INSTITUTE OF PHYSICS PUBLISHING NETWORK: COMPUTATION IN NEURAL SYSTEMS

Network: Comput. Neural Syst. **14** (2003) 391–412 PII: S0954-898X(03)53778-2

Statistics of natural image categories

Antonio Torralba[1] **and Aude Oliva**[2]

[1] Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA
[2] Department of Psychology and Cognitive Science Program, Michigan State University, East Lansing, MI 48824, USA

E-mail: torralba@ai.mit.edu and aoliva@msu.edu

Received 16 September 2002, in final form 30 January 2003
Published 12 May 2003
Online at stacks.iop.org/Network/14/391

Abstract

In this paper we study the statistical properties of natural images belonging to different categories and their relevance for scene and object categorization tasks. We discuss how second-order statistics are correlated with image categories, scene scale and objects. We propose how scene categorization could be computed in a feedforward manner in order to provide top-down and contextual information very early in the visual processing chain. Results show how visual categorization based directly on low-level features, without grouping or segmentation stages, can benefit object localization and identification. We show how simple image statistics can be used to predict the presence and absence of objects in the scene before exploring the image.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Figure 1 shows a collection of mean images created by averaging pictures from the same semantic category. According to the seminal work of Rosch and collaborators (1976), people recognize most of the objects at the same level of abstraction: the basic level (e.g. car, chair). It has been shown that objects of the same basic-level category share similar common features and usually they have a similar shape. This is illustrated in the averaged images or 'prototypes' shown in figure 1. In each prototype image, information about the level of spatial similarity existing between local features (e.g. distribution of coloured regions and intensity pattern) is demonstrated by the degree of sharpness. Object categories like faces and pedestrians are much more regular in terms of distribution of pixel intensities than other object groups (e.g. chairs).

Similarly to objects, natural images depicting environmental scenes can also be classified in basic-level categories (Tversky and Hemenway 1983), and as such, are expected to share common features (Jepson *et al* 1996). Despite the fact that the organization of parts and regions in environmental scenes is much less constrained than in objects, the resulting prototype images are not spatially stationary. A strong relationship exists between the category of a scene picture

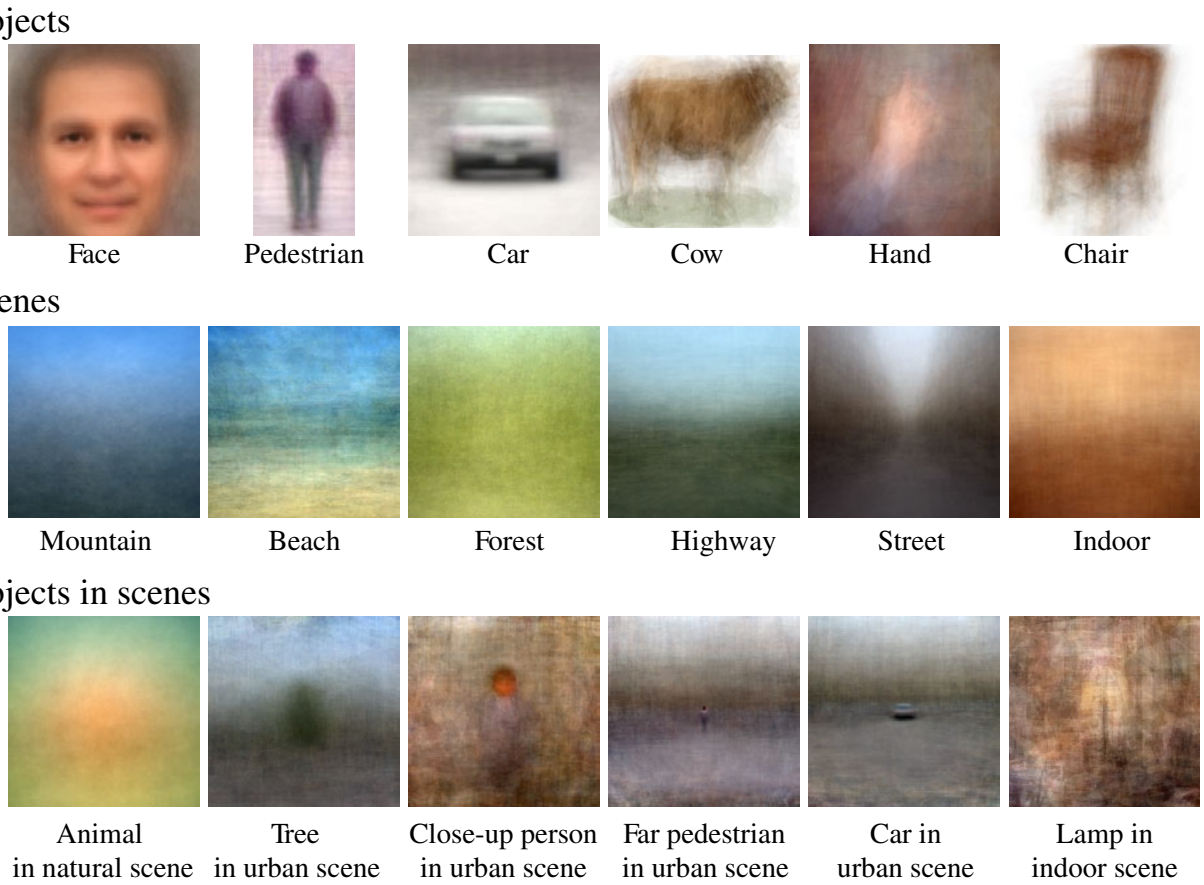

Figure 1. Averaged pictures of categories of objects, scenes and objects in scenes, computed with 100 exemplars or more per category. Exemplars were chosen to have the same basic level and viewpoint in regard to an observer. The group objects in scenes (third row) represent examples of the averaged peripheral information around an object centred in the image.

and the distribution of coloured regions in the image. In a similar vein, the distribution of structural and colour patterns of the background scene around an object is constrained. The third row of averaged prototypes shown in figure 1 (objects in scenes) have been created by averaging hundreds of images constrained to have a particular object at one scale present in the image (before averaging, the images were translated in order to have the object of interest in the centre). Because of the correlation existing between an object and its context the background does not average to a uniform field (Torralba 2003). On the contrary, the background exhibits the texture and colour pattern that is common to all environments where a specific object is usually found.

Figure 1 illustrates the degree of regularities found in natural image categories when ecological conditions of viewing are considered. The statistics of natural image categories depend not only on how the visual world is built to serve a specific function, but also the viewpoint that the observer adopts. Because different environmental categories are built with different objects and materials, and the point of view of the observer imposes its own constraints (such as its size and motion), we expect to find strong biases in the statistics distribution of the image information. Those statistics might have different biases for different animal species.

In this paper, we illustrate how simple statistics of natural images vary as a function of the interaction between the observer and the world. The paper is organized as follows: section 2 describes the statistical properties of natural images per scene category. Sections 3 and 4 introduce, respectively, the spectral principal components of natural images and scene-tuned filters, and describe how these methods could be used to perform simple scene categorization tasks. Section 5 summarizes computational approaches of scene categorization, and section 6 shows the robustness of simple statistics in performing object detection tasks.

2. Statistical properties of natural categories

2.1. $1/f$ *spectra of natural images*

Statistics of natural images have been found to follow particular regularities. Seminal studies (Burton and Moorhead 1987, Field 1987, 1994, Tolhurst *et al* 1992) have observed that the average power spectrum of natural images falls with a form $1/f^{\alpha}$ with $\alpha \sim 2$ (or $\alpha \sim 1$ considering the amplitude spectrum, see figure 2(a)).

Related studies found a bias in the distribution of orientations, illustrated in the power spectra of figure 2. In real-world images, including both natural landscapes and man-made environments, vertical and horizontal orientations are more frequent than obliques (Baddeley 1997, Switkes *et al* 1978, van der Schaaf and van Hateren 1996, Oliva and Torralba 2001). A more complete model of the mean power spectra (using polar coordinates) can be written as

$$E[|I(f,\theta)|^2] \simeq A_s(\theta)/f^{\alpha_s(\theta)} \tag{1}$$

in which the shape of the spectra is a function of orientation. The function $A_s(\theta)$ is an amplitude scaling factor for each orientation and $\alpha_s(\theta)$ is the frequency exponent as a function of orientation. Both factors contribute to the shape of the power spectra. The model of equation (1) is needed when considering the power spectra of man-made and natural scene images separately (cf figure 2 and table 1, also Baddeley1996, 1997). Table 1 shows that the values of the slope α and the amplitude A vary with orientation and also with the type of environment[3]. The anisotropic distribution of orientations is also compatible with

[3] The database used in this study contains about 12 000 pictures of scenes and objects. Images were 256×256 pixels in size. They come from the Corel stock photo library, pictures taken from a digital camera and images downloaded from the web.

