A Psychophysically Plausible Model for Typicality Ranking of Natural Scenes

ADRIAN SCHWANINGER
Max Planck Institute for Biological Cybernetics and University of Zurich
JULIA VOGEL
Max Planck Institute for Biological Cybernetics and University of British Columbia
FRANZISKA HOFER
University of Zurich
and
BERNT SCHIELE
Darmstadt University of Technology

Natural scenes constitute a very heterogeneous stimulus class. Each semantic category contains exemplars of varying typicality. It is, therefore, an interesting question whether humans can categorize natural scenes consistently into a relatively small number of categories, such as, coasts, rivers/lakes, forests, plains, and mountains. This is particularly important for applications, such as, image retrieval systems. Only if typicality is consistently perceived across different individuals, a general image-retrieval system makes sense. In this study, we use psychophysics and computational modeling to gain a deeper understanding of scene typicality. In the first psychophysical experiment, we used a forced-choice categorization task in which each of 250 natural scenes had to be classified into one of the following five categories: coasts, rivers/lakes, forests, plains, and mountains. In the second experiment, the typicality of each scene had to be rated on a 50-point scale for each of the five categories. The psychophysical results show high consistency between participants not only in the categorization of natural scenes, but also in the typicality ratings. In order to model human perception, we then employ a computational approach that uses an intermediate *semantic modeling step* by extracting local semantic concepts, such as, rock, water, and sand. Based on the human typicality ratings, we learn a psychophysically plausible distance measure that leads to a high correlation between the computational and the human ranking of natural scenes. Interestingly, model comparisons without a semantic-modeling step correlated much less with human performance, suggesting that our model is psychophysically very plausible.

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

General Terms: Experimentation, Algorithms

Additional Key Words and Phrases: Scene classification, typicality, image-retrieval systems

Authors' addresses: Adrian Schwaninger, Max Planck Institute for Biological Cybernetics, Dept. Bülthoff, 72076 Tübingen, Germany and Department of Psychology, University of Zurich, 8032 Zurich, Switzerland; Julia Vogel, Max Planck Institute for Biological Cybernetics, Dept. Bülthoff, 72076 Tübingen, Germany and Department of Computer Science, University of British Columbia, V6T 1Z4 Vancouver, Canada; Franziska Hofer, Department of Psychology, University of Zurich, 8032 Zurich, Switzerland; Bernt Schiele, Depto of Computer Science, Darmstadt University of Technology, 64289 Darmstadt, Germany.

1. INTRODUCTION

Natural scene categorization is a highly automated and rapid process [e.g., Potter 1976; Thorpe et al. 1996; VanRullen and Thorpe 2001]. In contrast to basic level object categorization, a natural scene can often be categorized into more than one category. In this context, one can distinguish between *overlapping* and *nonoverlapping* categories [e.g., Birnbaum 1998]. Overlapping categories have at least one stimulus that is sometimes categorized as being a member of one category and sometimes as a member of another. For nonoverlapping categories, this is never the case, i.e., a particular stimulus always belongs to either one or the other category. Natural scene images clearly represent stimuli belonging to overlapping categories. Consider, for example, an image depicting hills covered with forest in the foreground and silhouettes of mountains in the background. Such a scene could be categorized as belonging to the natural category mountains or to the category forest. Depending on the amount of visible foliage, rocks, or stones, such an image could represent a more or less typical member of one of the two categories.

Several studies have suggested that categorization and typicality are related. Typical members are learned faster, categorized more quickly, and recalled more readily than atypical members [Rosch 1975]. An overall high correlation between typicality and categorization has been found for many semantic categories [e.g., Hampton 1998]. When participants are asked to name members of a category, the order of item output can be predicted using typicality [Barsalou and Sewell 1985]. Moreover, typicality can predict reaction times in sentence-verification tasks [McCloskey and Glucksberg 1979; Rosch 1973].

Kalish [2002] suggests a continuum of category *gradedness* with some categories having more or less graded structure than others. Graded structure has been found in many category types [Barsalou 1987; Kalish 2002], such as in the domain of fruits [e.g., Rips et al. 1973; Rosch 1973], but graded structure and, in particular, the overlap of categories is very pronounced for natural scenes. Because overlapping and graded categories provide a test of categorization theories, they are theoretically very interesting.

Different theories on human categorization have been proposed [for an overview, see Birnbaum 1998]. The most prominent psychological theories are the *classical* theory [e.g., Bruner et al. 1956; Smith and Medin 1981], the *prototype* theory [Posner and Keele 1968, 1970; Rosch 1973, 1977], the *feature-frequency* theory [Estes 1986; Franks and Bransford 1971], the *exemplar* theory [Brooks 1978; Estes 1986; Hintzman 1986; Medin and Schaffer 1978; Nosofsky 1986], and the *decision-bound* theory, also called general recognition theory [Ashby 1992; Ashby and Gott 1988; Ashby and Lee 1991, 1992; Ashby and Maddox 1990, 1992, 1993; Ashby and Townsend 1986; Maddox and Ashby 1993].

The classical theory supposes that category membership is based on a set of necessary and sufficient conditions. The prototype theory proposes that a category is represented by a prototype and that categorization occurs by comparing the similarity of the input image to the prototype of each relevant category. The feature-frequency theory assumes that a category is represented by a list of different features present in all exemplars of the category plus their relative frequency of occurrence. According to the feature-frequency theory, the stimulus is decomposed into its components in order to categorize an image. Subsequently, the likelihood is calculated that the combination of the specific features was generated by each of the relevant categories. The exemplar theory assumes that the similarity between the internal stimulus representations to each stored exemplar of all relevant categories is calculated. The category is thus based on similarity calculations between the input image and each stored exemplar of all relevant categories. Finally, the decision-bound theory supposes that the perceptual space is divided into different response regions, which are partitioned by decision bounds, one for each relevant category. According to this theory, the human determines the region in which the image representation falls, which leads to the appropriate response.

Every theory assumes that the internal stimulus representation itself or properties or features are in some way compared to some internal category representations. The main difference between the theories is how the representation of the category is built. In the categorization process, some exemplar-specific information has to be ignored in order to match an image to an existing category. If, in contrast to this, not just the membership of an image to a category has to be judged, but instead typicality rankings of the scenes for each category have to be made, one could argue that more detailed information about the stimulus properties has to be processed and stored.

The information on how humans rate the typicality of scenes is not only interesting from a theoretical point of view, but could also be very valuable for applications, such as, image retrieval or video annotation systems for overviews of image-retrieval systems, see Smeulders et al. [2000] and Veltkamp and Tanase [2001]. In any domain where multiplicity exists, not only a rough categorization is of interest, but particularly the finer processing of typicality ranking. Certainly, *categorization* is a crucial first step for retrieving images, but the more detailed *typicality rankings* allow the development of more specialized image-retrieval systems. If humans would agree substantially regarding a typicality ranking of scenes and if a computational model would perform in a similar manner as humans do, it would be possible to use the model to develop general image-retrieval systems. In contrast, if only little agreement could be observed between individuals, then any general image-retrieval system would have to be adapted to the individual to assure efficiency. Knowledge on how humans agree in typicality rankings of natural scenes could help making an image-retrieval system more efficient and effective.

