Building the Gist of a Scene:
The Role of Global Image Features in Recognition

Aude Oliva (1) and Antonio Torralba (2)

(1) Department of Brain and Cognitive Sciences, MIT, Cambridge, USA
(2) Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA

In press, 2006. Progress in Brain Research

Running head: Scene Gist from Global Features

Keywords: scene recognition, gist, spatial envelope, global image feature, spatial frequency, natural image

Abstract

Humans can recognize the *gist* of a novel image in a single glance, independent of its complexity. How is this remarkable feat accomplished? Based on behavioral and computational evidence, this paper describes a formal approach to the representation and the mechanism of scene gist understanding, based on scene-centered, rather than object-centered primitives. We show that the structure of a scene image can be estimated by the mean of global image features, providing a statistical summary of the spatial layout properties (Spatial Envelope representation) of the scene. Global features are based on configurations of spatial scales and are estimated without invoking segmentation or grouping operations. The scene-centered approach is not an alternative to local image analysis but would serve as a feed-forward and parallel pathway of visual processing, able to quickly constrain local feature analysis and enhance object recognition in cluttered natural scenes.

Introduction

One remarkable aspect of human visual perception is that we are able to understand the meaning of a complex novel scene very quickly even when the image is blurred (Schyns & Oliva, 1994), or presented for only 20 msec (Thorpe et al., 1996). Mary Potter (1975, 1976, see also Potter et al., 2004) demonstrated that during a rapid presentation of a stream of images, observers were able to identify the semantic category of each image as well as a few objects and their attributes. This rapid understanding phenomenon can be experienced while looking at modern movie trailers which utilize many fast cuts between scenes: with a mere glimpse of each picture, you can identify each shot's meaning, the actors and the emotion depicted in each scene (Maljkovic and Martini, 2005) even though you will not necessarily remember the details of the trailer. The amount of perceptual and semantic information that observers comprehend within a glance (about 200 msec) refers to the *gist* of the scene (for a review, Oliva, 2005). In this paper, we discuss two main questions related to rapid visual scene understanding: what visual information is perceived during the course of a glance, and which mechanisms could account for the efficiency of scene gist recognition.

Figure 1: Illustration of the effect of a coarse layout (at a resolution of 8 cycles per image) on scene identification and object recognition. Despite the lack of local details in the left blurred scene, viewers are confident in describing the spatial layout of a street. However, the high resolution image reveals that the buildings are in fact furniture. This misinterpretation is not an error of the visual system. Instead, it illustrates the strength of the global spatial layout in constraining the identities of the local image structures (Navon, 1977).

Research in scene understanding has traditionally treated objects as the atoms of recognition. However, behavioral experiments on fast scene perception suggest an alternative view: that we do not need to perceive the objects in a scene to identify its semantic category. The semantic category of most real world scenes can be inferred from

their spatial layout (e.g. an arrangement of basic geometrical forms such as simple *Geons* clusters, Biederman, 1995; the spatial relationships between regions or blobs of particular size and aspect ratio, Oliva and Schyns, 2000; Sanocki & Epstein, 1997; Schyns & Oliva, 1994). Figure 1 illustrates the importance of the spatial arrangement of regions for scene and object recognition. When looking at the image on the left, viewers describe the scene as a street with cars, buildings and the sky. Despite the fact that the local information available in the image is insufficient for reliable object recognition, viewers are confident and highly consistent in their descriptions. Indeed, the blurred scene has the spatial layout of a street. When the image is shown in high resolution, new details reveal that the image has been manipulated and that the buildings are in fact pieces of furniture. Almost 30% of the image pixels correspond to an indoor scene. The misinterpretation of the low-resolution image is not a defect of the visual system. Instead, it illustrates the strength of spatial layout information in constraining the identity of the objects in normal conditions, which is especially evident in degraded conditions in which object identities cannot be inferred based only on local information (Schyns and Oliva, 1994).

In this paper, we examine what is the initial representation of a complex, real world scene image that allows for its rapid recognition. According to the global precedent hypothesis advocated by Navon (1977) and validated in numerous studies since (for a review see Kimchi, 1992), the processing of the global structure and the spatial relationships between components, precede the analysis of local details. The global precedence effect is particularly strong for images constituted of many element patterns (Kimchi, 1998), as it is the case of most real world scene pictures.

To clarify the terminology we will be using in this article, in the same way that “red” and “vertical” are local feature values of an object (Treisman & Gelade, 1980), a specific configuration of local features define a global feature value of a scene or an object. For instance, an image composed of vertical contours on the right side and horizontal contours on the left side could be estimated by one global feature receptive field tuned to respond to that specific “Horizontal -Vertical “ configuration. Global feature inputs are estimated by summations of local feature values but they have *holistic* properties of the scene as they encode the spatial relationships between components. Based on behavioral and computational experiments, we show the relevance of using a low dimensional code of the spatial layout of a scene, termed *global image features*, to represent the meaning of a scene. Global features capture the diagnostic structure of the image, giving an impoverished and coarse version of the principal contours and textures of the image that is still detailed enough to recognize the image’s *gist*. One of the principal advantages of the global image coding described here lies in its computational efficiency: there is no need to parse the image or group its components in order to represent the spatial configuration of the scene.

In this paper, we examine (1) the possible content of the global structure of a natural scene image, based on experimental results from the scene recognition literature; (2) how the global scene structure can be modeled and (3) how the global features could participate to real world scene categorization.

1- **The role of global image features on scene perception: experimental evidence**

Figure 2: -A- The two original images used to build the hybrid scenes shown above –B- A hybrid image combining the high spatial frequency (HSF, 24 cycles per image) of the beach and the low spatial frequency (LSF, 8 cycles per image) of the street scene. If you squint, blink, or defocus, the street scene should replace the beach scene (if this demonstration fails, step back from the image until your perception changes). B) The complementary hybrid image, with the street scene in HSF and the beach scene in LSF (cf. Oliva and Schyns, 1997; Schyns and Oliva, 1994).

There is considerable evidence visual input is processed at different spatial scales (from low to high spatial frequency), and psychophysical and computational studies have shown that different spatial scales offer different qualities of information for recognition purpose. On the one hand, the shape of an object is more precisely defined at high spatial frequencies but the object boundaries are interleaved by considerable noise, which requires extensive processing to be filtered out (among others, Marr and Hildreth, 1980; Shashua and Ullman, 1988). On the other hand, low scale resolution is more contrasted and might be privileged in terms of temporal processing than finer scale (Navon, 1977; Sugase, 1999), but this perceptual advantage might be offset by higher uncertainty about the identity of the blobs.

