Categorization of Natural Scenes: Local versus Global Information and the Role of Color

JULIA VOGEL
University of British Columbia
ADRIAN SCHWANINGER
Max Planck Institute for Biological Cybernetics and University of Zurich
and
CHRISTIAN WALLRAVEN and HEINRICH H. BÜLTHOFF
Max Planck Institute for Biological Cybernetics

Categorization of scenes is a fundamental process of human vision that allows us to efficiently and rapidly analyze our surroundings. Several studies have explored the processes underlying human scene categorization, but they have focused on processing global image information. In this study, we present both psychophysical and computational experiments that investigate the role of local versus global image information in scene categorization. In a first set of human experiments, categorization performance is tested when only local or only global image information is present. Our results suggest that humans rely on local, region-based information as much as on global, configural information. In addition, humans seem to integrate both types of information for intact scene categorization. In a set of computational experiments, human performance is compared to two state-of-the-art computer vision approaches that have been shown to be psychophysically plausible and that model either local or global information. In addition to the influence of local versus global information, in a second series of experiments, we investigated the effect of color on the categorization performance of both the human observers and the computational model. Analysis of the human data suggests that color is an additional channel of perceptual information that leads to higher categorization results at the expense of increased reaction times in the intact condition. However, it does not affect reaction times when only local information is present. When color is removed, the employed computational model follows the relative performance decrease of human observers for each scene category and can thus be seen as a perceptually plausible model for human scene categorization based on local image information.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Psychology; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Representations, data structures, and transforms*; I.4.8 [**Image Processing and Computer Vision**]: Scene analysis; I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*

General Terms: Human perception, Algorithms

Authors' addresses: Julia Vogel, University of British Columbia, Vancouver, British Columbia; Adrian Schwaninger, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, and Department of Psychology, University of Zurich, Zurich, Switzerland; Christian Wallraven and Heinrich H. Bülthoff, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Additional Key Words and Phrases: Scene perception, scene classification, computational modeling, semantic modeling, computational gist, local region-based information, global configural information

ACM Reference Format:

Vogel, J., Schwaninger, A., Wallraven, C., and Bülthoff, H. H. 2007. Categorization of natural scenes: Local versus global information and the role of color. ACM Trans. Appl. Percpt. 4, 3, Article 19 (November 2007), 21 pages. DOI = 10.1145/1278387.1278393 http://doi.acm.org/10.1145/1278387.1278393

1. INTRODUCTION

Categorization of scenes is a fundamental process of human vision that allows us to efficiently and rapidly analyze our surroundings. Since the early work by Biederman [1972] on the role of scene context in object recognition, much research has been devoted to describe and understand scene categorization processes (e.g., rapid scene categorization [Thorpe et al. 1996], categorization with little attention [Fei-Fei et al. 2005], categorization of blurred stimuli [Schyns and Oliva 1994]; see also the recent special issue on real-world scene perception [Henderson 2005b]).

Complementing this interest in human perception of scenes, computer vision research has recently focused on creating systems that enable automatic categorization of scenes. Although substantial progress has been made [Oliva and Torralba 2001; Vailaya et al. 2001; Szummer and Picard 1998; Fei-Fei and Perona 2005; Bosch et al. 2006; Vogel and Schiele 2007], the complexity of scenes continues to provide a challenge to computer vision research. We believe that it is valuable to pursue an interdisciplinary cognitive and computational approach to better understand human scene perception and to implement a perceptually plausible computational model (see also [Oliva and Torralba 2001; Walker Renninger and Malik 2004; McCotter et al. 2005]). Psychophysical experiments allow us, on one hand, to gain a deeper understanding of the processes and representations used by humans when they categorize scenes. This knowledge can help computer vision researchers to design more efficient computational systems. Computer vision, on the other hand, allows us to create algorithms with precisely defined features and classification schemes for processing and categorization of scenes. By comparing machine and human performance on the same image data, we can then try to validate and evaluate the degree with which these features and classifiers provide an accurate model of human scene perception. These results can again lead to experimentally testable predictions, closing the loop between human experiments and computer vision research.

In this paper, we follow Henderson [2005a] and define a scene as a semantically coherent, namable human-scaled view of a real-word environment. This view often is comprised of background and foreground elements that are arranged in a hierarchical spatial layout. This already implies a scale, a connection of local and global information that is at the core of scene processing. The role of local versus global information has received much attention in other areas such as face and object recognition (for recent reviews see Schwaninger et al. [2003]; Hayward [2003]). Using scrambling and blurring procedures, Schwaninger et al. [2002] showed that local part-based information and global configural information are processed separately and integrated in human face recognition (see also [Hayward et al. 2007]). A psychophysically plausible computational model of these processes and representations has been provided recently by Wallraven et al. [2005]. In object recognition, the role of rotation-invariant local parts (geons) versus more global view-based representations has been discussed extensively in the last 20 years (e.g. Hayward [2003]).

The goal of this paper is to examine the processing of local and global information in human scene categorization using psychophysics and to compare two computational models of scene categorization with human performance. In addition, we investigate the role of color and reaction times for scene categorization.

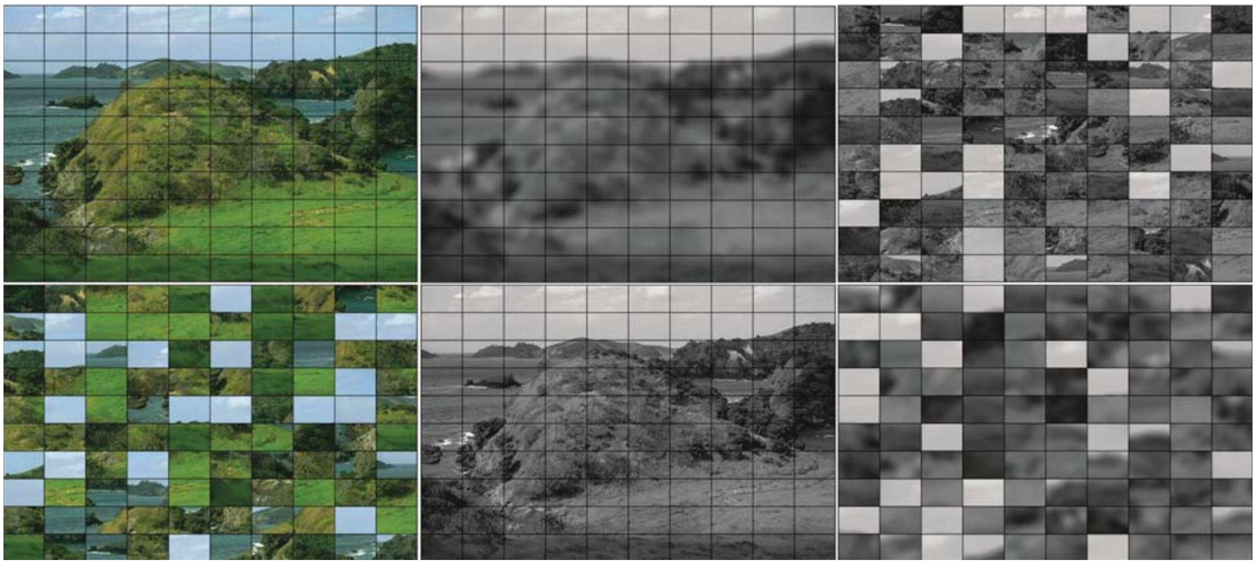


Fig. 1. Exemplary image in its six display conditions. Upper row: intact, blurred, scrambled gray-scaled. Lower row: scrambled, gray-scaled, blurred-scrambled.

In the following three sections, we present experiments and discussion of the processing of global and local information in human scene categorization (Experiments 1–3). Experiments 4 and 5 test the role of color in human categorization of intact scenes and of scenes with only local information. In Section 7, we compare and analyze the reaction times of the various categorization tasks. Section 8 presents two computational models that have been shown previously to be perceptually plausible for representing local and global image information and compares the computational performance to human scene categorization. These two computational models are combined in Section 9. The findings of this paper are summarized and discussed in Section 10.

2. EXPERIMENT 1: OBTAINING GROUND TRUTH

Natural scenes constitute a very heterogeneous and complex stimulus class. In contrast to basic level *object* categorization [Rosch et al. 1976], natural scenes often contain semantic details that can be attributed to more than one category. The goal of Experiment 1 was to determine ground truth and the benchmark for our employed scene database.

