]^+ \quad (10)$$

Stage 4 : Contrast-insensitive oriented filtering. This stage models contrast-invariant oriented V1 complex cells by pooling outputs from simple cells of opposite contrast polarities:

$$z_{ijk}^g = Y_{ijk}^{g^+} + Y_{ijk}^{g^-} \quad (11)$$

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 in orientation space:

$$\frac{d}{dt} Z_{ijk}^g = -Z_{ij}^g + (1 - Z_{ijk}^g) \sum_{\ell} z_{ij\ell}^g g_{\ell k}^+ - (1 + Z_{ijk}^g) \sum_{\ell} z_{ij\ell}^g g_{\ell k}^- \quad (12)$$

where $g_{\ell k}^+$ and $g_{\ell k}^-$ are 1D Gaussians:

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

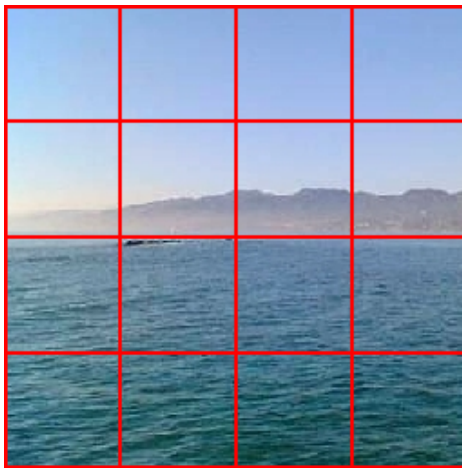
and

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

where $\sigma^+ = 0.5$ and $\sigma^- = 1$.

3.3 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 (Oliva & Torralba, 2001; Greene & Oliva, 2006). In our opinion, 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 example, the dominant orientation 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.

Boundary information such as edge orientation is not the only feature used in scene identification. Surface properties such as luminance or color are informative, although it has been argued that achromatic information is sufficient for scene recognition (Fei-Fei & Perona, 2005). The FACADE model (Grossberg, 1990; Grossberg, 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 & 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.



(a)



(b)

Figure 5: (a) The 4x4 partition used for the grid gist representation; five annotated regions used for texture representation. Compared with ‘sea water’, ‘sky’ and ‘mountain’, ‘houses occluded’ and ‘quay’ are relatively obscure.

To incorporate a composite of boundary, surface, and spatial information, we represent scene gist as an ensemble of sixteen evenly spaced local surfaces each of which is characterized by the average values of four orientation contrasts at four different scales and three RGB channels. Thus, the gist vector has 304 dimensions (see Figure 5). Mathematically, the 16-dimensional orientation vector and 3-dimensional color feature vector of a region π in the partition are:

$$O_k^{\pi g} = \frac{1}{|\pi|} \sum_{(i,j) \in \pi} Z_{ijk}^g \quad (15)$$

and

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

where $\omega = \{R, G, B\}$ and $|\pi|$ specifies the number of pixels in region π . The final gist feature vector G is a concatenation of normalized $O_k^{\pi g}$ and $C^{\pi \omega}$ values across all the region π .

$$G = \left(\frac{O_k^{\pi g}}{\sum_{\ell=1 \dots 4} O_\ell^{\pi g}}, \frac{C^{\pi \omega}}{\sum_{v=\{R,G,B\}} C^{\pi v}} \right), \quad (17)$$

where $\pi = 1, \dots, 16$. 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.

3.4 Texture feature vector: 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.

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*. This postulate is consistent with the observation that, on average, three principal textures together constitute 92.7% of the total area of a landscape image in the dataset that we studied, and appear 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. Bhatt, Carpenter, & Grossberg (2007) have shown how such an attentional spotlight can spread into a form-fitting shroud of spatial attention that selects the entire textured region, while down-regulating other scenic regions. We assume, as in Bhatt, Carpenter, & Grossberg, (2006), that such a shroud organizes texture-specific average quantities that comprise a texture feature vector.