In a series of behavioral experiments, Oliva and Schyns evaluated the role that different spatial frequencies play in fast scene recognition. They created a novel kind of stimuli, termed hybrid images (see Figure 2), by superimposing two images at two different spatial scales: the low-spatial scale is obtained by filtering one image with a low-pass filter (keeping spatial frequencies up to 8 cycles/image), the high spatial scale is obtained by filtering a second image with a high-pass filter (frequencies above 24 cycles/image). The final hybrid image is composed by adding these two different filtered images (the filters are designed in such a way that there is no overlapping between the two images in the frequency domain). The examples in figure 2 show hybrid images combining a beach scene and a street scene.

The experimental results using hybrid stimuli showed that for short presentation time (30 ms, followed by a mask, Schyns & Oliva, 1994), observers used the low spatial frequency part of hybrids (street in figure 2B) when solving a scene recognition task, whereas for longer (150 ms) durations of the same image, observers categorized the image based on the high spatial frequencies (e.g. beach in figure 2B). In both cases, participants were unaware that the stimuli had two interpretations. It is important to stress that this result is not a evidence for a preference of the low-spatial frequencies in the early stages of visual processing: additional experiments (Oliva and Schyns, 1997; Schyns and Oliva, 1999) showed that, in fact, the visual system can select which spatial scale to process depending on task constraints (e.g., if the task is determining the type of emotion of a face, participants will preferentially select the low spatial frequencies, but when the task is determining the gender of the same set of faces, participants used either low, either high spatial frequencies). Furthermore, priming studies showed that within a 30 msec exposure, both low and high spatial frequency bands from a hybrid image were registered by the visual system [1] (Oliva & Schyns, 1997, Exp.1; Parker et al., 1992, 1996) but that the requirements of the task determined which scale, coarse or fine, was preferentially selected for covert processing. This suggests that the full range of spatial frequency scales is available with only 30 msec of image exposure, although the resolution at which the local features are analyzed and pre-attentively combined, when embedded in cluttered natural images, is unknown.

However, hybrid images break one important statistical property of real-world natural images, i.e., the spatial scale contiguity. To the contrary of hybrid images, contours of a natural image are correlated across scale space: a contour existing at low spatial frequency exists also at high spatial frequency. Moreover, statistical analysis of the distributions of orientations in natural images has shown that adjacent contours tend to have similar orientations whereas segments of the same contour that are further apart tend to have more disparate orientations (Geisler et al., 2001). The visual system could take advantage of spatial and spectral contiguities of contours to rapidly construct a sketch of the image structure. Boundary edges that would persist across the scale space are likely to be important structures of the image (Linderberg, 1993), and would define an initial skeleton of the image, fleshed out later by finer structures existing at higher spatial frequency scales (Linderberg, 1993; Watt, 1987; Yu, 2005). Most of the contours in natural scenes need selective attention to be bound together to form a shape of higher complexity (Treisman and Gelade, 1980; Wolfe and Bennet, 1997; Wolfe et al., 2002), but contours persistent through the scale space might need fewer attentional (or computational) resources to be represented early on. Therefore, one cannot dismiss the possibility that the analysis of fine contours and texture characteristics could be performed at the very early stage of scene perception, either because low spatial frequency luminance boundaries bootstrap the perceptual organization of finer contours (Lindeberg, 1993), or because the sparse detection of a few contours is sufficient to predict the orientation of the neighborhood edges (Geisler

[1] A hybrid scene presented for 30 ms and then masked would prime the recognition of a subsequent related scene,matching either the low or the high spatial scale of the hybrid (Oliva and Schyns, 1997, Exp. 1).

et al., 2001), or because selective attention was attending to information at a finer scale (Oliva & Schyns, 1997).

Within this framework, the analysis of visual information for fast scene understanding proceeds in a global to local manner (Navon, 1977; Treisman and Gelade, 1980), but not necessarily from low to high spatial frequencies. In other words, when we say "global and local" we do not mean "low and high" spatial frequencies. All spatial frequencies contribute to an early global analysis of the scene layout information, but organized at a rather coarse layout. Fine image edges, like long contours, are available, but their spatial organization is not encoded in a precise way. In the rest of this section we discuss some of the possible mechanisms used for performing the global analysis of the scene.

A simple and reliable global image feature for scene recognition is obtained by encoding the organization of color blobs in the image (under this representation a view of a landscape corresponds to a blue blob on the top, a green blob on the bottom and a brownish blob in the center. e.g., Carson et al., 2002; Lipson et al., 1997; Oliva & Schyns, 2000). Despite the simplicity of such a representation, it is remarkable to note the reliability of scene recognition achieved by human observers when shown a very low-resolution scene picture. Human observers are able to identify most of real world scene categories based on a resolution as low as 4 cycles/images, but only when the blurred image is in color. If the images are presented in gray levels performance drop and participants need to see higher resolution images before achieving the same recognition performance: the same performance than a with a color image with 4 cycles/image is achieved at a resolution of 8 cycles/image for a gray scale image (Oliva & Schyns, 2000, Exp. 3).

However, color blobs are not equally important for all the scenes. The diagnosticity of colored surfaces in an image seems to be a key element of fast scene recognition (Goffaux et al., 2005; Oliva & Schyns, 2000). In order to study the importance of color information, color images were altered by transforming their colors modes (e.g red surfaces became green, yellow surfaces became blue). This provides a way of understanding if color is helping as a grouping cue (and therefore the specific color is not important) or if it is diagnostic for the recognition (the color is specific to the category). For presentation time as short as 30 msec, Oliva & Schyns (2000) observed that altering colors impaired scene recognition when color was a diagnostic feature of the scene category (e.g. forests are greenish, coasts are bluish) but it had no detrimental effect for the recognition of scenes for which color was no diagnostic (e.g., some categories of urban scenes). The naming of a colored scene, relative to a grey scale scene image, was faster if it belonged to a category from which the colors distributions did not vary greatly across exemplars (for natural scenes like forest, coast, canyons), than for scene categories where color distribution varied (for indoors scenes, urban environments, see also Rousselet et al., 2005). Colored surfaces, in addition to providing useful segmentation cues for parsing the image (Carson et al., 2003), also informs about semantic properties of a place, such as its probable temperature (Greene & Oliva, 2005). The neural correlates of the role of color layout has been recently investigated by Goffaux et al (2005), who have observed an ERP frontal signal 150 msec

after image onset (a well documented temporal marker of image categorization, Thorpe et al., 1996; Van Rullen & Thorpe, 2001), when observers identified normally colored scene pictures (e.g., a green forest, a red canyon) compared to their grayscale or abnormally colored version (e.g., a purple forest, a bluish canyon). In a similar vein, Steeves et al. (2004) have shown that an individual with a profound visual form agnosia (i.e., incapable of recognizing objects based on their shape) could still identify scene pictures from colors and texture information only. Their fMRI study revealed higher activity in the parahippocampal place area (Epstein & Kanwisher, 1997) when the agnostic patient was viewing normally colored scenes pictures than when she was viewing black and white pictures.