The selection of the natural scene categories follows the rationale of Vogel and Schiele [2007] and was strongly influenced by the work of Tversky and Hemenway [1983]. In their seminal work, the authors found indoors and outdoors to be superordinate-level categories, with the outdoors category being composed of the basic-level categories: city, park, beach, and mountains, and the indoors category being composed of restaurant, store, street, and home. In addition, Rogowitz et al. [1997] detected two main axes along which humans sort photographic images: human versus nonhuman and natural versus artificial. These semantic axes were further extended into 20 scene categories by Mojsilovic et al. [2004]. Human natural scene categorization should not be biased by the recognition of particular objects. Therefore, the images for our experiments were selected so that they did not contain specific objects or man-made material. Thus, the human/natural coordinate of Rogowitz et al. [1997] was selected as superordinate for the experiments. In addition, the natural, basic-level categories of Tversky and Hemenway [1983] and the natural scene categories of Mojsilovic et al. [2004] were combined and

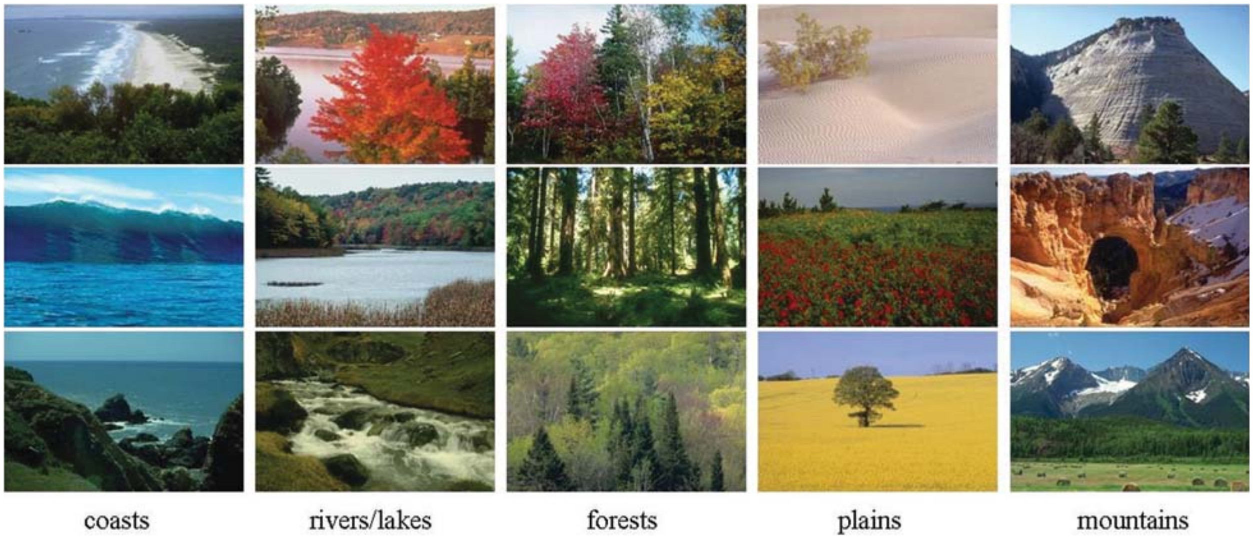


Fig. 2. Exemplary images of each category.

extended to the categories coasts, rivers/lakes, forests, plains, and mountains. See Figure 2 for exemplary images of each category.

2.1 Method

Participants. Eleven naive subjects were paid to participate in the study. All had normal or corrected to normal vision.

Stimuli. Of the Corel image natural image database (720 × 480 pixels), 250 in landscape format served as stimuli. The natural scenes were initially selected by one of the authors (JV) in the way that each of the five categories contained 50 images. Special care was taken to also include stimuli close to the category boundaries. Exemplary images of each category are displayed in Figure 2.

Experimental Conditions. The images were presented at 100 Hz on a Sony Trinitron 21-inch monitor with a resolution of 1280 × 960 pixels. The experiments were conducted in a dimly lit room. The viewing distance was maintained by a chin rest so that the center of the screen was at eye height. The length and width of displayed images covered a viewing angle of 24.6° and 16.5°, respectively. The 250 images were presented in random order. Display time was 4 s after which subjects were forced to make a choice. Below the images, five checkboxes labeled coasts, rivers/lakes, forests, plains, and mountains were displayed.[1] All images were superimposed with a regular 10 × 10 grid in order to contain the same high-order frequency components in all experiments. Participants were asked to categorize each displayed image into one of the five categories as fast and accurately as possible by clicking on the corresponding screen button using the mouse.

2.2 Results and Discussion

Ground truth for our database of 250 images was determined by assigning each image to the category that was selected by the majority of subjects. As a result, the database contains 57 coast, 44 rivers/lakes,

[1] Since the study was conducted at the Max Planck Institute of Biological Cybernetics in Tübingen, Germany, the following German category labels were used: Küste, Fluß/See, Wald, Ebene, and Berg/Gebirge.

Table I. Confusion Matrix for Categorization of Intact Images in Experiment 1

89.7%	Coasts	Rivers/lakes	Forests	Plains	Mountains
Coasts	**90.4%**	8.3%	0.3%	0.3%	0.6%
Rivers/lakes	6.0%	**82.9%**	2.1%	0.4%	8.7%
Forests	0.4%	1.6%	**91.5%**	4.7%	1.8%
Plains	0.4%	0%	0.8%	**92.7%**	6.1%
Mountains	0.2%	2.9%	1.4%	5.0%	**90.6%**

50 forest, 46 plains, and 53 mountain images. Based on this ground truth, the average categorization performance in Experiment 1 is 89.7%. Table I displays the average confusion matrix of the experiment. Disagreements mainly occur between rivers/lakes and coasts and between plains and mountains in both directions, as well as between rivers/lakes and mountains and between plains and mountains in only one direction. The rivers/lakes category also seems to be more ambiguous than the other categories. Experiment 1 confirms that the database consists of complex stimuli and ensures that no ceiling effects are present. The ground truth gained from Experiment 1 will be used as benchmark in the following experiments.

3. EXPERIMENT 2: CATEGORIZATION OF SCRAMBLED IMAGES

Experiment 2 investigated if human observers are able to categorize natural scenes when only local information is present and global configural information has been destroyed. In face and object recognition, local information has sometimes been defined in terms of local parts (e.g. Schwaninger et al. [2003] and Hayward et al. 2007). However, in this study we are interested in investigating the categorization of natural scenes that is not biased by objects in the scene or by diagnostic parts. This is inspired by a recent study of Schwaninger et al. [2006] in which a computational model using a semantic modeling step was compared to human perception of scene typicality. Based on the work by Vogel and Schiele [2007], the computational model implements an intermediate semantic modeling step by extracting local semantic concepts, such as rock, water, and sand. Schwaninger et al. [2006] found a high correlation between the computational and the human ranking of natural scenes regarding typicality. Interestingly, comparisons with a computational model without a semantic modeling step correlated much less with human performance, suggesting that a computational model based on local semantic concepts, such as, rock, water, and sand, is psychophysically very plausible. In this study, we further investigate the role of such local semantic information. Thus, instead of an object or part-based definition, we define local information as any information present in a small image region. In our case, these local regions cover 1% of the full image area (regular grid of 10 × 10 = 100 regions) and thus contain sufficient featural information for detecting higher-level information (e.g., the semantic concept class). In the experiment, global configural information was eliminated by cutting the scenes into local image regions and randomly relocating, i.e., scrambling, those local regions. If local image information is used for categorization, categorization performance should be above chance even if the scenes are scrambled.

3.1 Method

Participants. Eleven subjects were paid to participate in the study. None of them had participated in Experiment 1. All had normal or corrected to normal vision.

Stimuli. In Experiment 2, the 250 nature scenes used in Experiment 1 were scrambled. The scrambling was created by cutting the scenes into a regular grid of 10 × 10 = 100 regions of 72 × 48 pixels, and randomly repositioning the resulting regions. The scrambled image has the same size (720 × 480 pixels) as the original. The random scrambling of images was new for each participant in order to prevent any particular spatial configuration from influencing the results.

Experimental Conditions. See Experiment 1.

Table II. Confusion Matrix for Categorization of Scrambled Images in Experiment 2[a]

72.7%	Coasts	Rivers/lakes	Forests	Plains	Mountains
Coasts	**71.8%**	14.2%	2.6%	3.5%	8.0%
Rivers/lakes	18.8%	**36.8%**	16.3%	5.0%	23.1%
Forests	0.9%	1.5%	**91.3%**	5.3%	1.1%
Plains	0.8%	0.8%	2.8%	**87.0%**	8.7%
Mountains	4.6%	2.7%	6.9%	12.3%	**73.4%**

[a] Main diagonal: hit rate; off-diagonal: false alarm rate.

