Research Article

FROM BLOBS TO BOUNDARY EDGES: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition

Philippe G. Schyns[1,2] and Aude Oliva[2,3]

[1]*Massachusetts Institute of Technology,* [2]*University of Grenoble, and* [3]*Institut National Polytechnique de Grenoble*

Abstract—*In very fast recognition tasks, scenes are identified as fast as isolated objects. How can this efficiency be achieved, considering the large number of component objects and interfering factors, such as cast shadows and occlusions? Scene categories tend to have distinct and typical spatial organizations of their major components. If human perceptual structures were tuned to extract this information early in processing, a coarse-to-fine process could account for efficient scene recognition: A coarse description of the input scene (oriented "blobs" in a particular spatial organization) would initiate recognition before the identity of the objects is processed. We report two experiments that contrast the respective roles of coarse and fine information in fast identification of natural scenes. The first experiment investigated whether coarse and fine information were used at different stages of processing. The second experiment tested whether coarse-to-fine processing accounts for fast scene categorization. The data suggest that recognition occurs at both coarse and fine spatial scales. By attending first to the coarse scale, the visual system can get a quick and rough estimate of the input to activate scene schemas in memory; attending to fine information allows refinement, or refutation, of the raw estimate.*

Imagine a simple experiment. You are sitting in front of a screen on which slides of real-world scenes are projected in rapid succession (at a rate of, e.g., 125 ms/slide). In a period of a second, you see a living room, a city, a highway, a valley, and four other distinct scenes. Your instructions are to press a button as soon as you see a highway scene.

The difficulty of this task should not be underestimated. Each picture is presented very briefly on the display, immediately masked by another unrelated picture, and the exact appearance of the target stimulus cannot be predicted from its categorical membership. Naturally, the conjunction of trucks, cars, highway lamps, and road signs would strongly suggest a highway stimulus, but objects in real scenes are often difficult to identify in 125 ms. Furthermore, in a highway scene, cars and trucks may occlude one another; shadows cast by objects such as highway lamps, advertisement panels, or trees on the side of the highway may hide other objects from view or even darken complete portions of the highway. One might expect that these factors would make it difficult to recognize scenes in 125 ms, but results of such an experiment (see Potter, 1975) suggest that human subjects are able to perform the detection task with high efficiency (see also Thorpe, Beley, & Krupa, 1993). This finding illustrates a puzzling problem in scene analysis: How can a scene be so rapidly recognized despite its variability, large number of component objects, and multiple sources of interfering factors in the image?

In general, the *scene schema hypothesis* (Antes & Penland, 1981; Biederman, 1981; de Graef, 1992; Friedman, 1979; Henderson, 1992) suggests that fast scene recognition depends on the early activation of a few scene representations in memory to drive a top-down extraction of information in the noisy input. The nature of the information responsible for the early activation of scene schemas is the topic of this article. We suggest that scene recognition operates at different precision and time scales in information-specific pathways. To get a quick and rough estimate of the input scene–to activate scene schemas—very fast recognition processes could rely on coarse information that is readily extractable. Depending on the nature of the recognition task, this rough estimate could be fine-tuned by additional bottom-up processing in fine-grained channels.

TIME- AND SPATIAL-SCALE-DEPENDENT SCENE RECOGNITION

Modern approaches to vision often conceive of scene recognition as the penultimate result of a gradual bottom-up extraction of information from sense data. For example, in the computational approach (Hildreth & Ullman, 1990; Marr, 1982), early processing is decomposed into many modules dedicated to simple tasks (e.g., edge detection, motion perception, stereo and shading perception) whose outputs are integrated into more complicated modules such as object and scene recognition (Bülthoff & Mallot, 1988). Object-based scene recognition assumes a similar flow of information: A scene is recognized after a few diagnostic objects are recognized from local information such as object contours (Antes, Mann, & Penland, 1981). Even if some objects are more diagnostic than others of a particular scene category (a car, e.g., could indicate a highway scene, but also a parking lot, a commercial center, a city, or even a bookshelf scene), the conjunction of trucks, cars, highway lamps, and road signs strongly suggests a highway stimulus. Given enough time to freely explore a scene with multiple saccades, object-based recognition should quickly reduce the uncertainty of the input scene.