In addition to color, research has shown that the configuration of contours is also a key diagnostic cue of scene categories (Baddeley, 1997; McCotter et al., 2005; Oliva & Torralba, 2001; Torralba & Oliva, 2003) and can help to predict the presence or absence of objects in natural images (Torralba, 2003a; Torralba & Oliva, 2003). Basic-level classes of environmental scenes (forest, street, highway, coast, etc.) as well as global properties of the 3D space (e.g. in perspective, cluttered) can be determined with a high probability, from a diagnostic set of low level image features (Fei Fei & Perona, 2005; Oliva & Torralba, 2001; Walker Renninger & Malik, 2004). For instance in urban environments, an estimation of the volume that a scene subtends is well predicted by the layout of oriented contours and texture properties. As the volume of scene space increases, the perceived image on the retina changes from large surfaces to smaller pieces, increasing the high spatial frequency content (Torralba & Oliva, 2002). A different pattern is observed when looking at a natural scene: with increasing distance from the observer, natural surfaces becomes larger and smoother, so for a given region in the image, the texture becomes coarser.

In the following section, we suggest an operational definition of global image features. The global features proposed encode a coarse representation of the organization of low and high spatial frequencies in the image.

2- Building a scene representation from global image features

High-level properties of a scene such as the degree of perspective or the mean depth of the space that the scene subtends have been found to be correlated with the configuration of low-level image features (Torralba & Oliva, 2002, 2003). Evidence from the psychophysics literature suggest that our visual system analyzes global statistical summary of the image in a pre-selective stage of visual processing or at least, with minimal attentional resources (mean orientation, Parkes et al., 2001; mean of set of objects, Ariely, 2001; Chong and Treisman, 2003). By pooling together the activity of local low-level feature detectors across large regions of the visual field, we can build a holistic and low-dimensional representation of the structure of a scene that does not require explicit segmentation of image regions and objects (as in Oliva & Torralba, 2001) and therefore, require very low computational (or attentional) ressources. This suggests that a reliable scene representation can be built, in a feed-forward manner, from the same low-level features used for local neural representations of an image (receptive fields of early visual areas, Hubel & Wiesel, 1968).

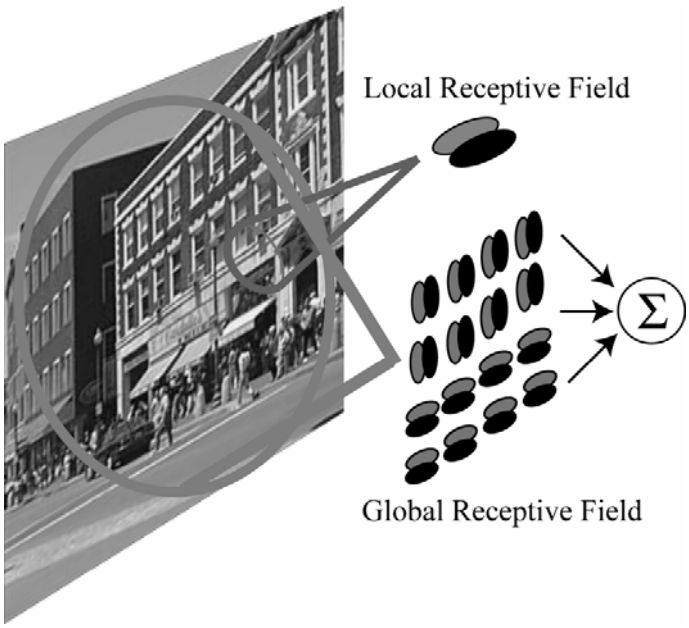

Figure 3: Illustration of a local receptive field and a global receptive field (RF). A local RF is tuned to a specific orientation and spatial scale, at a particular position in the image. A global RF is tuned to a spatial pattern of orientations and scales across the entire image. A global RF can be generated as a combination of local RFs and can, in theory, be implemented from a population of local RFs like the ones found in the early visual areas. Larger RFs, which can be selective to global scene properties, could be found in higher cortical areas (V4 or IT). The global feature illustrated in this figure is tuned to images with vertical structures at the top part and horizontal component at the bottom part, and will reply strongly to the scene street image.

For instance, in a forest scene picture, the shape of a leaf can be estimated by a set of local receptive fields (encoding oriented edges). The shape of the whole forest picture can be summarized by the configuration of many small oriented contours, distributed everywhere in the image. In the case of the forest scene, a global features encoding "fine-grained texture everywhere in the image" will provide a good summary of the texture qualities found in the image. In the case of a street scene, we will need a variety of global features encoding the perspective, the level of clutter, etc. Figure 3 illustrates a global receptive field which would respond maximally to scenes with vertical structures at the top part and horizontal components at the bottom part (as in the case of a street scene).

Given the variability of layout and feature distribution in the visual world, and given the variability of viewpoints that an observer can have on any given scene, most real world scene structures will need to be estimated not only by one, but by a collection of global features. The number of global features that can be computed is quite high. The most effective global features will be those that reflect the global structures of the visual world. Several methods of image analysis can be used to learn a suitable basis of global features (Fei Fei & Perona, 2005; Oliva & Torralba, 2001; Vailaya et al., 1998; Vogel et al, 2004) which capture the statistical regularities of natural scene images. In the modeling presented here, we only consider global features of receptive fields measuring orientations and spatial

frequencies of image components that have a spatial resolution between 1 and 8 cycles/image (see Figure 5). We employed a basis derived by principal component analysis perform on a database of thousands of real-world images.