In ARTSCENE, spatial attention is computed as an information window that masks out information outside the window. For gist, when attention is spread throughout the whole visual field, not all scenic information is available due to competitive normalization processes that prevent fine scales from being sufficiently activated. As illustrated by Equation (17), our

construction of gist uses a coarse-coding scheme that overcomes the ‘‘curse of dimensionality’’ by averaging orientations and colors over the regions π . Consequently, it is possible to further exploit scenic information by focusing a spatial attentional shroud on salient regions such as principal textures. For simplicity, we define the attention window, δ , to be the minimum bounding box of a principal texture in the ARTSCENE texture system (Figure 6b). 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 6a and Section 2.2). This approach relaxes the need for perfect texture segmentation. Our simulations show (see Section 4) that ARTSCENE classification works well with this segmentation scheme. In particular, as in the gist computation, the 16-dimensional orientation vector and 24-dimensional color feature vector for region δ are defined by:

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

and

$$C_b^{\delta \omega} = \frac{1}{|\delta|} n(S_b^{\delta \omega}) \quad (19)$$

where $\omega = \{R, G, B\}$, $b = 1, \dots, 8$, and $|\delta|$ specifies the number of pixels in region δ , and $n(S_b^{\delta \omega})$ is the number of elements in the set:

$$S_b^{\delta \omega} = \left\{ I_{pq}^{\omega} : \frac{b-1}{8} \leq I_{pq}^{\omega} < \frac{b}{8} \quad \forall (p,q) \in \delta \right\}. \quad (20)$$

We have tested bin numbers other than 8 for the color histogram $S_b^{\delta \omega}$ in Equation (20). Too many or too few bins both yielded comparable but less ideal classification rates.

In addition to orientation and color, we also incorporate spatial factors – notably, the region centroid $(P_x^{\delta}, P_y^{\delta})$, and the region area, A^{δ} – into the texture feature vector:

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

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

and

$$A^{\delta} = |\delta| = (\max_k x_k - \min_k x_k)(\max_k y_k - \min_k y_k), \quad (23)$$

where (x_k, y_k) are polygon vertex coordinates from the LabelMe database (see Section 2.2).

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 individually. In this case, texture position $(P_x^{\delta}, P_y^{\delta})$ 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.

We also simulated the more shroud-like case (see Figure 6c) where the attentional window conforms to the principal textures themselves (Fazl, Grossberg, & Mingolla, 2007). Here, the region centroid and area are calculated by:

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

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

and

$$A^\delta = |\delta| = \left| \sum_{k=0}^{N-1} (x_k y_{k+1} - x_{k+1} y_k) \right|, \quad (26)$$

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 6a), and N is the number of such vertices for a certain region label. The final 43-dimensional texture feature vector T^δ is a concatenation of normalized C_{ob}^δ , O_{gk}^δ , P_x^δ , P_y^δ and A^δ values:

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

If the δ are bounding rectangles of principal textures, we call the selected region *mix textures* to distinguish them from *pure textures* that are the exact polygons from the LabelMe database (see Figure 6).

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Figure 6: An example of mix textures vs. pure textures: (a) Red region was labeled as ‘sea water’ in LabelMe; (b) mix textures in the minimum bounding box of ‘sea water’; (c) the pure texture of ‘sea water’ after the non-annotated regions have been masked out.

3.5 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 w_j from feature vectors \mathbf{f} (see Appendix Equations (A1)-(A3), and (A12)), where $G = \mathbf{f}$ for gist features (see Equation (17)) and $T^\delta = \mathbf{f}$ for texture features (see 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 (see 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 (see Appendix 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:

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

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:

$$K^* = \underset{k=1, \dots, 4}{\operatorname{argmax}} (\psi_k^G + \psi_k^T) \quad (29)$$

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 Table 1 and Table 2, model categorization performance is summarized by mean, standard deviation, and range of overall percentage correct over these 100 simulations.

	Pure textures Mean±STD (Min-Max)	Mix textures Mean±STD (Min-Max)
1st principal texture → Scene	68.70±2.36% (62.50-73.64%)	70.91±2.24% (64.67-75.82%)
2nd principal texture → Scene	62.29±2.80% (55.43-69.29%)	66.14±2.50% (60.05-72.83%)
3rd principal texture → Scene	54.41±2.13% (49.46-60.05%)	59.98±2.34% (54.62-65.22%)
1st + 2nd textures → Scene	78.07±2.14% (72.28-83.70%)	78.70±2.25% (70.92-84.24%)
1st + 2nd + 3rd textures → Scene	80.41±1.83% (75.27-85.60%)	81.24±2.05% (76.09-86.68%)

Table 1: Predictive power of principal textures. Pure textures refer to principal textures and mix textures refer to the minimum bounding boxes of principal textures (see Section 3.4).

Table 1 summarizes the predictive power of principal textures and compares the performance difference between pure textures and mix textures. 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, $p < 0.025$). This trend is also reflected in the reduced gain when we incrementally combine smaller and smaller principal textures to make the final inference (one-tailed pairwise t-test, $p < 0.025$). Table 1 also shows that mix textures carry more scenic information than pure textures. All simulations using mix textures resulted in better classification performances than ones using pure textures (one-tailed pairwise t-test, $p < 0.025$). 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)).