The first objective of this study was to investigate the agreement between humans in categorization, as well as in the rating of typicality of natural scenes. Only if typicality is consistently perceived across different individuals a general image-retrieval system makes sense, at least for initial filtering of (images for possible query methods, see Smeulders et al. [2000]). If categorization of a scene is substantially driven by its perceived typicality, one would expect a significant correlation between categorization judgments and typicality rankings. Therefore, in the first experiment, we used a simple categorization task in which participants had to classify natural scenes into one of the five categories (coasts, rivers/lakes, forests, plains, and mountains). In the second experiment, the typicality of each scene had to be judged for each of the just-mentioned categories. The amount of agreement between individual participants was of special interest, as well as the correlation between measures obtained in both experiments.

The second objective of this study was to compare human perception with computational modeling. The importance of integrating psychophysics and computational modeling was already convincingly been put forward several years ago in the seminal work by [e.g., Edelman 1999; Ullman 1996]. Recently, psychophysical findings for gaining a deeper understanding of human scene perception are recognized as being essential for the design of content-based image retrieval (CBIR) systems [e.g., Cox et al. 2000]. Earlier work tried to incorporate the "user in the loop" through relevance-feedback algorithms [see Smeulders et al. 2000]. Oliva and Torralba [2001] used psychophysical experiments to find categories that describe the structure of a scene (e.g., openness and naturalness) and designed their vision system to order images along these perceptual dimensions. Mojsilovic et al. [2000b] did similar experiments to determine the perceptually relevant dimensions of color patterns and proposed a perceptually based system for pattern retrieval and matching [Mojsilovic et al. 2000a]. In contrast, Rogowitz et al. [1997] describe experiments to find perceptual semantic categories and an adequate similarity model.

Several computational models have been proposed for scene classification or global image annotation [e.g., Feng et al. 2003; Szummer and Picard 1998; Vailaya et al. 2001]. Most of these systems are based on global, low-level feature information only, which might not be detailed enough for typicality rating of natural scenes. Based on their psychophysical experiments, Mojsilovic et al. [2004] developed and implemented mapping from high-level category description via intermediate "semantic cues" to low-level image features and a query language that combines semantic (e.g., skin, sky, and background) with descriptive cues (e.g., number of objects and number of segmented regions). Their retrieval system generates very convincing retrieval results on a variety of scene categories. Fei-Fei and Perona [2005] propose an unsupervised approach to scene classification that is based on a modified LDA model and an image representation through codebooks of SIFT features. Although the authors stress the importance of an intermediate modeling step, their system does not incorporate any *semantic* feature information. Other approaches learn the correspondence of semantic labels and local regions [e.g., Barnard et al. 2002; Feng et al. 2004], but do not use this information for global image description or representation. For the objective of typicality rating of natural scenes, an image representation is necessary that includes sufficient local detail in order to detect fine differences between images. In addition, it could be beneficial if the image representation incorporates semantic information (e.g., grass, trunks, rocks, or flowers), since such information is often used by humans to describe natural scenes. For these reasons, the computational modeling used in this study employs an intermediate semantic modeling step. In the first stage of the system, local image regions are extracted and classified into semantic concept classes, such as, grass, trunks, rocks, and flowers. Subsequently, the frequency of occurrence of these local concepts is used as global image representation, which can be used for global image comparison and typicality rating.

A large similarity between the performance of a computational model and human ratings can provide further insights on how the human brain might perform the typicality rankings and what information could be relevant in the categorization process. Moreover, such a comparison could provide interesting insights for enhancing computational algorithms on scene classification and image retrieval and, therefore, help to close the loop between psychophysics and computer science.

2. EXPERIMENT 1

In experiment 1, we used a simple natural scene classification task in which the participants had to classify each scene into one of the following scene categories: coasts, rivers/lakes, forests, plains, and mountains. The selection of the categories for the categorization and typicality rating experiments was influenced by earlier work of others: The human/natural distinction of Rogowitz et al. [1997] is considered as super ordinate category for our experiments. In a second step, the natural outdoor 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. The aims of experiment 1 were (1) to assess the consistency between humans with regard to categorizing natural scenes into these five categories and (2) to provide ground truth for the subsequent experiments.

2.1 Method

2.1.1 *Participants.* Twenty undergraduates of the University of Zurich volunteered in this study. All had normal or corrected to normal vision.

2.1.2 *Materials and Procedure.* Natural digitalized scenes (250) of the Corel image database served as stimuli. The images of the natural scenes were selected by one of the authors (JV) in such a way that each of the five categories (coasts, rivers/lakes, forests, plains, and mountains) contained 50 images. A main goal when selecting the stimuli was to include a substantial number of images close to the category boundaries. The idea was to reflect a more "real-world" selection of images of each category, in contrast to "personal" image collections of natural scenes.

The experiments were conducted in a dimly lit room. The viewing distance was maintained by a head rest, so that the center of the screen was at eye height of participants and the length and width of displayed scenes covered 12 and 17° of visual angle. The displayed scenes were 720 × 480 pixels

Table I. Interrater Reliabilities for the Categorization Task in Experiment 1[a]

Interrater Reliabilities		
Scene Category	α	r_ϕ
Coasts	0.982	0.728
Rivers/lakes	0.972	0.640
Forests	0.992	0.857
Plains	0.988	0.813
Mountains	0.980	0.711

All that phicorrelations (r_ϕ) are significant with $p < 0.001$.

[a] Calculated for each category separately with the averaged phi correlation (r_ϕ) and cronbach's alpha (α) between participants.

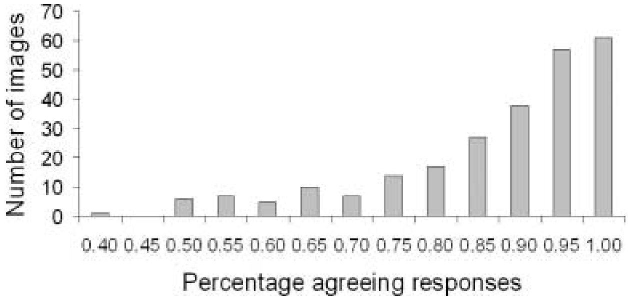

Fig. 1. Illustration of the percentage agreeing responses of the participants in the categorization experiment over all images; 1.00 means that all participants judged the scene into the same category.

(landscape format) and 480 × 720 pixels (portrait format), respectively. Prior to the experiment, all participants gave informed consent.

Stimuli were presented on a 15-inch TFT monitor. Each scene was presented to each participant once, in random order. Participants had to categorize each scene as fast and accurately as possible by pressing the corresponding response button on a serial five-button response pad. The five categories (coasts, rivers/lakes, forests, plains, and mountains[1]) were indicated at the bottom of the screen until the corresponding button was pressed. Scenes disappeared after 2 s. To prevent any label position effects, the arrangement of the names of the categories was counterbalanced across participants using a latin-square design.

2.2 Results and Discussion

The degree of interindividual consistency was estimated with Cronbach's alpha (α) between participants, as well as with the averaged phicorrelation (r_ϕ) between participants for each category (Table I). [See Appendix for the equations, references, and brief explanations of these statistics.]

We found for both interrater reliability measures high values: all α's were between 0.972 (rivers/lakes) and 0.992 (forests), and all $r\phi$ between 0.640 (rivers/lakes) and 0.857 (forests). This implies that the five categories used in this experiment are perceived consistently among different individuals and, thus, are psychologically plausible.