We summarize here the steps performed for learning a set of global features corresponding to the statistical configuration of orientation and spatial frequencies existing in the real world. Each global feature value is a weighted combination of the output magnitude of a bank of multiscale oriented filters. In order to set the weights, we use principal components analysis (PCA). Due to the high-dimensionality of images, applying PCA directly to the vector composed by the concatenation of the output magnitudes of all the filters will be very computationally expensive. Several regularization techniques can be used. Here, we decided to reduce the dimensionality of the vector of features by first downsampling each filter output to a size NxN (with N ranging from 2 to 16 in the computation performed here). All the filter outputs were downsampled to the same image size, independently of the scale of the filter. As a result, each image was represented by a vector of NxNxK values (K is the number of different orientation and scales, and NxN is the number of samples used to encode, in low-resolution, the output magnitude of each filter). This gives, for each image, a vector with a relatively small dimensionality (few hundreds of elements). The dimensionality of this vector space is then reduced by applying PCA to a collection of 22000 images (the image collection includes scenes at all ranges of views, from close-up to panoramic, for both man-made and natural environments, similar to Oliva & Torralba, 2001).

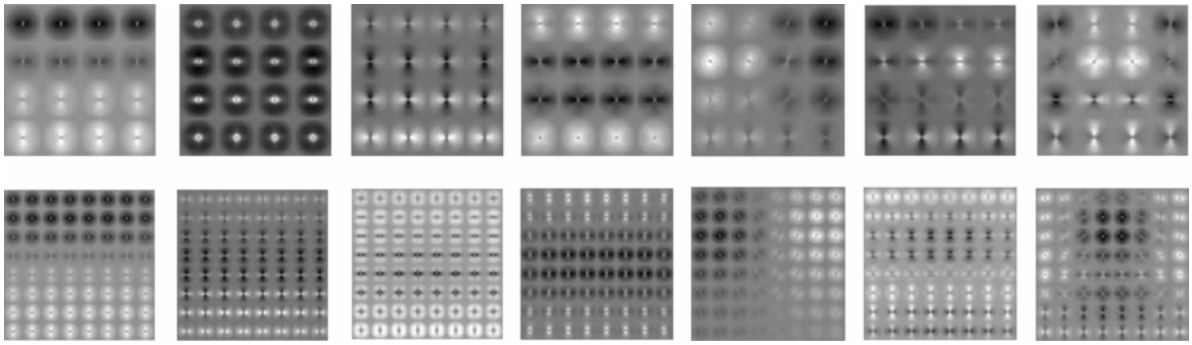

Figure 4. The Principal components of natural image statistics define the weights used to compute the global features. The set of weights are obtained by applying principal component analysis (PCA) to the responses of multiscale oriented filters to a large collection of natural images. The top row shows the 2nd to the 8th principal components for a spatial resolution of 2 cycles/image (4 x 4 regions). The first component behaves as a global average of the output of all orientations and scales and therefore it is not shown. The bottom row shows the PCs for a resolution of 4 cycles/image (8 x 8 regions). For each PC, each subimage shows, in a polar plot (low spatial frequencies are in the center of the plot), how the spatial scale and orientations are weighted at each spatial location. The white corresponds to positive value and the black to negative value. Here we refer to the PCs as global feature templates.

Figure 4 shows the first principal components of the output magnitude of multiscale oriented filters for the luminance channel for a spatial resolution of 2 and 4 cycles per image (this resolution refers to the resolution at which the magnitude of each filter output is reduced before applying the PCA. 4 cycles/image corresponds to averaging the output of each filter over NxN=8x8 non-overlapping windows, and 2 cycles/image corresponds to NxN = 4x4). Each principal component defines the weights used to compute each global feature. At each spatial location on the image, the polar plot shows the weighing of the spatial frequency at each orientation, with the lowest spatial frequencies in the center and the highest spatial frequencies along the maximum radius. In the following, we will refer to this visualization of the principal component weights (shown in figure 4) as a global feature *template*. In Figure 4, the first template responds positively for images with more texture (seen in the mid and high frequency range) in the bottom half than in the upper half of the image and responds negatively for images with more texture in the upper half than in the bottom (e.g. a landscape with trees in the background, with no view of the sky and snow on the ground). Beyond the first component, the global feature templates increase in complexity and cannot be easily described. Note that principal components are used here as an illustration of an orthogonal basis for generating global features, but they are not the only possibility. For instance, other bases could be obtained by applying independent component analysis (Bell and Sejnowski, 1997) or searching for sparse codes (Olshausen and Field, 1997).

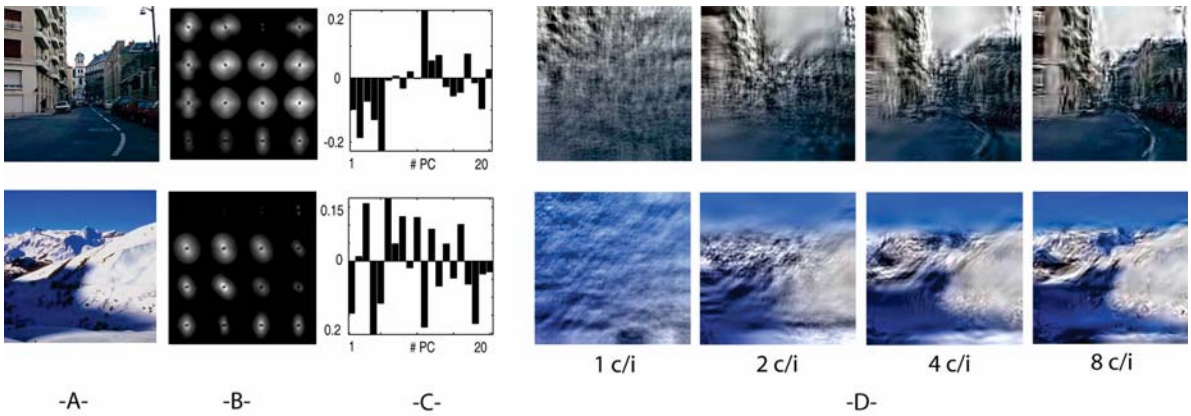

Figure 5. This figure illustrates the information preserved by the global features for two images. Fig. b) shows, on a polar plot, the average of the output magnitude of the multiscale oriented filters. Each average is computed locally by splitting the image into 4x4 non-overlapping windows. Fig. c) shows the coefficients (global features) obtained by projecting the averaged output filters into the first 20 principal components. In order to illustrate the amount of information preserved by this representation, Fig. d) shows noise images that are coerced to have the same color blobs and the same global features (N=100) than the target image. The very low frequency components (colored blobs) of the synthetic images are the same as from the original image. The high-spatial frequencies are obtained by adding noise with the constraint that the resulting image should have the same global features than the target image (this only affects the luminance channel). This constraint is imposed by an iterative algorithm. The algorithm starts from white noise. At each iteration, the noise is decomposed using the bank of multiscale oriented filters and the magnitude output of the filters is modified to match the global features of the target image. From left to right, the spatial resolution (number of windows used to average the filter outputs and the resolution of the color blobs) increases from 2x2, 4x4, 8x8, and