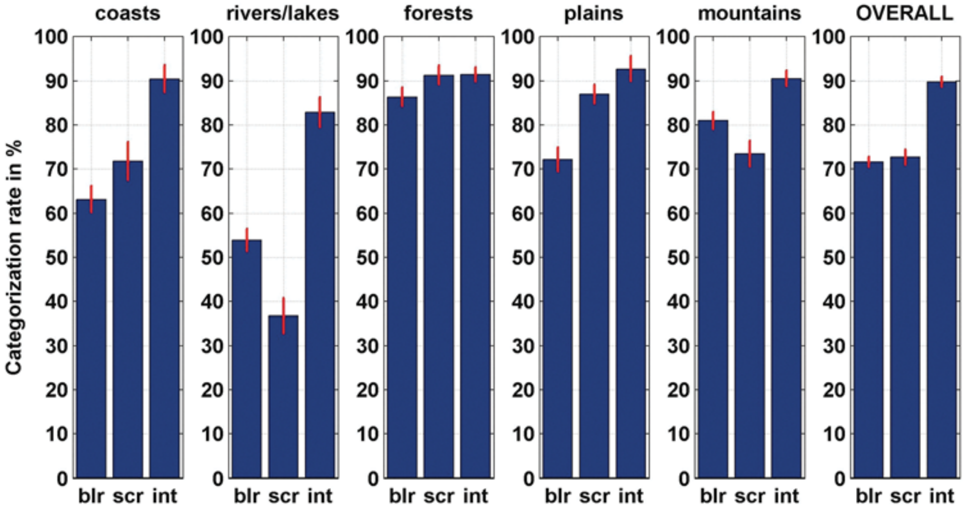


Fig. 3. Comparison of categorization rates between blurred (blr), scrambled (scr), and intact (int) display condition (in %; error bars represent the standard error of the mean).

3.2 Results and Discussion

Categorization performance was calculated relative to the ground truth determined in the previous experiment. Averaged over all subjects and all scene categories, the categorization rate was 72.7%. Table II shows the confusion matrix of the categorization (see also Figure 3). The categorization performance is surprisingly good, given that the important configural information has been eliminated.

One-sample t tests were carried out in order to test the per-category performance against chance performance (20%). All categories were recognized above chance with $p < 0.01$ for rivers/lakes and $p < 0.001$ for all other categories. This result shows that scene categorization relies on local information. In addition, a one-way analysis of variance (ANOVA) with the category as within-subjects factor was carried out. The analysis revealed a main effect of category ($F(2.551, 25.506) = 42.33$, $MSE = 187.225$, $p < 0.001$). We also measured the interaction between the display conditions using a two-factorial split-plot ANOVA with category as within-subjects factor and condition (intact versus scrambled) as between-subjects factor. There were main effects of condition ($F(1,20) = 78.301$, $MSE = 108.301$, $p < 0.001$) and category ($F(3.088,61.767) = 34.710$, $MSE = 130.302$, $p < 0.001$). There was also a significant interaction between condition and category ($F(3.088,61.767) = 17.169$, $p < 0.001$), implying that local region-based information is of different importance for different scene categories.

In summary, these results show that local, region-based information is an important factor in human scene categorization. This varies depending on the scene category. For instance, as can be seen in Figure 3, the categorization of forests and plains is hardly affected by the scrambled condition, while a large decrement is found for rivers/lakes. This suggests that forests and plains can be identified based on local region-based information, while identifying the rivers/lakes category requires also processing of more global information.

4. EXPERIMENT 3: CATEGORIZATION OF BLURRED IMAGES

Experiment 3 tested the influence of global, configural information on human scene categorization. We define global information as the overall "context" of a scene generated through the presence of large spatial structures (e.g., horizon lines) and the spatial arrangement of lighter and darker blobs in an image. Participants had to categorize the scenes of Experiment 1 when shown in a low-pass filtered and gray-scaled version. These image manipulations destroyed the main information carrier of the previous experiment, that is, local, region-based image information, while leaving global configural information intact. Low-pass filtering reduces the high-spatial frequency content, which is diagnostic for local texture features. Regarding color, one could imagine to scramble the image using smaller and smaller windows so that at the limit the image becomes scrambled at the level of single pixels. Although such an experimental condition was not included in this study, one could imagine that color could help for categorizing such extremely scrambled images. This would definitively be an effect of very local information. In Experiment 3, the aim was to eliminate local information. Therefore, we not only low-pass filtered the images, but also gray-scaled them to create stimuli that contain only global configural information.

4.1 Method

Participants. Eleven subjects were paid to participate in the study. None of them had participated in one of the previous experiments. All had normal or corrected to normal vision.

Stimuli and Procedure. In Experiment 3, the 250 nature scenes used in Experiment 1 were blurred using a 48-tap digital low-pass FIR filter with a cut-off frequency $f_{cutoff} = 0.07\, f_{nyquist} \equiv 16.8\, cycles/image$. The low-pass filter was applied to gray-scaled versions of the original images. This image manipulation destroys local information while leaving global information intact. (See Figure 1 for an exemplary blurred image.) However, note that the visual angle of the subjects was significantly larger than that of the reader. In addition, the displayed image was superimposed by a 10 × 10 grid in order to account for the same high-order frequency distortions as in the previous experiments. The blur level was determined in several pilot experiments using a similar procedure as Schwaninger et al. [2002]. In general, local featural information can be eliminated by blurring the stimuli. The blur level is determined by scrambling and blurring stimuli until categorization performance in the blurred–scrambled condition drops down to chance. This blur level by definition eliminates all local featural information. In the final pilot experiment for this study with 11 naive subjects, average categorization performance in the blurred–scrambled condition was 23%, which was in all but one category not significantly different to the chance level of 20%. Figure 1 also shows the exemplary image in the blurred-scrambled condition.

Experimental Conditions. See Experiment 1.

4.2 Results and Discussion

The average overall categorization performance in the blurred condition was 71.6%. As in the scrambled condition, categorization performance is relatively stable compared to the intact condition. Table III reveals that compared to Experiment 2, there are fewer confusions between rivers/lakes and coasts,

Table III. Confusion Matrix for Categorization of Blurred Images in Experiment 3[a]

71.6%	Coasts	Rivers lakes	Forests	Plains	Mountains
Coasts	**63.3%**	14.0%	3.8%	5.6%	13.4%
Rivers/lakes	8.7%	**53.9%**	8.7%	5.8%	22.9%
Forests	0.9%	4.9%	**86.4%**	2.4%	5.4%
Plains	4.0%	7.5%	3.8%	**72.1%**	12.6%
Mountains	2.6%	5.2%	5.1%	6.2%	**81.0%**

[a] Main diagonal, hit rate; off-diagonal: false alarm rate.

rivers/lakes and forests, and mountains and plains, but that there are more confusions between coasts and mountains, plains and mountains, and plains and rivers/lakes.

Also in the blurred condition, one-sample t tests revealed a significant difference to chance performance (20%) for all categories ($p < 0.001$). These results suggest that scene categorization also relies on global image information as proposed earlier [Schyns and Oliva 1994]. A one-way ANOVA indicated that there is a main effect of category also in the blurred condition ($F(1.989, 19.894) = 25.188$, $MSE = 151.273, p < 0.001$). In addition, data from Experiment 1 and 3 were subjected to a two-factorial split-plot ANOVA with category as within-subjects factor and condition as between-subjects factor. The analysis revealed main effects of condition (intact versus blurred) ($F(1, 20) = 129.666$, $MSE = 70.944$, $p < 0.001$), and of category ($F(2.839, 56.784) = 18.385$, $MSE = 110.640$, $p < 0.001$). There was also an interaction: $F(2.839, 56.784) = 7.853$, $p < 0.001$).), suggesting a different role of global configural information for identifying different scene categories. In order to compare the scrambled and blurred conditions with each other, a two-factorial split-plot ANOVA was carried out with the data from Experiments 2 and 3 with category as within-subjects factor and condition as between-subjects factor. There was no overall main effect of condition (scrambled versus blurred) ($F(1,20) = 5236{,}028$, $MSE = 107.921$, $p > 0.05$)), indicating that the two conditions are comparable in difficulty.[2] However, there was a main effect of category ($F(2.767, 55.336) = 61.997$, $MSE = 140.683$, $p < 0.001$), as well as an interaction between condition and category ($F(2.767, 55.336) = 9.415$, $p < 0.001$), suggesting that different types of information are used for different scene categories.

In summary, these results show that scene categorization relies not only on local, region-based information, but also on global, configural information. In the blurred condition, the categorization performance also depends on the particular scene category. Most interestingly, as Figure 3 shows, categorization performance in the blurred condition is better for those categories that did not score high in the scrambled condition: i.e., rivers/lakes and mountains. These results suggest that local and global information is integrated differently, depending on the category. Categories with many different local semantic concepts present in an image (such as mountains or rivers/lakes) require global context information for categorization. In contrast, categories, such as forests, plains, or coasts with local semantic concepts that are discriminant without global configural information, are categorized better using local information. It has to be noted that the performance for intact scenes was higher than the performance in the scrambled and blurred conditions. This is consistent with the view that processing of local and global information are integrated, resulting in higher categorization performance.

[2] This, of course, is related to the parameters of the manipulation. We aimed at producing comparable levels of difficulty, which was apparently achieved. Using a different level of blurring or scrambling could have resulted in a slightly different result, i.e., a main effect of condition. Note that this is not relevant for the main conclusions of this study.