Figure 1 illustrates the agreement between participants on the categorization of the images. A value of 1.00 denotes that all participants (100%) categorized an image into the same category. The fact that several images have been assigned to more than one category indicates that natural scenes are, indeed,

[1] Since the study was conducted at the University of Zurich, the category labels for coast, rivers/lakes, forests, plains, and mountains were presented in German as follows: Küste, Gewässer, Wald, Feld/Ebene, and Gebirge.

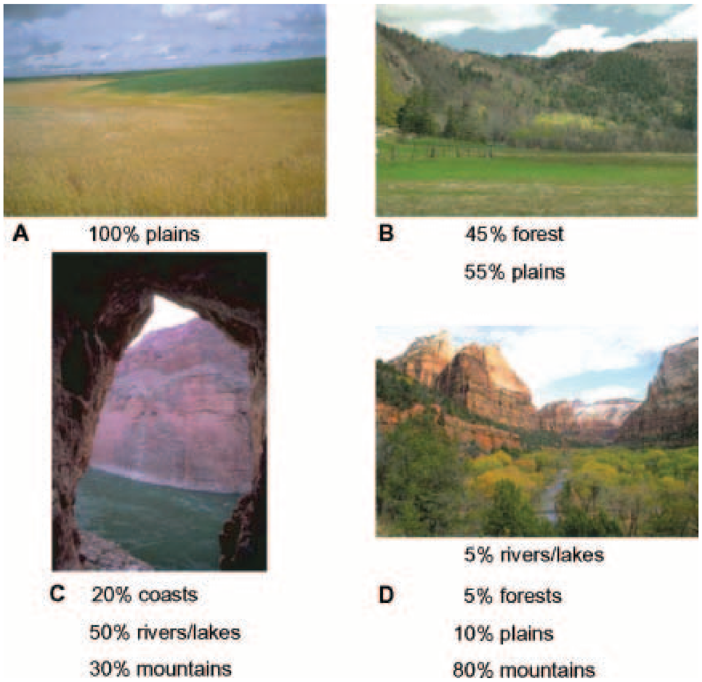

Fig. 2. Examples of human categorization data showing the semantic manifoldness of the employed scenes. The percentage numbers indicate the percentage of participants who assigned the scene to a particular category.

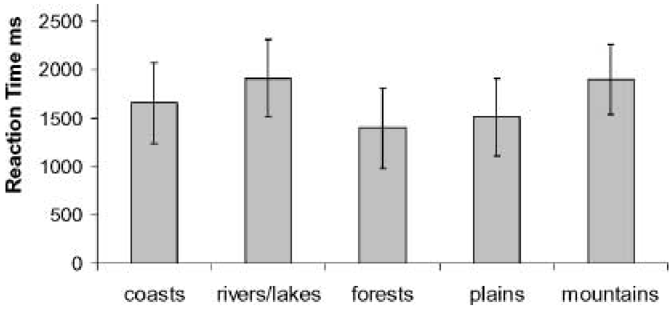

Fig. 3. Average correct response times broken up by category and averaged across participants. Error bars represent standard deviations.

overlapping categories [e.g., Birnbaum 1998]. Figure 2 shows examples of scenes that illustrate the varying degree of ambiguity.

Figure 3 shows response times of the scene images broken up by category. Reaction times longer than 6000 ms were discarded (2.6% of the data). Only correct response times are shown. A response was defined as correct if the selected category corresponds to ground truth. The latter is defined by the category label, which was selected by most participants for a given scene image.

In summary, experiment 1 showed that natural scene images are categorized quite consistently among different individuals, that natural scenes are overlapping categories, because several images have been assigned to more than one category, and that there are substantial variances in response times between different exemplars of each category.

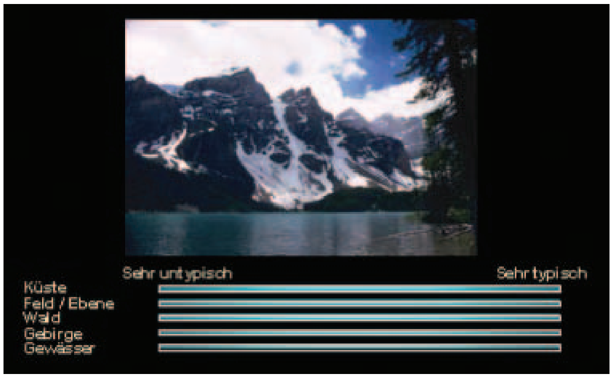

Fig. 4. Screenshot of experiment 2 (in German). Using the five bar sliders, participants had to rate the displayed scene image from "very atypical" ("sehr untypisch") to "very typical" ("sehr typisch") relative to each of the five categories ("Küste, Feld/ Ebene, Wald, Gebirge, Gewässer").

3. EXPERIMENT 2

In experiment 1, scenes were often classified into more than a single category. Moreover, scenes varied substantially regarding the response times for categorizations. This could be as a result of varying typicality. The aim of experiment 2 was to investigate this potential relationship and to examine whether different individuals consistently perceive the typicality of a scene. The latter is not only of theoretical importance, but also relevant for applications, such as, image-retrieval systems. The important question for an image-retrieval system is whether the *order* of typicality is consistently perceived among individuals. If this is not fulfilled, a time-consuming procedure in order to initially adapt an image-retrieval system to each individual is unavoidable. In contrast, if different individuals agree on the typicality of natural scenes, a general image-retrieval system would make sense, at least for initial filtering of images.

3.1 Method

3.1.1 *Participants.* Ten undergraduates of the University of Zurich participated in this experiment. None of them had participated in experiment 1. All had normal or corrected to normal vision.

3.1.2 *Materials and Procedure.* The same viewing distance, monitor, and scene images were used as in experiment 1. Scene presentation occurred in a random order with five bars labeled coasts, rivers/lakes, forests, plains, and mountains,[2] shown at the bottom of the screen (see Figure 4). The arrangement of the labels was counterbalanced across participants using a latin square. For each category, participants had to judge the typicality of the scene from 1 (very atypical) to 50 (very typical). Thus, for each of the 250 images, ratings relative to all five scene categories were obtained from each participant. After providing these five typicality ratings, participants could initiate the next trial by pressing the space bar. Scenes were displayed until the next trial was initiated. Prior to the experiment, all participants gave informed consent.

3.2 Results and Discussion

Figure 5 shows the histograms of the typicality ratings broken up by scene category (based on ground truth obtained in experiment 1). At least, for some scene categories, it appears that distributions were

[2] Again, the German names were used for the category labels.

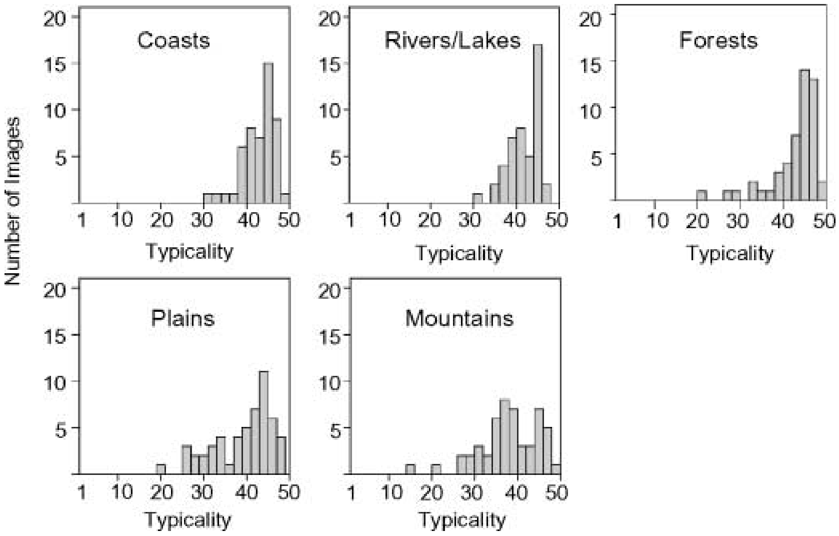

Fig. 5. Histograms of typicality ratings of the natural scenes broken up by category ground obtained in experiment 1.

