VISUAL COGNITION, 2005, 12 (6), 893–924

Psychology Press
Taylor & Francis Group

Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli

Li Fei-Fei

Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA

Rufin VanRullen

Centre de Recherche Cerveau et Cognition, Toulouse, France

Christof Koch

Division of Biology, California Institute of Technology, Pasadena, CA, USA

Pietro Perona

Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA

It was recently demonstrated that detecting target objects (e.g., animals) in natural scenes can be done in a dual-task paradigm, in the near absence of spatial attention. Under the same conditions, subjects were unable to perform apparently simpler tasks involving synthetic stimuli (e.g., discriminating randomly rotated Ts and Ls, or a bisected colour disc and its mirror image). Classical theories predict that attention is more critical for the recognition of complex stimuli that cannot be easily separated on a single feature dimension. Therefore, these puzzling results have raised a number of questions. If it is not the complexity of a stimulus, what then determines the recognition task's attentional requirements? How does this differ between natural and artificial stimuli? What can these observations tell us about the mechanism of natural scene processing as well as its relation to attention? Here we show that removing colour information, or doubling the amount of information to be analysed, failed to make the natural scene categorization tasks significantly more ''attention demanding''. Conversely, increasing discrimin-

Please address all correspondence to: Li Fei-Fei, Department of Electrical Engineering, Mail Code 136-93, California Institute of Technology, Pasadena, CA 91125, USA. Email: lifeifei@cs.bu.edu

This research was supported by grants from the NSF-sponsored Engineering Research Center at Caltech, the National Institutes of Health, the Keck Foundation and McDonnell Foundation. F.F.L. was supported by a Paul and Daisy Soros Fellowship for New Americans and a NSF Graduate Fellowship. R.V. was supported by a Caltech Fellowship.

© 2005 Psychology Press Ltd

http://www.tandf.co.uk/journals/pp/13506285.html DOI:10.1080/13506280444000571

ability or predictability did not diminish the need for attention in the case of synthetic stimuli. However, when the familiar letters, such as Ts or Ls, were presented upright, full attention was no longer required for discrimination. This suggests that familiarity and meaningfulness might be among the factors that determine attentional requirements for both natural and synthetic stimuli.

The visual system heavily relies on attentional resources for performing numerous discrimination, categorization, or recognition tasks (Broadbent, 1958; Deutsch & Deutsch, 1963; Kahneman, 1973; Neisser, 1967; Posner, Snyder, & Davidson, 1980). Previous research on attention has revealed that only some simple feature detection tasks seem to be free of attentional requirements, while increasingly complex tasks (e.g., feature conjunction discrimination, pattern recognition) all seem to involve some degree of attention (Braun, 1994; Treisman & Gelade, 1980; Wolfe, 1998). It has even been suggested that there could simply be no visual perception in the absence of attention (Joseph, Chun, & Nakayama, 1997; Mack & Rock, 1998). Most of these studies, however, have been performed using artificial, computer-generated stimuli; yet the visual system might be optimized for processing the stimuli that it most frequently encounters in its natural environment (Vinje & Gallant, 2000).

It was recently demonstrated that natural scene processing does not necessarily follow the same attentional constraints (Li, VanRullen, Koch, & Perona, 2002; see also Rousselet et al., 2002): Detecting the presence of an animal, or a vehicle, in a natural photograph, a genuinely challenging task for today's computer vision algorithms, can be carried out by human observers in the near absence of attention. Subjects could perform this task equally well alone or simultaneously with another attention-demanding task (i.e., deciding whether five randomly rotated Ts and Ls presented at fixation were all identical or whether one of them differed from the others). Under the same dual-task conditions, subjects could not perform apparently simpler discrimination tasks involving synthetic stimuli (discriminating between a single peripherally rotated T or L, or discriminating between a red–green and a green–red bisected disc). While this result implies that natural scenes probably hold a special status for our visual systems, it is unclear exactly what about these stimuli is responsible for this distinctiveness: Is it the mere fact that a picture is natural rather than synthetic, or are there some associated (or confounded) factors that could be responsible for determining attentional requirements? The purpose of the present study is to investigate a variety of such potential factors and establish how they affect the attentional requirements of recognition tasks using natural and artificial stimuli. We approach this question using two complementary lines of attack: On one hand, we attempt to increase the difficulty of the natural scene processing task (animal vs. nonanimal, or vehicle versus nonvehicle) and test whether these manipulations will also increase attentional requirements; on the other hand, we simplify the synthetic stimulus tasks (rotated Ls vs. Ts, or bisected two-colour discs) in various ways and test whether the attentional requirements concurrently diminish.

In Experiment 1, we control that the good dual-task performance observed for natural scene processing is not dependent on previous training on this specific task: Subjects were trained on a particular task (e.g., animal vs. nonanimal categorization) and later tested on a different one (e.g., vehicle vs. nonvehicle categorization). Compared to the large dependency of attention in the synthetic stimuli tasks (Li et al., 2002), much less attention seemed needed for categorizing the unfamiliar natural scene. We observed similar results even when the subjects were trained on a completely different dual-task paradigm without having ever practised natural scene categorization. In Experiment 2, we find that removing colour information from the natural scenes does not increase such attentional requirements. In Experiment 3, we double the amount of information to process by presenting two totally unrelated peripheral scenes instead of one, with one of them at most containing a target (animal) on 50% of the trials. Even though correct performance in this task often required both scenes to be analysed, we found no increase in attentional demands compared to the single scene categorization (Li et al., 2002).

In Experiments 4 and 5, we turn to synthetic stimuli. Our goal is to probe a variety of possibilities that might reduce the high attentional requirement for performing the synthetic stimuli tasks in Li et al. (2002). In Experiment 4 we ask whether increasing the total amount of useful signal can help dual-task performance: Four redundant peripheral stimuli were presented instead of one (i.e., four bisected two-colour discs of similar orientation). But this task remained highly attention demanding. In Experiment 5 we increased stimulus predictability by presenting the letter L or T with a fixed orientation, similar on each trial, either vertical, or diagonal. Although this manipulation greatly simplified the task, only in the case where the letters were systematically presented upright did the task become truly ''preattentive''. We thus conclude that the difference in the associated attentional requirements previously observed between natural and synthetic stimuli cannot be accounted for by a stronger ''signal-to-noise ratio'' in the case of natural scenes, but rather by the familiarity and meaningfulness of stimuli. Natural scenes are highly familiar to human observers, and their meaning can be used to determine the presence or absence of a particular object category (animal). Synthetic stimuli of the type used here (bisected discs, randomly rotated letters) are both unfamiliar and meaningless to human observers. Only when the letters were presented upright and could thus be interpreted as text (i.e., both familiar and meaningful) did the attentional load change dramatically.

GENERAL METHOD

Participants

Fifteen highly motivated California Institute of Technology undergraduate and graduate students (from 20 to 26 years' old) served as subjects in all or part of the following experiments. Two authors (LF-F, and RV) were among the

subjects for some of the experiments. Each subject enrolled for at least 15 daily sessions of 1 hour and received payment (excluding the two authors). Subjects reported normal colour vision and visual acuity (sometimes with corrective lenses or glasses), but underwent no tests in this respect. All subjects were right handed. Other than the two authors, all subjects were naïve about the purpose of the experiments.

Apparatus

Database. The pictures were complex colour scenes taken from a large commercially available CD-ROM library allowing access to several thousand stimuli. The animal category images included pictures of mammals, birds, fish, insects, and reptiles. The vehicle category images included pictures of cars, trucks, trains, airplanes, ships, and hot-air balloons. There was also a very wide range of distractor images, which included natural landscapes, city scenes, photos of food, fruits, plants, houses, and artificial objects.

Equipment. Subjects were seated in a dark room especially designed for psychophysics experiments. The seat was approximately 100 cm from a computer screen, connected to a Macintosh (OS9) computer. The refresh rate of the monitor was 75 Hz. All experimental software was programmed using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and Matlab™.

Procedure

Experimental paradigm. We use a dual-task paradigm in all of our experiments (Braun, 1994; Sperling & Dosher, 1986). Each experiment consists of three different conditions: The primary task—an attentionally demanding central task (identical in all experiments), a secondary peripheral task (in which the role of attention is investigated), and a dual-task condition in which both the central and peripheral tasks are performed concurrently. In each experiment, all trials are organized in the same way irrespective of the experimental condition (i.e., single-task condition or dual-task condition). Only the number of required responses varies between conditions.

Central letter discrimination task. In all experiments, each trial starts with a fixation cross 300 ± 100 ms before the onset of the central stimulus. At 0 ms, the central stimulus (a combination of five letters) is presented. The five letters (Ts and Ls, either all identical, or one differing from the other four), appear at nine possible locations within 1.2° eccentricity. Each letter is randomly rotated. After the central stimulus onset asynchrony (SOA; the time between the appearance of the central stimulus and the onset of the central mask), each stimulus letter is masked by the letter ''F'' rotated according to the random orientation of the stimulus letter. For a given subject, the central SOA is the same for both single-task and dual-task conditions. All trial types are presented

with equal probability. Subjects are instructed to respond by pressing ''S'' on the keyboard if the five letters are the same, or ''D'' if one of the letters differs from the other four. Figure 1 illustrates schematically the set-up of a sample trial of the dual-task paradigm. In earlier studies it was found that the central task performance is quite a sensitive measure to indicate the allocation of attentional resources (Braun, 1994). Subjects' performances on this central task decreased drastically if the SOA was slightly decreased (Li et al., 2002).

Peripheral task. In each peripheral task, the stimulus is always presented 53 ms after the central stimulus onset. Subjects respond to these tasks in a speeded fashion. They are instructed to continuously hold down the mouse button and release it as fast as possible (within 1000 ms) when they have detected the target. For a given trial, the location of the peripheral stimulus is randomly determined, keeping a distance of 6° eccentricity (Figure 1).