Table IV. Confusion Matrix for Categorization of Intact Gray-Scaled Images in Experiment 4[a]

83.7%	Coasts	Rivers/lakes	Forests	Plains	Mountains
Coasts	**83.8%**	10.4%	0.9%	0.4%	4.5%
Rivers/lakes	6.4%	**73.5%**	4.2%	1.7%	14.2%
Forests	0.3%	4.0%	**88.7%**	3.5%	3.5%
Plains	1.3%	0.5%	2.2%	**80.3%**	15.8%
Mountains	0.9%	2.2%	2.5%	3.9%	**90.4%**

[a] Main diagonal: hit rate; off-diagonal: false alarm rate.

5. EXPERIMENT 4: CATEGORIZATION OF GRAY-SCALED IMAGES

The following two experiments investigate the role of color in the categorization of intact and in the categorization of scrambled scenes. The goal was to quantify if and by how much categorization performance degrades when color is removed from the stimuli.

In Experiment 4, gray-scaled versions of the intact images from Experiment 1 were used to examine the role of color when local featural and global configural information is available in the image.

5.1 Method

Participants. Twelve subjects were paid to participate in the study. None of them had participated in one of the previous experiments. All had normal or corrected to normal vision.

Stimuli. For this experiment, the 250 intact scenes used in Experiment 1 were gray-scaled.

Experimental Conditions. See Experiment 1.

5.2 Results and Discussion

When removing color from the intact images, the average categorization performance drops from 89.7% in Experiment 1 to 83.7% in this experiment. Table IV displays the corresponding confusion matrix. It shows that compared to Table I additional confusions occur from all categories to the mountains category. Closer analysis of the image data reveals that especially images with rocks- and stonelike texture that can easily be confused with rocks in gray-scaled images are wrongly categorized as mountains. A two-factorial split-plot ANOVA on the data from Experiments 1 and 4 with category as within-subjects factor and condition as between-subjects factor was carried out. The analysis revealed main effects of condition (intact versus intactGS) ($F(1, 21) = 13.834, MSE = 81.837, p < 0.01$), and of category ($F(2.981, 62.595) = 5.704$, $MSE = 132.817, p < 0.01$). There was no interaction, suggesting that the influence of color is about the same for all scene categories.

In summary, the removal of color from intact images leads to a drop in categorization performance of about 6%. This drop mainly results from assigning images with rockslike texture to the mountains category. The impact of removing color is very similar in all scene categories, but mountains.

6. EXPERIMENT 5: CATEGORIZATION OF SCRAMBLED GRAY-SCALED IMAGES

The goal of this experiment was to investigate the effect of color in scene categorization when on only local featural information is available in the image. Here, the images of Experiment 1 were converted to gray scale prior to random relocation of the local image regions. The experiment tests whether color is the only information needed for categorization based on local image information. If that was the case, categorization performance would drop to chance performance. If color is a cue, in addition to other cues such as texture, categorization should be above chance, even if the scenes are gray scaled and scrambled.

Table V. Confusion Matrix for Categorization of Scrambled Gray-Scaled Images in Experiment 5[a]

63.9%	Coasts	Rivers lakes	Forests	Plains	Mountains
Coasts	**57.3%**	16.8%	7.3%	6.1%	12.4%
Rivers/lakes	19.3%	**27.3%**	24.6%	6.6%	22.2%
Forests	1.5%	1.0%	**93.3%**	1.8%	2.3%
Plains	8.3%	2.0%	11.4%	**67.8%**	10.5%
Mountains	6.9%	2.7%	13.2%	6.8%	**70.4%**

[a] Main diagonal: hit rate; off-diagonal: false alarm rate.

6.1 Method

Participants. Twelve subjects were paid to participate in the study. None of them had participated in one of the previous experiments. All had normal or corrected to normal vision.

Stimuli. In order to obtain gray-scaled scrambled images, the scenes of Experiment 1 were first converted to gray-scale and subsequently scrambled by randomly repositioning image regions as in Experiment 2 (see Figure 1 for an exemplary image). As in Experiment 2, the random scrambling was repeated for each participant in order to control for particular spatial arrangements of the image regions.

Experimental Conditions. See Experiment 1.

6.2 Results and Discussion

The average categorization performance for scrambled gray-scaled images is 63.9%. That is a drop of 8.8% compared to the performance in categorizing scrambled color images in Experiment 2. The amount of the performance drop is comparable to the drop from intact to gray-scaled intact images in Experiment 4. Table V displays the confusion matrix for the scrambled gray-scaled condition. Detailed analysis of the confusion matrix shows that more confusions into the forests and into the mountains categories occur than in Experiment 2. Newly miscategorized coasts images often contain choppy water and when shown in gray-scaled patches might no longer be recognized as water, thus leading to the image being confused with either forests or mountains. Similarly, newly miscategorized plains images show strongly textured terrain or soil that may be confused with water or with vegetation in a gray-scaled display and thus lead to the image being confused with coasts or forests.

In the scrambled gray-scaled condition, one-sample t test revealed significant difference to chance performance (20%) in all categories ($p < 0.001$), but rivers/lakes ($p = 0.101$). This result shows that all scenes but rivers/lakes can be reliably categorized based on only local, gray-scaled information. A one-way ANOVA with the categories as within-subject factor revealed a main effect of category also in the scrambled gray-scaled condition ($F(1.903, 20.929) = 46.812$, $MSE = 310.775$, $p < 0.001$). Using a two-factorial split-plot ANOVA, we measured the interaction between the display conditions intactGS versus scrambledGS and scrambled versus scrambledGS. In both cases, there were main effects of condition (intactGS versus scrambledGS: $F(1, 22) = 68.346$, $MSE = 177.248$, $p < 0.001$, scrambled versus scrambledGS: $F(1, 21) = 10.774$, $MSE = 206.995$, $p < 0.01$) and of category (intactGS versus scrambledGS: $F(2.48, 54.56) = 41.05$, $MSE = 211.599$, $p < 0.001$, scrambled versus scrambledGS: $F(2.366, 49.695) = 85.44$, $MSE = 227.681$, $p < 0.001$). There was a significant interaction between intactGS and scrambledGS ($F(2.48, 54.56) = 15.933, p < 0.001$), but only a weakly significant interaction between scrambled and scrambledGS ($F(2.366, 85.44) = 3.127$, $p < 0.05$) suggesting that scrambling of gray-scaled images affects the five categories differently, but that the removal of color of already scrambled images affects all categories similarly as is the case for removal of color of intact images (see Experiment 4).

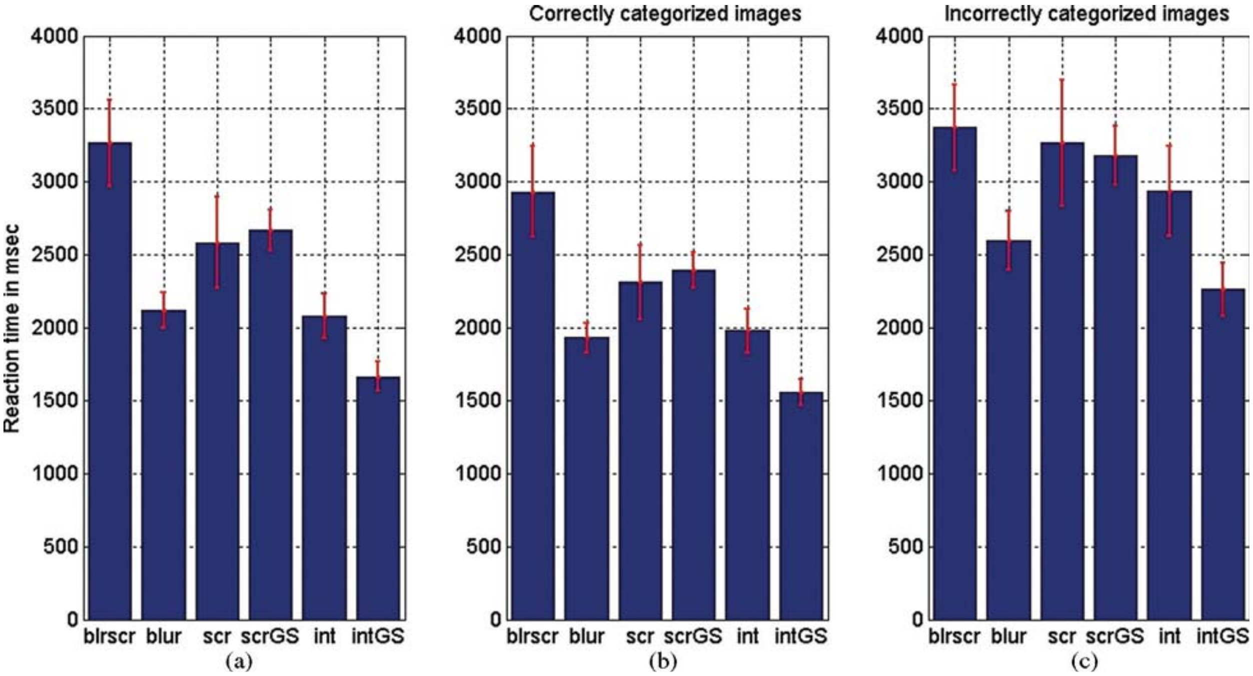


Fig. 4. Reaction times per display condition (blurred–scrambled, blur, scrambled, scrambled gray-scaled, intact, intact gray-scaled): (a) RT over all images, (b) RTs averaged over correctly categorized images, and (c) RTs averaged over incorrectly categorized images. Error bars represent the standard error of the mean.