Table II. Interrater Reliabilities for the Typicality Task in Experiment 2[a]

Interrater Reliabilities		
Scene Category	α	r_s
Coasts	0.981	0.693
Rivers/lakes	0.981	0.780
Forest	0.971	0.805
Plains	0.968	0.679
Mountains	0.943	0.645

All Spearman rank correlations (r_s) are significant with $p < 0.001$.

[a] Calculated for each category separately with the averaged Spearman rank correlation (r_s) and Cronbachs alpha (α) between participants.

skewed toward the upper bound. However, a one-sample Kolmogorov–Smirnov test (two-tailed) showed that none of the distributions are significantly different from a normal distribution (for forests $p = 0.09$; for all other categories $p >= 0.20$).

In Table II, measures of consistency between participants are summarized (see Appendix for the equations, references, and brief explanations of these statistics). Cronbach's alpha (α) values were between 0.943 (mountains) and 0.981 (coasts, rivers/lakes). Averaged rank correlation between participants was between 0.645 (mountains) and 0.805 (forests). These results show that there is a large agreement between participants concerning the perceived typicality of the scenes used in this experiment.

Another aim of experiment 2 was to examine whether typicality and categorization of natural scene images are related. To this end, we correlated categorization data (percentage agreeing categorizations and response times) from experiment 1 with typicality ratings from experiment 2 (averaged across participants for each scene image).

As can be seen in Table III (top row), categorization responses of experiment 1 were significantly correlated with typicality ratings obtained in experiment 2. The more typical a scene, the more often it was correctly categorized. The response times of categorization responses from experiment 1 were also

Table III. Correlations Between Categorization Data (Experiment 1) and Typicality Ratings (Experiment 2)[a]

Correlations Between the Categorization Experiment (Experiment 1) and Typicality Ratings (Experiment 2)					
	Coasts	Rivers/Lakes	Forest	Plains	Mountains
r % Agreeing Categorizations_typicality ratings	0.459**	0.535**	0.321*	0.766**	0.368**
r Reaction times_typicality ratings	−0.538**	−0.600**	−0.713**	−0.840**	−0.248

[a] Top row: Pearson correlations r between the categorizations (percent agreeing categorizations, experiment 1) and typicality ratings (experiment 2) for each category separately. Bottom row: Pearson correlations r between the reaction times (experiment 1) and typicality ratings (experiment 2) for each category separately.
*$p < 0.05$;** $p < 0.01$.

correlated with the typicality ratings of Experiment 2 (Table III, bottom row). As mentioned above, reaction times longer than 6000 ms were disregarded (2.6% of the data). Indeed, the more typical a scene, the faster it was categorized. The medium-to-large effect size of the correlations [according to Cohen, 1988] thus provide evidence for a relationship between categorization and typicality. This is consistent with earlier findings mentioned above, which showed that typical members are categorized more quickly and recalled more readily than atypical members [Rosch 1975]. Typicality can predict reaction times in sentence-verification tasks [McCloskey and Glucksberg 1979; Rosch 1973] and an overall high correlation between typicality and categorization has been found for many semantic categories [e.g., Hampton 1998].

4. EXPERIMENT 3

The computational modeling applied in experiments 3 and 4 is based on earlier work by Vogel and Schiele [2004]. We present here a short overview of the computational model and, especially, a psychophysically plausible distance measure.

Influenced by the studies of Mojsilovic et al. [2004] and through the analysis of the semantic similarities and dissimilarities of the images used in our studies, we determined nine local semantic concepts that are discriminant for the employed scene categories.

These local semantic concepts are sky, water, grass, trunks, foliage, field, rocks, flowers, and sand. Images are analyzed on a regular grid of 10 × 10 local regions, each of which is classified as one of the nine local concepts. Using these nine local semantic concepts, about 99.5% of the image regions can be consistently annotated [Vogel 2004]. A main advantage of this annotation method is that the local semantic concepts are not ambiguous and can easily be labeled consistently. A strong consensus across several observers is, thus, to be expected. Regions that contain two concepts in about equal amounts are doubly annotated.

In a second step, the local image information is summarized in a concept-occurrence vector (COV), which is essentially a histogram over the frequency of occurrence of the local semantic concepts (see Figure 6). Doubly annotated regions count 0.5 toward the COV. The COVs also be computed on several horizontally layered image areas (see Figure 7 for one, two, and three image areas) and, subsequently, concatenated. This allows evaluating whether concepts appearing at the top or bottom of an image are important for the image representation.

Each scene category is represented by the mean over the COVs of all images belonging to the respective category. This leads to a prototypical representation $\mathbf{p}^c$ of the scene categories, where the semantic concepts act as attributes and their occurrences as attribute scores. The typicality of a scene relative to a category c is computed by a distance measure that compares the scene representation $\mathbf{x}$ to the prototype $\mathbf{p}^c$ of the respective category:

$$\mathbf{d}^c(r) = \sum_{j=1}^{N(r)} w_j^f \left(x_j - p_j^c\right)^2$$

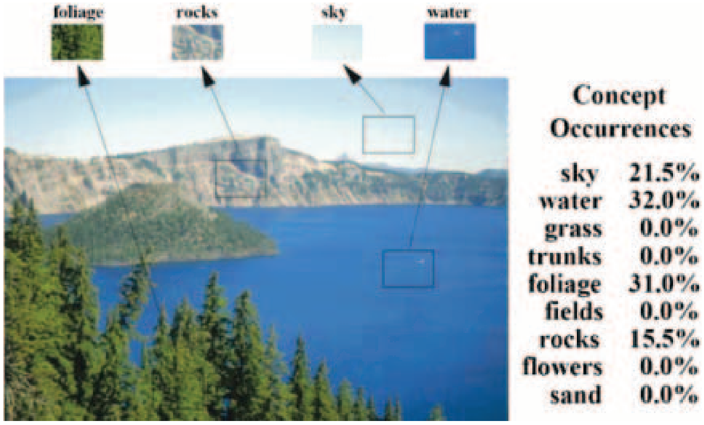

Fig. 6. Image analysis through concept-occurrence vectors.

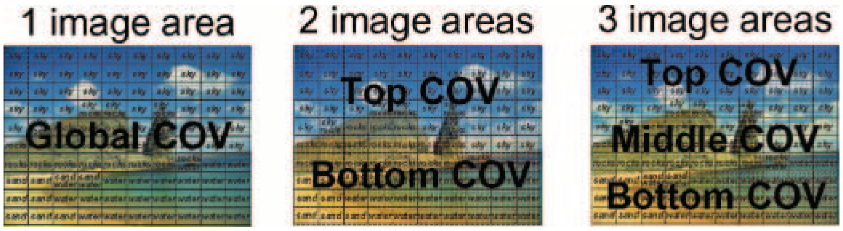

Fig. 7. Computation of one, two, and three COVs per image. The resulting vectors are subsequently concatenated and used for image representation.