Training procedure. Each novel subject to the dual-task paradigm underwent a training process. It usually took more than 10 hours for a new subject to coordinate his/her motor responses well enough to answer both a speeded peripheral task and the central task. The central SOA, starting at 500 ms, was decreased after each block where the performance of this task exceeded 85%

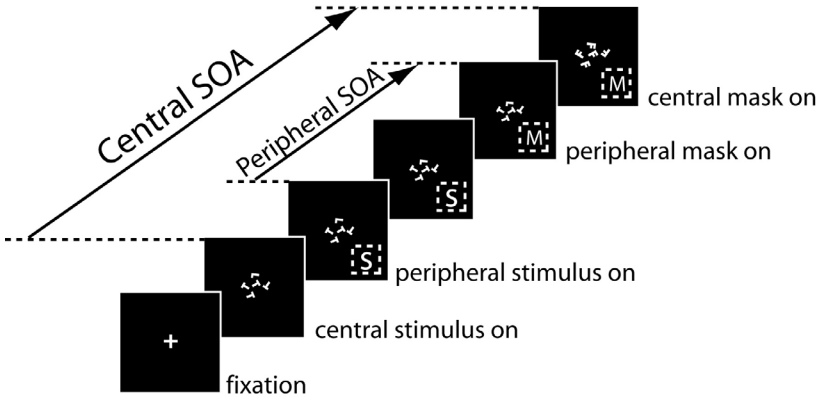

Figure 1. Dual-task experimental set-up for a single trial. A fixation cross of 1° visual angle is flashed for 300 ms at the onset of each trial, started by the subject. After that, the central letter discrimination task stimuli are presented for a central SOA amount of time. The central stimuli are then masked by an appropriate perceptual mask. Central SOA is determined individually for each subject so that the performances of the central letter discrimination task centre around 80% correct. Then, 53 ms after the onset of the central stimuli, a second stimulus is presented randomly at a peripheral location at 6° eccentricity. For different experiments reported in this paper, peripheral stimuli vary. Each peripheral stimulus is then masked by its corresponding perceptual mask, after peripheral SOA amount of time. It is important to note that to ensure that spatial attention is properly withdrawn by the central task under the dual-task condition, the peripheral mask always onsets earlier than or at the same time as the central stimulus mask.

correct. The training procedure was terminated after the subject's performance had stabilized and the central SOA was below 250 ms. This value is chosen to limit the possibility of switching attention or eye movement during stimulus presentation. Central task and peripheral task always received the same amount of training.

Data analysis. For each subject in a given experiment, we obtained two baseline performances: Central letter discrimination with attention (single-task condition) and peripheral recognition task with attention (single-task condition). Each of these two performances consists of performances of 9–15 blocks (depending on the experiment) of 96-trial experiments (Experiment 3 has 48 trials per block). We also obtain the corresponding performances for the central letter discrimination task with attention (primary task) and peripheral recognition task without attention under the dual-task condition. Each of these two performances are based on 9–15 blocks (depending on the experiment) of 96 trials each (Experiment 3 has 48 trials per block). *T*-tests are computed for each experiment to compare single- and dual-task performances. A statistical significance *p*-value of .05 is used for all statistical tests.

To visualize results, we summarize each of the experiment using a ''normalized performance'' figure. The ''normalized performance'' for each task is obtained in the following way. The averages of the two baseline performances are linearly scaled to 100% such that chance level performance remains at 50%. Then the same scaling factor for each subject is used to obtain normalized performance levels of the two tasks under the dual-task condition. In other words,

$$\text{Normalized performance} = 0.5[(P_d—0.5) / (P_s—0.5)] + 0.5$$

where P_d and P_s refer to performance in the dual-task and single-task conditions, respectively. It is important to point that since the 100% baseline is the average performance of a given task under the single-task condition (with attention), it is possible that the normalized performance, same task's performance under the dual-task condition (without attention) might sometimes be larger than 100%. This simply means that the actual performance has a higher average under the dual-task condition than under the single-task condition. Statistical tests will determine whether this difference is significant or not.

EXPERIMENT 1: NATURAL SCENE CATEGORIZATION IN THE NEAR ABSENCE OF ATTENTION DOES NOT REQUIRE TRAINING

In our study reported in Li et al. (2002), we had demonstrated the human visual system's amazing efficiency in natural scene categorization with little or no spatial attention. In these experiments, an average training period of 10–15 hours on the dual task was necessary for each subject. It is likely that this training

helps sharpen the executive control necessary for performing different tasks, particularly when they are carried out simultaneously (Pashler, 1998; Shapiro, 2001). However, training sometimes also decreases the attentional demand on perceptual processing (Joseph, Chun, & Nakayama, 1998). So, could it be that this superb efficiency in natural scene processing is mainly due to the training process that each subject had received in these experiments? We had argued that this is unlikely since the same amount of training was applied both to the natural scene categorization and the seemingly simpler synthetic stimuli tasks (rotated single T versus L, bisected disc versus its mirror image). Our data showed, however, a large discrepancy between the attentional requirements of these two types of tasks. This is difficult to explain by the same training process.

Here we further investigated the effects of training (or lack of it), particularly its influence on the natural scene stimuli. If training indeed helped performing natural scene categorization with little attention, this might be achieved through learning specific visual features critical for performing this task. It is important to note that all the data collected in the testing phase was from a set of novel images that the subjects were never trained on. Therefore simple image-based learning or memorization cannot account for the observed results. The testing stimuli, however, were drawn from the same set of images as the training images. Could subjects have, therefore, learned to categorize animal versus nonanimal (or vehicle versus nonvehicle) images because the same image types were presented repetitively? In a computational framework, it is conceivable that a specific set of ''animal filters'' (or ''vehicle filters'') were sharply tuned and enhanced during this training period. But if this were the case, training on a specific task would only help to tune the specific ''filters'' for that particular categorization task. If we tested on a different natural scene categorization task, we should be able to observe a difference in performance.

Method

We tested this hypothesis in two separate experiments. In the first experiment, we divided a group of six subjects into two groups. Both groups received the same amount of training in all tasks. Specifically, both groups were trained on the central letter discrimination task (see Procedure section) and a natural scene categorization task under both single- and dual-task conditions. Recall that the ''single-task'' condition refers to the situation where attention is available to perform the current peripheral task, while the ''dual-task'' condition refers to the situation where attention is drawn to the centre, leaving the peripheral task in the near absence of spatial attention. There are 96 trials in any given block of task under all conditions. In Group I, the three subjects were trained on the ''animal vs. nonanimal'' categorization task. They were instructed to categorize natural scenes with or without animals in a go/no-go fashion by releasing a mouse button. The task was speeded so that any response after 1000 ms was automatically registered as a ''no target'' answer. The test data was then

acquired by having them perform ''vehicle vs. nonvehicle'' categorization without any additional training. Similarly to the animal categorization task, subjects responded by releasing the mouse button when a target, i.e., vehicle(s), was detected. In Group II, the three subjects were trained on vehicle categorization and tested on animal categorization (Figure 2). Previous experiments (Li et al., 2002) have already established that a trained natural scene categorization task requires little attention. We are, therefore, interested in seeing whether such performance can be transferred from one type of categorization (e.g., animal) to another (e.g., vehicle). Namely, will the performance of vehicle categorization without attention be comparable to the performance of animal categorization without attention for Group I subjects, and vice versa for Group II?

One might argue that in the above manipulation, even though the tested natural scene category was not trained, it was nevertheless learned during training because natural scene photographs shared much commonalities (Olshausen & Field, 1996). When one is trained on one type of natural scene categorization, say ''animal scenes'', it is possible that similar image statistics help to tune the ''filters'' on other types of natural scene categories (e.g., ''vehicle scenes''). If this were the case, however, such performance should not hold for a recognition task that does not share similar stimulus statistics. For the second experiment in this section, we tested this hypothesis with another four subjects who were previously trained on the dual-task paradigm in a task that did not involve natural scene photographs. Specifically, these subjects performed a face gender discrimination task in a dual-task paradigm (Reddy, Wilken, & Koch, 2004). The gender discrimination task utilizes very different stimuli that bore little commonalities with the natural scene images (Troje & Bulthoff, 1996). After the subjects completed their training on this task, we tested them on both animal and vehicle categorization tasks (Figure 3).

Results

The bottom panels of Figure 2 illustrate the normalized performance from the cross-training experiment between two groups of subjects. In Group I, three subjects were trained on animal categorization and then tested on vehicle categorization. Their central performances during the training phase show that under the dual-task condition, they had successfully maintained their attention at the central task (single central task, average over nine blocks: 75.9 $\pm$ 3.5%, 85.4 $\pm$ 3.9%, and 83.2 $\pm$ 2.5% for each subject respectively; dual central task, average over nine blocks: 70.3 $\pm$ 5.5%, 86.1 $\pm$ 2.7%, and 80.8 $\pm$ 4.1% respectively); t-test results: $t(16) < 1.75$, $p > .05$ for each subject. During the testing condition, only one subject has a slight drop in central task performance under the dual-task condition (single central task, average over nine blocks: 75.9 $\pm$ 3.6%, 85.4 $\pm$ 3.9%, and 83.2 $\pm$ 2.5% for each subject respectively; dual central task, average over nine blocks: 71.2 $\pm$ 2.9%, 87.6 $\pm$ 3.7%, and 81.3 $\pm$