These results show that humans can clearly categorize scenes when only local texture information is present. The categorization performance depends on the scene category. Forests, mountains, and plains are, on average, better categorized than coasts and rivers/lakes. In comparison with the data from Experiment 2, the removal of color seems to affect all categories similarly, as is the case when removing color from intact images (see comparison between Experiment 1 and Experiment 4).

7. ANALYSIS OF THE REACTION TIMES

The analysis of the categorization accuracies in the previous sections showed that display condition, i.e., removal of local and removal of global information, as well as color influence human categorization performance. The goal of this section is to analyze the corresponding reaction times. The reaction times for the various display conditions are averaged over all scene categories and presented in Figure 4a. In addition, the figure shows reaction times per category split into correctly (Figure 4b) and incorrectly (Figure 4c) categorized scenes. The comparison of Figures 4b and c reveals that the reaction times for incorrectly categorized scenes are higher than for correctly classified scenes, but that the general pattern (e.g., intactGS is categorized faster than intact) is repeated in both graphs. Maximum display time was 4 s. Repeated-measures ANOVAs were performed with the scene category as within-subjects factor and the respective condition as between-subjects factor.

7.1 Global versus Local Image Information

The overall reaction time for scrambled in Figure 4 seems longer than for the blurred condition, however, statistical analysis revealed no main effect of condition when comparing the blurred versus scrambled conditions ($F(1, 20) = 1.968$, $MSE = 3084754.0$, $p = 0.176$). However, we found a main effect of scene category ($F(2.796, 55.923) = 19.89$, $MSE = 99666.631$, $p < 0.001$) and an interaction between display condition and scene category ($F(2.796, 55.923) = 12.88$, $p < 0.001$) suggesting that only local and only global image information influence the reaction times of the five categories differently.

7.2 Influence of Color on Categorization of Intact Scenes

Scene categorization is faster for gray-scaled intact images than for the original color images. Statistical analysis on the reaction times for intact versus intactGS revealed a main effect of condition ($F(1, 21) = 5.191$, $MSE = 952721.457$, $p < 0.05$) and a main effect of category ($F(2.129,44.703) = 3.465$, $MSE = 298715.319$, $p < 0.05$), but no interaction. The removal of color seems to influence all scene categories similarly and leads to faster categorization.

7.3 Influence of Color on Categorization of Scrambled Scenes

The speed of scene categorization is not influenced by the removal of color from scrambled images. Statistical analysis of scrambled versus scrambledGS yielded only a main effect of category ($F(3.176,65.688) = 44.884$, $MSE = 344258.147$, $p < 0.001$) and a weakly significant interaction ($F(3.176,66.688) = 2.740$, $p < 0.05$). As for the categorization performance, the reaction time for the categorization of images with only local information depends on the particular scene category. However, the removal of color from the local image region does not influence the overall reaction time (see Figure 4).

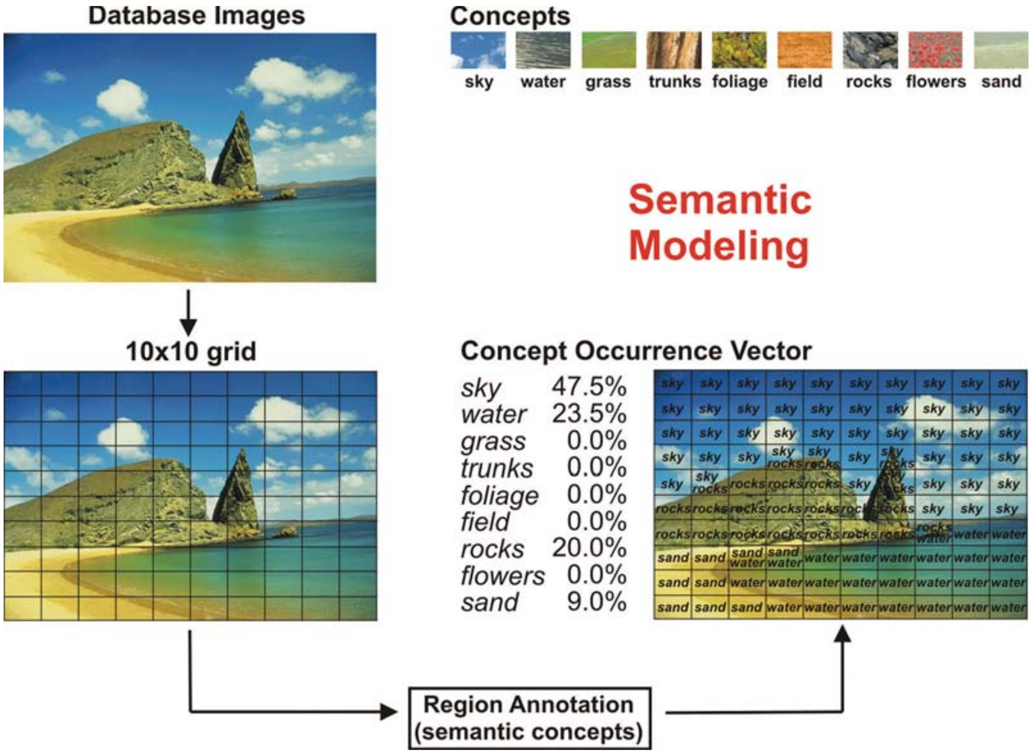
In summary, the removal of color seems to have different influence on intact scenes and on scenes with only local information. Gray-scaled intact scenes are categorized faster than their colored counterparts, but at the expense of categorization accuracy, as shown in Section 5. The removal of color from images with only local information does also result in lower categorization accuracy. However, it does not influence categorization speed.

8. COMPUTATIONAL SCENE CATEGORIZATION

In the previous sections, we analyzed human performance in categorizing natural scenes when only local or only global information is present, as well as the influence of color on the categorization process. The experiments showed that humans use both local and global information and that this information seems to be integrated for the final categorization decision. The removal of color leads to a decrease in categorization accuracy. The goal of the following experiments is to evaluate computational categorization performance for the same tasks. In particular, we compare a local, region-based approach proposed by Vogel and Schiele [2007] and Schwaninger et al. [2006] and an approach that models global context information proposed by Oliva and Torralba [2001] with the human performance. The approaches have been selected because they have been shown to be psychophysically plausible models of human scene perception: one modeling local image information Schwaninger et al. [2006] and one modeling global image information [Oliva 2005].

8.1 Modeling Local, Region-Based Information: Semantic Modeling

For modeling local, region-based image information, we employ the semantic modeling approach of Vogel and Schiele [2007] that makes use of an intermediate modeling step for categorization and ranking of natural scenes. Images are divided into a regular grid of 10 × 10 local regions, which are classified into one of nine local-concept classes. In a subsequent step, this local information is summarized and used for image categorization. The concepts that were determined as being discriminant for the employed scene categories are sky, water, grass, trunks, foliage, field, rocks, flowers, and sand. All database images have been annotated manually with these nine concepts in order to obtain training and benchmark data. For automatic *concept* classification when color is present, the image regions are represented by a concatenation of 84-bin HSI color histograms, 72-bin edge-direction histograms, and 24 features of the



sky	47.5%
water	23.5%
grass	0.0%
trunks	0.0%
foliage	0.0%
field	0.0%
rocks	20.0%
flowers	0.0%
sand	9.0%

sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	sky	sky rocks	sky rocks	sky	sky rocks	sky	sky	sky
sky	sky rocks	rocks	rocks	rocks	sky	sky rocks	sky	sky	sky
rocks	rocks	rocks	rocks	rocks	rocks	rocks	sky	sky	sky
rocks	rocks	rocks	rocks	rocks	rocks	rocks	rocks water	water	water
sand	sand	sand water	sand water	water	water	water	water	water	water
sand	sand	water	water	water	water	water	water	water	water
sand	sand	sand	water	water	water	water	water	water	water

Fig. 5. Image representation through semantic modeling.

gray-level cooccurrence matrix [Jain et al. 1995]. When color has been removed, the image regions are represented by a concatenation of a 32-bin intensity histogram, a 72-bin edge-direction histograms, and 24 features of the gray-level cooccurrence matrix. Using this low-level feature information, two support vector machine (SVM) classifier [Chang and Lin 2001] were trained. The classification performance on *image region level* with color present is 71.7%. When classifying gray-scaled regions, the classification performance drops to 65.7%. In a subsequent step, the regionwise information of the concept classifiers is combined to a global image representation: the frequency of occurrence of each local semantic concept is counted, leading to the so-called concept occurrence vectors (see Figure 5). The concept-occurrence vector can be computed both using the information of the manual region annotation and using the automatic region classifications where the former serves as benchmark for the approach.