Since only the typicality ranking is of interest, the images are ranked according to $\mathbf{d}^c(\mathrm{r})$. The w_i^c are a set of weights that model the relative importance of the semantic concepts in each category. The distance measure $\mathbf{d}^c(\mathrm{r})$ becomes perceptually meaningful, because the weight vector $\mathbf{w}^c$ is learned from the averaged human typicality scores that were obtained in experiment 2. In particular, we are solving a constrained minimization problem using Matlab's fmincon function where Spearman's ranking correlation $r_s(\text{typrating}_{\text{human}}, \text{typrating}_{\text{machine}})$ serves as the objective function for the optimization of the weights w_i^c.

4.1 Method

The goal of the computational experiments is to evaluate the capacity of the semantic image-retrieval system to rank natural scenes in a similar way as humans. The database consists of the same 250 Corel images used in the psychophysical experiments 1 and 2. As explained above, ground truth was obtained from the human categorization results of experiment 1 (50 coasts, 46 rivers/lakes, 50 forests, 53 plains and 51 mountains images). In order to obtain a maximally achievable benchmark, all 25,000 local regions (10 × 10 regions × 250 images) have been annotated manually with the mentioned nine semantic concepts. Since local image regions exhibit much less ambiguity than full images, this annotation can be consistently done by one observer [Vogel 2004]. All experiments have been fivefold cross-validated, meaning that in each round, 4/5 of each category has been used as training set for the computation of the prototype and the weight optimization. The remaining images were ranked using the learned prototype and the optimized weights and correlated with the corresponding human typicality ranks. The reported Spearman's rank correlation coefficient is the average over all cross-validation rounds.

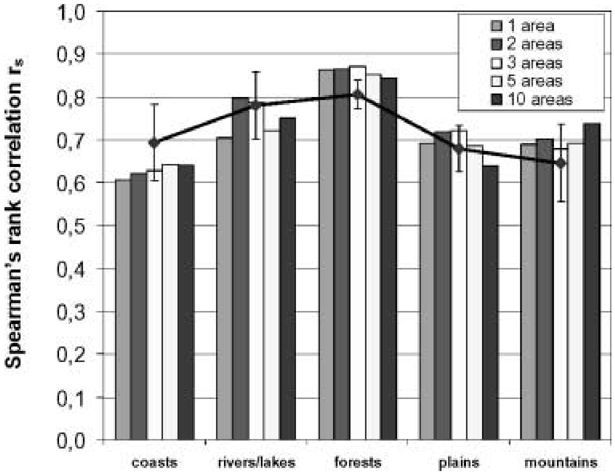

Fig. 8. Spearman's rank correlation between human and computational ranking: semantic modeling, prototype approach with perceptually plausible distance measure. Annotated images regions: image areas: 1, global image; 2, top/bottom; 3, top/middle/bottom 5, top/upper middle/middle/lower middle/bottom, 10, ten equally sized rows.

As visualized in Figure 7, the COVs was computed globally (1 image area) as well as on 2 (top/bottom), 3 (top/middle/bottom), 5 (top/upper middle/middle/lower middle/bottom), and 10 horizontally layered image areas, and subsequently concatenated. Using more than one image area includes spatial information in the image representation and might thus lead to better typicality ranking results.

4.2 Results and Discussion

Figure 8 shows the obtained correlation between the human typicality ranking and the machine typicality ranking using annotated image regions as input. Each group of bars belongs to one category. In each category, 1, 2, 3, 5, and 10, horizontally layered image areas have been tested. The black line displays the average inter individual correlation from the psychophysical experiment 2 (see Table II, column 2) and its standard deviation broken up by category. The machine ranking performs in all categories, except coasts, at least as similar to averaged human typicality rankings as the average correlation between the different human observers. Moreover, all categories lie inside the 1σ interval. In other words, the very consistent typicality judgments of the participants in experiment 2 can be modeled with our system. In addition, the achieved human–machine correlations follow closely the varying interindividual correlations when comparing different categories. That is, forestscenes are ranked very consistently by humans and also by the machine, whereas the performance for mountains is a little worse (lower average correlation and higher variance) both for humans and for the machine. The number of image areas seems to have varying influence on the ranking performance.

Figure 9 shows mean and standard deviation over the five cross-validation rounds of the weights learned during the optimization for one image area. The graphs visualizes that different semantic concepts are relevant depending on the scene category: For plains, grass, field, and, especially flowers, are most important. The weight of foliage is high when the goal is to rank relative to rivers/lakes. Sand can also help to detect "coast-ness." The partly high standard deviation of the weights can be explained by the small number of scenes that contain, e.g., sand or flowers. Note that the absolute value of the weights is meaningful only within a category, because each set of category weights is optimized independently.

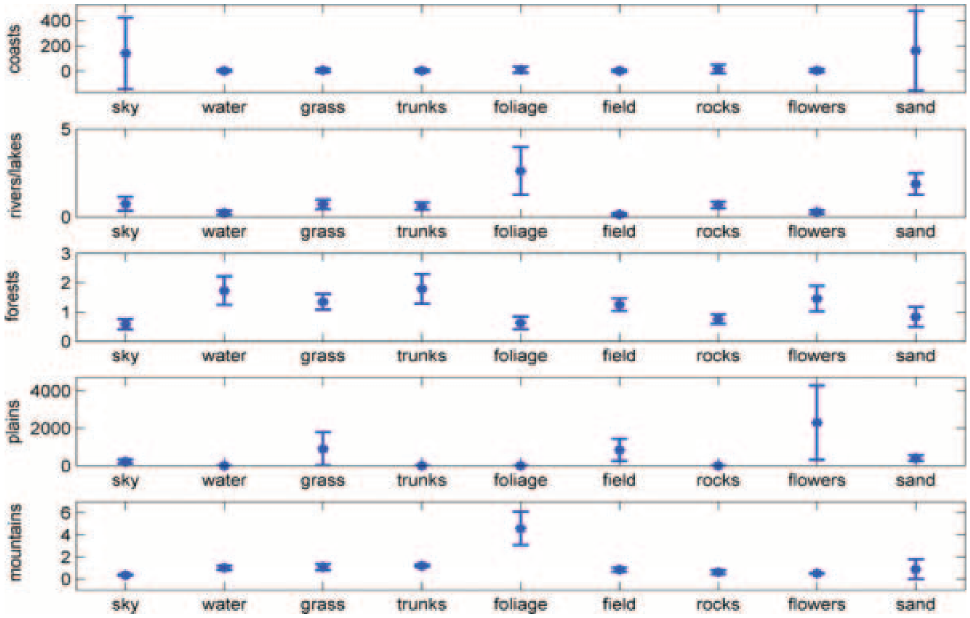

Fig. 9. Mean and standard deviation of the weights learned during optimization (1 image area).