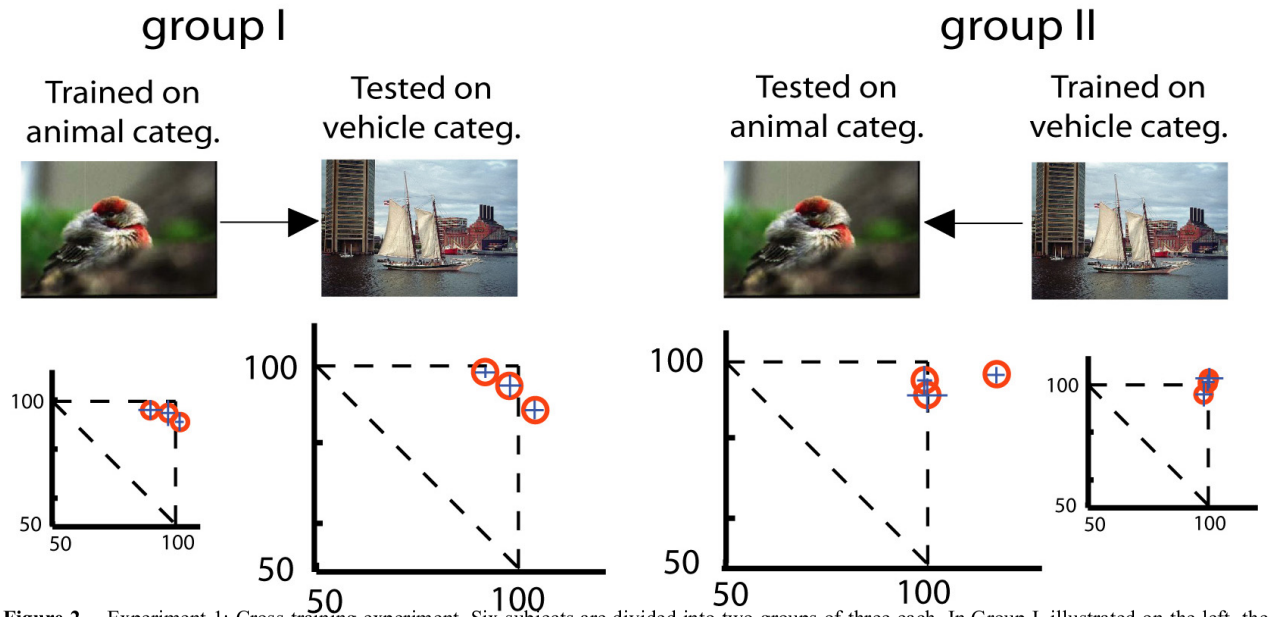

Figure 2. Experiment 1: Cross-training experiment. Six subjects are divided into two groups of three each. In Group I, illustrated on the left, the subjects are trained with central letter discrimination and animal categorization task, as well as the dual-task condition using these two tasks. After training, they are immediately tested on the vehicle categorization task in both the single-task condition (with attention) and the dual-task condition (without attention). Results are shown in normalized performance plots. In Group II, illustrated on the right, the subjects are trained with central letter discrimination and vehicle categorization task, as well as the dual-task condition using these two tasks. After training, they are immediately tested on the animal categorization task in both the single-task condition (with attention) and the dual-task condition (without attention). Results are shown in normalized performance plots. Our results illustrate that subjects need not be trained for the specific natural scene categorization task in order to perform it without spatial attention, suggesting category-specific training is not necessary to carry out this high-level task without spatial attention.

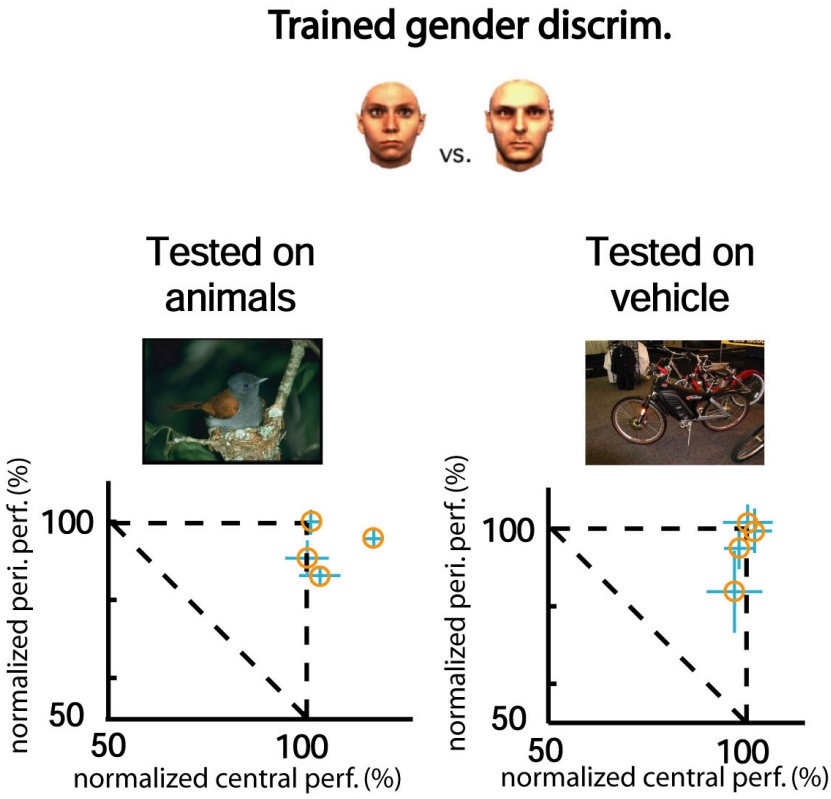

Figure 3. Experiment 1: Gender-training dual-task experiment. In this experiment, four subjects are trained on central letter discrimination task, gender discrimination as well as the dual-task condition using these two tasks. Two examples of the stimuli of the gender discrimination task are shown here (Troje & Bulthoff, 1996). Subjects are instructed to respond whether a hairless face, briefly presented in the periphery and then masked, is female or male (Reddy et al., 2004). After the training process is completed, subjects are tested on two natural scene categorizations without attention: Animal and vehicle. Normalized performances of the single- and dual-task conditions are presented for each natural scene categorization. Our results show that little training on natural scenes is needed to perform natural scene categorization.

4.1%); *t*-test results: $t(16) = 2.40$, $p = .01$ for the first subject; $t(16) < 1.75$, $p > .05$ for the rest. During the training phase, the subjects performed the animal categorization task without any interference when comparing the performances under the dual-task condition with the single-task condition (single peripheral task, average over nine blocks: 79.5 $\pm$ 0.6%, 84.3 $\pm$ 0.6%, and 77.1 $\pm$ 1.8%; dual peripheral task, average over nine blocks: 77.3 $\pm$ 5.2%, 78.1 $\pm$ 5.6%, and 74.3 $\pm$ 6.7%); *t*-test results: $t(16) < 1.75$, $p > .05$ for each subject. A similar

performance pattern is observed for these three subjects during the testing phase, in which they were directly tested on the vehicle categorization task without any prior training (single peripheral task, average over nine blocks: 83.2 $\pm$ 4.5%, 87.3 $\pm$ 4.4%, and 80.2 $\pm$ 2.6%; dual peripheral task, average over nine blocks: 82.2 $\pm$ 2.1%, 79.3 $\pm$ 3.9%, and 77.3 $\pm$ 3.7%); *t*-test results: $t(16) = 3.22$, $p = .003$ for the second subject; $t(16) < 1.75$, $p > .05$ for the rest. Note that there is a slight drop in the vehicle categorization task under dual-task condition for the second subject. This small decrease, although significant, should be viewed in the light of the results of Lee, Koch, and Braun (1999) and Li et al. (2002): When attention is taken away, performances for a simple rotated T versus L task (or red–green disc versus green–red disc) dropped much more dramatically, often to chance level (50%).

Compared to Group I, Group II subjects went through reversed training and testing categorization tasks. During the training stage, three subjects were trained on vehicle categorization only. All of them have successfully allocated attention at the centre under both the single- and dual-task condition (single central task, average over nine blocks: 68.6 $\pm$ 4.6%, 78.5 $\pm$ 6.1%, and 88.9 $\pm$ 3.2%; dual central task, average over nine blocks: 67.9 $\pm$ 2.8%, 78.6 $\pm$ 7.2%, and 88.8 $\pm$ 3.5%); *t*-test results: $t(16) < 1.75$, $p > .05$ for each subject. Their vehicle categorization task results also show that they were able to perform this task without attention (single peripheral task, average over nine blocks: 72.2 $\pm$ 7.9%, 77.4 $\pm$ 1.2%, and 81.6 $\pm$ 1.6%; dual peripheral task, average over nine blocks: 70.0 $\pm$ 4.8%, 78.3 $\pm$ 5.0%, and 81.4 $\pm$ 4.7%); *t*-test results: $t(16) < 1.75$, $p > .05$ for each subject. During the testing stage, where subjects were tested directly on animal categorization without any prior training (training was done using vehicle categorization), they maintained good performances on central task under both conditions just as they did during the training sessions (single central task, average over nine blocks: 68.6 $\pm$ 4.6%, 78.5 $\pm$ 6.1%, and 88.9 $\pm$ 3.2%; dual central task, average over nine blocks: 68.6 $\pm$ 4.2%, 88.3 $\pm$ 2.7%, and 88.4 $\pm$ 3.4%); *t*-test results: $t(16) = 3.79$, $p = .001$ for the second subject; $t(16) < 1.75$, $p > .05$ for the rest. Similarly, all three subjects performed the animal categorization task under the dual-task condition as well as under the single-task condition (single peripheral task, average over nine blocks: 74.8 $\pm$ 5.8%, 84.2 $\pm$ 1.8%, and 81.1 $\pm$ 4.3%; dual peripheral task, average over nine blocks: 70.4 $\pm$ 4.5%, 81.5 $\pm$ 3.0%, and 77.8 $\pm$ 5.1%); *t*-test results: $t(16) < 1.75$, $p > .05$ for each subject.

Figure 3 illustrates the results from the second experiment in this section. Four new subjects were trained on the dual-task paradigm with the same central letter discrimination task but a peripheral face gender discrimination task (Reddy et al., 2004). Testing of whether animal categorization and vehicle categorization require attention followed after the training phase. Three of the four subjects showed natural scene categorization performances (both animal and vehicle) in the near absence of spatial attention statistically indistinguish-

able from the same tasks performed with attention available (single peripheral animal categorization task performance, averaged over nine blocks for each: 83.0 ± 3.2%, 81.9 ± 0.6%, 83.3 ± 3.4%, and 76.9 ± 4.3%; dual peripheral animal categorization task performance, averaged over nine blocks: 79.9 ± 4.3%, 77.4 ± 3.7%, 74.3 ± 3.3%, and 77.1 ± 3.4%); *t*-test results: $t(16) = 4.46$, $p = .0002$ for the third subject; $t(16) < 1.75$, $p > .05$ for the rest. Single peripheral vehicle categorization task performance, averaged over nine blocks for each: 81.8 ± 3.9%, 83.0 ± 4.5%, 85.1 ± 2.6%, and 78.1 ± 7.3%; dual peripheral vehicle categorization task performance, averaged over nine blocks: 79.6 ± 2.5%, 80.0 ± 2.7%, 73.6 ± 14.8%, and 77.8 ± 6.4%); *t*-test results: $t(16) < 1.75$, $p > .05$ for each subject. Note that one subject showed a decrease in animal categorization performance when attention was drawn away (single task performance, average over nine blocks: 83.3 ± 3.4%; dual-task performance, average over nine blocks: 74.3 ± 3.3%); *t*-test results: $t(16) = 4.46$, $p = .0002$. Here again, this decrease is much smaller than those observed on known ''attentionally demanding'' tasks (Lee et al., 1999; Li et al., 2002). Note that the same subject's performance on vehicle categorization was not significantly different in the single- and dual-task conditions.