Tachistoscopic studies have shown that scenes can be recognized in a single glance—in less than 250 to 300 ms (Biederman, 1981, 1988; Biederman, Mezzanotte, & Rabinowitz, 1982; Potter, 1976). Is this type of fast recognition object-based, or does it depend on some other kind of scene-based information that could bypass the expensive bottom-up hierarchy and activate scene schemas early in processing?

In the space of real-world scenes, distinct categories tend to have distinct and typical spatial organizations of their major

Address correspondence to Philippe G. Schyns, Department of Psychology, C.P. 6128, Succursale A, University of Montréal, Montréal, Québec H3C 3J7, Canada; e-mail: schyns@ai.mit.edu, schyns@ere.umontreal.edu.

components. This global organization could provide scene-specific information sufficient for fast recognition. Consider the highway category. Most highways tend to go in straight lines, with long, sweeping curves frequently "attached" to the ground. Of course, there might be atypical exemplars (such as a very twisted highway crossing a city above ground level), but most highways tend to be straight and close to the ground. By design and because of physical constraints, the global structure of city scenes is quite different. The two-dimensional projection of a typical city has a dense vertical organization, with most of the components in the lower part of the image. Even if many objects are difficult to recognize from the image (e.g., because they are partially occluded), the global organization of the image should be more characteristic of the city category than of the highway category.

Our belief is that the regularity of spatial organization of scene categories might provide the information for a mechanism by which scene schemas are activated. There is considerable psychophysical and neurophysiological evidence that the bottom-up flow of visual information is vertically organized into scale-specific channels (spatial frequency channels), with coarse channels capturing low-level qualities of images and fine channels extracting finer details (Campbell & Robson, 1968; Marr & Hildreth, 1980; Schiller & Logothetis, 1992; Watson & Nichmias, 1977).

What concerns us here is not so much the spatial frequencies themselves, but the information they convey for scene recognition. Presumably, recognition depends on blobs, edges, texture, and other higher order statistics that are vehiculated by scale-specific channels. To illustrate, the bottom pictures of Figure 1 represent the coarse information of the top scenes. At a coarse spatial scale, scenes are represented as clusters of oriented blobs of specific sizes and aspect ratios, organized in particular graphs of spatial relationships. Although no isolated object can be identified precisely at this level of resolution, the overall spatial organization of the blobs carries relevant information about the scene category. Such spatial graphs may convey scene-based information sufficient to constrain the selection of allowable scene schemas, thereby facilitating later object-based recognition. Object-based recognition might require information of another nature: It has been argued that the precise identification of objects uses primarily the object edges

Fig. 1. Examples of stimuli used in Experiment 1. The bottom pictures show the coarse information of the two scenes in the top pictures. The procedures used for computing the bottom images involved standard filtering techniques of signal processing. They are explained in more detail in the appendix.

provided by fine-grained channels (Biederman & Ju, 1988; Marr & Hildreth, 1980).

The studies reported in this article examined the respective roles of coarse blobs and fine boundary edges in scene recognition at a glance. We investigated whether these two sources of information were used in a particular sequence over the course of fast and leisurely scene processing. We then tested whether a coarse-to-fine (CtF) sequence accounted for fast scene categorization.