5. EXPERIMENT 4

The tests with the manually annotated image regions serve as a benchmark for the maximum correlation performance that can be reached with our model. In a second computational experiment, we determined the correlation performance when using classified image regions as input. The employed concept classifier is a multiclass support-vector machine classifier [libSVM; Chang and Lin 2001] that is based on a combined color and texture feature. Around 61,000 singly annotated image regions were used for training and testing the concept classifier. We extracted a 84-bin HSI color histogram, a 72-bin edge direction histogram, and 24 features of the gray-level cooccurrence matrix [see Jain et al. 1995] from the image regions. The obtained feature vectors of length 180 served as input to a support-vector machine classifier (10-fold cross-validation, RBF kernel). The resulting classifier has an overall classification performance of 71.7%. Note that the employed region features are very similar to the global color features and the local edge features of Vailaya et al. [2001]. For details on the parameter selection and also on a comparison between two classifiers, refer to Vogel [2004].

5.1 Method

The same methods were used as in experiment 3 except for the fact that classified image regions were used instead of annotated image regions. The training, that is, the learning of the prototypes and the weight optimization, is based on the annotated image regions and, thus, the same as in experiment 3. The experiments have also been performed with a training phase based on the automatically classified image regions. This variant is a feasible alternative, but led to slightly inferior results and is, therefore, not discussed here. As before, all tests have been fivefold cross-validated.

5.2 Results and Discussion

The results are displayed in Figure 10. On average, the obtained correlations using classified regions are 0.1 below the correlations obtained with annotated regions. This is surprisingly robust given the

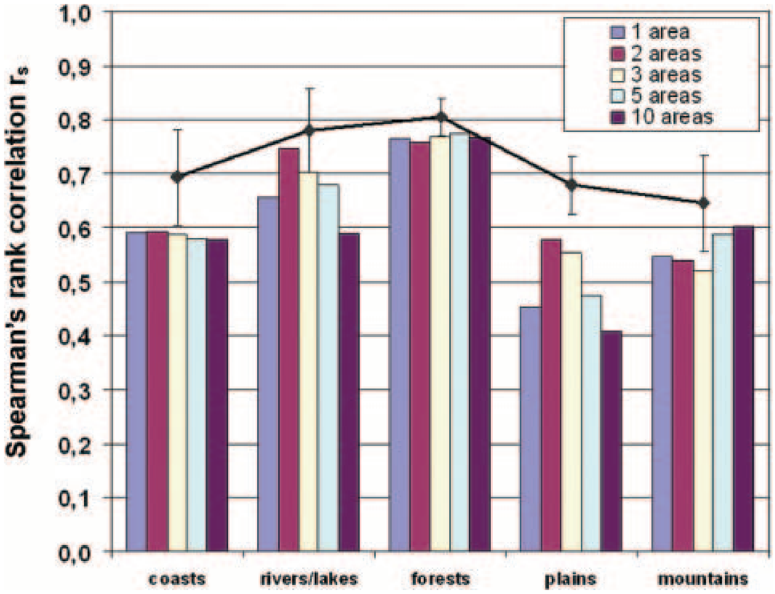

Fig. 10. Spearman's rank correlation between human and computational ranking: semantic modeling, prototype approach with perceptually plausible distance measure. Automatically classified images regions: image areas: 1, global image; 2, top/bottom; 3, top/middle/bottom; 5, top/upper middle/middle/lower middle/bottom; 10, ten equally sized rows.

fact that the concept classifiers are right only with a probability of 71.7%. Per category, the average loss is 0.05 for coasts and for rivers/lakes, 0.1 for forests, and 0.14 for plains and for mountains. In all cases, Spearman's rank correlation is close to the 1σ interval of the interindividual correlation. In addition, the correlations again follow the variations of the interindividual correlations between categories. This indicates that the computational model, the full image analysis, and typicality ranking system are, indeed, psychophysically plausible. Again, the number of image areas is not affecting the performance in a consistent way. The averaged variance of Spearman's rank correlation over the cross-validation rounds is small with 0.0026 for the annotated image regions and 0.0076 for the classified image regions.

Figure 11 Visually illustrates the typicality ranking performance of our system. The test set of one cross-validation round (50 images, about 10 per category) was picked randomly. The typicality ranking was performed once relative to each category using two image areas. Figure 11 displays, in the topmost row, exemplary coasts scenes and their rank relative to the coasts category. In the second row, rivers/lakes scenes ranked relative to the rivers/lakes category, etc. All rows visualize the typicality decrease from left to right. Usually from rank 10–12 onward, the image no longer unambiguously represents the particular category, but contains visual elements of one or more other categories. This is also consistent with the psychophysical results obtained in experiment 1 in which we found that also, in human perception, natural scenes are, indeed, overlapping categories. In addition, Figure 11 shows that the database, indeed, contains images at the category borders and that those images also assigned to be "medium typical" for the respective categories.

6. EXPERIMENT 5

In the remaining two experiments, the ranking performance of two other methods of image representation is compared to the ranking performance of the semantic modeling. In this section, we evaluate the ranking performance when directly employing extracted image features. Compared to the semantic

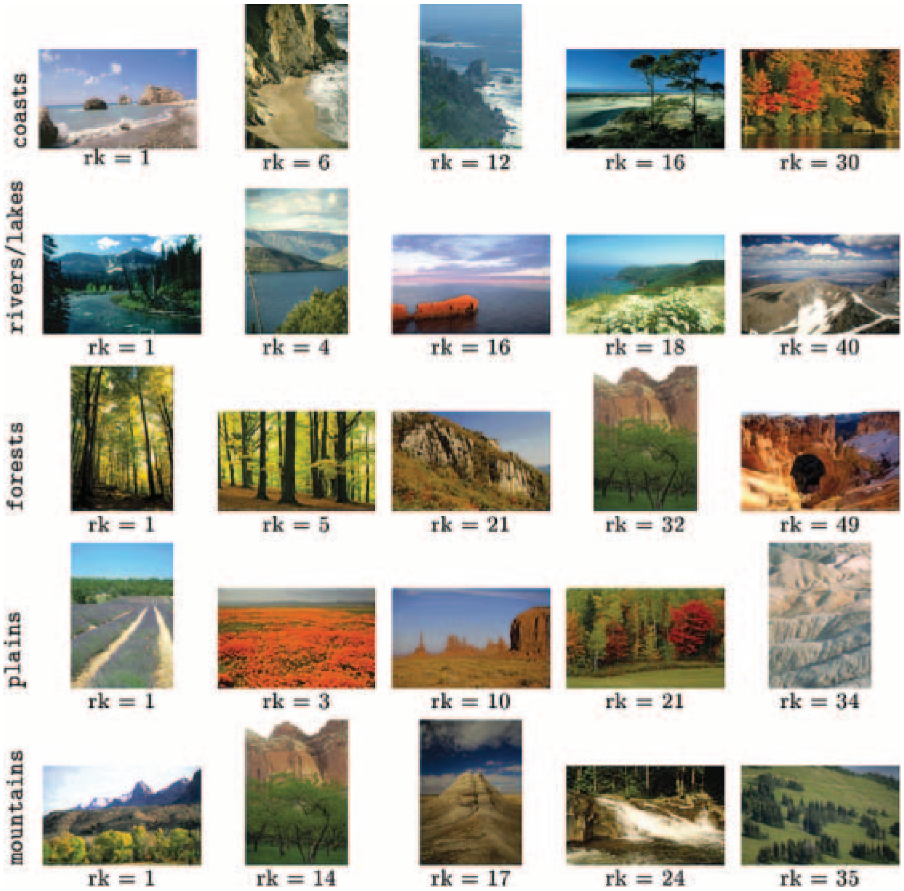

Fig. 11. Examples of ranked scenes (maximum rank = 50). From top to bottom row: coasts, rivers/lakes, forests, plains, mountains.

modeling, this approach does not contain an intermediate modeling step or any semantic annotation. This leads to a nonlocal image representation, especially in the case of only one image area.