Both experiments in this section demonstrate that little training is needed for natural scene categorization without attention. Subjects are able to perform natural scene categorization tasks in the near absence of spatial attention without previous training on the specific task or type of stimuli.

Discussion

The first experiment in this section has shown that when trained on animal categorization (or vehicle) in the near absence of attention, subjects can perform vehicle categorization (or animal) with no further training. This effect is even more dramatically demonstrated by the second experiment, in which a totally different type of stimulus is used during training. But as soon as subjects have learned to perform this recognition task under the dual-task paradigm, they can apply this ability to a natural scene categorization task. On the contrary, our previous experiments have shown that single letter discrimination or colour disc discrimination cannot be performed without attention given the same amount of training as the natural scene categorization task (Li et al., 2002). Hochstein and colleagues argue that ''higher level'' tasks can be more easily transferred than ''lower level'' tasks (Hochstein & Ahissar, 2002). Could it be that less attentional resource is needed because natural scene categorization is carried out in ''higher'' areas of the visual system than the other synthetic tasks? We will revisit this point in both Experiments 3 and 5.

EXPERIMENT 2: NATURAL SCENE CATEGORIZATION WITHOUT ATTENTION CAN BE PERFORMED WITHOUT COLOUR

We set out to explore different factors that might contribute to the fast recognition of natural scene categories with little or no attention. A simple question to ask is whether some low-level features might have been useful cues to the categorization task. For example, it has been shown that colour histograms are very informative for natural scene recognition in both human and computer vision (Oliva & Schyns, 2000; Ostergaard & Davidoff, 1985; Price & Humphreys, 1988). In today's computer vision field, some image retrieval algorithms have utilized colour information to categorize different images (Rubner, Tomasi, & Guibas, 1998). Delorme, Richard, and Fabre-Thorpe (2000) have shown that colour is not a critical component in fast categorization, under conditions where attention was not explicitly controlled. In the present experiment, we ask the role of colour information in natural scene categorization with little or no attention, by changing the stimuli to greyscale.

Method

We use the same dual-task paradigm for this experiment as the previous experiment. The central task is an attentionally demanding letter discrimination. The peripheral task is natural scene categorization, using novel greyscale images (examples of the greyscale images are shown in Figure 4). Five subjects participated in this experiment. They were instructed to respond as fast as possible when they detected the presence of an animal in the image shown at a random peripheral position. Subjects performed 15 blocks of dual task and 12 blocks of single task. Each block consisted of 96 trials.

Results

It is first interesting to observe that individual SOAs for the natural scene categorization task are not much different from the ones observed in Li et al. (2002) where colour information was included in the images (average natural scene categorization SOA of five subjects in Li et al., 2002: 61 ms; average SOA of five subjects in current experiment: 85 ms); *t*-test result: $t(8) < 1.86$, $p > .05$. Figure 4 illustrates the normalized performances of the subjects' dual-task performances. Note that each subject has achieved a central-task performance at his/her baseline level in dual-task (single central task, average over 12 blocks for each subject: 76.9 $\pm$ 6.1%, 76.0 $\pm$ 3.2%, 74.2 $\pm$ 4.9%, 75.0 $\pm$ 4.4%, and 74.3 $\pm$ 6.6%; dual central task, average over 15 blocks for each subject: 75.4 $\pm$ 5.7%, 78.4 $\pm$ 4.7%, 71.3 $\pm$ 5.5%, 70.8 $\pm$ 5.5%, and 72.5 $\pm$ 6.3%); *t*-test results: $t(25) < 1.71$, $p > .05$ for each subject. This result assures us that much of attentional

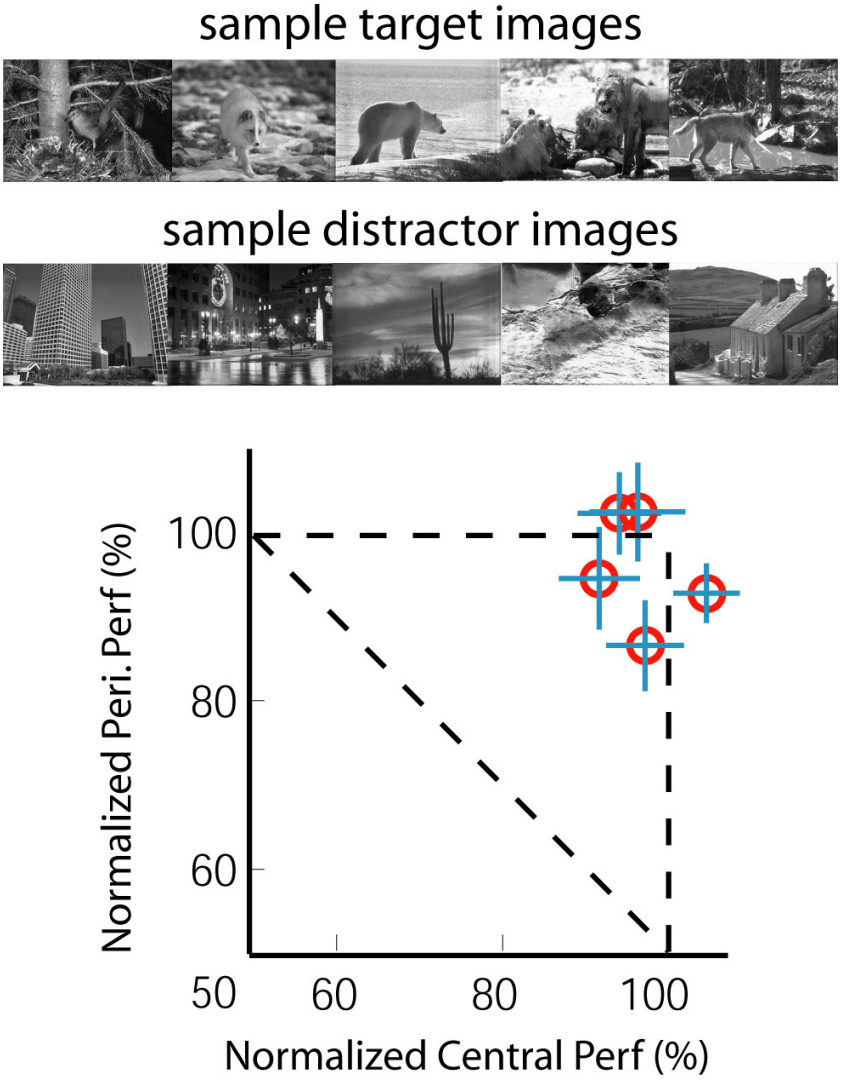

Figure 4. Experiment 2: Natural scene categorization without colour. In this experiment, the peripheral task is animal categorization with greyscale natural scene images. All experimental conditions remain identical to the ones introduced in Li et al. (2002). The only difference is that all peripheral task stimuli as well as the masks are presented in greyscale. The top two rows show some samples of the target stimuli and distractor stimuli respectively. The bottom panel indicates subjects' normalized performances of this task, showing that there is little cost in greyscale natural scene categorization when spatial attention is withdrawn.

resources are allocated to the demanding letter discrimination task under the dual-task condition. Four of the five subjects' peripheral natural scene categorization task performances remain comparable to their respective baseline performances (single peripheral task, average over 12 blocks for each subject: 79.3 $\pm$ 2.7%, 73.2 $\pm$ 7.1%, 76.9 $\pm$ 6.5%, and 74.2 $\pm$ 7.4%; dual peripheral task, average over 15 blocks for each subject: 75.2 $\pm$ 4.7%, 74.4 $\pm$ 5.1%, 71.7 $\pm$ 5.9%, and 72.9 $\pm$ 5.8%); *t*-test results: $t(25) < 1.71$, $p > .05$ for each subject. Only one subject's natural scene categorization task performance decreases slightly while attention is allocated elsewhere (single peripheral task, average over 12 blocks: 71.9 $\pm$ 5.6%; dual peripheral task, average over 15 blocks: 66.1 $\pm$ 5.3%); *t*-test results: $t(25) = 3.82$, $p = .0004$. Overall, greyscale natural scene images can be categorized rather efficiently in the near absence of spatial attention.

Discussion

Our results indicate that, when much of attention is engaged elsewhere, subjects can perform a rapid natural scene categorization task without colour. This result suggests that colour information is not critical in performing such a task in the near absence of attention. Dunai, Castiello, and Rossetti (2001) have also found that colour cues are only more informative at a longer timescale in an attentionally demanding detection task. In addition, Delorme et al. (2000) have already shown that colour information is not critical in the initial recognition of the same set of natural scenes that we are using. These findings, together with ours, suggest that natural scene categorization might be carried out by a rapid and efficient process that does not require much of the slower colour information. Torralba and Oliva (2003) suggested that some natural scene images can be categorized based on second-order statistics derived from power spectral analysis. It would be fruitful to test their hypothesis on object recognition such as animal or vehicle categorizations. To conclude, removal of colour information failed to make natural scene categorization an ''attentionally demanding'' task. Thus, if this type of natural scene processing only relies on low-level features, colour cannot be counted as one such feature.

EXPERIMENT 3: EVIDENCE FOR PARALLEL PROCESSING FOR NATURAL SCENE CATEGORIZATION WITHOUT ATTENTION

We have established so far the amazing robustness of the human visual system when categorizing natural scenes with little attention. In an attempt to search for a ''breaking point'' of this ability, we investigate the effects of natural scene categorization when the number of peripheral images is increased to two. In other words, instead of searching for a possible target in just one image, the subjects have to now search for a possible single occurrence of the target in two images. We ask whether by effectively halving the ''signal to noise ratio'' the

efficiency of this task decreases? Our rationale is that by identifying the condition in which such natural scene processing is no longer doable, we can start comparing and contrasting different conditions in order to understand the underlying neuronal mechanisms.

Method

The attentionally demanding central letter discrimination task remained the same as in the previous experiments. Five subjects performed a letter discrimination task at the central SOAs individually adjusted for each of them. The peripheral task was a coloured animal scene categorization. Each block consisted of 48 trials. In half of the trials, there were two peripheral natural scene images (''double-image'' condition), with two equally likely configurations—either one of the two images contained a scene with animal(s), or neither image contained an animal. In the other half of the trials, there was only one image, just like in the previous experiments on scene categorization (''single-image'' condition). Subjects are told to respond by lifting the mouse button when they detect the presence of an animal (or animals) in both conditions. These two types of trials were intermixed randomly throughout the experiment. Figure 5 shows the schematic set-up of this experiment. For the double-image condition, the separation between the two images varied randomly between 4° and 12° (each maintaining an eccentricity of 6°, just as in the previous experiments). Subjects were informed before the experiments of the two different possible conditions. No one reported any confusion or difficulty with the instructions. There were a total of 15 blocks for the dual-task condition and 15 blocks for each of the single-task condition.