EXPERIMENT 1

This experiment tested how coarse and fine information contribute to scene processing. Our strategy was to provide scale-specific channels with different information by using hybrid sample stimuli in a yes/no matching task. The hybrids were a mixture of the blobs of one scene and the boundary edges of objects of another scene (e.g., the coarse information of a highway added to the boundary edges of a city scene, or vice versa; see Fig. 2). We expected to find that variation in the presentation time of hybrid scenes would change their interpretation, thereby showing distinctive uses of information at different stages of processing, compatible with a CtF analysis. For very brief presentations, we expected subjects to match the blobs of a hybrid (the scene-based information) with the target stimulus, disregarding the fine-edge structure (the object-based information). Conversely, at longer presentations of the same hybrid, we expected subjects' judgments to depend on boundary edges.

Methods

Twenty adult subjects (10 per condition) volunteered their time to participate in a yes/no matching task. A sample was presented on a computer monitor for one of two durations (30 ms in the short condition and 150 ms in the long condition), followed by a mask and a target stimulus. Subjects were instructed to indicate whether the sample matched the target by pressing the "yes" or "no" key on a computer keyboard. Stimuli were 256 gray-level pictures of four scenes: a highway, a city, a living room, and a valley. The four scenes were chosen with the constraint that their overall contrast was similar (more formally, their Fourier amplitude spectra were highly correlated with one another). Samples were either normal (N), low-passed (LF), high-passed (HF), or hybrid pictures of scenes. (See the appendix for a more detailed explanation of the filtering procedure.) A hybrid could match either the low frequencies (LF-hybrid) or the high frequencies (HF-hybrid) of the target (see Fig. 2). The target was always an N picture. The experiment consisted of a random presentation of 240 trials divided equally into matching trials (the "yes" trials) and nonmatching trials (the "no" trials). In the latter case, sample and target were pictures of different scenes.

Results and Discussion

To establish a CtF use of information, we must show (a) that the control LF and HF stimuli were perceived correctly in the short and long conditions plus (b) that the hybrid stimuli were interpreted differently in the two conditions (despite availability of coarse and fine information). Table 1 summarizes performance on matching trials for LF-hybrid, HF-hybrid, LF, HF, and N samples.