	Frame gist Mean±STD (Min-Max)	Grid gist Mean±STD (Min-Max)
Gist → Scene	77.14±2.15% (70.38-82.07%)	85.08±1.72% (80.98-90.22%)
Gist + 3 pure textures → Scene	81.61±1.95% (76.90-87.50%)	86.10±1.65% (82.61-91.58%)
Gist + 3 mix textures → Scene	81.73±1.85% (76.09-86.41%)	86.50±1.69% (82.88-91.30%)

Table 2: Categorization performance of gist and texture integration. Grid gist refers to the gist with spatial partition and frame gist is the one without it (see Section 3.3).

Table 2 summarizes how well gist predicts a scene and how much the texture information improves this prediction. 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. In addition, for all gist and texture representations, the gist-plus-texture predictions outperform predictions from either gist or texture alone (one-tailed pairwise t-test, $p < 0.025$). This performance boost due to textures is more pronounced when gist is less sure of scene identity, as in the case of frame gist, which agrees with the notion that active vision helps to minimize expectation uncertainty. Finally, it should be noted that, after gist-texture integration, the performance advantage of using mix textures over pure textures is less marked although still significant (one-tailed pairwise t-test, $p < 0.05$). The performance gain from mix texture interfaces in Table 1 is diminished during gist-texture integration because gist also includes texture interfaces by definition (see Section 3.3).

Truth\Predicted	Coast	Forest	Mountain	Countryside
Coast	79.94%	0.55%	1.02%	18.50%
	80.91%	0.55%	0.79%	17.75%
Forest	0%	87.83%	7.52%	4.65%
	0%	88.78%	7.16%	4.06%
Mountain	0.62%	1.19%	88.33%	9.86%
	0.44%	1.87%	90.03%	7.65%
Countryside	9.26%	1.86%	4.47%	84.41%
	9.22%	2.05%	2.37%	86.35%

Table 3. Confusion table before and after model eye movements. 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.

To understand where misclassification happens, Table 3 breaks down overall performance into its component categorical performances. Table 3 is the confusion table of gist and gist-plus-texture predictions, which separately simulates human scene recognition before and after model scanning eye movements. The table is constructed from 100 simulations on different testing sets. ‘Forest’ and ‘mountain’ are easy scenes to tell apart, whereas ‘countryside’ is often confused with the other three categories, especially ‘coast’. After model eye movements, almost all diagonal elements increase and off-diagonal elements decrease, which indicates that the gist-plus-texture approach is generic and does not favor any specific scene category. Figure 7 shows some misclassified images in the best simulation. Significantly, these images are also ambiguous to humans and the model well captures that ambiguity.

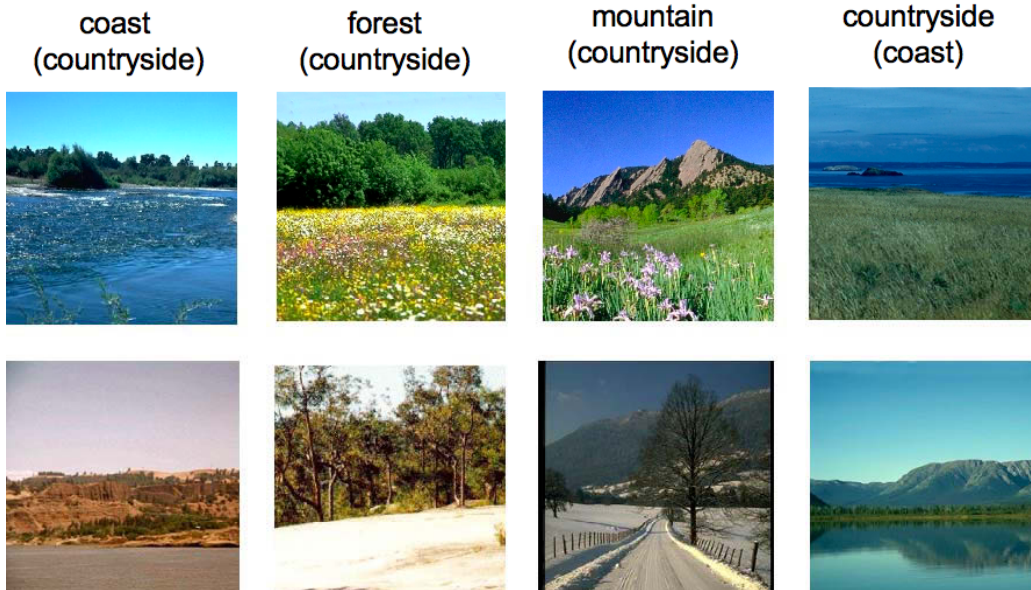


Figure 7. Examples of misclassified images in the dataset. Class labels are model predictions followed by the ground-truth categories in parenthesis.