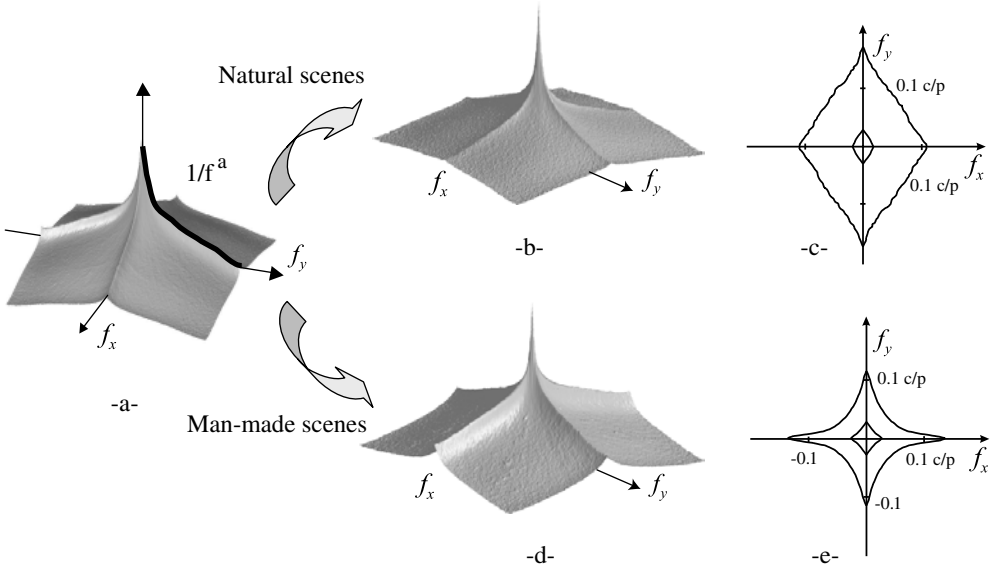

Figure 2. (a) Mean power spectrum averaged from 12 000 images (vertical axis is in logarithmic units). Mean power spectra computed with 6000 pictures of man-made scenes (b) and 6000 pictures of natural scenes (d); (c) and (e) are their respective spectral signatures. The contour plots represent 50 and 80% of the energy of the spectral signature. The contour is selected so that the sum of the components inside the section represents 50% (and 80%) of the total. Units are in cycles per pixel (cf also Baddeley 1996).

Table 1. Average α and A values for images representing man-made and natural environments. The values α and A were obtained by fitting the model A/f^{α} to the power spectrum of each single image at three orientations (horizontal, f_x, oblique and vertical, f_y). The fit was performed in the frequency interval [0.02, 0.35] cycles/pixel. The amplitude factor A is normalized so that the maximum averaged value is unity. Averages were computed with more than 3500 images per category (cf also figures 2(b), (c)). Similar values are obtained when the fit is performed on the averaged power spectrum. The numbers in parenthesis give the standard deviation.

		H	O	V
Natural	α	1.98 (0.58)	2.02 (0.53)	2.22 (0.55)
	A	0.96 (0.40)	0.86 (0.38)	1 (0.35)
Man-made	α	1.83 (0.58)	2.37 (0.45)	2.07 (0.52)
	A	1 (0.32)	0.49 (0.24)	0.88 (0.29)

neurophysiological data showing that the number of cells in early cortical stages varies in regard to the spatial scale tuning and the orientation (e.g. more vertical and horizontal tuned cells than oblique in the fovea, DeValois and DeValois 1988).

2.2. Spectral signatures of image categories

Different categories of environments also exhibit different orientations and spatial frequency distributions, captured in the averaged power spectra (Baddeley 1997, Oliva *et al* 1999, Oliva and Torralba 2001). Figure 3 shows that the differentiation among various man-made categories resides mainly in the relationship between horizontal and vertical contours at different scales, while the spectral signatures of natural environments have a broader variation in spectral shapes.

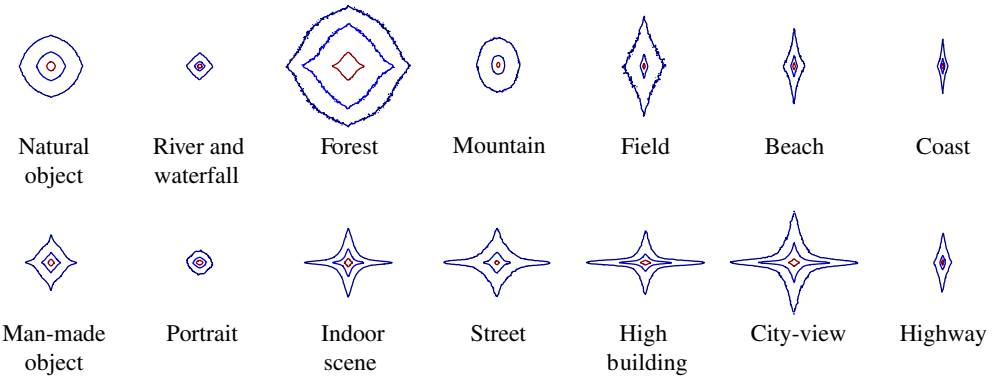

Figure 3. Spectral signatures of 14 different image categories. Each spectral signature is obtained by averaging the power spectra of a few hundred images per category. The contour plots represent 60, 80 and 90% of the energy of the spectral signatures (energy is obtained by adding the square of the Fourier components). The size of the spectral signature is correlated with the slope (α). A large value of α produces a fast decay of the energy at high spatial frequencies, which produces a smaller contour. The overall shape is a function of both $\alpha(\theta)$ and $A(\theta)$.

The particular spectral signature per scene category is even more striking when considering basic-level classes of environmental scenes such as streets, highways, buildings, forests etc. From the contour plots of figure 3, we can see that the dominant spatial scales and dominant orientations are typical of classes of scenes representing different volumes or depth ranges. The spectral signatures of pictures of large-scale scenes (e.g., beach, coast, field) are dominated by the horizon. When the scene background becomes closer to the observer (from mountains to enclosed landscapes and natural objects), the spectral signatures become isotropic and denser in high spatial frequencies. The shape of the spectral signatures is correlated with the scale (e.g. size) at which the main components of the image should be found (e.g. finer texture in forest, coarser texture in waterfalls).

2.3. Scene scale and image scale

Image statistics also vary when considering scenes at different scales. Figure 4 shows the spectral signatures of scenes sharing similar depth range. These signatures have been obtained from a database of images for which four subjects were asked to provide the mean depth or volume of the environment represented in the image (Torralba and Oliva 2002). Scene scale is measured in metres. Each spectral signature was computed by averaging the power spectra of scene pictures within a similar distance range.

When considering large changes in scale (a factor larger than 10), significant differences exist between the spatial and the spectral statistics of pictures depicting scenes and objects at different scales. There are at least two factors that can explain the dependence between structure and depth range. First, the point of view that any given observer adopts on a specific scene is constrained by the volume of the scene. Many real world objects can be observed from an infinite number of viewpoints as long as the observer is directly capable of manipulating the object. However, as distance and scale increase, the possible viewpoints of a human observer become increasingly limited and predictable. For instance, tall buildings are usually observed from the ground, or a window of another building. Second, the parts or objects that compose one scene differ strongly from one scale to another scale, due to functional constraints and to the physical processes that shape the space at each scale.

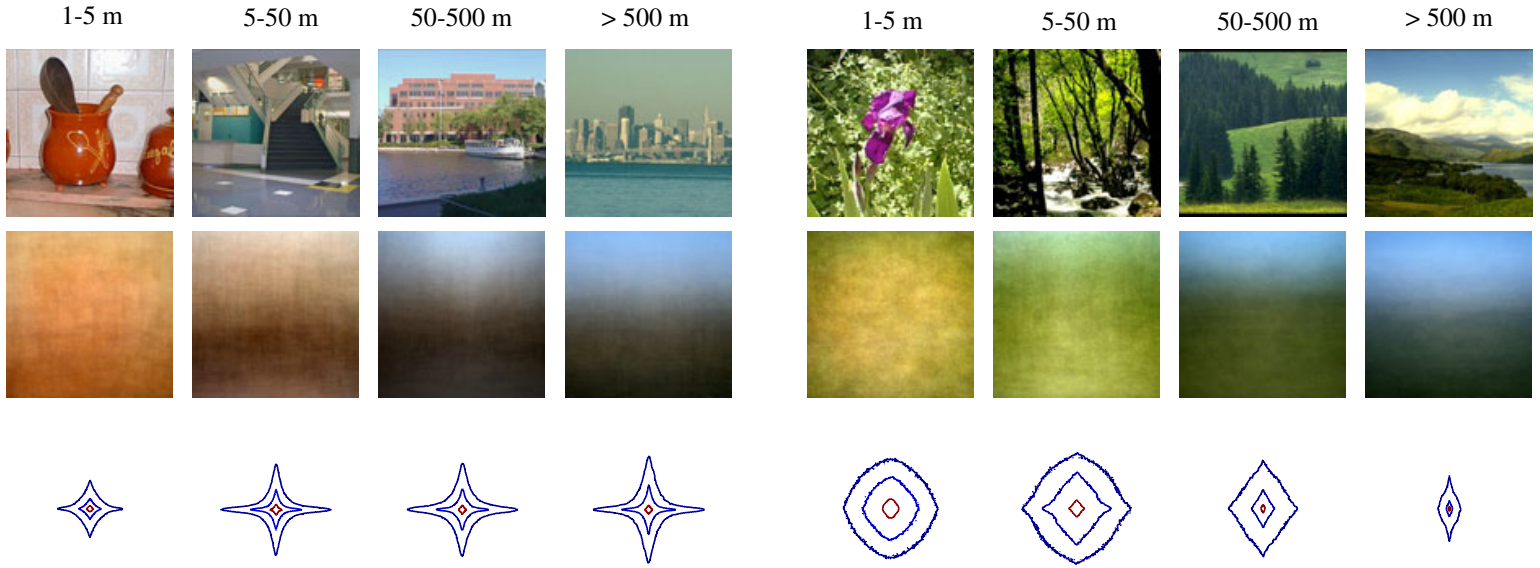

Figure 4. Averaged spatial images and spectral signatures as a function of scene scale. Scene scale refers to the mean distance between the observer and the principal elements that compose the scene. Each image average and spectral signature was calculated with 300–400 images.