For this experiment, we employed the same color and texture features that were used in experiment 4 for the classification of local image regions. The important difference compared to experiment 4 is that the features were extracted directly from the images without making use of the semantic or any other intermediate modeling step. As done before with the COVs, the direct features are also extracted on 1, 2, 3, 5, or 10 horizontally layered image areas. Note that when using only one image area, the extracted features are similar to the global color and edge features used in Vailaya et al. [2001]. The main difference is that Vailaya et al. [2001] use color and edge features for independent tasks, while we concatenate the resulting color and texture feature vectors.

6.1 Method

Similar methods were used as in experiment 3 and 4, except for the fact that the image representation consists of directly extracted color and texture features instead of the COV. The images are ranked according to their distance to the respective category prototype. Because of the length of the feature vector (180 bins) and the given amount of data (250 images), the weight optimization is not feasible and

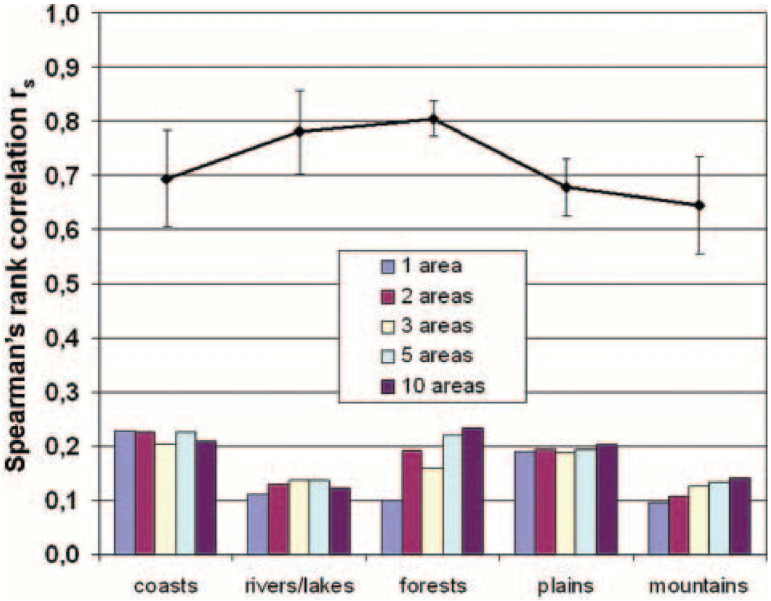

Fig. 12. Spearman's rank correlation between human and computational ranking: no semantic modeling step; features extracted directly from images. Prototype approach with Euclidean distance. Image areas: 1, global image; 2, top/bottom; 3, top/middle/bottom; 5, top/upper middle/middle/lower middle/bottom; 10, ten equally sized rows.

would have led to overfitting. The distance to the category prototypes was thus computed with weights $w_i^c = 1$, which corresponds to the Euclidean or sum-squared distance. The experiments have also been five fold cross-validated here.

6.2 Results and Discussion

The ranking performance based on the directly extracted low-level features is displayed in Figure 12. As before, the black line corresponds to the average interindividual correlation from experiment 2.

The results show a large drop in correlation between human and machine compared to experiments 3 and 4 (compare Figure 12 with Figures 8 and 10). One even could say that the machine ranking hardly correlates at all with the human ranking. These results clearly show that directly extracted low-level image features fail in catching the relevant details that are necessary for our (semantic) ranking task.

7. EXPERIMENT 6

In the final experiment, we reintroduce an intermediate modeling step, but do not make use of any semantic interpretation. The goal of the experiment is to compare the ranking performance of a local, but nonsemantic method to the performance of the semantic modeling employed in experiments 3 and 4, i.e., the classification of image regions into concept classes, such as, grass, water, and rocks.

As in experiment 4, the images are divided into 10×10 local image regions. From each region, a 84-bin HSI color histogram, a 72-bin edge direction histogram, and 24 features of the gray-level cooccurrence matrix [see Jain et al. 1995] are extracted. In contrast to experiment 4, the resulting feature vectors are now clustered using the k-means algorithm with cluster centers $k = 9, 18, 27, 45, 90$. The number of cluster centers corresponds to the varying length of the COV in experiments 3 and 4. The k-means algorithm was randomly initialized several times and the lowest overall distance to cluster centers was selected for further computation. After convergence of the k-means algorithm, all image regions were

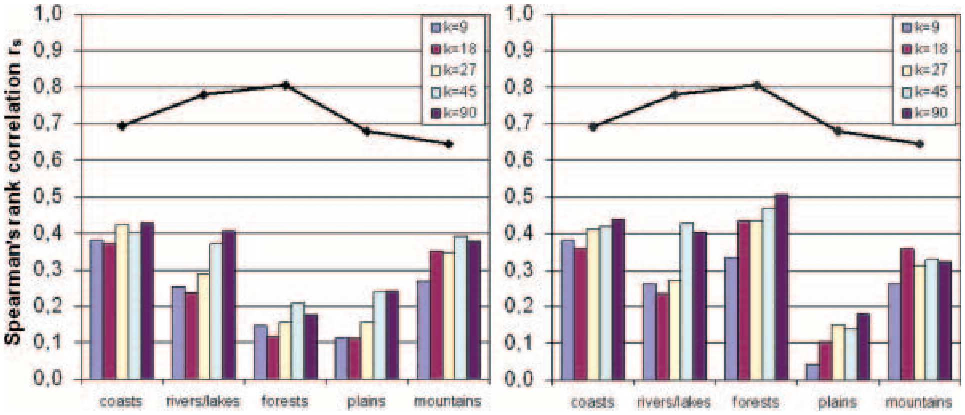

Fig. 13. Spearman's rank correlation between human and computational ranking: no semantic modeling and images represented through histograms over cluster memberships. Prototype approach with Euclidean distance (left) and with optimized weights (right). k, number of cluster centers.

assigned to one of the cluster center and a membership histogram for all regions within an image could be computed. This membership histogram was used in the same way as the COV in experiments 3 and 4 for image representation and image ranking.

As in the previous experiments, each scene category is represented by a category prototype obtained by computing the mean over the membership histograms belonging to the respective category. In contrast to experiment 5, weight optimization as described in experiment 3, has been conducted here.

7.1 Method

The same methods were used as in experiment 3, except for the fact that the COVs have been replaced by the cluster membership histograms. As before, the experiments have been fivefold cross-validated. The training sets were used for weight optimization and the computation of the prototypes.

7.2 Results and Discussion

The ranking performance using the membership histograms is displayed in Figure 13. The left plot shows the performance when using the Euclidean distance for computing the distance to the prototypes; the right plot shows the performance using optimized weights. When comparing Figure 13 with Figures 8 and 10, it becomes evident that an intermediate modeling step without any semantic information is not able to catch the relevant information for the ranking task. Even in the case of weight optimization, the ranking performance does not reach the performance of experiment 4. Note that the method is very similar to experiment 4, except for the feature clustering instead of feature classification in the first stage of the system. In particular, the lengths of the feature vectors for scene ranking are the same in both sets of experiments.