5. Discussion and Conclusions

The best ARTSCENE classification rate for 4 landscape scene categories is 91.58%, which is better than the 90.28% (Bosch, Zisserman, & Muñoz, 2006) and 89% (Oliva & Torralba, 2001) reported in the literature using the same image dataset. These results derive from the use of locally-computed multi-scale boundary and surface information that has proved to be necessary for explaining a wide range of visual phenomena (e.g., Grossberg & Swaminathan, 2004; Cao & Grossberg, 2005; Grossberg & Yazdanbakhsh, 2005; Grossberg & Hong, 2006; Grossberg, Kuhlmann, & Mingolla, 2007; Bhatt, Carpenter, & Grossberg, 2007; Fazl, Grossberg, & Mingolla, 2007). In contrast, the gist in Oliva & Torralba (2001) was represented in terms of three spectral templates – openness, ruggedness, and roughness – which requires the brain to parse a scene in the frequency domain and to operate non-locally. Although the texton approach in Fei-Fei & Perona (2005) guarantees locality, some textons exhibited irregular complex patterns that do not resemble those oriented receptive fields of visual neurons. For learning, we use an ARTMAP classifier that is capable of fast, incremental, stable learning of recognition categories and predictions in response to non-stationary data streams, and to automatically discover the proper degree of category generalization in response to changing environmental statistics. It is of interest that all the major predictions of ART since its introduction in Grossberg (1976a) and Grossberg (1976b) have received increasing support from psychological, neuropsychological, and neuroanatomical data over the years. See reviews in Grossberg (2003) and Raizada & Grossberg (2003). The use of ART as a gist and texture classifier is thus also compatible with a biological account of scene understanding. In terms of information fusion, we illustrate how spatial attention shifts that control eye movements can revise and improve initial large-scale gist predictions.

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 & Torralba (2001) and Vogel & Schiele (2007). Although the search can be replaced by machine learning schemes (e.g., Fei-Fei & Perona, 2005; Bosch, Zisserman, & Muñoz, 2006), 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 allows ARTSCENE to naturally perform multi-category classification and information integration across scales. In contrast to Vogel et al. (2006) and Vogel & Schiele (2007) who used 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. Finally, since our system builds upon perceptual and cognitive processes that are common to many visual tasks, it can be integrated into a multi-purpose machine vision system.

A weakness of the current implementation is the use of LableMe polygon coordinates (see Section 2.2 and 3.4). However, simulations in Table 2 show that minimum bounding rectangles (i.e., mix 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 LableMe 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, Grossberg, & Mingolla, 2007) 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.

Appendix – Default ARTMAP 2 (Amis & Carpenter, 2007):

The Default ARTMAP2 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 categories (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)).

Training, with Distributed Next-Input Test:

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

$$\mathbf{f}^c = 1 - \mathbf{f}. \quad (\text{A1})$$

The total $2M$ -dimensional input vector

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

is said to be *complement coded*. The L^1 norm of \mathbf{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 $\mathbf{F} = (F_1, F_2, \dots, F_{2M})$ to category for all $i = 1, \dots, 2M$ and $j = 1, \dots, C$. 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 \mathbf{F} . Associate it with the output class label K .

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

$$\mathbf{w}_C = \mathbf{F}, \quad (\text{A3})$$

and

$$W_{CK} = 1. \quad (\text{A4})$$

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

$$\rho = \bar{\rho}. \quad (\text{A5})$$

6. Reset all category activities:

$$\mathbf{y} = 0. \quad (\text{A6})$$

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 $j = 1, \dots, C$ using the choice-by-difference signal function (Carpenter, 1997):

$$T_j = |\mathbf{F} \wedge \mathbf{w}_j| + (1 - \alpha)(M - |\mathbf{w}_j|) \quad (\text{A7})$$

In Equation (A7), \exists denotes the fuzzy intersection: $(\mathbf{F} \exists \mathbf{w}_j)_k = \min(F_k, w_{jk})$, $|\cdot|$ denotes the L^1 norm, $(\mathbf{w}_j)_i = w_{ij}$ 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 sorted category ($j = J$) that meets the matching criterion:

$$\left(\frac{|F \wedge w_j|}{M} \geq \rho \right), \quad (\text{A8})$$

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

(b) Output class prediction:

$$\psi_k = \sum_{j=1}^C y_j W_{jk} = W_{Jk}. \quad (\text{A9})$$

(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:

$$\rho = \frac{|F \wedge w_j|}{M} + \varepsilon, \quad (\text{A10})$$

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):

$$C = C + 1. \quad (\text{A11})$$

Return to Step (4).