Results

All five subjects were able to categorize novel natural scenes without spatial attention. The top two panels of Figure 6 illustrate the normalized dual-task performances for the single-image condition and double-image condition respectively. First we observe that all subjects maintained their central letter discrimination task performances under single task and dual-task conditions (single central task performance, average over 15 blocks for each subject: 75.4 $\pm$ 3.3%, 71.5 $\pm$ 1.3%, 67.7 $\pm$ 5.0%, 77.6 $\pm$ 4.9%, and 77.1 $\pm$ 3.8%; dual central task performance, average over 15 blocks for each subject: 73.8 $\pm$ 5.4%, 71.5 $\pm$ 5.5%, 67.3 $\pm$ 4.3%, 75.0 $\pm$ 6.0%, and 77.3 $\pm$ 5.9%); *t*-test results: $t(28) < 1.70$, $p > .05$ for each subject. These results indicate that spatial attention was successfully allocated to the central task for all these subjects. Now we are interested in comparing subjects' natural scene categorization

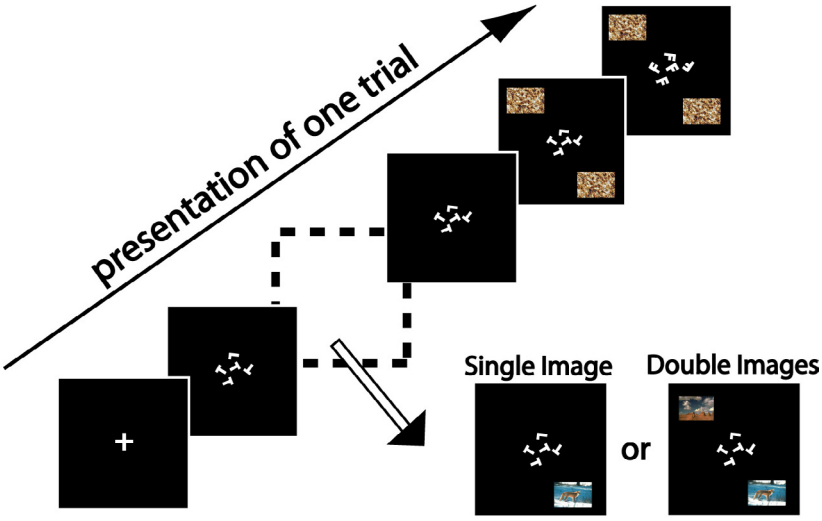

Figure 5. Experimental set-up for single-image versus double-image experiment. We illustrate here the set-up of a single trial for this experiment. The basic procedure is the same as in Figure 1. For a given trial, there are two possible peripheral stimulus presentation set-ups. For 50% of the trials, there are two unrelated natural scene images presented randomly in the periphery. The centre of each image is located at 6° eccentricity with respect to the centre of the screen. Their mutual separation varies randomly. Two similar perceptual masks follow the images after the peripheral SOA. For the other 50% of the trials, there is only one natural scene image presented randomly in the periphery, which is exactly the same case as in Li et al. (2002). At the end of the presentation, a random perceptual mask follows the image stimulus.

performances with or without attention under single-image and double-image conditions. Figure 6 shows the performance pattern of double-image categorization (single task performance, average over 15 blocks for each subject: 64.6 $\pm$ 3.1%, 71.7 $\pm$ 5.8%, 74.7 $\pm$ 7.6%, 75.4 $\pm$ 6.1%, and 73.6 $\pm$ 8.4%; dual-task performance, average over 15 blocks for each subject: 64.0 $\pm$ 6.4%, 68.1 $\pm$ 6.5%, 70.2 $\pm$ 5.8%, 72.3 $\pm$ 6.3%, and 70.7$\pm$6.0%); *t*-test results: $t(28) < 1.70$, $p > .05$ for each subject, as well as the single-image case (single task performance, average over 15 blocks for each subject: 77.1 $\pm$ 5.4%, 80.1 $\pm$ 7.1%, 87.5 $\pm$ 6.9%, 81.9 $\pm$ 4.1%, and 79.2 $\pm$ 11.0%; dual-task performance, average over 15 blocks for each subject: 76.3 $\pm$ 4.9%, 74.0 $\pm$ 4.3%, 79.0 $\pm$ 8.3%, 83.8 $\pm$ 5.9%, and 70.4 $\pm$ 9.0%); *t*-test results: $t(28) = 2.48$, $p = .01$ for the second subject, $t(28) = 1.98$, $p = .03$ for the third subject, $t(28) < 1.70$, $p > .05$ for the rest. All subjects' results show that when there are two images to process, the categorization performances with attention (single-task condition) are statistically no different from the performances without attention (dual-task

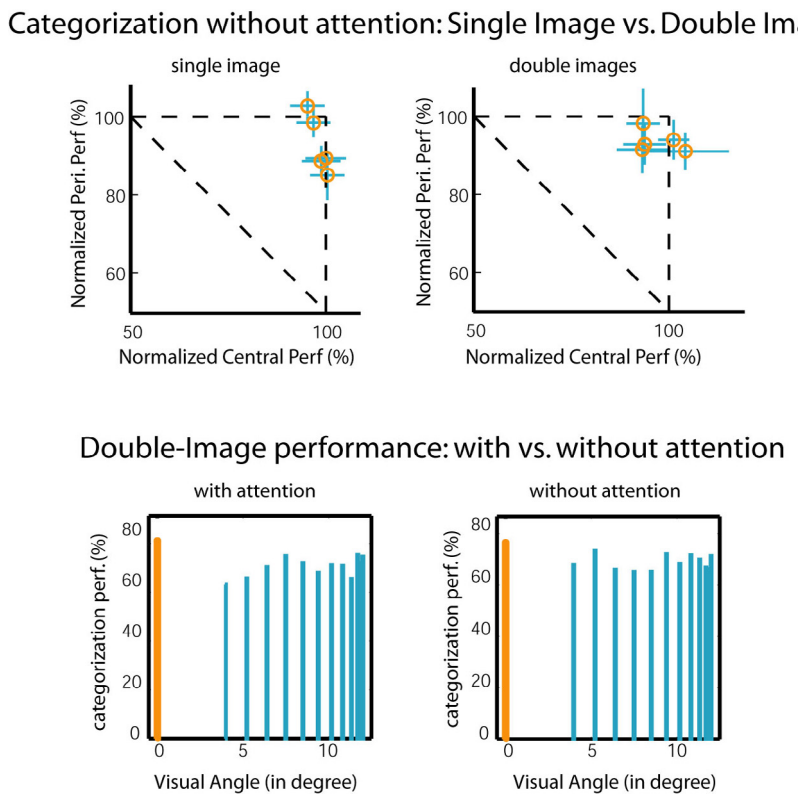

Figure 6. Experiment 3: Results of the single-image versus double-image experiment. The top two panels illustrate normalized performances of the single-image dual-task and the double-image dual-task respectively. Five subjects participated in this experiment. The results show that there is little difference between the one-image case and the two-image case, suggesting that natural scene categorization without attention is a highly parallel process. The bottom two panels break down the performances of the double-image case by the distance separating the two image centres. The two leftmost bars in each panel, placed at the 0° angle, indicate the average single-image performances of the subjects with or without attention respectively. The left panel shows the result when attention is available; whereas the right panel shows when attention is withdrawn. For each attentional condition, there is no apparent pattern of performance difference as a function of visual angle separation.

condition). There is, however, a small but significant drop for one subject when categorizing the single-image without attention compared to with attention. Similarly to the previous arguments, we think this is a rather small effect in the light of the comparative results obtained from synthetic stimuli (Lee et al., 1999; Li et al., 2002). The average baseline performances (i.e., single-task condition when attention is available) show an overall decrease in the double-image categorization condition (single-task condition for single-image case:

81.2 ± 3.9%; single-task condition for double-image case: 72.0 ± 4.4%); *t*-test result: $t(8) = 3.04$, $p = .008$. This set-size effect, different from previous results by Rousselet, Fabre-Thorpe, and Thorpe (2002), might simply be attributed to our keeping SOAs constant between the single- and double-image conditions. In addition, stimulus location was totally unpredictable in our study, whereas it was fixed in the experiments of Rousselet et al. But the main result is that, when attention is taken away, subjects were able to perform double-image categorization just as well as they did when attention was available. This result suggests that by halving the ''signal-to-noise ratio'' of the stimuli, natural scene categorization can still be efficiently carried out in the near absence of attention.

Since we randomly varied the visual angle distance between the two images under the double-image condition, we can ask whether the subjects' performances differ for different visual angle separations. The bottom two panels of Figure 6 show double-image condition performances sorted by the visual angle separation. The left panel corresponds to the condition where attention is available (single-task condition), whereas the right panel corresponds to a reduced attention condition (dual-task condition). The two leftmost bars in each panel, placed at the 0° angles, indicate the average single-image performances of the subjects with or without attention respectively. We investigated the effects of visual angle separation between the two stimulus images with a two-way ANOVA (attentional condition vs. interstimulus separation). In accordance with our previous results, there was no main effect of attention on performances: Single-task condition vs. dual-task condition, $F(1, 96) = 1.2112$, $p > .05$. Furthermore, there was no main effect of the visual angle separations between stimuli either, $F(11, 96) = 0.5817$, $p > .05$. Additionally, there was no significant interaction between these two factors, $F(11, 96) = 1.1654$, $p > .05$. Note that due to the size of the image itself, the minimal separation distance between two images is 4°.

Discussion

Our finding suggests that a natural scene categorization task not only demands little attentional resource, but is also highly parallel. When two images are presented simultaneously, subjects are able to process both of the images in search of a target object in a comparable fashion as when there is only one image. Physiological data from ERP recording also supports this finding. Rousselet et al. (2002) found that subjects are as fast for animal categorization with two images as with one image. Together our results suggest that high-level information can be accessed by the visual system in a parallel fashion with little attentional assistance. It suggests that some categorical information might be able to reach higher areas of the visual hierarchy rather efficiently, without much serial focal-attention selection.

