Categorization of natural scenes: local vs. global information

Julia Vogel[1,2] * Adrian Schwaninger[2] † Christian Wallraven[2] ‡ Heinrich H. Bülthoff[2§]

[1]University of British Columbia
Vancouver, Canada

[2]Max Planck Institute for Biological Cybernetics
Tübingen, Germany

Figure 1: Exemplary image in its four display conditions: intact, scrambled, blurred, blurred-scrambled.

Abstract

Understanding the robustness and rapidness of human scene categorization has been a focus of investigation in the cognitive sciences over the last decades. At the same time, progress in the area of image understanding has prompted computer vision researchers to design computational systems that are capable of automatic scene categorization. Despite these efforts, a framework describing the processes underlying human scene categorization that would enable efficient computer vision systems is still missing. In this study, we present both psychophysical and computational experiments that aim to make a further step in this direction by investigating the processing of local and global information in scene categorization. In a 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 model either local or global information.

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

Keywords: scene perception, scene classification, computational modeling, semantic modeling, gist, local region-based information, global configural information

*e-mail: vogel@cs.ubc.ca
†e-mail: adrian.schwaninger@tuebingen.mpg.de
‡e-mail: christian.wallraven@tuebingen.mpg.de
§e-mail: heinrich.buelthoff@tuebingen.mpg.de

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 characterizing and understanding scene categorization processes (e.g. rapid scene categorization [Thorpe et al. 1996], categorization with little attention [Fei-Fei et al. 2005], categorization in blurred condition [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; Vogel and Schiele ; Fei-Fei and Perona 2005], the complexity of scenes continues to provide a challenge to computer vision research. Because of the difficulty of the problem, in this paper we follow a combined cognitive and computational approach to understanding and implementing scene categorization (see also [Oliva and Torralba 2001; Walker Renninger and Malik 2004; McCotter et al. 2005]): On the one hand, psychophysical experiments allow us 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. On the other hand, computer vision 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

APGV 2006, Boston, Massachusetts, July 28–29, 2006.
© 2006 ACM 1-59593-429-4/06/0007 $5.00

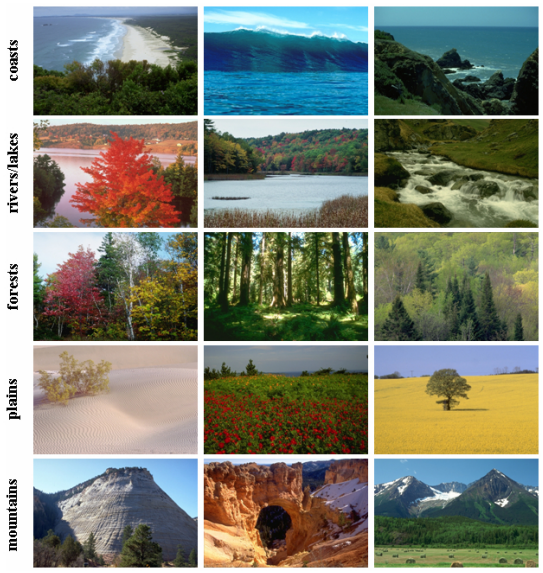

Figure 2: Exemplary images of each category.

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 vs. 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. A psychophysically plausible computational model of these processes and representations has been provided recently by [Wallraven et al.]. In object recognition, the role of rotation-invariant local parts (geons) vs. more global view-based representations has been discussed extensively in the last 20 years (e.g. [Hayward 2003]). The goal of this paper is 1) to examine the processing of local and global information in human scene categorization using psychophysics, and 2) to compare two computational models of scene categorization with human performance.

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 might be attributed to more than one category. The goal of Experiment 1 was to determine the ground truth and the benchmark for our employed scene database. This information is the basis for all succeeding experiments.

The selection of the natural scene categories follows the rationale of [Vogel and Schiele] 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 vs. non-human and natural vs. 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 to be used our experiments were to not contain any 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 extended to the categories coasts, rivers/lakes, forests, plains, and mountains.

2.1 Method

Participants 11 naive participants were paid to participate in the study. All had normal or corrected to normal vision.