Figure 4 emphasizes the differences between man-made and natural environments at different scale ranges. Close-up views on man-made objects tend to produce images that are composed of flat and smooth surfaces. Consequently, the energy of the power spectra for close-up views is concentrated mainly in low spatial frequencies. As distance between the observer and the scene background increases, the visual field comprehends a larger space, that is likely to encompass more objects. The perceived images of man-made scenes appear as a collection of surfaces that break down into smaller pieces (objects, walls, windows etc). Thus, the spectral energy corresponding to high spatial frequencies increases as the scene becomes more cluttered due to the increase, with distance, in the area covered by the visual field. In contrast, spectral signatures of natural environments behave differently while increasing depth. Figure 4 shows that when the distance between the observer and the background grows, natural structures become larger and smoother (small grain disappears due to the spatial sampling of the image). Therefore, on average, with an increment of distance, the level of clutter decreases, as does the energy in the high spatial frequencies. In addition, the pattern of orientation varies with the scale. Close-up views on natural structures have a tendency to be isotropic in orientations (and the point of view of the observer is unconstrained). As distance grows, there is an increased bias towards vertical and horizontal orientations, together with the point of view of the observer becoming more constrained. As distance continues to increase, energy is concentrated mainly in vertical spatial frequencies, as very large environmental scenes are organized along horizontal layers. In order to recognize the scene or to navigate such panoramic environments, faced with point of view limitations, an observer might consider looking towards the horizon to visually embrace the whole scene.

Several studies have examined the scale invariance properties of natural image statistics (e.g. Field 1987, Ruderman 1997, among others). These studies have focused on the similarity between the statistics of wavelet outputs at different image scales. Results indicate that some image statistics are scale invariant. Here, we differentiate between *image scale*, that refers to scales in terms of spatial frequencies, and *scene scale*, that refers to the mean distance between the observer and the elements that compose the scene. Note that for the range of distances we consider (from 1 m to several kilometres) the problem of distance cannot be modelled as a scaling factor. With each change on an order of magnitude in distance, images perceived also belong to different scene semantic categories (single objects, rooms, places, large outdoors and panoramic scenes). Therefore, we can expect that the statistics of images might evolve when changing scene scale and provide useful categorical information about the probable depth range of the scene (see sections 5 and 6, and also Torralba and Oliva 2002).

Figure 5 shows how the output energy of oriented wavelets may vary when considering both image scale and scene scale. The wavelets used are oriented Gabor filters tuned at a radial frequency of 1/4 cycles/pixel at 12 different orientations. Changes in image scale are obtained by subsampling the images from 256^2 to 32^2 pixels by factors of two. Changes in scene scale are obtained by averaging the outputs obtained from scene pictures with different depth ranges (from close-up views of objects and textures to panoramic views and natural landscapes). The most important modification of the spectral signatures is observed across scene scales. When modifying scene scale, the shape of the polar plots evolves by changing the amount of energy at each orientation. However, across image scales, there is little variation. This observation is more striking for natural environments than man-made scenes.

2.4. Non-stationary statistics

Another important characteristic of natural images is how the image statistics change with spatial location. When considering all the possible directions of the eye-camera, statistics of

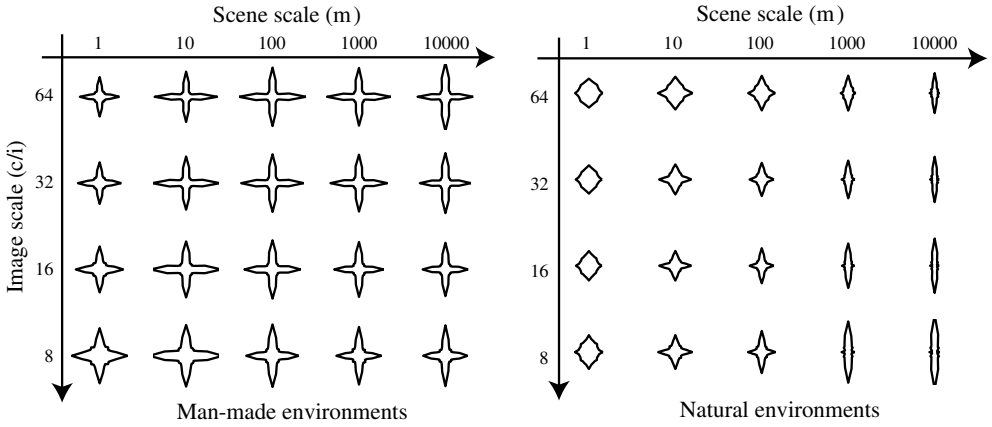

Figure 5. Polar plots of responses of multiscale oriented Gabor filters. The magnitude of each orientation corresponds to the total output energy averaged across the entire image. The energies are normalized across image scale by multiplying by a constant so that noise with $1/f$ amplitude spectrum has the same polar plots at all image scales.

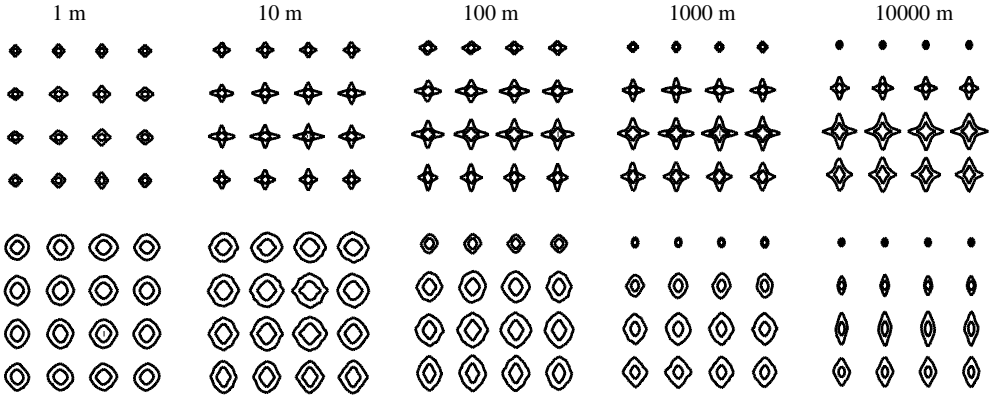

Figure 6. Illustration of the non-stationarity of image statistics in groups of man-made (top) and natural (bottom) environments at different depth scales (from left to right, close-up views to panoramic views). The spectral signatures were obtained by averaging the windowed power spectra at 4×4 locations in the images. As scene scale increases, the image statistics become non-stationary.

natural images are expected to be scale invariant (Field 1994, 1999, Ruderman 1994, 1997) and stationary (the features are equally distributed in regard to locations, Field (1994), (1999)). This is the case indeed with the statistics of images of close-up views of objects that are, on average, stationary, as there is no preferred point of view for the camera (cf figure 6, scene scale of 1 metre). However, for images of scenes that embrace a large volume, the probable points of view that a human observer will adopt become much more restrained, because of its height and its probable location (on the floor). If the task of the observer is to recognize the identity of a large space scene, most of the useful information will be given while looking towards the horizon. Therefore, different image statistics will characterize the top and bottom half of the image (e.g. smooth texture of the sky, long vertical contours of skylines at the top, cluttered forms at the bottom). Figure 7 shows an example of how image inversion affects the

Figure 7. Top–down effect on depth judgments. The image on the left is generally recognized as close-up view on bushes and maybe a spider's web on top. On the right-hand side, the image is categorized as the inside of a forest, corresponding to a larger distance than the image on the left. The image on the left corresponds to the image on the right after inversion upside down and left–right. The upside-down inversion affects the perception of concavities and convexities due to the assumption of light from above, and, therefore, modifies the perceived relative 3D structure of the scene. But, moreover, the wrong recognition affects the absolute scale of the perceived space.

perception of the absolute depth of a scene. Note that it is not just the relative shape of the scene that is misperceived but also the absolute scale. The image on the left appears as a closer structure than the image on the right for most observers.