EXPERIMENT 4: MULTIPLE COPIES OF SYNTHETIC STIMULI DO NOT HELP RECOGNITION WITHOUT ATTENTION

We have so far probed in a number of ways to what extent natural scene categorization can be carried out by the human visual system without attention. Our results tell us that such categorization is highly efficient and robust to the lack of attentional resource. In contrast, seemingly much simpler tasks involving synthetic stimuli do not enjoy this freedom of attention (Lee et al., 1999; Li et al., 2002). In the following two experiments, we turn to the question of these synthetic stimuli: What type of manipulation would decrease attentional requirements for these stimuli? In other words, through which dimension of manipulation can we make the synthetic stimuli task ''easier'' to process in the near absence of attention?

One possible hypothesis of this contrast between natural scene images and synthetic stimuli is that an object target (e.g., an elephant) in a natural image might carry multiple ''diagnostic features'' for its detection or recognition. The exact nature of these ''diagnostic features'' is unknown. But it is conceivable that many of the body parts of an animal in the animal categorization task, for example, are potential cues for the detection of the object. The synthetic stimuli, on the other hand, do not enjoy the luxury of multiple potential ''diagnostic features'' (Li et al., 2002). In that study, there were two types of synthetic stimuli, a rotated T versus a rotated L, and a vertical red–green bisected disc versus another vertical green–red disc. In the case of T versus L, the only obvious ''diagnostic feature'' is the T-junction versus the L-junction in the letters respectively. A failure to detect such junction would result in an ambiguous decision. Similarly, for the bisected discs, since the stimuli position is random from trial to trial, it is not possible to determine whether it is a red–green disc or a green–red disc by detecting the colour on half of the disc. The only ''diagnostic feature'' is the junction between the two coloured semicircles. Hence we predict that if the advantage of scene categorization without attention lies in the higher probability of detecting one or more of the possible ''diagnostic features'', then increasing the number of stimuli in the synthetic stimulus task could result in an increase of performance.

Method

We test this hypothesis using the bisected colour discs. Four subjects participated in this experiment. The basic set-up remained the same. The attentionally demanding central task was letter discrimination. In the periphery, subjects were instructed to respond when the bisected disc(s) was(were) arranged in a red–green fashion, as opposed to an equally likely green–red pattern. In half of the trials, there were four identical discs. Subjects were assured of the fact that all

stimuli were redundant. In the other half of the trials, there was only one such bisected disc (which is the same condition as in Li et al., 2002). These two types of trials were intermixed randomly within a block of trials. All discs were the same size. Figure 7 illustrates the arrangement of the stimuli. In the single-disc condition, the disc was centred at 6° eccentricity. In the four-disc condition, the centre of the four-disc array was located at 6° eccentricity. Each block consisted of 96 trials. Subjects performed 18 blocks of experiments.

Results

The hypothesis described above predicts that an increased number of peripheral stimuli might result in an increase of dual-task performance for recognition of the colour discs. The intuition is that there are more potential ''diagnostic features'' to be sampled by the visual system when the number of redundant stimuli is greater. Contrary to our prediction, we observe no improvement in dual-task performances for the trials where there are four discs rather than one. For both the single-disc and the four-disc conditions, subjects' dual-task performances of the peripheral colour disc recognition are not significantly better than chance (one-disc performance under dual-task condition, averaged over 18 blocks for each subject: 50.4 $\pm$ 7.63%, 49.3 $\pm$ 6.6%, 48.9 $\pm$ 8.0%, and 52.4 $\pm$ 6.2%; four-disc performance under dual-task condition, averaged over 18 blocks for each subject: 47.9 $\pm$ 8.1%, 49.7 $\pm$ 6.1%, 50.5 $\pm$ 3.4%, and 46.9 $\pm$ 7.4%); t-test results: $t(17) < 1.74$, $p > .05$ for each subject and each task. Note that when the peripheral colour disc recognition task is carried out with attention available, subjects' performances centre around 85% at their individual SOAs (one-disc performance under single-task condition, average over 18 blocks for each subject: 88.8 $\pm$ 8.5%, 81.3 $\pm$ 5.0%, 89.6 $\pm$ 3.3%, and 77.0 $\pm$ 7.5%; four-disc performance, averaged over 18 blocks: 91.7 $\pm$ 6.3%, 87.1 $\pm$ 5.1%, 90.5 $\pm$ 5.4%, and 74.5 $\pm$ 7.5%). It seems that the subjects cannot take advantage of the increased number of possible ''diagnostic features'', even when attention is fully available.

Discussion

We test the hypothesis that independent ''diagnostic features'' might contribute to recognition without attention. The assumption was natural scene categorization might be potentially ''easier'' than the synthetic stimuli recognition due to the multitude of ''diagnostic features''. In other words, different body parts of an animal (or vehicle) might increase the chance of detection while the synthetic stimuli tend to have a very localized, nearly singular point of ''diagnostic feature'' (e.g., bisecting line of the double colour discs). We therefore increased the probable number of features by replicating the number of stimuli from one to four. It is important to point out that this manipulation is not comparable to the one that we did in Experiment 3. In Experiment 3, we left the amount of ''target

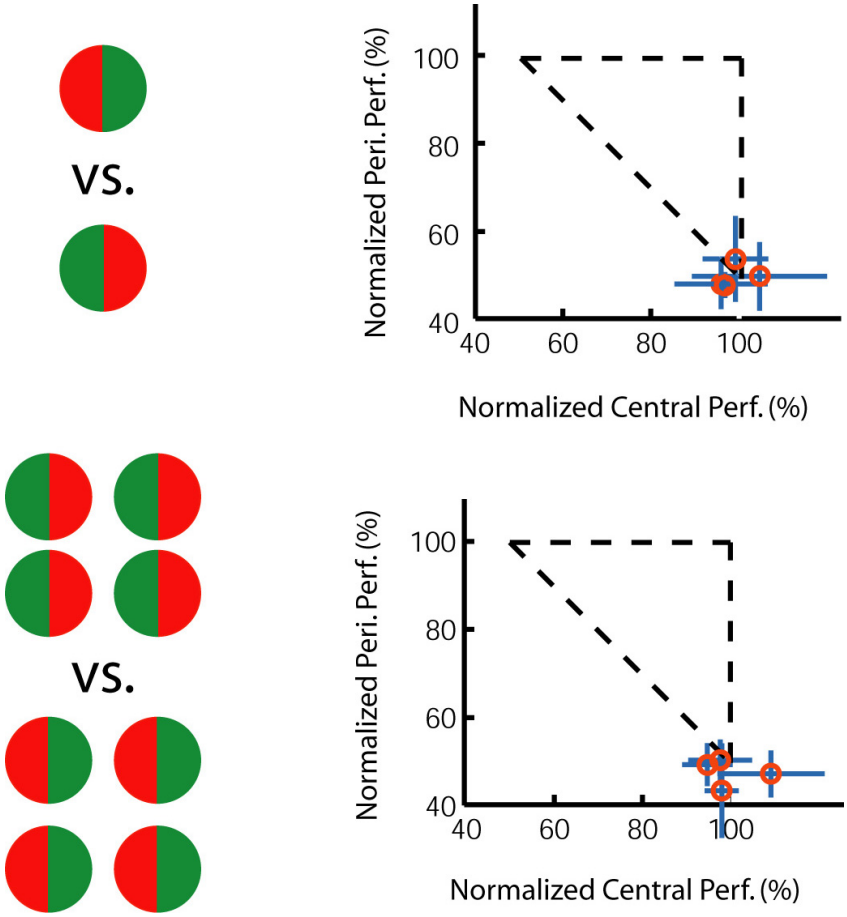

Figure 7. Experiment 4: Multiple-stimuli experiment. The basic set-up of this experiment is also a dual-task paradigm. A colour disc discrimination task is used peripherally. Two types of peripheral stimuli are mixed randomly during the experiment. The top row shows the first peripheral stimulus and the normalized dual-task performances. Peripheral recognition task is a red–green colour disc versus its mirror image, green–red colour disc. Subjects' performances of this task without attention is at chance level compared to their baseline performances, centred around 80% before normalization. The bottom row shows the second type of peripheral stimulus and the corresponding normalized dual-task performances. In this case, the recognition task remains the same as the top row, with the exception that there are four copies of the same stimuli arranged in the indicated pattern. We show here that subjects' performances of this task without attention is also at chance level, no better than the case with one copy of the stimulus.

signal'' (or probability of the presence of an ''animal'' scene) constant while doubling the amount of ''distracting noise'' (or probability of the presence of a ''nonanimal'' scene). Here the absolute number of potential ''diagnostic features'' is increased through having multiple, redundant copies of stimuli. Our results show clearly that such increase of potential diagnostic features did not help at all in recognition of synthetic stimuli with little or no spatial attention. This observation implies that it is unlikely that the bottleneck of such synthetic stimuli recognition without attention is the number of available ''diagnostic features''. Natural scenes might have an overall advantage over the synthetic stimuli used here due to the intrinsic image statistics or different processing mechanisms.

An alternative explanation also deserves further investigation. Assume that features related to the targets and distractors for the stimuli live in a high dimensional ''feature space''. Then it is possible that ''diagnostic features'' in the synthetic stimuli case might lie too closely to the ''distractor features'' of the synthetic stimuli in the ''feature space''. On the other hand, in the rich natural scene stimuli case, the ''diagnostic features'' of the targets might be much more easily isolated from the distractors than the synthetic stimuli case. If this hypothesis were true, simply repeating the number of targets in the synthetic stimuli task would not increase the discriminability of the target from the distractor, just as observed here.

EXPERIMENT 5: EVIDENCE FOR WELL-LEARNED CATEGORIES OF OBJECTS ENTAILING LESS ATTENTIONAL LOAD DURING RECOGNITION

So far our attempt to ''increase'' the difficulty of natural scene recognition without attention by reducing the amount of signal or decreasing the amount of training have not broken down the system dramatically. Similarly, adding copies of stimuli to the synthetic recognition task does not ''ease'' the task difficulty either. We hence want to test whether task ''predictability'' can be an influential factor in the recognition task without attention.