Each scene category is represented by the mean over the concept-occurrence vectors (length: $Ncov = 9$) of all images belonging to the respective category. This leads to a prototypical representation of the scene categories where the semantic concepts act as attributes and their occurrences as attribute scores. For each scene, the Euclidean distance between the concept-occurrence vector of the scene and the five prototypes is computed. The scene is assigned to the category with the shortest distance.

In Schwaninger et al. [2006], the authors show that the semantic modeling approach is psychophysically very plausible. They gathered human typicality ratings of natural scenes and learned a psychophysically plausible distance measure that lead to a high correlation between the computational and the human ranking of natural scenes even without an optimized distance measure. This correlation decreases significantly in control experiments using global or nonsemantic image information, showing that the semantic modeling approach is consistent with scene processing by humans.

Fei-Fei and Perona [2005] and Bosch et al. [2006] also propose approaches for scene and object classification that use local features (maximum size 15 × 15 pixels) and an intermediate level representation. Semantic modeling differs from those methods in that the local patches have a size of 48 × 72 pixels and thus contain "recognizable" semantic image information leading to the psychophysical plausibility of the method as shown in Schwaninger et al. [2006]. The intermediate representation of both Fei-Fei and Perona [2005] and Bosch et al. [2006] have no semantic meaning. In addition, the methods have not been shown to be perceptually relevant.

8.2 Modeling Global Information: Computational Gist

Several studies in scene perception have shown that humans are able to understand the general context of novel scenes even when presentation time is very short (<100 ms) [Thorpe et al. 1996], when images are not fully attended to Fei-Fei et al. [2005], or are presented blurred [Schyns and Oliva 1994]. This overall meaning of a scene is often referred to as "gist" and is most commonly associated with low-level global features, such as color, spatial frequencies, and spatial organization, although the full definition of gist also includes objects as well as higher-level perceptual and conceptual information (see [Wolfe 1998; Oliva 2005]). Very recently, Fei-Fei et al. [2007] proposed to use the term "gist" only to denote the content of a scene after a certain amount of viewing time.

In this paper, we want to model the global, i.e., the spatial, configural information in a scene, since that is the main image information that remains when blurring the scenes in Experiment 3. Thus, for modeling the global information of a scene, we use the computational approach of Oliva and Torralba [2001], and Torralba et al. [2004] that analyzes the spatial information in gray-scale images and has been shown to be psychophysically relevant [Oliva and Torralba 2006].

The authors propose a low-dimensional representation of the scene structure, based on the output of filters tuned to different orientations and scales. We tested two different implementations of the method provided by the authors. Oliva and Torralba [2001] employ a bank of Gabor filters in the frequency domain tuned to different orientations and scales. Torralba et al. [2004] use a wavelet image decomposition through a steerable pyramid tuned to several orientations and scales. The second method, based on the approach in Torralba et al. [2004], resulted in significantly better performance so that we will only discuss this method in the following. The representation resulting from multiple-scale image filtering is projected onto the first *Npc* principal components computed on the full database. The number of orientations (*Nori*= 6), scales (*Nsc*= 5), and principal components (*Npc*= 50) was selected so as to maximize performance. The resulting feature vector of length *Npc*= 50, often referred to as computational gist, represents the global, configural image information and is used for scene categorization. Each scene category is represented by the mean over all gists belonging to the respective category. For each scene, the Euclidean distance between the gist of the scene and the five prototypes is computed. The scene is assigned to the category with the shortest distance.

8.3 Experiments

The following experiments test the categorization performance of the representations through semantic modeling and of the computational gist representation. Category ground truth was obtained from the human categorization results of Experiment 1. All experiments have been ten-fold cross-validated, meaning that in each round, 9/10 of each category has been used as training set for the computation of the prototype. The remaining images were categorized using the learned prototype. In the case of the semantic modeling, all 25,000 local regions (10 × 10 regions × 250 images) have been annotated manually with the nine local semantic concepts in order to obtain a maximally achievable benchmark and for the training of the concept classifiers. The experiment has then been performed three times: *SemMod anno* refers to the benchmark experiment with the concept–occurrence vector based on manually labeled

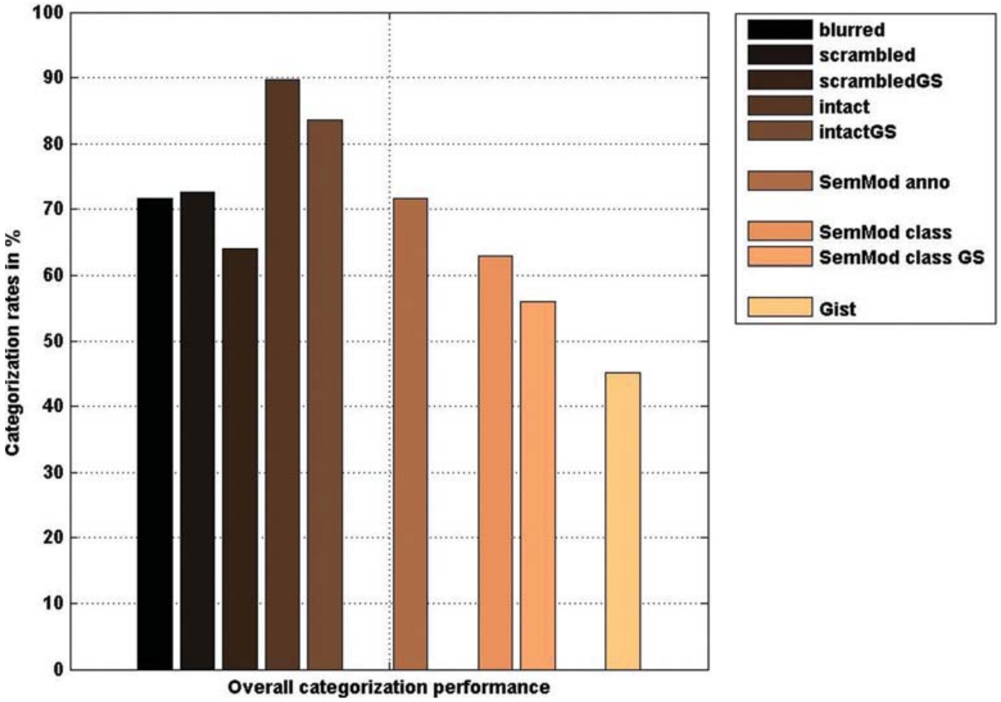


Fig. 6. Averaged categorization performance of human subjects (group of bars on the left), semantic modeling based on region annotations (SemMod anno), semantic modeling based on classified image regions (SemMod class and SemMod class GS), and of gist.

data. *SemMod class* refers to the fully automatic categorization when local image regions in color have been *classified* using a SVM classifier. *SemMod class GS* refers to the fully automatic categorization when *gray-scaled* local image regions have been classified using an SVM classifier.

8.4 Results and Discussion

Figure 6 shows the categorization performance of the computational approaches compared to the human performance averaged over all categories and Figure 7 per category. The legend for both figures is shown in Figure 6.

8.4.1 *SemMod Anno.* When looking at the overall performance, the semantic modeling, based on annotated concepts, performs with 72.8% as well as humans in both the scrambled and in the blurred condition. That means, given perfect (i.e. annotated) information from the concept classifiers, semantic modeling is able to reproduce the overall categorization accuracy of humans in the scrambled condition. In the scrambled case, both humans and the computer have the same amount of information. The per-category performance follows the human performance pattern in the blurred condition for coasts and rivers/lakes. For forests, plains, and mountains, semantic modeling performs in a similar fashion to humans in the scrambled condition.