16x16. Note that despite the fact that the 2x2 image provides a poor reconstruction of the detailed structure of the original image, the texture contained in this representation is still relevant for scene categorization (e.g. open, closed, indoor, outdoor, natural or urban scenes).

Figure 5c shows the values of the 20 first global features (according to the ordering of principal components) for coding the structure of the street and the mountain scene. By varying the spatial resolution of the global features, we can manipulate the degree to which local features will be appropriately localized in the image. In order to illustrate the amount of information preserved by a set of global features at various resolution, Figure 5d shows noise images that are coerced to have the same color blobs (here the color information is added by projecting the image into the principal components of the color channels, and retaining only the first 32 coefficients) and the same global features (N=100) as the street and the mountain scenes. The global feature scene representation looks like a sketch version of the scene in which most of the contours and spatial frequencies from the original image have been conserved, but their spatial organization is only loosely preserved: a sketch at a resolution of 1 cycle/image (pulling local features from a 2 x 2 grid applied on image) is not informative of the spatial configuration of the image, but keeps the texture characteristics of the original scene so that we could probably decide whether the scene is a natural or man-made environment (Oliva & Torralba, 2001). For higher resolution, we can define the layout of the image and identify regions with different texture qualities, and recognize the probable semantic category of the scene (Oliva and Torralba, 2001, 2002).

3- Building the gist of the scene from global features: the Spatial Envelope model

How can we infer the semantic *gist* of a scene from the representation generated by the global image features? The *gist* refers to the meaningful information that an observer can identify from a glimpse at a scene (Oliva, 2005; Potter, 1975). The gist description usually includes the semantic label of the scene (e.g. a kitchen), a few objects and their surface characteristics (Rensink, 2000), as well as the spatial layout (e.g. the volume the scene subtends, its level of clutter, perspective) and the semantic properties related to the function of the scene. Therefore, a model of scene gist should go beyond representing the principal contours or objects of the image or classifying an image into a category: it should include a description of semantic information that human observers comprehend and infer about the scene (Oliva, 2005).

In Oliva and Torralba (2001), we introduced a holistic approach to scene recognition permitting to categorize the scene in its superordinate (e.g. urban, natural scene) and basic level categories (e.g. street, mountain), but also describing its spatial layout in a meaningful way. There are many interesting properties of a real world scene that can be defined independently of the objects. For instance, a forest scene can be described in terms of the degree of roughness and homogeneity of its textural components. These properties are in fact meaningful to a human observer who may use them for comparing similarities between

two forest images (cf. Heaps and Hendel, 1999; Rao and Lohse, 1993 for a similar account in the domain of textures).

Because a scene is inherently a three dimensional entity, Oliva & Torralba (2001) proposed that fast scene recognition mechanisms might initially be based on global properties diagnostic of the space that the scene subtends and not necessarily the objects that the scene contains. A variety of spatial properties like “openness” or “perspective” (e.g., a coast is an "open" environment) have indeed a direct transposition into global features of two-dimensional surfaces (e.g., a coast has a long horizon line). This permits to evaluate the degree of openness or mean depth of an image by measuring the distribution of local image features (Torralba & Oliva, 2002, 2003). To determine a vocabulary of spatial layout properties useful for scene recognition, we asked observers to describe real world scene images according to spatial layout and global appearance characteristics. The vocabulary given by observers (naturalness, openness, expansion, depth, roughness, complexity, ruggedness, symmetry) served to establish an initial *scene-centered description* of the image (based on spatial layout properties, Oliva and Torralba, 2002) offering an alternative to object-centered description (where a scene is identified from labeling the objects or regions, Barnard and Forsyth, 2001; Carson et al., 2002). Similar to the vocabulary used in architecture to portray the spatial properties of a place, we proposed to term the scene-centered description the *Spatial Envelope* of a scene.