Our observations, however, also point to the direction that it could be the different level of processing that result in such different performances between natural scenes and synthetic stimuli. It has been long known that object categories are encoded in higher level visual areas such as the inferior temporal lobe (IT; Logothetis & Sheinberg, 1997; Tanaka, 1996). The most prominent object category is face for the human visual system (Epstein & Kanwisher, 1998; Rossion & Gauthier, 2002). Haxby, Gobbini, Furey, Ishai, Schouten, & Pietrini (2001) have shown differentiable fMRI patterns in IT and related areas when responding to different type of stimuli of a wide range of visual categories. So could it be that existing neuronal representations of natural scene categories are responsible for such efficient and fast recognition of natural images with little

attention? If this is the case, can we find meaningful categories of objects in synthetic stimuli to test this hypothesis?

We test here in this experiment two independent hypotheses by using the letter discrimination task as the peripheral task. The first hypothesis is that stimulus predictability might affect the attentional requirement in recognition. It is conceivable that less attention is required when subjects know beforehand the exact shapes of the stimuli to be discriminated. The second hypothesis is that well-learned object categories can be recognized with significantly less attentional load. Evidence from visual search paradigm using familiar and unfamiliar letter-like patterns indicates that visual search speed is strongly facilitated by more familiar objects (Shiffrin & Schneider, 1977; Wang, Cavanagh, & Green, 1994). Peripheral letter discrimination task in dual-task paradigm has previously been set up in such a way that the single peripheral letter stimulus is randomly rotated on each trial (Braun, 1998; Li et al., 2002). Though these letters can be considered as well-learned categories of objects, letter recognition is better trained for upright letters for obvious reasons (try reading this text upside-down). Hence we might observe some performance difference under the dual-task condition between upright letter discrimination and the original, rotated letter discrimination.

Method

We use the dual-task paradigm to test these hypotheses. As usual, we use the central letter discrimination task as the attention-demanding central task. Three conditions are tested for the peripheral letter discrimination task: Randomly rotated letter, fixed rotation, and upright positions. For all conditions, the letter discrimination task is between T and L (L is the target in a go/no-go set-up where subjects have to release a mouse button when the target is detected). The letter and its mask are located at a random position at 6° eccentricity. For a given block of 96 trials, one of the three different tasks is run and subjects are informed beforehand about the block task. There are 10 blocks tested for each of the three conditions. Figure 8 depicts the three different conditions. Four subjects participated in this experiment. A short training period of 3–4 hours proceeded the actual testing. During this period only the randomly rotated letter discrimination was trained concurrently with the central discrimination task. All three conditions were presented for an equal amount of time in the subsequent real data collection. For each subject, the peripheral letter task SOA was determined based solely on their single-task performance on the randomly rotated letter discrimination task.

Results

When attention is available, the single-task performances for all three different letter discrimination tasks are highly comparable (randomly rotated, average over four subjects and 10 blocks: $77.4 \pm 6.6\%$; fixed rotation, average over four

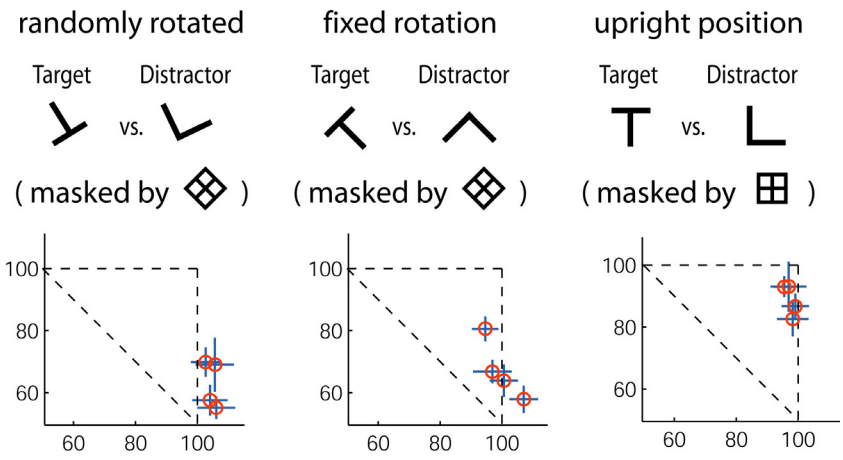

Figure 8. Experiment 5: Rotated versus fixed rotation versus upright letter experiments. The basic set-up of this experiment is also a dual-task paradigm. A letter discrimination task is used peripherally. Three different peripheral conditions and their corresponding performances are shown in three columns. In the first column, the peripheral task is a randomly rotated T versus randomly rotated L, masked by a perceptual mask. Subjects perform barely at or above chance when attention is not available. This result confirms Braun & Julesz (1998) and Li et al. (2002). The second column shows the peripheral task of a fixed rotation of T versus L, indicated in the figure. Subjects' performances of this task without attention are slightly better than the randomly rotated letter condition. In the last column, we show results of subjects' performances for the peripheral task in which the letter T and L are in their upright position, a configuration that is familiar and well learned for all of our subjects. The normalized performances indicate a clear advantage of this condition, showing a near-baseline performance when attention is withdrawn. This result suggests that tasks that are meaningful and well learned can be carried out by the visual system in a much more efficient manner.

subjects and 10 blocks: 78.8 $\pm$ 7.0%; upright, average over four subjects and 10 blocks: 84.9 $\pm$ 4.8%); pair-wise *t*-test shows no statistically significant difference between each pair of the three different tasks, $t(9) < 1.83$, $p > .05$ for all cases. This suggests that when there is abundant attentional resource, carrying out the letter discrimination task for all rotation conditions is similar. It is important to point out that the fixed rotation and upright conditions have similar performances as the randomly rotated condition. This indicates that the SOA that was determined for each subject for the peripheral task was effective for all condition. We observed, however, different performance results under the dual-task paradigm. For the randomly rotated letter task, subjects' performances were congruent with what is reported in previous studies (Braun & Julesz, 1998; Li et al., 2002). Two subjects' performances were not significantly different from chance (randomly rotated letter task performance under single-task condition, average over 10 blocks for each of the two subjects: 74.2 $\pm$ 3.3% and 85.9 $\pm$ 4.3%; randomly rotated letter task performance under dual-task condition, average over 10 blocks for each of the two subjects: 53.7 $\pm$ 4.8% and 53.7 $\pm$

5.2%); *t*-test results comparing the dual-task condition with chance: $t(9) < 1.83$, $p > .05$ for each subject. The other two subjects performed slightly better than chance, but significantly lower than their baseline performances obtained when attention was available (randomly rotated letter task performance under single-task condition, average over 10 blocks for each of the two subjects: 70.6 $\pm$ 4.4% and 78.9 $\pm$ 4.1%; randomly rotated letter task performance under dual-task condition, average over 10 blocks for each of the two subjects: 57.8 $\pm$ 7.2% and 61.5 $\pm$ 5.5%); *t*-test results comparing dual-task condition performances with single-task condition performances: $t(18) > 3.61$, $p < .001$ for each subject. For the fixed but uncommon rotation (nonupright), subjects' performances varied more than the random rotation condition. Three subjects' dual-task performances were comparable to chance level or just slightly higher (fixed-rotation letter task performance under dual-task condition, average over 10 blocks for each of the three subjects: 56.0 $\pm$ 4.5%, 56.0 $\pm$ 6.7%, and 59.1 $\pm$ 4.2%); *t*-test results: $t(9) < 1.83$, $p > .05$ for the first two subject; $t(9) > 4.30$, $p < .01$ for the third subject. One subject performed at 81% of his baseline level (single-task condition: 78.1 $\pm$ 8.1%; dual-task condition: 67.2 $\pm$ 4.6%). We should then compare this result to the last condition: Upright letter discrimination. Here all subjects performed above 80% of their baseline performance level (single-task condition performance, average over 10 block for each subject: 85.2 $\pm$ 4.4%, 87.2 $\pm$ 5.7%, 78.1 $\pm$ 8.1%, and 89.1 $\pm$ 4.3%; dual-task condition performance, average over 10 blocks for each subject: 72.9 $\pm$ 7.9%, 77.3 $\pm$ 6.1%, 74.2 $\pm$ 9.1%, and 83.6 $\pm$ 5.3%). One subject's performance was not significantly different from her performance when attention was available (single-task condition: 78.1 $\pm$ 8.1%; dual-task condition: 74.2 $\pm$ 9.1%); *t*-test result: $t(18) < 1.73$, $p > .05$. There is thus a clear pattern that the least amount of attention resource is needed when letters are in their upright positions.

Discussion

Two independent hypotheses were tested in this set of experiments. We found subjects could not recognize rotated letters without attention even when the rotation is fixed to one angle throughout the entire testing period of several hours. This result indicates that stimulus predictability alone cannot reduce much of the attentional load required for such a task.

On the other hand, while keeping everything else exactly the same, recognizing upright letters shows a clear advantage over randomly rotated letters when the attentional resource is scarce. One might argue that a fixed upright rotation is easier for a mental template matching algorithm because the stimuli are much more predictable, whereas such method is less useful for a random rotation. This is why we observe a significantly improved performance for the upright letter condition compared to the random letter condition. But this argument does not support our observation for the fixed but uncommon rotation

condition. If the template-matching theory works, it should work for any fixed rotation, not just upright. In fact a modern computer vision algorithm can easily implement such a template-matching method to carry out this task. Our results suggest that there is a clear bias for people to recognize upright letters better than other rotated versions, even highly predictable ones. This is not surprising if we consider how much more learning one receives on upright letter recognition over a lifetime. Similar evidence has also been found in the visual search tasks. Wang et al. (1994) have reported that familiar letters take less time to search than unfamiliar ones. This observation supports our hypothesis that familiar and meaningful categories of objects can be recognized with much less attentional load.

GENERAL DISCUSSION

Natural scene categorization is one of the most evolutionarily relevant tasks of the human visual systems. The superb efficiency of this task has stimulated much research across the fields of neural psychophysics, physiology, and modelling (Biederman, 1972; Li et al., 2002; Potter & Levy, 1969; Thorpe, Fize, & Marlot, 1996; Torralba & Oliva, 2003; VanRullen & Thorpe, 2001). Contrary to the daily experience of the effortlessness of natural scene recognition, it is one of the hardest tasks that modern state-of-art computer vision algorithms have yet to accomplish. This difficulty is mostly due to the vast variability across similar categories of the natural scenes. Unlike low-level tasks such as orientation discrimination or texture recognition, in which much of the tasks can be accomplished through filtering of the primary visual cortex (Malik & Perona, 1990), understanding the categorical information across different examples of natural scenes is usually thought to be a high-level visual task. In an effort to understand the processing of natural scene categorization, we have chosen an approach to study the efficiency of this process in the near absence of attention.