Stimuli and procedure 250 nature images of the Corel image database (720x480 pixels) 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. The images were presented at 100 Hz on a Sony Trinitron 21" monitor with resolution 1280x960 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. T he 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 seconds 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 10x10 grid in order to simulate similar high-order frequency distortions as in the subsequent experiments (see Figure 1). Participants were asked to categorize the displayed image into one of the five categories as fast and accurately as possible by checking the corresponding box 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 as result, the database contains 57 coast-, 44 rivers/lakes-, 50 forest-, 46 plains-, and 53 mountain-images. Based on this ground truth, the average categorization performance in Experiment 1 is 89.7%. Table 1 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

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

and mountains in only one direction. Also, the rivers/lakes category 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; Hayward 2003]). 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.] 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], [Schwaninger et al.] implement an intermediate semantic modeling step by extracting local semantic concepts such as rock, water, sand, etc.. The authors found a high correlation between the computational and the human ranking of natural scenes regarding typicality. Interestingly, computational model comparisons 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, sand, etc. is psychophysically very plausible. In this study we further investigate the role of such local semantic information. Thus, instead of a 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 10x10=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 11 participants were paid to participate in the study. None of them had participated in Experiment 1. All had normal or corrected to normal vision.

Stimuli and procedure 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 10x10=100 regions of 72x48 pixels, and randomly repositioning the resulting regions. The scrambled image has the same size (720x480 pixels) as the original. The random scrambling of images was new for each participant to prevent any particular spatial configuration from influencing the results. As before, the images were superimposed by a 10x10 grid to control for high-frequency distortions (see Figure 1). Regarding monitor, room, viewing distance, and display times, the experimental conditions were the same as in Experiment 1. As before, the task was to categorize the displayed image as fast and as accurately as possible into one of the five given scene categories.

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%

Table 1: Confusion matrix for categorization of intact images in Experiment 1.

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%

Table 2: Confusion matrix for categorization of scrambled images in Experiment 2.

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 2 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 $p < .01$ for rivers/lakes and $p < .001$ 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 ($F(2.551, 25.506) = 42.33$, $MSE = 187.225$, $p < .001$). 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 vs. scrambled) as between-subjects factor. There were main effects of condition ($F(1,20) = 78.301$, $MSE = 108.301$, $p < .001$) and category ($F(3.088,61.767) = 34.710$, $MSE = 130.302$, $p < .001$). There was also a significant interaction between condition and category ($F(3.088,61.767) = 17.169$, $p < .001$), 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

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%

Table 3: Confusion matrix for categorization of blurred images in Experiment 3.

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 completely scrambled. 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 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 11 participants 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. In addition, the displayed image was superimposed by a 10x10 grid in order to account for the same high-order frequency distortions as in the previous experiments (See Figure 1 for an exemplary image. However, note that the visual angle of the subjects was significantly larger than that of the reader.). The experimental setup concerning monitor, room, viewing distance, and display time was the same 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 reduced by blurring of the stimuli. The blur level is adjusted by scrambling and blurring stimuli until categorization performance in the blurred-scrambled condition drops down to chance which indicates that both local and global information has been eliminated. In the final pilot experiment for this study with 11 naive subjects, average categorization performance in the blurred-scrambled condition was 23% which is close to the chance level of 20%. Figure 1 shows also the exemplary image in the blurred-scrambled condition.

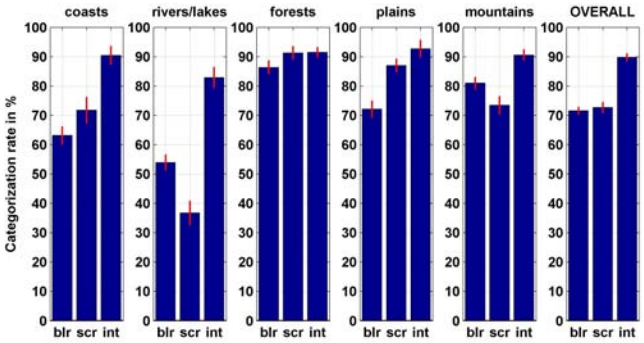

Figure 3: Comparison of categorization rates between blurred (blr), scrambled (scr), and intact (int) display condition (in %).