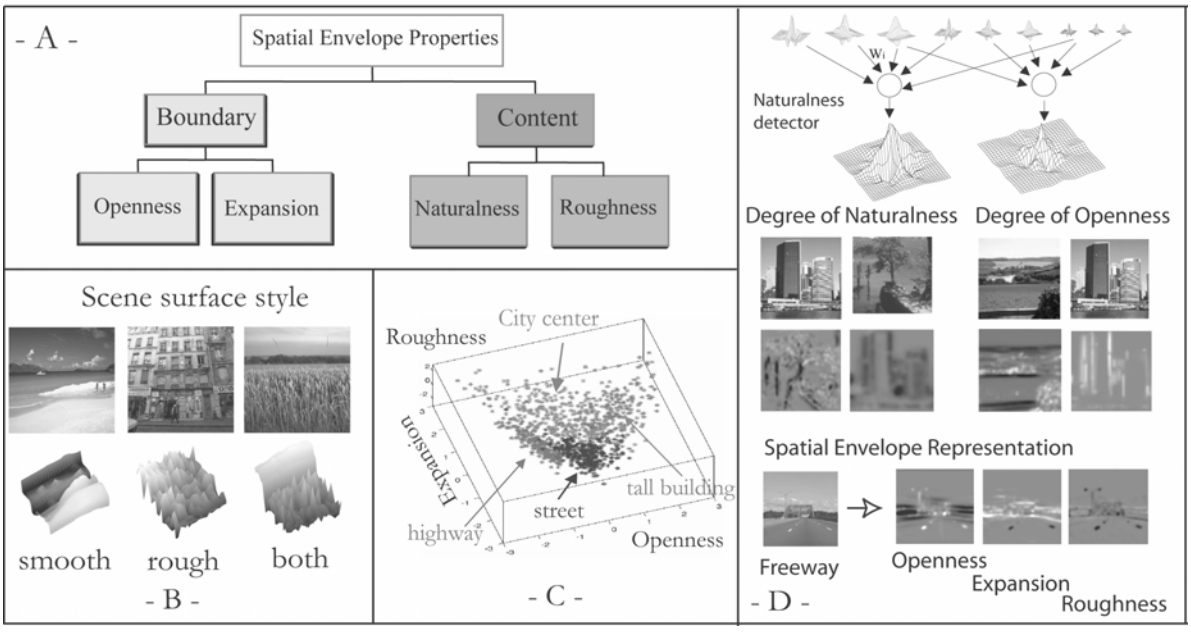

Figure 6: Schematic representation of the Spatial Envelope model as in Oliva and Torralba (2001). A- Spatial envelope properties can be classified into properties of boundaries and properties of content. For simplicity, only four properties are represented. B- Illustration of a scene as a single surface, with different “roughness” qualities. The spatial envelope does not explicitly represent objects; therefore “roughness” refers to the surface quality of the whole scene. C- Projection of ~1,200 pictures of typical urban scenes onto three spatial envelope axes (openness, roughness, expansion) as in Oliva & Torralba (2001). Semantic categories emerge, showing

that the spatial envelope representation carries information about the semantic class of a scene. D- Illustration of an implementation of the Spatial Envelope model in the form of "scene filters" applied onto the image. A complex "scene filter" can be computed as a linear combination of Gabor-like filters, and as a combination of global feature templates. Features of *openness* are shown in black and features of *closedness* are shown in white.

Figure 6 illustrates the framework of the Spatial Envelope model (details can be found in Oliva & Torralba, 2001). For simplicity, the Spatial Envelope model is presented here as a combination of four global scene properties (Fig. 6A). Object identities are not represented in the model. Within this framework, the structure of a scene is characterized by the properties of the boundaries of the space (e.g. the size of the space, its degree of openness and perspective) and the properties of its content (e.g. the style of the surface, natural or man-made, the roughness of these surfaces). Any scene image can be described by the values it takes along each spatial envelope property. For instance, to describe the degree of openness of a given environment, we could refer to a "panoramic", "open", "closed" or "enclosed" scene. A forest would be described as "an enclosed environment, with a dense isotropic texture" and a street scene would be a "man-man outdoor scene, with perspective, and medium level of clutter" (Oliva & Torralba, 2001, 2002). This level of description is meaningful to observers who can infer the probable semantic category of the scene, by providing a conceptual summary of the gist of the scene.

Computational modeling demonstrated that each spatial envelope property (naturalness, openness, expansion, etc.) could be estimated from a collection of global features templates (Figure 4) measuring each how natural, open, expanded, rough, the scene image is (Oliva and Torralba, 2001). The principal structure of a scene image is initially represented by a combination of global features, on the basis of which the spatial envelope properties can be estimated: each scene is described as a vector of meaningful values indicating the image's degree of naturalness, openness, roughness, expansion, mean depth, etc. This description refers to the spatial envelope representation of the scene image. Therefore, if spatial envelope properties capture the diagnostic structure of a scene, two images with similar spatial envelopes should also belong to the same scene semantic categories. Indeed, Oliva & Torralba observed that scenes images judged by observers to have the same categorical membership (street, highway, forest, coastline, etc.) were projected close together in a multi-dimensional space whose axes correspond to the Spatial Envelope dimensions (Fig. 6c). Neighborhood images in the spatial envelope space corresponded to images with similar spatial layout and similar conceptual description (cf. Figure 7 for exemplars of scenes and their nearest neighbors in a spatial envelope space of urban environments. Note that the spatial envelope properties (e.g. openness, naturalness, expansion, symmetry) are implemented here as a weighted combination of global features, but spatial envelope properties could also be derived from other basis of low or intermediate level features (Ullman et al., 2002). By providing semantic classification at both super-ordinate (e.g. open, natural scene) and basic levels (e.g. beach, forest) of description, the Spatial Envelope model provides a theoretical and computational framework for the representation of a

meaningful global scene structure, and a step towards understanding the representation and mechanisms of the *gist* of a scene.

Figure 7: Examples of urban scenes sharing the same spatial envelope representation (for a resolution of global features of 2 c/i). Similar scenes were retrieved as the nearest neighbors of the first image of each row, in a 5 dimensional spatial envelope representation (naturalness, openness, mean depth, expansion and roughness). On the left, the scenes on each row pertain clearly to the same semantic category. On the right, the spatial envelope similarities are less representative of basic level categories per se, however the global structure of the image (coarse layout organization and levels of details) is very similar. There are other important global scene properties that are not shown here (for instance, visual complexity is not represented here, Oliva et al, 2004) and color is not taken into account neither.

Conclusion

Research over the last decade has made substantial progress toward understanding the brain mechanisms underlying human object recognition (Grill-Spector and Malach, 2004; Kanwisher, 2003) and its modeling (Reisenhuber and Poggio, 1999; Serre et al. 2005; Torralba et al., 2004; Ullman et al., 2002). Converging evidence from behavioral, imaging and computational studies suggest that, at least in early stages of processing, mechanisms involved in natural scene recognition may be independent from those involved in recognizing objects (Fei Fei & Perona, 2004; Li et al., 2002; Marois et al., 2004; McCotter et al., 2005; Oliva & Torralba, 2001; Schyns & Oliva, 1994). Based on a review of behavioral and computational work, we argue that fast scene recognition does not need to be built on top of the processing of objects, but can be analyzed in parallel by scene-centered mechanisms. In our framework, a scene image is initially processed as a single entity and local information about objects and parts comes into play at a later stage of visual processing. We propose a formal basis of global features permitting to estimate quickly and in a feedforward manner, a meaningful representation of the scene structure. Global image feature values provide a summary of the layout of real world images that may precede and

constrain the analysis of features of higher complexity. Based on a global spatial representation of the image, the Spatial Envelope model (Oliva and Torralba, 2001) provides a conceptual framework for the representation and the mechanisms of fast scene gist interpretation. Global image features and the spatial envelope representation are not meant to be an alternative to local image analysis but serve as a parallel pathway that can, on the one hand, quickly constrain local analysis, narrowing down the search for object in cluttered, real world scenes (global contextual priming, Torralba 2003a) and, on the other hand, provide a formal instance of a feed-forward mechanism for scene context evaluation, for the guidance of attention and eye movements in the scene (Oliva et al., 2003; Torralba et al., submitted; Torralba, 2003a,b).