As image statistics are a function of the observer, images of large-scale scenes taken from a bird's-eye view should elicit almost stationary distribution of features as the point of view becomes totally unconstrained in regard to the possible orientation of the perceived images. In the case of a human observer, standing, the viewpoint is strongly constrained, producing images with non-stationary statistics, as shown in figure 6. The property of non-stationarity is very relevant to sensory and cognitive systems, as it may provide an invariant signature of a specific kind of environment.

3. Principal components of real-world images and power spectra

Principal component analysis (PCA) has been extensively used in vision problems for coding and recognition. In the face recognition domain, when faces are correctly aligned and scaled, PCA gives a reduced set of orthogonal functions able to reconstruct faces (Craw and Cameron 1991, Turk and Pentland 1991). This operation facilitates the recognition procedure that is performed in a low-dimensional space (Sirovich and Kirby 1987, Swets and Weng 1996). PCA has also been used for obtaining efficient codes of the visual input, adapted to the statistics of natural stimuli (e.g. Hancock *et al* 1992, Liu and Shouval 1994). Image principal components (IPCs) decompose the image as

$$i(x, y) = \sum_{n=1}^{P} v_n \mathrm{IPC}_n(x, y) \tag{2}$$

where $i(x, y)$ is the intensity distribution of the image along spatial variables x and y. $P \leqslant N^2$ is the number of total IPCs and $N^2 = 256^2$ is the number of pixels of the image. The $\mathrm{IPC}_n(x, y)$

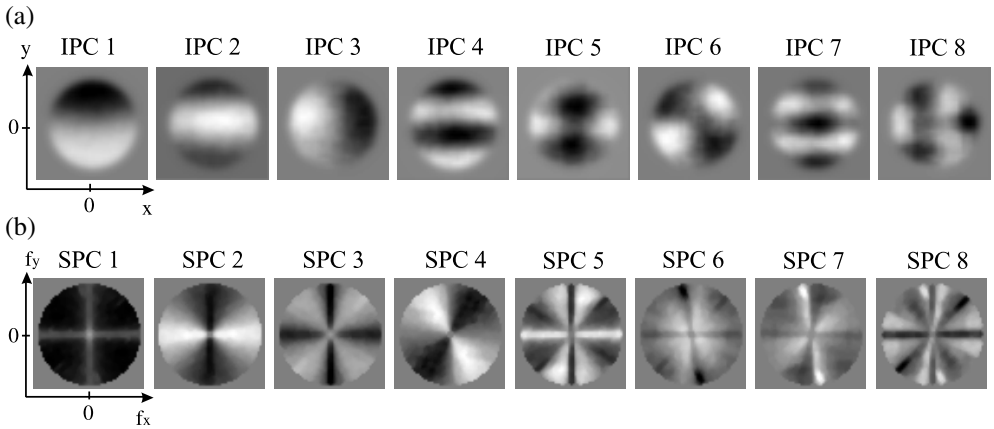

Figure 8. (a) The first eight principal components of images (IPCs) and the (b) energy spectra of images (SPCs). The frequency $f_x = f_y = 0$ is located at the centre of each image (SPC). Frequencies are defined in the interval $[-1/2, 1/2]$. The amplitude at each frequency is normalized with respect to its standard deviation before applying the PCA.

are the eigenvectors of the covariance matrix: $\boldsymbol{T} = E[(\boldsymbol{i}-\boldsymbol{m})(\boldsymbol{i}-\boldsymbol{m})^{\mathrm{T}}]$, where $\boldsymbol{i}$ are the pixels of the image rearranged in a column vector. E is the expectation operator. $\boldsymbol{m} = E[\boldsymbol{i}]$ is the mean of the images. v_n are the coefficients for describing the image $i(x, y)$. Figure 8(a) shows the IPCs computed from 5000 pictures of real-world scenes. As discussed by Field (1994), stationarity of natural images is responsible for IPC shape (figure 8(a)) which corresponds to the Fourier basis.

Here, we compute the power spectrum of an image by taking the squared magnitude of its discrete Fourier transform (DFT):

$$\Gamma(k_x, k_y) = \frac{1}{N^2}|I(k_x, k_y)|^2 \tag{3}$$

where

$$I(k_x, k_y) = \frac{1}{N^2}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} i(x, y)\exp\left(-\frac{\mathrm{j}2\pi}{N}(xk_x + yk_y)\right) \tag{4}$$

$f_x = k_x/N$ and $f_y = k_y/N$ are the discrete spatial frequencies. The power spectrum[4], $\Gamma(k_x, k_y)$, encodes the energy density for each spatial frequencies and orientations over the whole image.

PCA applied to power spectra gives the main components that take into account the structural variability between images. First, the power spectrum is normalized with respect to its variance for each spatial frequency: $\Gamma'(k_x, k_y) = \Gamma(k_x, k_y)/\mathrm{std}[\Gamma(k_x, k_y)]$, with $\mathrm{std}[\Gamma(k_x, k_y)] = \sqrt{E[(\Gamma(k_x, k_y) - E[\Gamma(k_x, k_y)])^2]}$. This normalization compensates for the $1/f^{\alpha}$ shape of the power spectrum. The spectral principal components (SPCs) decompose the normalized power spectrum of an image as

$$\Gamma'_s(k_x, k_y) = \sum_{n=1}^{P} u_n \mathrm{SPC}_n(k_x, k_y). \tag{5}$$

P is the number of SPCs.

[4] Although not reflected in equation (4), for computing the spectral signatures of the precedent sections we have applied a circular Hamming window to the images in order to avoid boundary artifacts.

Figure 9. Projection of images into the space represented by the second and the third principal components of the power spectra. Images are organized according to spectral properties: SPC_2 puts images with dominant energy in the f_y axis on top of the figure opposed to images with dominant energy in the f_x axis which are at the bottom. SPC_3 opposes images with energy in the f_x and f_y axis (cross shape) with respect to images with energy at oblique orientations. A coarse organization of scenes emerges: man-made versus natural scenes and open versus closed environments.

Figure 8(b) shows the resulting SPCs. In accordance with results of section 2, the three first principal components exhibit vertical and horizontal spectral component dominance. Figure 9 shows a set of scene images organized according to the projections of their power spectra along the second and the third spectral principal components. Along the SPC_2 axis, scenes with a dominant horizon, such as open landscapes and suburban open scenes, stand at the top of the figure, in opposition to scenes defining closed and enclosed environments (with isotropic power spectra). Along the SPC_3 axis, images having a cross shape power spectrum (mostly urban areas) stand at one extreme whereas natural landscapes stand at the other extreme. An organization of broad environmental categories emerges, suggesting that the variability in the second-order statistics of natural images may be relevant for natural scenes classification tasks. A linear combination of the first three SPCs is able to separate man-made scenes from natural scenes with an accuracy of 80%.

4. Receptive fields for scene recognition

The level of organization achieved by the SPCs suggests that the variability in the second-order statistics of natural images may be relevant for categorization purposes. As illustrated in figures 2–6, we have observed that second-order image statistics vary along the *naturalness* dimension (man-made versus natural landscapes) and *openness* dimension (related to the depth, Baddeley 1997, Oliva and Torralba 2001). This suggests that the categorical status of an environmental view, along those two perceptual dimensions, could be computed in a feedforward manner, from a set of low-level detectors encoding information similar to the one provided by the power spectrum (see also section 6).

In this section, we look for the best spectral statistics providing discrimination between man-made, natural, open and closed scene categories. Linear discrimination between two scene categories, using the normalized image power spectra $\Gamma'(k_x, k_y)$, can be achieved as

$$w = \sum_{k_x=0}^{N-1}\sum_{k_y=0}^{N-1} \Gamma'(k_x, k_y)\mathrm{DST}(k_x, k_y) = \sum_{k_x=0}^{N-1}\sum_{k_y=0}^{N-1} \Gamma(k_x, k_y)\mathrm{DST}'(k_x, k_y) \qquad (6)$$

with $\mathrm{DST}'(k_x, k_y) = \mathrm{DST}(k_x, k_y)/\mathrm{std}[\Gamma(k_x, k_y)]$.