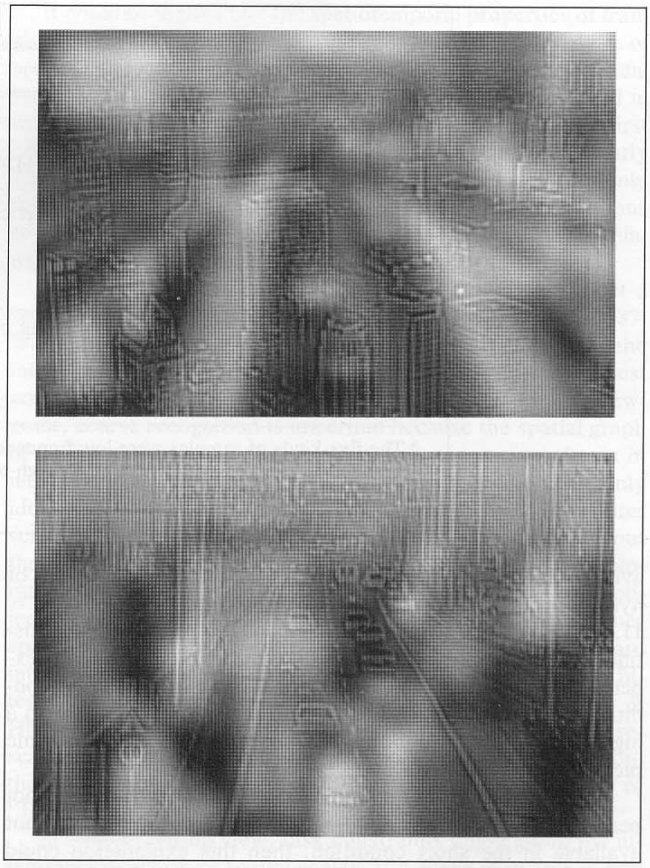

Fig. 2. Examples of the hybrid stimuli used in Experiment 1 and Experiment 2. Hybrids were constructed by combining the low-frequency (LF) components of the amplitude and phase spectra of one scene (e.g., a highway) with the high-frequency (HF) components of another scene (e.g., a city; see the appendix for more details on the procedure). Twelve hybrids were constructed by taking each possible combination of the four scenes. The top picture mixes the LF components of a highway and the HF components of a city. This picture would qualify as an LF-hybrid (vs. HF-hybrid) if the target was a highway (vs. a city). The bottom picture mixes the LF components of a city with the HF components of a highway. These stimuli were displayed on the color monitor of an Apple Macintosh Quadra. In Experiment 2, the hybrids shown in the figure were combined in an animated sequence: The top picture was presented first, and the bottom hybrid followed immediately. Each stimulus subtended 6.27 × 4.38° of visual angle on the display (subjects stood 150 cm from the screen).

A two-way analysis of variance for exposure duration (short vs. long) by sample type (N vs. LF vs. HF vs. LF-hybrid vs. HF-hybrid) indicates the significance of the main effect of condition ($F[1, 18] = 8.254$, $p = .01$), the main effect of sample type ($F[4, 72] = 103.933$, $p < .001$), and the interaction of condition and sample type ($F[4, 72] = 58.74$, $p < .001$). The

Table 1. *Hit rates, false alarm rates, and* d′ *for the yes/no matching task of Experiment 1*

Statistic	Sample[a] LF-hybrid	HF-hybrid	LF	HF	N
		Short condition			
Hit rate	.63	.28	.90	.75	.94
False alarm rate	.04	.05	.08	.05	.02
d'	2.08	1.06	2.68	2.32	>3.0
		Long condition			
Hit rate	.18	.86	.97	1.0	.98
False alarm rate	.01	.02	.01	.01	.01
d'	1.4	>3.0	>3.0	>3.0	>3.0

[a] The five kinds of samples were low-frequency hybrids (LF-hybrid), high-frequency hybrids (HF-hybrid), low-passed samples (LF), high-passed samples (HF), and normal samples (N). See the text for further explanation.

average difference of correct matches between the two types of hybrids in each duration shows a significant interaction, $t(18) = 11.18$, $p < .001$; this interaction means that subjects gave distinct interpretations of the same hybrid stimuli in the two experimental conditions. For example, subjects in the short condition preferentially matched the top picture of Figure 2 with a highway, but subjects in the long condition matched the same picture with a city.

Could it be that the fine-grained information was simply not perceived in the short condition? If boundary edges were not available in the short condition, then this explanation could partially account for the interaction. However, this interpretation is ruled out by the control samples, LF and HF, whose high d' values confirm that both low and high frequencies were detected in each condition. Still, despite the availability of all frequency information for matching, the hit-rate differences between the HF-hybrid and HF samples in the short condition (.47) and between the LF-hybrid and LF samples in the long condition (.79) indicate clearly that each condition solicited preferentially one type of recognition information. These results suggest a decoupling of coarse and fine information in fast scene analysis: Spatial relationships of sized and oriented blobs dominate very fast processing, but boundary edges and outputs of other low-level modules take over when more processing time is allowed. This CtF mode of processing suggests that there is enough information at coarse scales to initiate scene categorization.

EXPERIMENT 2

Results of Experiment 1 demonstrate CtF processing in a scene-matching task. A matching task could, however, trigger processes and representations atypical of spontaneous categorization. A test of the CtF recognition hypothesis must tap into categorization processes. Experiment 2 expanded the hybrid methodology to test whether fast scene categorization is coarse-to-fine. Our strategy was to present the visual system simultaneously with two distinct sequences of scene information (coarse-to-fine and fine-to-coarse) and record which sequence was preferentially categorized. For example, consider an animated sequence of the two hybrids of Figure 2: The top hybrid is presented first, immediately followed by the bottom hybrid. This animated sequence is inherently ambiguous: A CtF reading integrates the blobs of the top hybrid with the boundary edges of the bottom hybrid in a highway interpretation, and a fine-to-coarse (FtC) reading combines the boundary edges of the top hybrid with the blobs of the bottom hybrid in a city interpretation.