8.4.2 *SemMod Class.* When based on classified image regions, the overall performance drops from 72.8% to 62%. This performance decrease is mainly as a result of a large drop in the plains category and a smaller drop in the coasts category. The main reason for this is that concepts that are very important

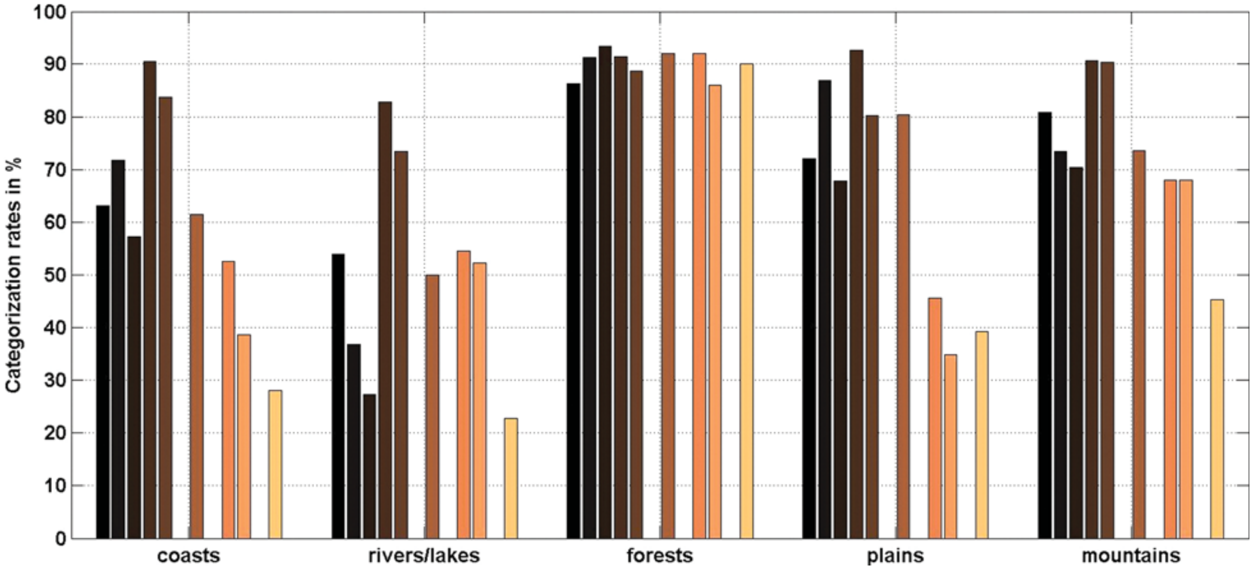


Fig. 7. Averaged categorization performance per category for human subjects and for computational models (for legend see Figure 6).

for the categorization of these categories, such as sand, flowers, and field have a fairly low classification rate in the SVM classification. This issue might be solved by improving the concept classifier and the low-level feature representation of the image regions. In all other categories, the performance of the fully automatic categorization is very close to the benchmark and, thus, surprisingly stable given that the SVM concept classifier has only a performance of 71.7%.

8.4.3 *SemMod ClassGS.* The removal of color information from the local image regions leads to a performance drop of the SVM *region classifier* to 65.7%. This yields a lower performance of 56% for computational scene categorization only based on local texture and intensity information. Figure 7 shows that as with SemMod class additional miscategorizations mainly occur in the coasts and in the plains category. From Figure 7, we also see that the relative performance drop due to color removal between scrambled and scrambledGS for humans and between SemMod class and SemMod classGS for the computer are very similar and occur in the same categories. In addition to the results in Schwaninger et al. [2006], this is another indication that semantic modeling is a good model for human scene categorization performance.

8.4.4 *Gist.* The categorization performance based on the gist at 52% is inferior to semantic modeling based on colored-image regions and slightly inferior to semantic modeling based on gray-scaled image regions. In all categories except for forest, gist performance is significantly lower than human performance compared to both blurred and scrambled display conditions. This outcome is surprising given the good results for similar categories reported in [Oliva and Torralba 2001]. It seems that the categories in our database do not exhibit consistent properties that are well detected by gist, such as openness, expansion, or roughness. In addition, the support for computing the principal components is much smaller in our experiment since the database with only 50 images per category is smaller.

In all computational experiments, we also tested a sparse multinomial logistic regression classifier instead of the prototype classifier. In all cases, this classifier did not lead to a higher categorization

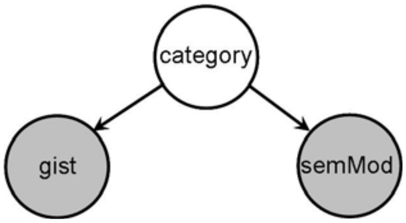


Fig. 8. Naive Bayes classifier for the combination of local and global information (shaded nodes refer to observed nodes).

Table VI. Result of Categorizer Combination Using Naive Bayes for Different Inputs of the Semantic Modeling Categorizer

		Gist	Naive Bayes Combination
SemMod anno:	72.8%	52.0%	73.6%
SemMod class:	61.2%	52.0%	66.0%
SemMod class GS:	56.4%	52.0%	61.2%

performance suggesting that not the classifier, but the image representation is the weak point of the categorization procedure.

9. GLOBAL AND LOCAL INFORMATION: CLASSIFIER COMBINATION

In a final computational experiment, we combined the outcomes of the global and the local classifiers using a naive Bayes classifier with five states per node representing the five scene categories (see Figure 8). Using a Bayes classifier is one choice for information fusion. Note that is not necessarily psychophysically plausible. However, the primary goal of the experiment is to test whether any combination of the local and the global classifier lead to a improvement in computational categorization accuracy.

Since the two image representations model different aspects of the image, they can be assumed to be independent. The latent variable is the label of the scene category and observed variables are the result of the gist and of the semantic modeling classification. The prior probability $P(category)$ is the average of the category priors over the cross-validation rounds of the previous section. The confusion matrices per classifier are as well averaged over the cross-validation rounds. They can thus be employed as the conditional probabilities $P(gist = c|category = c')$ and $P(semMod = c|category = c')$. Input to the graphical model are the observations of both the gist and the semantic modeling classifier.

Table VI summarizes the overall categorization rates for each of the classifiers separately and for the naive Bayes combination of SemMod anno and gist, of SemMod class and gist, and of SemMod class GS and gist. In all three cases, i.e., with annotated and with classified, local concepts in the semantic modeling, the combined classifier outperforms both single classifiers. In the annotated case, the performance increase results in only two additional images being recognized compared to the performance of semantic modeling alone. However, the performance increase is nearly 5% in the cases of semantic modeling, based on classified image regions.

The classifier combination results in a moderate performance increase. Though, even with combined classifiers, the performance in the *fully automatic* case does not reach human performance in either the blurred or the scrambled condition. It seems that the computational models do no to pick up all relevant details that humans use in scene categorization. In the case of the semantic modeling, this information is most likely local semantic concepts that are important, but not well classified, such as sand, flowers, or field (see Section 8.4). In the case of gist, the global information per category in our database might be too inconsistent to be modeled successfully. Here, a larger database might help. In our case, the database of

250 images, i.e., 50 images per category, is very small for training such complex computational models. The assumption that the two classification approaches are orthogonal and can thus be integrated using a simple combination scheme seems not to be fully valid. Instead, the two classifiers agree on the categorization of many scenes. It has to be noted, however, that for a fair comparison with the human data, both computational models have been implemented without the inclusion of any spatial information. Both semantic modeling and the computational gist can be computed using additional layout image information that leads in both cases to greatly improved performance (see Vogel and Schiele [2007] and Oliva and Torralba [2001]).

10. DISCUSSION AND CONCLUSION

In this paper, we took a closer look at the influence and interaction of local versus global information in scene categorization, as well as the role of color. In recent years, much evidence was presented that humans are able to catch the global, general idea of scene very rapidly, with little attention, and in blurred or color-transformed conditions (for an overview, see Oliva [2005]). However, little research has been done covering the impact of local, nonobject-centered information, also in the case of longer presentation times.

The human experiments in the first part of this paper show clearly that humans use both local, region-based, and global, configural information for scene categorization. When images contain either only local or only global information, categorization performance is lower than when intact images are presented. This is consistent with the view that humans, in fact, integrate these two kinds of image information. Most interestingly, the experiments showed that the categorization performance depends on the scene category: rivers/lakes and mountains are categorized better using global information, whereas coasts, forests, and plains are categorized better using local information. Intuitively, this result makes sense: for recognizing a mountain or a rivers/lakes scene global information, such as horizon lines or the outline of a lake are very important. In contrast, the identification of local regions containing water or foliage helps to recognize coasts or forests. A good example for this phenomenon is the coast image displayed in Figure 1. In the blurred condition, the image is a reminder of a mountain scene because of the global structure, whereas in the scrambled condition the local water regions can be recognized based on texture and color information. Given these observations, humans seem to integrate global and local information. Thus, modeling and integration of global as well as local information could be of vital importance for any automatic categorization system.

It should be noted that, of course, we do not assume or claim that all information is either local or global. However, we are interested in the relevance of local versus global information in the stimulus for human scene categorization. Scrambling of images destroys global configural information. Blurring gray-scale versions eliminates local featural or textural information. As explained above, the blur level was determined in pilot experiments by blurring scrambled images until chance performance was achieved. This blur level, therefore, by definition, eliminates all local information in the stimulus. By using these scrambled and blurred stimuli, we can, therefore, selectively test how relevant local and global information in the stimulus are for human scene categorization. The fact that humans can recognize both scrambled and blurred gray-scale images shows that both local and global information in the stimulus is important for human scene processing, irrespective of the sensitivity of the human visual system for different spatial frequencies.