$\mathrm{DST}'(k_x, k_y)$ is the weighing of the spectral components needed to discriminate between the two classes (DST standing for discriminant spectral template). w is the most discriminant feature and is obtained as a weighted sum of the power spectrum of the image. As SPCs define a complete orthogonal basis for describing the normalized power spectra, we can write the DST as

$$\mathrm{DST}(k_x, k_y) \simeq \sum_{n=1}^{P} a_n \,\mathrm{SPC}_n(k_x, k_y). \qquad (7)$$

The coefficients a_n indicate how to weight of each SPC in order to build a specific template DST. Here, we used the first $P = 16$ SPCs. The coefficients a_n are determined by a supervised learning stage. In the learning stage, each image is represented by a column vector of features $\boldsymbol{u} = \{u_n\}$, u_n being the projection of the power spectrum of the image into the nth SPC. Then, we define two groups of images of different scene categories (e.g., pictures of man-made and natural environments, the same image database as described in section 3). The parameters of the DST, a_n, can be learnt by applying Fisher linear discriminant analysis (e.g. Ripley 1996, Swets and Weng 1996) that looks for the coefficients a_n giving the best classification rate. After training, the classification rate for man-made scenes versus natural landscapes goes to 93% (versus 80% when using the SPCs only). Training and testing are respectively done on different set of images, of thousands of exemplars each. When applying the discrimination analysis to the differentiation between open and closed and enclosed environments, the correct classification rate reaches 94%.

Although more complicated classifiers could be used, the linear classifier allows for a simple analysis. Due to the linearity of equation (6), the discrimination performed in the spectral domain (DST') can be written in the spatial domain. It is then possible to compute *receptive fields* tuned to global scene statistics that discriminate between the categories of natural scenes.

The output energy of a discrete filter with transfer function $H(k_x, k_y)$ can be computed as

$$E = \sum_{k_x=0}^{N-1}\sum_{k_y=0}^{N-1} \Gamma(k_x, k_y)|H(k_x, k_y)|^2. \qquad (8)$$

This expression is similar to equation (6) used to compute the structural feature w. However, as the squared magnitude of the transfer function of a filter cannot have negative values, the DST'

can be implemented by computing the difference between the output energies of two filters. In such a case, we can compute w as the difference between two energies as $w = E_+ - E_-$, where E_+ and E_- are respectively the output energy of two filters with transfer functions H_+ and H_-. In such a case, we obtain

$$w = \sum_{k_x=0}^{N-1} \sum_{k_y=0}^{N-1} \Gamma(k_x, k_y)(|H_+(k_x, k_y)|^2 - |H_-(k_x, k_y)|^2). \tag{9}$$

With this expression, it is possible to obtain positive and negative values for DST$'$. Several functions H_+ and H_- give the same resulting DST$'$. Here, we use

$$|H_+(k_x, k_y)|^2 = \text{rect}[\text{DST}'(k_x, k_y)] \tag{10}$$

and

$$|H_-(k_x, k_y)|^2 = \text{rect}[-\text{DST}'(k_x, k_y)] \tag{11}$$

where rect(x) is a half rectifying function: rect$(x) = x$ if $x > 0$ and rect$(x) = 0$ if $x < 0$. These equations give the magnitude of the two filters. As the phase can be freely chosen, we chose null phase filters as this allows us to obtain filters with spatial localized impulse response.

The impulse responses of these two filters are receptive fields that best discriminate between two groups of images. Figure 10 shows the impulse responses of both filters for the *naturalness* DST and the *openness* DSTs, respectively, for man-made and natural scenes.

The DST computed in the Fourier domain is equivalent to the convolution of the image with two spatial invariant filters and, then, computing the difference of their total output energies. The two impulse responses $h_+(x, y)$ and $h_-(x, y)$ reveal the spatial features that are discriminant between the two opposite sets of images to be considered. For man-made versus natural scenes, we see a cross impulse response versus an isotropic (slightly oblique) impulse response. For open versus closed natural scenes, we find a horizontal edge detector versus an isotropic impulse response. For man-made scenes, the impulse responses for measuring the openness of the space are similar to a horizontal versus a vertical edge detector.

If we compute the output of the two filters by convolution of the image with the respective impulse responses, $o_+(x, y) = i(x, y) * h_+(x, y)$ and $o_-(x, y) = i(x, y) * h_-(x, y)$, then the structural feature can be also obtained as

$$w = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (|o_+(x, y)|^2 - |o_-(x, y)|^2). \tag{12}$$

We define an opponent energy channel as the image obtained by the difference $\text{d}(x, y) = |o_+(x, y)|^2 - |o_-(x, y)|^2$. This signal gives the spatial locations that contribute to the discriminant feature w. Figure 10 show the opponent energy channel for different images. The city picture shows in black the vertical and horizontal components typical of man-made environments. Natural image components are represented in white. The farm scene shows natural textured components of the grass and the trees (in white) and man-made components (in black). Opponent energy channels corresponding to *openness* oppose horizon lines of panoramic scenes (negative values in black) to other spectral components (in white).

5. Semantic categorization

In seminal approaches to vision, visual processing is depicted as a hierarchical organization of modules of increasing complexity, the last of which derives the category of the scene (Barrow and Tenenbaum 1978, Marr 1982, Biederman 1987). Recent computational approaches render

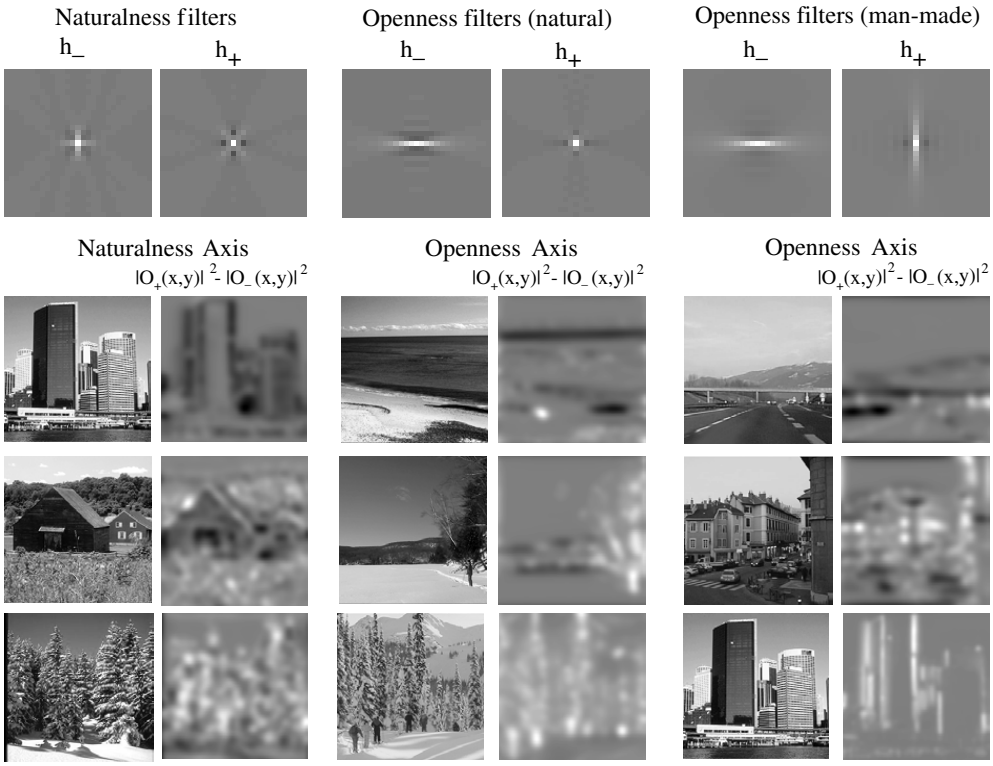

Figure 10. Illustration of the opponent filters h_- and h_+ for the naturalness and openness dimensions. For each dimension, we show examples of opponent energy channel images.

scene recognition as a process that combines a set of image-based features (e.g. colour, orientation, texture) to form higher-level representations such as regions (*blobword* from Carson *et al* 2002), simple blocks (*geons* from Biederman 1987) or objects (Barnard and Forsyth 2001). This approach is founded on biological evidence suggesting that the visual encoding mechanisms are grounded on a multi-scale and multi-orientation representation (Hubel and Wiesel 1968). At the earliest stages of visual processing (from the retina to V1), the image is represented by local features that encode lines and edges (e.g., Atick and Redlich 1992). Independent component models: Bell and Sejnowski (1997), van Hateren and van der Schaaf (1998). Sparse coding models: Field (1994), Olshausen and Field (1996), Olshausen *et al* (2001), Vinje and Gallant (2000), Simoncelli and Olshausen (2001). At the next processing stage, families of more complex features are encoded in visual cortex V4 and TEO. For example, cells in these areas have been shown to be selective to curve contours (e.g. Gallant *et al* 1993) or tuned to 3D orientation (Hinkle and Connor 2002). Cells responding to complex object patches (Tanaka 1993, Fujita *et al* 1992, Logothetis *et al* 1995) are found later in the anterior regions of the infero-temporal cortex (IT, see Gallant 2000, for a review, and Ullman *et al* 2002). Finally, a cortical representation of the layout of real world scenes has been suggested to reside in the PPA (a region of the parahippocampal cortex, Epstein and Kanwisher (1998)).