4.2 Results and discussion

The average overall categorization performance in the blurred condition was 71.6%. As in the scrambled condition, the categorization performance is relatively stable compared to the intact condition. Table 3 reveals that compared to Experiment 2 there are fewer confusions between rivers/lakes and coasts, rivers/lakes and forests, and mountains and plains, but that there are now 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 ($p < .001$). 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 ($F(1.989, 19.894) = 25.188$, $MSE = 151.273$, $p < .001$). 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 vs. blurred) ($F(1, 20) = 129.666$, $MSE = 70.944$, $p < .001$), and of category ($F(2.839, 61.767) = 18.385$, $MSE = 110.640$, p<.001). There was also an interaction: $F(2.839, 61.767) = 7.853$, $p < .001$).), 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 vs. blurred) ($F(1,20) = 5236,028$, $MSE =$

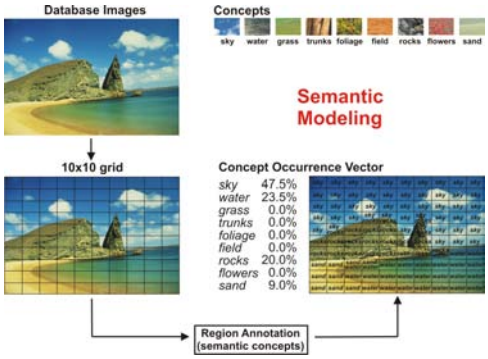

Figure 4: Image representation through semantic modeling

107.921, $p > .05$)), indicating that the two conditions are comparable in difficulty[2]. However, there was a main effect of category ($F(2.767,55.336) = 61.997$, $MSE = 140.683$, $p < .001$) as well as an interaction between condition and category ($F(2.767,55.336) = 9.415$, $p < .001$), suggesting that different types of information are used for different categorization judgements.

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. Interestingly, 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.

5 Computational scene categorization

In the previous sections, we analyzed human performance in categorizing natural scenes when only local information or only global information is present. The experiments showed that humans use both local and global information, and that this information seems to be integrated for the final category decision. The goal of the following experiments is to evaluate computational categorization performance for the same task. In particular, we compare a local, region-based approach proposed by [Vogel and Schiele ; Schwaninger et al.] and an approach that models global context information proposed by [Oliva and Torralba 2001] with the human performance. Both approaches have been shown to be psychophysically plausible models of human scene perception [Schwaninger et al. ; Oliva 2005].

[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. However, this is not relevant for the main conclusions of this study.

5.1 Modeling local, region-based information: Semantic modeling

For modeling local, region-based image information, we employ the semantic modeling approach of [Vogel and Schiele] that makes use of an intermediate modeling step for categorization and ranking of natural scenes. Images are divided into a regular grid of 10x10 local regions, and the local regions 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, the image regions are represented by a concatenation of 84-bin HSI color histograms, 72-bin edge direction histograms, and 24 features of the gray-level co-occurrence matrix [Jain et al. 1995]. Using this low-level feature information, a support-vector-machine (SVM) classifier [Chang and Lin 2001] was trained. Its classification performance on *image region level* is 71.7%. In a subsequent step, the region-wise 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 4). The concept-occurrence vector can be computed both using the information of the manual region annotation and using the automatic region classification 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.], the authors show that the semantic modeling approach is psychophysically very plausible. They recorded 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 non-semantic image information, showing that the semantic modeling approach is consistent with scene processing by humans.

5.2 Modeling global information: Gist of a scene

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 msec) [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 or gist of a scene 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 higher-level perceptual and conceptual information (see [Oliva 2005]). [Wichmann et al. 2002; Fei-Fei et al. 2005] have suggested that color as is not particularly informative in scene categorization. Thus, for modeling the global information of a scene, we use the computational approach of [Oliva and Torralba 2001].

These 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. [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], however, 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 $Npc = 50$, i.e. the gist, represent 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.

5.3 Experiments

The following experiments test the categorization performance of the representation through semantic modeling and of the gist representation. Category ground truth was obtained from the human categorization results of Experiment 1. All experiments have been 10-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 x 10 regions x 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 twice. *Anno* refers to the benchmark experiment with the concept-occurrence vector based on manually labeled data. *Class* refers to the fully automatic categorization when the local image regions have been *classified* using the SVM classifier.

5.4 Results and discussion

Figure 5 shows the categorization performance of the computational approaches compared to the human performance per category and overall (on the far right).

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. 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.