Evidence in favor of distinct neural mechanisms supporting scene and object recognition, at least at an earlier stage of visual processing, comes from the pioneer work of Epstein and Kanwisher (1998). They found a region of cortex referred as the parahippocampal place area (PPA) that responds more strongly to pictures of intact scenes (indoors, outdoors, close-up views), than to objects alone (Epstein et al., 2000). Furthermore, the PPA seems to be sensitive to holistic properties of the scene layout, but not to its complexity in terms of quantity of objects (Epstein and Kanwisher, 1998). The neural independence between scenes and object recognition mechanisms was recently strengthened by Goh and collaborators (2004). They observed activation of different parahippocampal regions when pictures of scenes were processed alone compared to pictures containing a prominent object, consistent within that scene. In a related vein, Bar (2004; Bar and Aminoff, 2003) found specific cortical regions (a network relating regions in the parahippocampal region and the retrosplenial cortex) involved in the analysis of the context of objects. The neural underpinnings of the global features, the spatial envelope properties or the gist of a scene, remain open issues: the global features are originally built as combinations of local low-level filters of the type found in early visual areas. Lateral and/or feedback connections could combine this information locally to be read out by higher visual areas. Receptive fields in the inferior temporal cortex and parahippocampal region cover most of the useful visual field (20-40 degrees) thus are also capable, in theory, of encoding scene layout information like the global features and the spatial envelope properties. Clearly, the mechanisms by which scene understanding occurs in the brain remain to be found.

Acknowledgments

The research was funded by an NIMH grant (1R03MH068322-01) and an award from NEC Corporation Fund for Research in Computers and Communications to A.O. Thanks to Michelle Greene, Barbara Hidalgo-Sotelo, Naomi Kenner, Talia Konkle and Thomas Serre for comments on the manuscript. Correspondence can be addressed to A.O

References

Ariely, D. (2001) Seeing sets: Representation by statistical properties. *Psychol. Sci.,* 12: 157- 162.

Baddeley, R. (1997) The correlational structure of natural images and the calibration of spatial representations. *Cogn. Sci.,* 21: 351-372.

Bar, M. (2004) Visual objects in context. *Nat. Neurosci. Rev.,* 5: 617-629.

Bar, M., and Aminoff, E. (2003) Cortical analysis of visual context. *Neuron,* 38: 347-358.

Barnard, K., and Forsyth, D.A. (2001) Learning the semantics of words and pictures. *Proc. Int. Conf., Comp. Vis*., Vancouver, Canada (pp 408-415).

Bell, A., J., and Sejnowski, T. J. (1997) The 'Independent components' of natural scenes are edge filters, *Vision Res.,* 37: 3327-3338.

Biederman, I. (1995) Visual object recognition. In *An Invitation to Cognitive Science: Visual Cognition (2nd edition).* M. Kosslyn and D.N. Osherson (eds.), 2: 121-165.

Carson, C., Belongie, S., Greenspan, H. and Malik, J. (2002) Blobworld: image segmentation using expectation-maximization and its expectation to image querying. *IEEE Trans. in PAMI*., 24:1026-1038.

Chong S. C., and Treisman, A. (2003) Representation of statistical properties. *Vision Res*., 43: 393-404.

Epstein, R., and Kanwisher, N. (1998) A Cortical Representation of the Local Visual Environment. *Nature,* 392: 598-601.

Epstein, R., Stanley, D., Harris, A., and Kanwisher, N. (2000) The Parahippocampal Place Area: Perception, Encoding, or Memory Retrieval? *Neuron***,** 23: 115-125.

Fei-Fei, L., & Perona, P. (2005) A Bayesian Hierarchical Model for Learning Natural Scene Categories. *IEEE Proc. Comp. Vis. Patt. Rec.,* 2: 524-531.

Geisler, W.S., Perry, J.S., Super, B.J., & Gallogly, D.P. (2001) Edge co-occurrence in natural images predicts contour grouping performances. *Vision Res.,* 41: 711-724.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P.G., and Rossion, B. (2005) Diagnostic colours contribute to early stages of scene categorization: behavioural and neurophysiological evidence. *Visual Cogn.,* 12: 878-892.

Goh, J.O.S., Siong, S.C., Park, D., Gutchess, A., Hebrank, A., and Chee, M.W.L. (2004) Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *J. Neurosci.,* 24: 10223-10228.

Greene, M. R., and Oliva, A. (2005) Better to run than to hide: The time course of naturalistic scene decisions [Abstract]. *J. Vis., 5,* 70a.

Grill-Spector, K., and Malach, R. (2004) The Human Visual Cortex. *Annu. Rev. Neurosci*., 27: 649-677.

Heaps, C., and Handel, C.H. (1999) Similarity and features of natural textures. *J. Exp. Psychol. Hum. Percept. Perform.*, 25: 299-320.

Hubel., D.H., and Wiesel, T.N. (1968) Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, 195: 215-243.

Kanwisher, N. (2003) The Ventral Visual Object Pathway in Humans: Evidence from fMRI. In *The Visual Neurosciences*. Ed. Chalupa, L. and Werner, J. MIT Press (pp. 1179-1189).

Kimchi, R. (1992) Primacy of wholistic processing and global/local paradigm: a critical review. *Psychol. Bull.*, 112: 24-38.

Kimchi, R. (1998) Uniform connectedness and grouping in the perceptual organization of hierarchical patterns. *J. Exp. Psychol. Hum. Percept. Perform.*, 24, 1105-1118.

Li, F. F., VanRullen, R., Koch, C. & Perona, P. (2002) Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. U. S. A,* **99**: 9596-9601.

Lindeberg, T. (1993) Detecting salient blob-lie images structures and their spatial scales with a scale-space primal sketch: a method for focus of attention. *Int. J. Comp. Vis*., 11: 283-318.

Lipson, P., Grimson, E., and Sinha, P. (1997) Configuration-based scene classification and image indexing. In *IEEE Comp. Vis. Patt. Recogn.* (pp. 1007-1013).

Maljkovic, V., and Martini, P. (2005) Short-term memory for scenes with affective content. *J. Vis.*, 5: 215-229.