An alternative approach to the progressive reconstruction scheme proposes to build semantic information directly from a pool of low-level features. This approach takes advantage of the regularities found in the statistical distribution of features *per scene category* described in

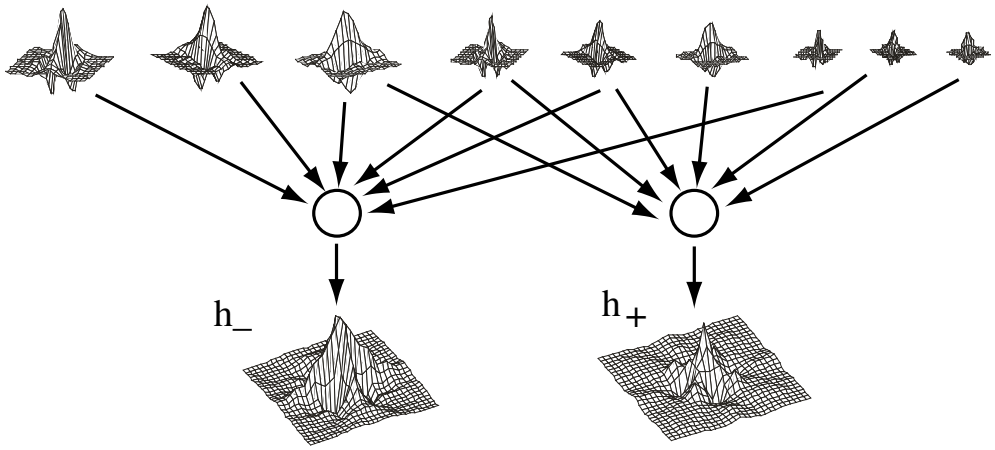

Figure 11. The figure illustrates how a linear combination of simple filters tuned to multi-orientations and multi-scales could compose a complex filter representing a naturalness detector.

the preceding sections. Using this view, high-level mechanisms could categorize a real world scene image without being exclusively grounded in stages of region and object segmentation (Schiele and Crowley 2000, Vailaya *et al* 1998, Oliva and Torralba 2001, 2002, Torralba and Oliva 2002, Torralba 2002). This approach is in agreement with experimental studies suggesting that human observers are able to identify pictures of complex scenes within a single fixation (Potter 1975). Observers can also recover the identity of a scene image when the objects have been degraded so much that they become unrecognizable in isolation (Oliva and Schyns 1997, 2000, Torralba 2003). Computational approaches, in line with this direct scene categorization scheme, have proposed simple methods for successful classification of real world scenes at different levels of abstraction. Superordinate classes of real world images can be distinguished based on the statistics of multi-scale oriented features (Gorkani and Picard 1994, Szummer and Picard 1998, Guerin-Dugue and Oliva 2000, Oliva and Torralba 2001, Vailaya *et al* 1998, 1999). Scene categorization in basic-level classes (as precise as street, buildings, highways) may also be directly performed by a selection of relevant spatial frequencies and orientation-tuned filters (Oliva and Torralba 2001, 2002, Torralba and Oliva 2002). The simple emergence of a meaningful organization of images from the first spectral principal components (figure 9) and the feasibility of coding complex scene properties (such as *naturalness* and *openness*) suggest that scene categorization could arise from a set of dimensions matched to statistical properties of the visual world in a biological system. In Oliva and Torralba (2001) it is shown how 3D properties inherent to the space that the scene subtends (e.g. depth range, openness, expansion, roughness and ruggedness, among others) can be estimated by a linear combination of low-level features. Figure 11 shows the two potential receptive fields for discriminating between natural and man-made scenes, as a linear combination of Gabor-type cells. Cells tuned to complex features of the sort used to categorize images along the *naturalness* and *openness* dimension could be located at different levels of the visual processing architecture, albeit their shape may be quite similar to receptive fields found in V4 (Gallant *et al* 1993, Hinkle and Connor 2002).

The consequence of a system that would compute 'high-level scene' primitives directly from a basis of low-level features is appealing to theoretical and computational issues of recognition. If robust scene categorization could be computed in such a feedforward manner (Van Rullen and Thorpe 2001), it would provide top-down and contextual information very early in the visual processing chain that could prove beneficial to object localization, segmentation and identification processes. It has been shown that human observers are using a top-down mechanism to find regions of interest when looking for an object independently of the presence or absence of the physical features of the object (Henderson *et al* 1999). The next section is dedicated to the effects of such a contextual effect on object processing.

6. Image statistics and contextual object processing

As suggested in the introduction, the global image statistics are also correlated with the objects present in the scene. This is not just because the object shape has an impact on the global image statistics (when the object is small this effect is negligible), but also because there exists strong correlation between the objects present and their context.

Figure 12 shows a superposition of scene pictures with the only constraint that the averaged scenes should contain a particular object in the scene. The average pictures show the mean intensity values and the spectral signature of the scene pictures that contain each object. Although no other constraints are imposed, the different average pictures correspond to the signatures of the environments that are the typical context of the objects selected. For instance, animals are expected to be in natural environments. Cars are found in roads and urban environments. People can be found in many different indoor and outdoor places. The spectral signature will also vary when constraining the object to have a particular image size (which is equivalent to fixing the distance between the observer and the object and therefore is related to the scene scale). Scenes that produce images where people occupy only a few pixels are outdoor scenes (large streets, roads). Images with close-up views on faces can be both indoors and outdoors, which will produce a different spectral signature. Similarly, there exists correlation between global image properties and the location of objects in the scene. Global image statistics can be predictors of the presence of objects, their scale and location in the image. And this is true even when the object of interest is so small that it does not contribute directly to the statistics of the image. For instance, small cars are typical of large street scenes and these scenes are characterized by particular global image statistics.

Using image statistics provides a way of introducing contextual information in object detection approaches that does not require the detection of other objects. For the prediction of the presence/absence of an object of interest we use a statistical framework more reliable than the linear discrimination performed in section 4. The function $P(O|\vec{v}_C)$ gives the probability of presence of the object class O given a set of global image statistics $\vec{v}_C$ that represent the contextual information.

The training set used for learning the probability $P(O|\vec{v}_C)$ is a set of annotated pictures. We use Bayes rule to write $P(O|\vec{v}_C) = P(\vec{v}_C|O)P(O) + P(\vec{v}_C|\overline{O})P(\overline{O})$. We learn the PDF $P(\vec{v}_C|O)$ from a set of images in which the object is present. We approximate the PDF using mixtures of Gaussians and the learning is performed using the EM algorithm (Torralba and Sinha 2001). In the same way, we learn the PDF $P(\vec{v}_C|\overline{O})$ using a set of images in which the object is absent. We assume that $P(\overline{O}) = P(O) = 1/2$. The contextual features $\vec{v}_C$ are the features obtained by projecting the image power spectra into the first 16 SPCs. Once the learning is done, the function $P(O|\vec{v}_C)$ evaluates the consistency of the object O with the context $\vec{v}_C$ and, therefore, provides information about the probable presence/absence of one object category before scanning the picture looking for the object.