To the unbiased observer, both readings and both interpretations should be equally likely. The results of Experiment 1 suggest, however, that fast scene analysis introduces a CtF bias in processing. If fast scene categorization processes are also coarse-to-fine, the CtF highway interpretation should be used more frequently than the FtC city interpretation.

Methods

Twenty adult subjects with normal or corrected vision were paid to participate in a categorization task. Great care was taken to measure performance in conditions of fast categorization of unknown scenes. Each of four scene categories (living room, bedroom, city, and highway) was represented by four distinct picture exemplars. The 16 pictures were combined to synthesize 48 hybrid stimuli with the constraint that any particular picture could appear only three times in the hybrids. A trial consisted of a pair of hybrids presented in rapid succession on a computer monitor (45 ms/hybrid without interval, to allow for retinal fusion of the stimuli). Each pair was constructed such that the blobs of one hybrid and the boundary edges of the other hybrid were of the same scene. Thus, the CtF and the FtC readings of a pair always corresponded to two different category names. There are 12 possible combinations of 4 categories. Each trial required two categories, so the 48 hybrids were combined into 24 trials, with an equivalent number of CtF and FtC readings for each category. Subjects were instructed to categorize each scene they perceived by naming it. We recorded the

subjects' categorization responses and monitored their reaction times with a vocal key.

Results and Discussion

Subjects correctly reported either the CtF or the FtC interpretation of the stimuli on average 96% of the time, with an average reaction time of 972 ms. The 4% categorization errors were confusions between the bedroom and living room categories. Although the CtF and the FtC interpretations were equivalently available at each trial, subjects systematically reported the CtF interpretation more frequently than the FtC interpretation (67% vs. 29%, respectively; $t[19] = 6.83$, $p < .001$).

These results generalize the findings of Experiment 1 to the realm of categorization processes: A CtF use of information is preferred in fast scene identification. If blobs seem to be attended first in conditions of fast categorization, it should be noted that the opposite, FtC, sequence of information was also categorized. In Experiment 1, the boundary edges expressed by high spatial frequencies were detected and therefore available for recognition from the very first stages of processing (see data for the HF sample in Table 1). The 29% FtC interpretations in this experiment confirm this finding and demonstrate that boundary edges are sometimes the basis of fast categorization, although they are less relied on than are blobs when scenes are unknown.

It should be emphasized that subjects could not learn the particular scenes of the experiment. The pictures were not shown to subjects prior to the experiment, and each possible combination of pictures was experienced only once. The absence of repetition of trials means this was a realistic testing of natural categorization processes of real-world scenes.

GENERAL DISCUSSION

The aim of these studies was to investigate the nature of the information used at different stages of very fast scene recognition—the kind of recognition that occurs in a single eye fixation, below 250 to 300 ms. Results of Experiment 1 suggest a CtF processing strategy: The same hybrid stimulus was given a coarse-scale interpretation or a fine-scale interpretation depending on whether the hybrid was seen briefly or not. The second experiment showed that in spontaneous categorization of ambiguous sequences of stimuli, the CtF interpretation was systematically preferred over the FtC reading. Together, these results suggest a time- and spatial-scale-dependent scene recognition process in which the very first stages rely on scene-specific information and the later stages are object-based. Our findings are consistent with the idea that a regular spatial organization of major blobs in a scene could be responsible for the early activation of scene schemas in memory, but that object-based recognition dominates the later stages of processing.

These results are in line with the tachistoscopic scene research of Biederman et al. (1982), which also suggests that coarse edges of the global scene structure can activate scene schemas. Our research, however, explicitly contrasts the respective roles of coarse and fine information (as defined in a Fourier scale space) over the course of scene recognition.