Marois, R., Yi, D.J., and Chun, M. (2004) The neural fate of consciously perceived and missed events in the attentional blink. *Neuron*, 41: 465-472.

Marr, D., and Hildreth, E.C. (1980) Theory of edge detection. *Proc. Roy. Soc. Lond: B.*, 207:187-217.

McCotter, Gosselin, F., Sowden, P., and Schyns, P.G. (2005) The use of visual information in natural scenes, *Vis. Cogn.*, 12: 938-953.

Navon, D. (1977) Forest before trees: the precedence of global features in visual perception. *Cognit. Psychol*., 9: 353-383.

Oliva, A. (2005) Gist of the Scene. In Neurobiology of Attention, L. Itti, G. Rees and J. K. Tsotsos (Eds.), Elsevier, San Diego, CA (pp 251-256).

Oliva, A., and Schyns, P.G. (1997) Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognit. Psychol*., 34: 72-107.

Oliva, A., and Schyns, P.G. (2000) Diagnostic colors mediate scene recognition. *Cognit. 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. Comp. Vis*.,42: 145-175.

Oliva, A., and Torralba, A. (2002) Scene-Centered Description from Spatial Envelope Properties. Lecture Note in Computer Science Series. *Proc. 2nd Workshop on Biologically Motivated Computer Vision, Tubingen, Germany*.

Oliva, A., Mack, M.L., Shrestha, M., & Peeper, A. (2004) Identifying the perceptual dimensions of visual complexity in scenes. *Proc. of the 26th Annual Meeting of the Cogn. Sci. Soc.*, Chicago, August.

Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. (2003) Top-Down control of visual attention in object detection. *Proc. IEEE Int. Conf. Image Proc.*, 1: 253-256.

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**:** 607–609 (1996).

Parker, D.M., Lishman, J.R., and Hughes, J. (1992) Temporal integration of spatially filtered visual images. *Perception*, 21:147-160.

Parker, D.M., Lishman, J.R., and Hughes, J. (1996) Role of coarse and fine information in face and object processing. *J. Exp. Psychol. Hum. Percept. Perform.*, 22:1448-1466.

Parkes, L., Lund, J., Angelucci, A., Solomon, J.A., and Morgan, M. (2001) Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4: 739-744.

Potter, M.C. (1975) Meaning in visual scenes. *Science*, 187, 965-966.

Potter, M.C. (1976) Short-term conceptual memory for pictures. *J. Exp. Psychol. [Hum. Learn.]*, 2: 509-522.

Potter, M.C., Staub, A., and O' Connor, D.H. (2004) Pictorial and Conceptual Representation of Glimpsed Pictures. *J. Exp. Psychol. Hum. Percept. Perform.*, 30: 478-489.

Rao, A.R., and Lohse, G.L. (1993) Identifying high-level features of texture perception. *Graphical Models and Image Processing*, 55: 218-233.

Rensink, R.A. (2000). The dynamic representation of scenes. *Vis. Cogn.*, 7: 17-42.

Riesenhuber, M., and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci*.,2: 1019–1025 .

Rousselet, G.A., Joubert, O.R., and Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Vis. Cogn.*, 12: 852-877.

Sanocki, T., and Epstein, W. (1997) Priming spatial layout of scenes. *Psychol. Sci.*, 8: 374-378.

Serre, T., Wolf, L., and Poggio, T. (2005) Object Recognition with Features Inspired by Visual Cortex. *Proc. IEEE CVPR*, IEEE Computer Society Press, San Diego, June.

Shashua, A., and Ullman, S. (1988) Structural saliency: the detection of globally salient structures using a locally connected network. *Proc. Int. Conf. Comp. Vis*, 321-327.

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

Schyns, P.G., and Oliva**,** A. (1999) Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, 69: 243-265.

Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999) Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400: 869-873.

Thorpe, S., Fize, D., and Marlot, C. (1996) Speed of processing in the human visual system. *Nature*, 381: 520-522.

Torralba, A. (2003a) Modeling global scene factors in attention. *J. Opt. Soc. Am. A*, 20: 1407-1418.

Torralba, A. (2003b) Contextual priming for object detection. *Int. J. Comp. Vis*., 53: 153-167.

Torralba, A., and Oliva, A. (2002) Depth estimation from image structure. *IEEE Patt. Analys. Machine Intell.*, 24: 1226-1238.

Torralba, A., and Oliva, A. (2003) Statistics of Natural Images Categories. *Network: Computation in Neural Systems*, 14: 391-412.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2004) Sharing features: efficient boosting procedures for multiclass object detection. *Proc. IEEE CVPR*, 762- 769.

Torralba, A., Oliva, A., Castelhano, M.S., and Henderson, J.M. (2006) Contextual guidance of attention in natural scenes: the role of global features on object search. Submitted manuscript.

Treisman, A., and Gelade, G. A. (1980) Feature integration theory of attention. *Cognit. Psychol*., 12: 97-136.

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002) Visual features of intermediate complexity and their use in classification. *Nat. Neurosci*., 5: 682–687.

Vailaya, A., Jain, A., and Zhang, H.J. (1998) On image classification: city images vs. landscapes. *Patt. Recogn.*, 31:1921-1935.

Van Rullen, R., and Thorpe, S.J. (2001) The time course of visual processing: from early perception to decision making. *J. Cogn. Neurosci*., 13, 454-461.

Vogel, J., and Schiele, B. (2004) A Semantic typicality measure for natural scene categorization. *Proc. of Pattern Recognition Symposium DAGM*, Tubingen, Germany.

Yu, S.X. (2005) Segmentation Induced by Scale Invariance**.** *IEEE Conf. on Comp. Vis. Pat. Rec*., San Diego.

Walker Renninger, L., and Malik, J. (2004) When is scene identification just texture recognition? *Vision Res*., 44: 2301-2311.

Watt, R.J. (1987) Scanning from coarse to fine spatial scales in the human visual system after onset of a stimulus. *J. Opt. Soc.Am:A*, 4:2006-2021.

Wolfe, J. M., and Bennett, S.C. (1997) Preattentive object files: shapeless bundles of basic features. *Vision Res.*, 37: 25-44

Wolfe, J.M., Oliva, A., Butcher, S., and Arsenio, H. (2002) An unbinding problem: the desintegration of visible, previously attended objects does not attract attention. *J. Vis.*, 2: 256-271.