Visual attention is considered to be one of the first and foremost means of controlling the flow of information between different levels of visual processing. Numerous studies have probed the function of attention, demonstrating much attentional control over stimuli with complex and conjugate features (Treisman & Gelade, 1980; Wolfe, 1998; and others). Needless to say, the function of attention is tightly associated with the computational function of recognition in the human visual system. We hope that by manipulating the attentional condition of various natural scene categorization tasks, as well as comparing it with other recognition tasks, much light can be shed in the understanding of the fundamental mechanisms that enable us such a rapid and fast ability of scene categorization.

In this study, we present a series of experiments exploring the mechanisms of attention and recognition for natural scenes, especially in comparison with synthetic stimuli. Based on a previous study (Li et al., 2002), we specifically

want to probe the puzzling observation that natural scene categorization seems much more efficient than simple, synthetic stimuli recognition in the near absence of attention. We explored this set of questions by increasing the potential ''difficulty'' of the natural scene categorization task through decreasing signal or lack of training. Meanwhile, we attempted to ''ease'' the synthetic stimuli recognition tasks by increasing the amount of signal, the predictability, or the meaningfulness of the stimuli. Interestingly, most of these manipulations were ineffective, i.e., did not affect the attentional requirements of the tasks.

It is important to point out that we are certainly not claiming that (1) natural scene processing would resist ''all'' attempts to make it attentionally demanding (asking subjects to respond only when there are eight trees in a scene would probably require attention), or that (2) no synthetic stimuli can ever be discriminated without attention (oriented bars can, for example). What we are after is some form of general understanding of the factors that determine attentional requirements for natural scene categorization and its relationship with synthetic stimuli recognition. The experiments that we have chosen are certainly not exhaustive but did allow us to reach interesting conclusions in this respect.

Natural scene categorization is an ''easy'' task to learn and to perform

Perceptual learning is closely linked to the mechanisms of recognition and attention. Our results show that contrary to common belief, certain seemingly much simpler stimuli are harder to learn to discriminate than complex natural scenes, when attention is not available. In fact, little stimulus specific training is necessary for subjects to perform the natural scene categorization task. Given the current models of visual recognition, this result is highly counterintuitive. Today's start-of-the-art computer algorithms take much more training to recognize natural scenes than simple geometric configurations (Viola & Jones, 2001; Weber, Welling, & Perona, 2000). Hochstein and Ahissar (2002) have coined the term ''easy'' for task conditions where considerable learning transfer occurs. Under this definition, natural scene categorization is a much ''easier'' task given the results of our first set of experiments. They hypothesized that ''easier'' tasks involve higher cortical level processing than lower ones. We will revisit this point in our discussion of ''meaningful categories'', in which our further findings with differently rotated letters also give hint to this possible architecture of the visual processing.

Another piece of indirect evidence of the ''easiness'' for natural scene categorization is its performance pattern without colour information. In Experiment 2 we found that there is virtually no cost in removing colour information of the natural scenes. Again, state-of-the-art computer algorithms for image retrieval often utilize colour features as one of the most informative

cues for categorization. Our results show a clear discrepancy between such algorithms and the actual properties of the human visual system. It is suggested that rapid natural scene categorization might take advantage of coarser, achromatic information from the magnocellular pathway earlier than the finer, chromatic parvocellular pathway (Delorme et al., 2000). Several studies on face recognition (Epstein & Kanwisher, 1998) also suggest that there is a response of IT neurons for early phasic component of the stimuli rather than more fine-tuned information. This suggest that categorical recognition might be achieved in higher level visual areas using early waves of neuronal information, where more detailed features such as colours and fine edges have not yet been computed or incorporated.

Parallel processing

The robustness of natural scene categorization is further confirmed through the experiment of double-image recognition (Experiment 3). Our human visual system is surprisingly parallel in processing the complex stimuli of natural scenes (Rousselet et al., 2002). In contrast, Experiment 4 shows that such ability does not apply to stimuli that are defined by their low-level differences, such as the configuration of the red semicircle and a green semicircle. Attention has long been considered to be deployed preferentially to tasks that require much scrutiny and processing. This experiment confirms further that the seemingly much simpler stimuli requires more attention to be categorized.

It has been long suggested that the more a recognition task requires feature conjunctions and binding, the more attention will be needed for this task (Treisman, 1993; Treisman & Gelade, 1980). Therefore only ''elementary features'' are processed in a parallel fashion (i.e., under visual search, where serial focal attentional scan is not required). Our results suggest the possibility that natural scene categories might belong to the set of ''elementary features'' while the colour discs or rotated letters do not. But this type of feature is unlikely to be encoded in lower level visual pathways where receptive fields are small and neurons tend to respond to primitive features such as orientations and brightness.

Meaningful categories

In an attempt to understand the efficiency and robustness of natural scene categorization, so far we have gathered much indirect evidence that visual tasks involving higher cortical levels are recognized easier, learned faster, and deploy less attentional resource. In the last set of experiments, we find strong evidence suggesting that meaningfulness and familiarity might participate in determining attentional load and more efficient recognition and learning of the natural scene categorization task. Everything else being equal, an upright letter is discriminated much better than one rotated to a fixed, but uncommon orientation.

There is little low-level difference between these two sets of stimuli, but they do differ in terms of familiarity and meaningfulness. A similar result was recently discovered by Reddy et al. (2004). In their study, they contrast gender discrimination of hairless upright faces versus inverted faces in the near absence of attention. They find that little attention is required to perform the task with upright faces (which are both familiar and meaningful) while the attentional cost is rather high with inverted faces.

This observation is also consistent with the recent development of change blindness studies. Change blindness has shown that attention is critical for our visual awareness (Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1997). Changes of large patches of the visual world can escape our awareness without attending to them. But the amount of attention needed to discern such changes seems to depend also on the meaningfulness of the stimuli. Semantically relevant information is less likely to be neglected in change blindness than less relevant information (Hollingworth & Henderson, 2000).

In short, we hypothesize that natural scene categorization requires little or no attentional cost because of the familiarity and ''meaningfulness'' of the stimuli and task. Attention acts as a gauge for information processing. When the task or stimuli are unfamiliar, hence are not directly associated with previous neuronal representations, attention helps to select and process features for the recognition task. When there are preexisting neuronal representations for a given task or stimulus, for example natural scene categorization, little attention is needed.

CONCLUSION

We have presented five different experiments in this report. Our results show that natural scene categorization without attention requires little stimulus-specific training. It is robust to lack of colour information or increasing the set size of the stimuli presented. In contrary, multiple redundant copies of synthetic stimuli do not improve the performances of recognition without attention. Some simple tasks, such as single letter discrimination, require much attentional assistance unless the letters are presented in a familiar, upright position. We hypothesize that attention is particularly important for tasks that do not have neuronal representations in the visual pathway. Natural scene categorization, a well-learned and familiarized task for most human observers, does not require much attention.

REFERENCES

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*, 77–80.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Braun, J. (1994). Visual search among items of different salience: Removal of visual attention mimics a lesion in extrastriate area V4. *Journal of Neuroscience*, *14*, 554–567.

Braun, J. (1998). Vision and attention: The role of training. *Nature*, *393*, 424–245.

Braun, J., & Julesz, B. (1998). Withdrawing attention at little or no cost: Detection and discrimination tasks. *Perception and Psychophysics*, *60*(1), 1–23.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Rapid categorisation of natural scenes is color blind: A study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200.

Deutsch, J., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*, 80–90.

Dunai, J., Castiello, U. & Rossetti, Y. (2001). Attentional processing of colour and location cues. *Experiment Brain Research*, *138*(4), 520–526.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Haxby, J., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse heirarchies in the visual system. *Neuron*, *36*, 791–804.

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, *7*(1–3), 213–235.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a ''preattentive'' feature search task. *Nature*, *387*, 805–807.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1998). Reply to Braun. *Nature*, *387*, 805–807.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliff, NJ: Prentice-Hall.

Lee, D. K., Koch, C., & Braun, J. (1999). Attentional capacity is undifferentiated: Concurrent discrimination of form, color, and motion. *Perception and Psychophysics*, *61*(7), 1241–1255.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601.

Logothetis, N. K., & Sheinberg, D. L. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences*, *94*, 3408–3413.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America, A*. *7*(5), 923–932.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Croft.

Oliva, A., & Schyns, P. G. (2000). Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*, *41*, 176–210.

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, *7*(2), 333–339.

Ostergaard, A. L., & Davidoff, J.B. (1985). Some effects of color on naming and recognition of objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 579–587.

Pashler, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection and signals. *Journal of Experimental Psychology: General*, *109*, 160–174.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15.

Price, C. J., & Humphreys, G. W. (1989). The effects of surface detail on object categorization and naming. *Quarterly Journal of Experimental Psychology*, *41A*, 797–828.

Reddy, L., Wilken, P., & Koch, C. (2004). Face–gender discrimination is possible in the near-absence of attention. *Journal of Vision*, *4*(2), 106–117.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). *Psychological Science*, *8*(5), 368–373.

Rossion, B., & Gauthier, I . (2002). How does the brain process upright and inverted faces. *Behavioral and Cognitive Neuroscience Reviews*, *1*(1), 63–75.

Rousselet, G., Fabre-Thorpe, M., & Thorpe, S. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, *5*, 629–630.

Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). Adaptive color-image embeddings for database navigation. In *Proceedings of the Third Asian Conference on Computer Vision* (Vol. 1, pp. 104–111). London, UK: Springer-Verlag.

Shapiro, K. (Ed.). (2001). *The limits of attention*. Oxford, UK: Oxford University Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127–188.

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

Sperling, G., & Dosher, B. (1986). Strategy and optimization in human information processing. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (pp. 1–65). New York: Wiley.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139.

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

Torralba, A., & Oliva, A. (2003). Statistics of natural images categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Treisman, A. (1993). The perception of features and objects. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control. A tribute to Donald Broadbent* (pp. 5–35). Oxford, UK: Clarendon Press.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Troje, N. F., & Bulthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, *36*, 1761–1771.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition*, *1*, 511–518.

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception and Psychophysics*, *56*, 495–500.

Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. In *Proceedings of the sixth European conference of Computer Vision* (Vol. 1842, pp. 18–32). London, UK: Springer-Verlag.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–74). Hove, UK: Psychology Press.