12. Learning: Update coding weights:

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

where β is the learning fraction ($\beta = 1$ denotes fast learning), and w_j^{old} is the previously learned weight vector for category j .

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 F using distributed activation, as prescribed for testing (compare with Equation (29)):

$$K^* = \arg \max_k \psi_k. \quad (\text{A13})$$

(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:

$$\rho = \frac{|F \wedge w_j|}{M} + \varepsilon. \quad (\text{A14})$$

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

Default ARTMAP Testing (Distributed Code):

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

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

3. Reset the category activities:

$$\mathbf{y} = 0. \quad (\text{A15})$$

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

$$T_j = |\mathbf{F} \wedge \mathbf{w}_j| + (1 - \alpha)(M - |\mathbf{w}_j|) \quad (\text{A16})$$

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

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

$$\Lambda = \{ \lambda = 1, \dots, C: T_\lambda > \alpha M \}, \quad (\text{A17})$$

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

$$\Lambda' = \{ \lambda = 1, \dots, C: T_\lambda = M \} = \{ \lambda = 1, \dots, C: \mathbf{w}_\lambda = \mathbf{F} \}. \quad (\text{A18})$$

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, Milenova, & Noeske, 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 $\Lambda' \neq \emptyset$ (i.e., $\mathbf{w}_j = \mathbf{F}$ for some j), set

$$y_j = \frac{1}{|\Lambda'|} \quad (\text{A19})$$

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:

$$y_j = \frac{\left[\frac{1}{M - T_j} \right]^p}{\sum_{\lambda \in \Lambda} \left[\frac{1}{M - T_\lambda} \right]^p} \quad (\text{A20})$$

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:

$$\psi_k = \sum_{j=1}^C y_j W_{jk}. \quad (\text{A21})$$

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

Reference

- Amis, G., & Carpenter, G. (2007). Default ARTMAP 2. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Orlando, Florida.
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617-629.
- Bhatt, R., Carpenter, G.A., & Grossberg, S. (2007). Texture segregation by visual cortex: perceptual grouping, attention, and learning. *Vision Research*, in press.
- 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(5), 515-578.
- Carpenter, G.A. (1997). Distributed Learning, Recognition, and Prediction by ART and ARTMAP Neural Networks. *Neural Netw*, 10(8), 1473-1494.
- 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(3), 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(5), 793-813.
- Fazl, A., Grossberg, S., & Mingolla, E. (2007). View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *CAS/CNS-TR-07-011*. Submitted for publication.
- Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *Proceedings of the 2005 IEEE Computer Vision and Pattern Recognition (CVPR'05)*, 2, 524- 531.
- 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(3), 121-134.
- Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, 23(4), 187-202.
- 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(1), 48-121.
- Grossberg, S. (2000). The complementary brain: unifying brain dynamics and modularity. *Trends in Cognitive Sciences*, 4(6), 233-246.
- Grossberg, S. (2003). 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(1), 47-76.
- Grossberg, S., & Hong, S. (2006). A neural model of surface perception: Lightness, anchoring, and filling-in. *Spatial Vision*, 19(2), 263-321.

- 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(11), 1147-1187.
- Grossberg, S., & Yazdanbakhsh, A. (2005). Laminar cortical dynamics of 3D surface perception: Stratification, transparency, and neon color spreading. *Vision Research*, 45(13), 1725-43.
- Grossberg, S., Kuhlmann, L., & Mingolla, E. (2007). A neural model of 3 D shape-from-texture: Multiple-scale filtering, boundary grouping, and surface filling-in. *Vision Research*, 47(5), 634-672.
- Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3), 353-383.
- Oliva, A., & Schyns, P.G. (2000). Diagnostic Colors Mediate Scene Recognition. *Cognitive Psychology*, 41(2), 176-210.
- Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3), 145-175.
- Potter, M. (1975). Meaning in visual search. *Science*, 187(4180), 965-966.
- Raizada, R.D.S., & Grossberg, S. (2003). Towards a Theory of the Laminar Architecture of Cerebral Cortex: Computational Clues from the Visual System. *Cerebral Cortex*, 13(1), 100-113.
- 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(4), 195-200.
- Vogel, J., & Schiele, B. (2007). Semantic Scene Modeling and Retrieval for Content-Based Image Retrieval. *International Journal of Computer Vision*, 72(2), 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.