In an additional set of experiments, we explored the influence of color on the categorization process and evaluate reaction times. Our experiments showed that we use color as an additional information channel for a scene categorization task. The effect of color, however, depends on the information that is already present in the image (whether it is a globally consistent, intact scene, or a local, scrambled scene). Analysis of the reaction times revealed that the processing of this additional channel can lead to higher categorization performance albeit at the cost of increased response times. This result differs from experiments reported in Oliva and Schyns [2000] and Goffaux et al. [2005], where color has been found to lead to faster scene recognition. It is important to note here that the test paradigm and the display times are different from the experiments reported in Oliva and Schyns [2000] and Goffaux et al. [2005]. We are using a five-alternative forced-choice paradigm and a display time of up to 4 s instead of express categorization with only one single fixation as in Goffaux et al. [2005] or for 150 ms as in Oliva and Schyns [2000]. Oliva and Schyns [2000] conducted scene categorization experiments using ”color-diagnostic” (such as desert, canyon) and “noncolor-diagnostic” categories (such as, city, restaurant). Although, by their definition, our categories would be “color-diagnostic” and would, therefore, presumably result in faster categorization, in our database, color is by no means diagnostic for the scene categories. The scenes in each category were taken in a variety of seasons and weather conditions so that priors such as “water is blue”, and “forest is green” do not apply. This might explain why categorization based on intensity information alone is faster than when stimuli are presented in color. In the case of scrambled images, most likely additional cognitive processes are active so that color does not influence the reaction times. Furthermore, the absolute increase in performance that is provided by color, while significant, seems small (6–8%). This confirms earlier studies, who also found small or even no effects of color on tasks, including rapid scene categorization [Delorme et al. 2000], scene categorization without attention [Fei-Fei et al. 2005], as well as scene memory [Wichmann et al. 2002]. More specifically, the gain we found is consistent with the one reported in Oliva and Schyns [2000], who conducted scene categorization experiments across several resolution levels and found a consistent gain for color images of 10 to 15% compared to gray-level images. Taken together, these results suggest that color, as opposed to the role of other cues such as local texture information, for example, plays only a minor role in scene categorization.

We tested two state-of-the-art computational approaches for scene categorization: semantic modeling analyzing local, region-based information [Vogel and Schiele 2007] and computational gist modeling global, configural information [Oliva and Torralba 2001]. The experiments show that in the benchmark condition, semantic modeling reaches the same performance as humans in the degraded display conditions (scrambled/blurred). Because of the imperfection of the concept classifier, the performance of semantic modeling drops slightly in the fully automatic case when based on color and more when based on gray-scale information. However, the relative performance drop between color and gray-scale information follows the human pattern in each category, indicating that given better concept classifiers semantic modeling is a very good model of human categorization performance. Categorization based on the computational gist representation exhibits lower performance compared to semantic modeling. A reason for this low performance might be the intracategory variations of the images: all categories contain images with varying depth which poses a challenge for gist. Computational gist is particularly strong in modeling images with similar spatial layout. Furthermore, the training set of 50 images per category is very small for categorization based on computational gist features.

In a final experiment, the local and the global classifier were combined using a Bayesian framework. Categorization results with the combined classifier outperformed both single classifiers in each case. This is a promising step in the direction of integrating local and global information for scene classification. However, the combined performance remains below the ultimate goal of scene classification, which is human performance in the intact condition. Therefore, the development of more sophisticated and perceptually plausible methods for information integration remains an interesting area for future research.

Further manipulations that will need to be done in order to investigate the perceptual parameters of scene categorization include shortening the presentation time (this will address cognitive influences on categorization) as well as exploring different scrambling and blurring levels (this will address the scale and frequency content of global and local information). In general, research in both human perception and in computer vision remains challenged in the future. Research in human perception needs to determine what is the important *semantic* or *context* information for human scene recognition, while research in computer vision needs to develop mainly features, but also algorithms and methods for modeling this information and for building automatic scene recognition systems.

ACKNOWLEDGMENT

JV was supported by research fellowships of the Max Planck Society and the German Research Foundation (DFG). All intact images used in this work: ©2006 MPI for Biological Cybernetics and its licensors. All rights reserved.

REFERENCES

Biederman, I. 1972. Perceiving real-world scenes. *Science 177,* 43, 77–80.

Bosch, A., Zisserman, A., and Munoz, X. 2006. Scene classification via pLSA. In *European Conference on Computer Vision ECCV'06*. Graz, Austria.

Chang, C.-C. and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines*. Software available at: http://www.csie.ntu.edu.tw.

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

Fei-Fei, L. and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conf. on Computer Vision and Pattern Recogntion CVPR'05*. San Diego, CA.

Fei-Fei, L., Van Rullen, R., Koch, C., and Perona, P. 2005. Why does natural scene categorization require little attention? exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition 12*, 6, 893–924.

Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision 7*, 1, 1–29.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P., and Rossion, B. 2005. Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition 12*, 6, 878–892.

Hayward, W. 2003. After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences 7*, 10, 425–427.

Hayward, W., Rhodes, G., and Schwaninger, A. 2007. An own-race advantage for components as well as configurations in face recognition. *Cognition*. DOI: 10.1016/j.cognition.2007.04.002

Henderson, J. 2005a. Introduction to real-world scene perception. *Visual Cognition: Special Issue on Real-World Scene Perception 12*, 849–851.

Henderson, J., Ed. 2005b. *Visual Cognition: Special Issue on Real-World Scene Perception,* Vol. 12.

Jain, R., Kasturi, R., and Schunck, B. 1995. *Machine Vision*. McGraw-Hill, New York.

McCotter, M., Gosselin, F., Sowden, P., and Schyns, P. 2005. The use of visual information in natural scenes. *Visual Cognition 12*, 6, 938–953.

Mojsilovic, A., Gomes, J., and Rogowitz, B. 2004. Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues. *International Journal of Computer Vision 56*, 1/2 (Jan.), 79–107.

Oliva, A. 2005. Gist of a scene. In *Neurobiology of Attention.*, L. Itti, G. Rees, and J. Tsotsos, Eds. Academic Press, and Elsevier, New York. 251–256.

Oliva, A. and Schyns, P. 2000. Diagnostic color blobs mediate scene recognition. *Cognitive Psychology 41*, 176–210.

Oliva, A. and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision 42*, 3 (Mar.), 145–175. Matlab code from http://people.csail.mit.edu/torralba/code/spatialenvelope/.

Oliva, A. and Torralba, A. 2006. Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research 155*, 23–39.

Rogowitz, B., Frese, T., Smith, J., Bouman, C., and Kalin, E. 1997. Perceptual image similarity experiments. In *SPIE Conference on Human Vision and Electronic Imaging*. San Jose, CA. 576–590.

Rosch, E., Simpson, C., and Miller, R. 1976. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance 2*, 491–502.

SCHWANINGER, A., LOBMAIER, J. S., AND COLLISHAW, S. M. 2002. Role of featural and configural information in familiar and unfamiliar face recognition. In *2nd Conference on Biologically Motivated Computer Vision BMCV*. Springer, Lecture Notes in Computer Science, 2525, Tübingen, Germany.

SCHWANINGER, A., CARBON, C., AND LEDER, H. 2003. Expert face processing: Specialization and constraints. In *Development of face processing*, G. Schwarzer and H. Leder, Eds. Hogrefe & Huber Publishers, Inc., Cambridge, MA. 81–97.

SCHWANINGER, A., VOGEL, J., HOFER, F., AND SCHIELE, B. 2006. A psychophysically plausible model for typicality ranking of natural scenes. *Transactions of Applied Perception 3*, 4 (Oct.), 333–353.

SCHYNS, P. AND OLIVA, A. 1994. From blobs to boundary edges: evidence for time- and spatial-scale dependent scene recognition. *Psychological Science 5*, 195–200.

SZUMMER, M. AND PICARD, R. 1998. Indoor-outdoor image classification. In *Workshop on Content-based Access of Image and Video Databases*. Bombay, India.

THORPE, S., FIZE, D., AND MARLOT, C. 1996. Speed of processing in the human visual system. *Nature* (London) *381*, 520–522.

TORRALBA, A., MURPHY, K. P., AND FREEMAN, W. T. 2004. Contextual models for object detection using boosted random fields. Tech. Rep. AIM-2004-008, MIT, AI Lab. (Apr.).

TVERSKY, B. AND HEMENWAY, K. 1983. Categories of environmental scenes. *Cognitive Psychology 15*, 121–149.

VAILAYA, A., FIGUEIREDO, M., JAIN, A., AND ZHANG, H. 2001. Image classification for content-based indexing. *IEEE Transactions on Image Processing 10*, 1 (Jan.), 117–130.

VOGEL, J. AND SCHIELE, B. 2007. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision 72*, 2 (Apr.), 133–157.

WALKER RENNINGER, L. AND MALIK, J. 2004. When is scene identification just texture recognition? *Vision Research 44*, 4 (Apr.), 2301–2311.

WALLRAVEN, C., SCHWANINGER, A., AND BÜLTHOFF, H. 2005. Learning from humans: computational modeling of face recognition. *Network: Computation in Neural Systems 16*, 4, 401–418.

WICHMANN, F., SHARPE, L., AND GEGENFURTNER, K. 2002. The contribution of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory and Cognition 28*, 3, 509–520.

WOLFE, J. 1998. Visual memory: What do you know about what you saw? *Current Biology 8*, R303–R304.

Received December 2006; revised June 2007; accepted June 2007