It could be argued that the spatiotemporal properties of transient and sustained channels could account for a CtF analysis of spatial frequencies and explain our data. Neurophysiological studies suggest, however, that this dichotomy does not hold in the primate cortex.[1] Furthermore, the HF samples of the first experiment and the FtC categorizations of Experiment 2 clearly demonstrate that high spatial frequencies are detected (but only sometimes used for recognition) after very brief presentations of stimuli. This finding rules out a simple "hard-wired" explanation based on differences of conduction rates.

Attentional studies have shown that attention can select a spatial scale for preferred processing (Shulman & Wilson, 1987; Wong & Weisstein, 1983). This selection could arise from the interplay between task constraints and the information best conveyed at each spatial scale. From an informational viewpoint, coarse recognition is uncertain because the spatial graph of an input scene may trigger more than one scene schema in memory (e.g., the spatial graph of a city may be mistakenly identified as the one of a desk cluttered with many computer screens). Nonetheless, blobs reveal salient information about the global scene structure. Given enough time for visual exploration, a scene can also be recognized after a few criterial objects are recognized from fine-grained object contours. At finer spatial scales, however, the edges useful for recognition are interleaved with a considerable noise level, which can be filtered out only by extensive processing (Marr, 1982; Marr & Hildreth, 1980; Shashua & Ullman, 1988). In short, if coarse-scale information is more salient than fine-scale information, this saliency advantage is offset by the higher uncertainty of "blob recognition."

In the soft-wiring view, task constraints might guide the attentional selection of spatial scales. If a scene is unknown and one must categorize it very quickly, highly salient—but uncertain—information may be more efficient for a first rough estimate of the scene's identity. To paraphrase Navon (1977), forest-before-trees is a better strategy than trees-before-forest. But if one already knows what the scene might be before seeing it, the informational constraints of a fast verification task might lead to a selection of trees-before-forest. Grice, Graham, and Boroughs (1983) showed that an advantage for the global interpretations of larger letters made of smaller letters (see Navon, 1977) could be overcome when subjects could attend to and fixate the local constituent letters. Although these results do not explicitly address task constraints and do not test categorization of real scenes, they show an effect similar to the FtC categorizations in our second experiment: Sometimes fine-before-coarse is preferred in processing. We see these data as encouraging evidence to support the soft-wiring interpretation of the CtF hypothesis. This interpretation, together with the new hybrid methodology, opens new perspectives to study the

1. "There is no evidence, either within the simple cell population or within the complex cells or within the population as a whole, for a bimodal distribution of temporal properties such as would justify a dichotomy into sustained versus transient cell types. Furthermore a comparison of the temporal properties of simple versus complex cells also indicates little evidence for any significant temporal difference between these two classes of cells, which differ so drastically in their spatial properties" (de Valois & de Valois, 1990, p. 111).

interplay between spatial scale information and attention in processes of fast and leisurely real-world scene recognition.

Acknowledgments—The authors wish to thank Pierre Demartines, Jeanny Hérault, and Christian Jutten from LTIRF at the Institut National Polytechnique de Grenoble for useful discussions about stimulus computation. Many thanks also to Peter Eimas, William Estes, Peter de Graef, Gregory Murphy, Mike Paradiso, Paul Quinn, Amnon Shashua, Guy Tiberghien, Gregory Zelinsky, and two anonymous reviewers for helpful comments on an earlier version of the manuscript.

REFERENCES

Antes, J.R., Mann, S.M., & Penland, J.G. (1981, November). *Local precedence in picture naming: The importance of obligatory objects*. Paper presented at the 22nd meeting of the Psychonomic Society, Philadelphia.

Antes, J.R., & Penland, J.G. (1981). Picture context effect on eye movement patterns. In D.F. Fisher, R.A. Monty, & J.W. Sanders (Eds.), *Eye movements: Cognition and visual perception* (pp. 157–170). Hillsdale, NJ: Erlbaum.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J.R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–253). Hillsdale, NJ: Erlbaum.

Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z.W. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary approach* (pp. 370–428). Norwood, NJ: Ablex.

Biederman, I., & Ju, G. (1988). Surface vs. edge-based determinants of visual recognition. *Cognitive Psychology, 20,* 38–64.

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177.

Bülthoff, H.H., & Mallot, H. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America, 5A,* 1749–1758.

Campbell, F.W., & Robson, J.G. (1968). Application of the Fourier analysis to the visibility of gratings. *Journal of Physiology London, 88,* 551–556.

de Graef, P. (1992). Scene-context effects and models of real-world perception. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 243–259). New York: Springer-Verlag.

de Valois, R.L., & de Valois, K.K. (1990). *Spatial vision*. New York: Oxford University Press.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108,* 316–355.

Grice, G.R., Graham, L., & Boroughs, J.M. (1983). Forest before trees? It depends where you look. *Perception and Psychophysics, 33,* 121–128.

Henderson, J.M. (1992). Object identification in context: The visual processing of natural scenes. *Canadian Journal of Psychology, 46,* 2.

Hildreth, E.C., & Ullman, S. (1990). The computational study of vision. In M. Posner (Ed.), *Foundations of cognitive science* (pp. 581–630). Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Marr, D., & Hildreth, E.C. (1980). Theory of edge detection. *Proceedings of the Royal Society of London, 207B,* 187–217.

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

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

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509–522.

Schiller, P.H., & Logothetis, N.K. (1992). The color-opponent and broad-band channels of the primate visual system. *Trends in Neuroscience, 13,* 392–398.

Shashua, A., & Ullman, S. (1988). Structural saliency: The detection of globally salient structures using a locally connected network. In *Proceedings of the Second International Conference on Computer Vision* (pp. 321–327). Washington, DC: Computer Society Press.

Shulman, G.L., & Wilson, J. (1987). Spatial frequency and selective attention to local and global information. *Perception, 16,* 89–101.

Thorpe, S.J., Beley, T., & Krupa, M. (1993). Identification of rapidly presented natural images: Effects of image size. *Perception, 22,* 110.

Watson, B.W., & Nichmias, J. (1977). Patterns of temporal interaction in the detection of gratings. *Vision Research, 17,* 893–902.

Wong, E., & Weisstein, N. (1983). Sharp targets are detected better against a figure and blurred targets are detected better against a background. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 194–202.

(RECEIVED 8/3/93; REVISION ACCEPTED 11/23/93)

APPENDIX

The Fourier transform of an image produces amplitude and phase spectra in the spatial frequency domain. The amplitude spectrum denotes the strength of each frequency in the image, and the phase spectrum indicates how the spatial frequencies interact with each other to create the spatial structure of the image. In order to produce nonbiased stimuli, the amplitude spectrum of each scene was modified as follows. We equalized the spectral density energy of each scene stimulus in order to fit a model of spectral density profile, which was established by averaging over all scene stimuli. The spectral density energy can be thought of as a histogram of the energy of each spatial frequency. The equalization procedure consisted of constraining the histogram of each scene to fit the model histogram, so that the energy difference between any two spatial frequencies was the same in all images.

The inverse Fourier transform produces an image from a phase spectrum and an amplitude spectrum. All the stimuli of Experiment 1 were computed from the low-frequency (LF) components (below 2 cycles/°) and the high-frequency (HF) components (above 6 cycles/°) of the four scenes. The LF and HF components were band passed with a Butterworth filter of order two in order to avoid the problem of Gibbs ondulations of hard-limiter filters. Normal (N) stimuli were simply the addition of LF and HF components of the same scene (see Fig. 1, top pictures). Hybrids had LF and HF components from different scenes (see Fig. 2). LF stimuli had pure LF components, and HF stimuli had pure HF components. (The bottom pictures of Fig. 1 are LF stimuli.)