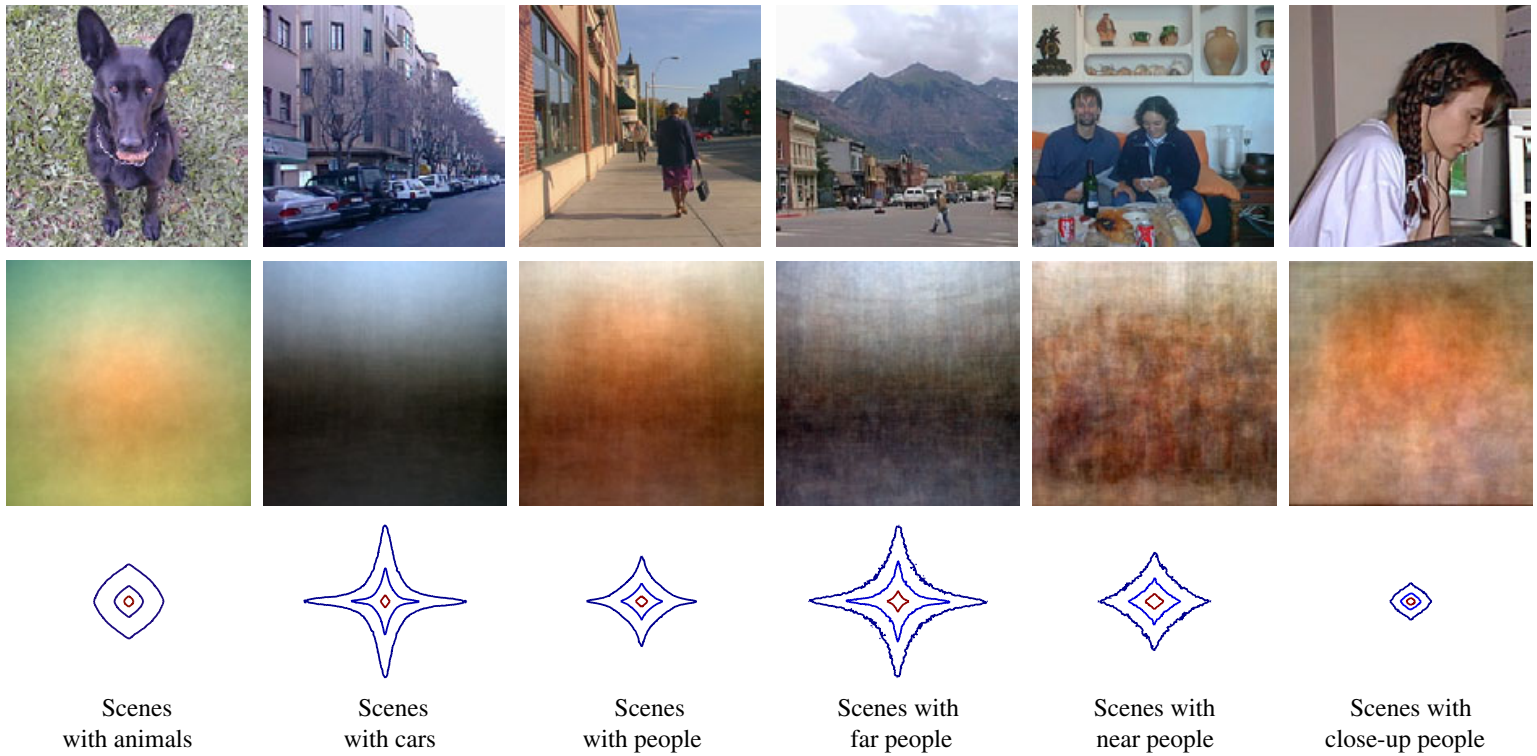

Figure 12. Average intensity and spectral signatures of sets of images constrained to contain specific objects. Image statistics can be predictors of the presence/absence of particular objects in the scene.

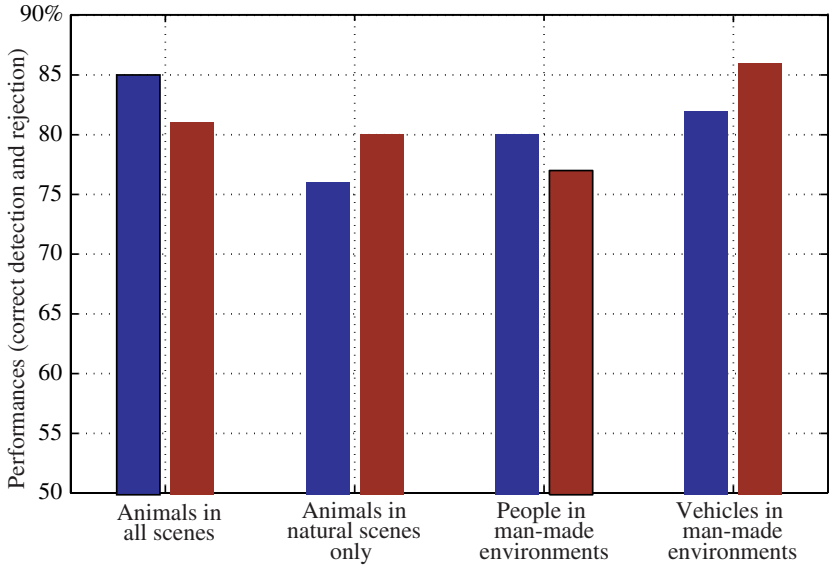

Figure 13. Performance in object prediction. For each object category we show performance for prediction of presence (left bar) of the objects and prediction of absence (right bar).

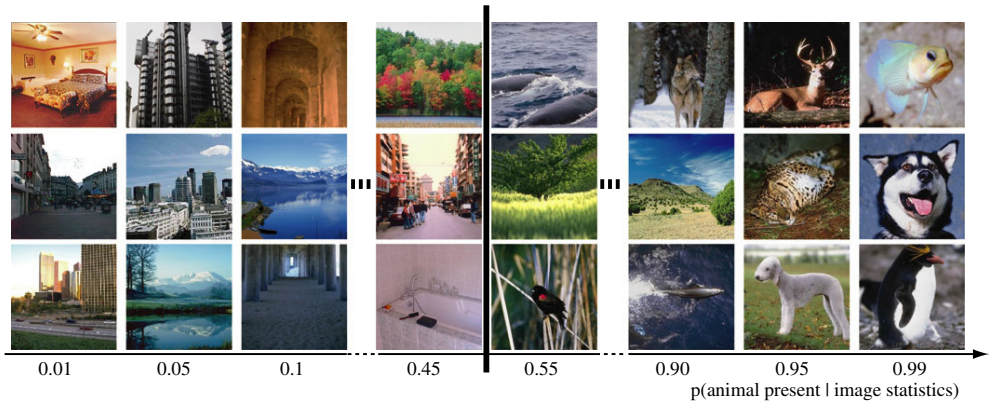

Figure 14. Illustration of images organized according to the predicted likelihood of presence of animals using image statistics. Images in the centre are ambiguous in terms of image statistics and do not produce reliable predictions. Images located at the extremes provide reliable predictions of absence of animals or presence of animals.

Figure 13 shows the performance in predicting the presence/absence of animals, people and vehicles in real-world images using the second-order statistics of the images. For each object, we used about 1000 images for the training and 1000 new images for the test. For all the objects, performance is clearly above chance (averaging 80%) and shows that global image statistics provide relevant information for object detection. Figure 14 illustrates the values of the PDF $P(O|\vec{v}_C)$ when the task is to predict whether there is an animal or not in the scene. When PDF $P(animal|\vec{v}_C) \simeq 0$ the system can reliably decide the absence of animals in the scene even before scanning it. When PDF $P(animal|\vec{v}_C) \simeq 1$ then the system can predict the presence of an animal and use more specialized mechanisms to detect it. Images in the centre,

Figure 15. Examples of images and the selected regions that are expected to contain faces based on contextual features. The regions are selected according to global image statistics and not to the actual presence of the object of interest.

$P(animal|\vec{v}_C) \simeq 0.5$, do not provide reliable predictions when using these simple features. These results corroborate studies by Thorpe and collaborators (Thorpe *et al* 1992, Rousselet *et al* 2002) showing that a cognitive task such as animal versus non-animal categorization could be performed in a feedforward way and without the need of sequential focus of attention or segmentation stages.

Image statistics can also predict other object properties such as scale (due to the correlation between scene scale and image statistics) or its location in the scene. For instance, we can learn the statistical relationship between the location of faces in an image ($\vec{x}$) and the global image statistics ($\vec{v}_C$):$P(\vec{x}|\vec{v}_C)$. In order to be able to have access to spatial information we include in the image features $\vec{v}_C$ the spectral features computed at several locations in the image. Specifically, we divide the image in 4×4 patches and we compute the power spectra at each patch. This allows us also to encode in $\vec{v}_C$ information about the spatial organization of the image (Oliva and Torralba 2001).

The learning of the probability density function $P(\vec{x}|\vec{v}_C)$ provides the relationship between the context and the more typical locations of the object of interest in the image. For modelling this PDF we use a mixture of Gaussians (Gershnfeld 1999):

$$P(\vec{x}, \vec{v}_C) = \sum_{i=1}^{M} b_i G(\vec{x}; \vec{x}_i, \boldsymbol{X}_i) G(\vec{v}_C; \vec{v}_i, \boldsymbol{V}_i). \tag{13}$$

We learn the parameters of this model using the EM algorithm and a training database of annotated images (Torralba and Sinha 2001, Torralba 2002). This PDF formalizes one aspect of the contextual control of the focus of attention. When looking for faces (or any other object of interest), attention will be directed into the candidate regions with the highest likelihood $P(\vec{x}|\vec{v}_C)$ of containing the target object, based on the past experience of the system. Note that although the contextual features encode the image using 4×4 patches, the variable $\vec{x}$ is a continuous variable. Given $\vec{v}_C$ from an image, the PDF $P(\vec{x}, \vec{v}_C)$, as a function of location $\vec{x}$, will be a mixture of Gaussian blobs (figure 15).

Figure 15 shows two examples of images and the selected regions that are expected to contain faces based on contextual features. The regions are selected according to image statistics $\vec{v}_C$. Note that in the left-hand example, pedestrians are small and do not contribute to the global image statistics. When considering the full test database, 90% of the faces were within a region of 35% of the size of the image defined by the largest values of the function $P(\vec{x}|\vec{v}_C)$.

7. Conclusion

The statistics of natural images strongly vary as a function of the interaction between the observer and the world. The paper shows how the second-order statistics of images are

correlated with scene scale and scene category and provide information to perform fast and reliable scene and object categorization. Statistical regularities might be a relevant source for top-down and contextual priming, very early in the visual processing chain. Results show how visual categorization based directly on low-level features, without grouping or segmentation stages, can benefit object localization and identification and can be used to predict the presence and absence of objects in the scene before exploring the image.

Acknowledgments

The authors would especially like to thank David Field and Bill Freeman for fruitful discussion and advice, as well as two anonymous reviewers. Thanks also to Monica Castelhano, Dan Gajewski, Aaron Pearson and Michael Mike for useful comments about the final manuscript. Both authors contributed equally to this research and authorship was arbitrarily determined. Correspondence may be sent to either author.