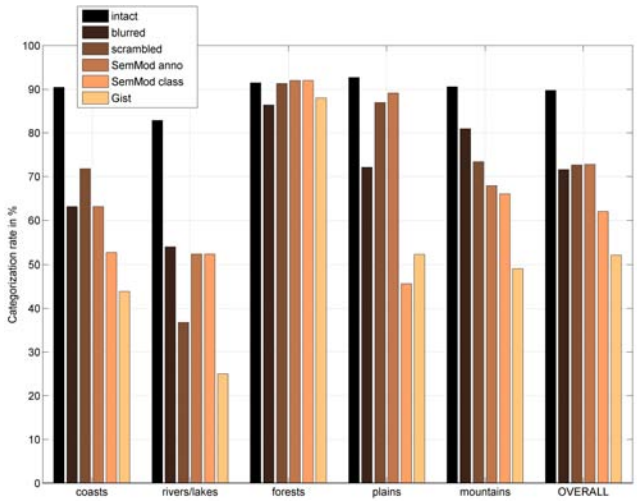

Figure 5: Comparison categorization performance human-computational

Class: When based on classified image regions, the overall performance drops from 72.8% to 62%. This performance decrease is mainly due to 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 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%.

Gist: The categorization performance based on the gist at 52% is far inferior to semantic modeling. 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 is smaller.

In an effort to understand the results, we repeated the computational experiments with only those images that at least 10 of the 11 subjects in Experiment 1 agreed on. The question was whether the image close to category boundaries have a high impact on the categorization results. However, the improvement by using only images with high agreement was marginal ($<3\% \equiv 7$ images in the case of gist).

We also tested a sparse multinomial logistic regression classifier in place of the prototype classifier. In all cases, this classifier did not lead to a higher categorization performance suggesting that not the classifier, but the image representa-

intact	89.7%
blurred	71.6%
scrambled	72.7%
SemMod anno	72.8%
SemMod class	61.2%
Gist	52.0%
Naive Bayes (SemMod anno + Gist)	73.6%
Naive Bayes (SemMod class + Gist)	66.0%

Table 4: Overall categorization rates for human, local, global, and local+global, naive Bayes categorization

tion is the weak point of the categorization procedure.

6 Global and local information: classifier combination

In a final 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. Since the two image representations model different aspects of the image, they can be assumed independent. 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 4 summarizes the overall categorization rates for the various systems. The first three rows repeat the categorization rates of the human experiments. Rows four and five show the categorization rate of the semantic modeling approach alone with annotated as well as with classified local concepts. Row six shows the performance of the gist categorizer. The last two rows display the categorization performance of the simple integration of global and local information. In both 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 is equivalent to an additional two images that are recognized compared to the performance of semantic modeling alone. The performance increase is nearly 5% in the fully automatic classification.

The classifier combination results in a moderate performance increase. However, even with combined classifiers, the performance in the fully automatic case does not reach human performance in either the blurred or the scrambled, nor in the intact display condition. It seems that the computational models, especially the gist, 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 5.4). 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. Finally, the assumption that the two classification approaches are orthogonal and can thus be integrated using a simple combination scheme might not be fully valid.

7 Discussion and Conclusion

In this paper, we took a closer look at the influence and interaction of local vs. global information in scene categorization. In recent years, much evidence was presented that humans are able to catch the gist, that is 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, non-object 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 reminds of a mountain scene due to 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.

We tested two state-of-the-art computational approaches for scene categorization: semantic modeling analyzes local, region-based information [Vogel and Schiele] and gist models 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 condition. Due to the imperfection of the concept classifier, the performance of semantic modeling drops slightly in the fully automatic case. Categorization based on the gist representation exhibits significantly lower performance compared to semantic modeling. A reason for this low performance might be the intra-category variations of the images: all categories contain images with varying depth which poses a challenge for gist. Gist is particularly strong in modeling images with similar spatial layout.

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, that is human performance in the intact condition. Therefore, the development of a sophisticated or even perceptually plausible

methods for information combination 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.

Acknowledgments JV was supported by research fellowships of the Max Planck Society and the German Research Foundation.

References

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

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

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*.

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.

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

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

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

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

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 (January), 79–107.

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 (March), 145–175.

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

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*, 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., Vogel, J., Hofer, F., and Schiele, B. A psychophysically plausible model for typicality ranking of natural scenes. *Transactions of Applied Perception*. Under revision.

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. 81–97.

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*.

Thorpe, S., Fize, D., and Marlot, C. 1996. Speed of processing in the human visual system. *Nature 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, April.

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 (January), 117 – 130.

Vogel, J., and Schiele, B. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*. In press.

Walker Renninger, L., and Malik, J. 2004. When is scene identification just texture recognition? *Vision Research 44*, 4 (April), 2301–2311.

Wallraven, C., Schwaninger, A., and Bülthoff, H. Learning from humans: computational modeling of face recognition. *Network: Computation in Neural Systems*. In press.

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.