References

Atick J J and Redlich N A 1992 What does the retina know about natural scenes? *Neural Comput.* **4** 196–210

Baddeley R 1996 Searching for filters with 'interesting' output distributions: an uninteresting direction to explore? *Network* **7** 409–21

Baddeley R 1997 The correlational structure of natural images and the calibration of spatial representations *Cogn. Sci.* **21** 351–72

Barnard K and Forsyth D A 2001 Learning the semantics of words and pictures *Proc. Int. Conf. on Computer Vision (Vancouver)* (Los Alamitos, CA: IEEE Computer Society Press) pp 408–15

Barrow H G and Tenenbaum J M 1978 Recovering intrinsic scene characteristics from images *Computer Vision Systems* ed A Hanson and E Riseman (New York: Academic) pp 3–26

Bell A and Sejnowski T J 1997 The 'independent components' of natural scenes are edges filters *Vis. Res.* **37** 3327–38

Biederman I 1987 Recognition-by-components: a theory of human image interpretation *Psychol. Rev.* **94** 115–48

Burton G J and Moorhead I R 1987 Color and spatial structure in natural scenes *Appl. Opt.* **26** 157–70

Carson C, Belongie S, Greenspan H and Malik J 2002 Blobworld: image segmentation using expectation-maximization and its application to image querying *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1026–38

Craw I and Cameron P 1991 Parameterising images for recognition and reconstruction *British Machine Vision Conf.* ed P Mowforth (London: Springer) pp 367–70

DeValois R L and DeValois K K 1988 *Spatial Vision* (New York: Oxford)

Epstein R and Kanwisher N 1998 A cortical representation of the local visual environment *Nature* **4** 598–601

Field D J 1987 Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am.* **4** 2379–94

Field D J 1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601

Field D J 1999 Wavelets, vision and the statistics of natural scenes *Phil. Trans. R. Soc.* A **357** 2527–42

Fujita I, Tanaka K, Ito M and Cheng K 1992 Columns for visual features of objects in monkey inferotemporal cortex *Nature* **360** 343–6

Gallant J L 2000 The neural representation of shape *Seeing* ed K K DeValois and R L DeValois (San Diego, CA: Academic)

Gallant J L, Braun J and Van Essen D C 1993 Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex *Science* **259** 100–3

Gershnfeld N 1999 *The Nature of Mathematical Modeling* (Cambridge: Cambridge University Press)

Gorkani M M and Picard R W 1994 Texture orientation for sorting photos 'at a glance' *Proc. Int. Conf. on Pattern Recognition (Jerusalem)* vol 1 (New York: IEEE) pp 459–64

Guerin-Dugue A and Oliva A 2000 Classification of scene photographs from local orientations features *Pattern Recogn. Lett.* **21** 1135–40

Hancock P J, Baddeley R J and Smith L S 1992 The principal components of natural images *Network* **3** 61–70

Henderson J M, Weeks P A and Hollingworth A 1999 Effects of semantic consistency on eye movements during scene viewing *J. Exp. Psychol. Hum. Percept. Perform.* **25** 210–28

Hinkle D A and Connor C E 2002 Three-dimensional orientation tuning in macaque area V4 *Nat. Neurosci.* **5** 665–81
Hubel D H and Wiesel T N 1968 Receptive fields and functional architecture of monkey striate cortex *J. Physiol. (Lond.)* **195** 215–43
Jepson A, Richards W and Knill D 1996 Modal structures and reliable inference *Perception as Bayesian Inference* ed D Knill and W Richards (Cambridge: Cambridge University Press) pp 63–92
Liu Y and Shouval H 1994 Localized principal components of natural images—an analytic solution *Network: Comput. Neural Syst.* **5** 317–25
Logothetis N K, Pauls J, Bulthoff H H and Poggio T 1995 Shape representation in the inferior temporal cortex of macaque *Curr. Biol.* **5** 552–63
Marr D 1982 *Vision* (San Francisco, CA: Freeman)
Oliva A and Schyns P G 1997 Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli *Cogn. Psychol.* **34** 72–107
Oliva A and Schyns P G 2000 Diagnostic color blobs mediate scene recognition *Cogn. Psychol.* **41** 176–210
Oliva A and Torralba A 2001 Modeling the shape of the scene: a holistic representation of the spatial envelope *Int. J. Comput. Vis.* **42** 145–75
Oliva A and Torralba A 2002 Scene-centered representation from spatial envelope descriptors *Proc. Biologically Motivated Computer Vision (Springer Lecture Notes in Computer Science vol 2525)* ed H H Bulthoff *et al* (Berlin: Springer) pp 263–72
Oliva A, Torralba A, Guerin-Dugue A and Herault J 1999 Global semantic classification using power spectrum templates *Proc. Challenge of Image Retrieval (Electronic Workshops in Computing Series)* (Newcastle: Springer)
Olshausen B A and Field D J 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9
Olshausen B A, Sallee P and Lewicki M S 2001 Learning sparse image codes using a wavelet pyramid architecture *Adv. Neural Inform. Process. Syst.* **12** 887–93
Potter M C 1975 Meaning in visual search *J. Exp. Psychol.* **2** 509–22
Ripley B D 1996 *Pattern Recognition and Neural Networks* (Cambridge: Cambridge University Press)
Rosch E, Mervis C B, Gray W D, Johnson D M and Boyes-Braem P 1976 Basic objects in natural categories *Cogn. Psychol.* **8** 382–439
Rousselet G A, Fabre-Thorpe M and Thorpe S J 2002 Parallel processing in high-level categorization of natural images *Nat. Neurosci.* **5** 629–30
Ruderman D L 1994 The statistics of natural images *Network* **5** 517–48
Ruderman D L 1997 Origins of scaling in natural images *Vis. Res.* **37** 3385–98
Schiele B and Crowley J L 2000 Recognition without correspondence using multidimensional receptive field histograms *Int. J. Comput. Vis.* **36** 31–50
Simoncelli E P and Olshausen B A 2001 Natural image statistics and neural representation *Annu. Rev. Neurosci.* **24** 1193–216
Sirovich L and Kirby M 1987 Low-dimensional procedure for the characterization of human faces *J. Opt. Soc. Am.* **4** 519–24
Swets D L and Weng J J 1996 Using discriminant eigenfeatures for image retrieval *IEEE Trans. Pattern Anal. Mach. Intell.* **18** 831–6
Switkes E, Mayer M J and Sloan J A 1978 Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis *Vis. Res.* **18** 1393–9
Szummer M and Picard R W 1998 Indoor–outdoor image classification 1998 *IEEE Int. Workshop on Content-Based Access of Image and Video Databases* (New York: IEEE) pp 42–51
Tanaka K 1993 Neuronal mechanisms of object recognition *Science* **262** 685–8
Thorpe S, Fize D and Marlot C 1992 *Nature* **381** 520–2
Tolhurst D J, Tadmor Y and Tang C 1992 The amplitude spectra of natural images *Ophthalmic Physiol. Opt.* **12** 229–32
Torralba A 2002 Contextual modulation of target saliency *Advances in Neural Information Processing Systems* vol 14, ed T G Dietterich, S Becker and Z Ghahramani (Cambridge, MA: MIT Press)
Torralba A 2003 Contextual priming for object detection *Int. J. Comput. Vis.* **53** 169–91
Torralba A and Oliva A 2002 Depth estimation from image structure *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1226–38
Torralba A and Sinha P 2001 Statistical context priming for object detection *Proc. Int. Conf. on Computer Vision (Vancouver)* (Los Alamitos, CA: IEEE Computer Society Press) pp 763–70
Turk M and Pentland A 1991 Eigenfaces for recognition *J. Cogn. Neurosci.* **3** 71–86
Tversky B and Hemenway K 1983 Categories of environmental scenes *Cogn. Psychol.* **15** 121–49

Ullman S, Vidal-Naquet M and Sali E 2002 Visual features of intermediate complexity and their use in classification *Nat. Neurosci.* **5** 682–7

Vailaya A, Figueiredo M, Jain A and Zhang H-J 1999 Content-based hierarchical classification of vacation images *Proc. IEEE Multimedia Systems'99 (ICMCS'99, Proc. Int. Conf. on Multimedia, Computing and Systems, Florence, June 1999)*

Vailaya A, Jain A and Zhang H-J 1998 On image classification: city images versus landscapes *Pattern Recognit.* **31** 1921–36

van der Schaaf A and van Hateren J H 1996 Modeling of the power spectra of natural images: statistics and information *Vis. Res.* **36** 2759–70

van Hateren J H and van der Schaaf A 1998 Independent components filters of natural images compared with simple cells in the primary visual cortex *Proc. R. Soc.* B **265** 359–66

Van Rullen R and Thorpe S J 2001 Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex *Neural Comput.* **13** 1255–83

Vinje W E and Gallant J L 2000 Sparse coding and decorrelation in primary visual cortex during natural vision *Science* **297** 1273–6