8. GENERAL DISCUSSION

The first aim of this study was to examine categorization and perceived typicality of natural scenes. Natural scenes constitute a very heterogeneous stimulus class. Some scenes are very easy to classify whereas others are quite ambiguous. This was found for human categorization (experiment 1) as well as for typicality ratings by humans (experiment 2) and by our computational model (experiments 3 and 4). For applications, such as image-retrieval systems, the important question is whether the difference in typicality is perceived consistently among different individuals. Only if this is the case, a general image-retrieval system can be developed. The results of the psychophysical experiments showed high interindividual consistency for both categorization (experiment 1) and perceived typicality (experiment 2). This not only means that our categories are perceptually plausible, but also that a general image-retrieval system makes sense, at least for initial filtering of images [for possible query methods, see Smeulders et al. 2000]. Interestingly, categorization responses were significantly correlated with typicality ratings. Typical scenes could be categorized easier and faster, which is consistent with earlier studies using other stimuli than natural scenes. As mentioned above, significant correlations between typicality and categorization have been found for many semantic categories. Moreover, Rosch [1975] reported that typical members are categorized more quickly than atypical ones, and it was found that typicality can predict reaction times in sentence-verification tasks [McCloskey and Glucksberg 1979; Rosch 1973].

The second objective of this study was to compare human perception with computational modeling. The computational model used is based on earlier work by Vogel and Schiele [2004] using an intermediate *semantic modeling step* by extracting local semantic concepts (sky, water, grass, trunks, foliage, field, rocks, flowers, and sand). The local image information is summarized in a COV, which is a normalized histogram over the frequency of occurrence of the local semantic concepts. Each scene category is represented prototypically where the semantic concepts act as attributes and their occurrences as attribute scores. For each category, the relative importance of the semantic concepts is modeled by a set of weights learned from the averaged human typicality scores that were obtained in experiment 2. This model correlated very well with human performance. Experiment 3 showed that machine ranking performs in all categories, except coasts, at least as similar to averaged human typicality rankings as the average correlation between the different human observers. Even more interesting is the fact that the achieved human–machine correlations closely follow the varying interindividual correlations when comparing different categories. More specifically, forest scenes are ranked very consistently by humans and machine, whereas for mountains a lower average correlation and higher variance was again found both for humans and machine. Converging evidence for these results were found in experiment 4 in which a computational model with classified image regions as input (based on a multiclass support-vector machine classifier; Chang and Lin [2001]) instead of annotated image patches was used.

We tested two other methods of image representation in experiments 5 and 6. In experiment 5, we evaluated the ranking performance of a model that uses much more information by employing directly extracted image features instead of using the semantic modeling step (which actually reduces the amount of available information drastically by using nine semantic concepts instead of all input pixels). This results in a nonlocal and nonsemantic image representation. Our semantic model outperformed the direct feature extraction by far, which provides further support for its psychophysical plausibility.

In experiment 6, region-wise image features were clustered and the distribution over cluster membership was used for image ranking. This results in a local, but nonsemantic image representation. In this case, our semantic modeling also outperforms the alternative image representation, even in the case of weight optimization. Thus, we can deduct that the key to a psychophysically plausible model for typicality ranking is the combination of a local, region-based plus a semantic image representation.

The large similarity between the performance of our computational model and human ratings can provide further insights into human typicality perception and categorization. The graded structure found in human and computational experiments is incompatible with the classical theory. This theory predicts that all members of a category are treated equally, because they all share the same set of necessary and sufficient features. For scenes, this is definitively not the case, neither for humans nor for our computational model. Our computational model can be related to the feature-frequency theory, since our concept occurrence can be interpreted as feature frequency. Note, however, that our concept occurrence is used for calculating the distance to the category prototype and no likelihood is calculated.

Our implementation of a category prototype in a multidimensional semantic concept space is also compatible with the prototype theory. Most importantly, the comparison of different computational models with human performance (experiments 3–6) provided converging evidence for the psychophysical plausibility of a semantic modeling step based on the extraction of local semantic concepts, such as rock, water, and sand, as the basis for typicality ranking of natural scenes.

APPENDIX

This appendix contains the equations, a brief explanation, and references of the statistics Cronbach's alpha, Phicorrelation, and Spearman's rank correlation.

Cronbach's Alpha

$$\alpha = (N * \overline{r})/(1 + (N - 1) * \overline{r})$$

This measure was first described by Cronbach (1951) and it is a coefficient for the internal consistency of a scale. This means that Cronbach's alpha measures how well a set of variables measures a single unidimensional latent construct. It can also be used to measure the consistency between participants, i.e., the interrater reliability between participants.

When measuring internal consistency of a scale, N is equal to the number of items. Because, in this study, Cronbach's alpha was used as an interrater reliability measure, N equals the number of participants. The R bar is the average interitem correlation among the items or participants, respectively.

Phi Correlation

The phi coefficient measures the degree of association between two *binary* variables. This measure is similar to the correlation coefficient in its interpretation [e.g., Kline, 2000].

A positive association between two binary variables is given if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal.

The formula for phi is

$$\phi = (ad - bc)/\sqrt{efgh}$$

Phi compares the product of the diagonal cells ($a * d$) to the product of the off-diagonal cells ($b * c$). The denominator is an adjustment that ensures that phi is always between -1 and $+1$.

	X^-	X^+	*Total*
Y^-	a	b	e
Y^+	c	d	f
Total	g	h	n

Spearman's Rank Correlation:

$$rs = 1 - \left(6 * \sum_{i=1}^{n} d_i^2\right) / (n * (n^2 - 1))$$

The Spearman's rank correlation is a nonparametric (distribution-free) rank statistic proposed by Spearman (1904) and measures the strength of the associations between two variables. It is used when the distribution of the data makes Pearson's correlation coefficient undesirable or misleading. Variable d is the difference between the ranks of corresponding values of X and Y and n is the number of pairs of values.

Pearson Correlation

The formula for Pearson correlation is:

$$r = \mathrm{cov}(x,\, y)/s_x * s_y$$

where

$$\mathrm{cov}(x,\, y) = \left(\sum_{i=1}^{n}(xi - \overline{x}) * (yi - \overline{y})\right)/n$$

Pearson correlation determines the extent to which two variables are related linearly to each other [e.g., Kline, 2000]. It assumes that the two variables are measured on at least interval scales. The value of correlation does not depend on the specific measurement units used. The x bar and y bar stand for the mean of the variables x and y, n is the number of cases, and $s_x$and s_y stands for the standard deviations of x and y.

ACKNOWLEDGMENTS

We would like to thank Stefan Ryf for valuable comments on an earlier version of the manuscript. Thanks also to three anonymous referees for helpful comments.

REFERENCES

Ashby, F. G. 1992. Multidimensional models of categorization. In *Multidimensional Models of Perception and Cognition,* F. G. Ashby, Ed. Lawrence Erlbaum Associates, Hillsdale, NJ, 449–483.

Ashby, F. G. and Gott, R. 1988. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition 14*, 33–53.

Ashby, F. G. and Lee, W. W. 1991. Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General 120*, 150–172.

Ashby, F. G. and Lee, W. W. 1992. On the relationship between identification, similarity, and categorization: Reply to Nosowsky and Smith (1992). *Journal of Experimental Psychology: General 121,* 385–393.

Ashby, F. G. and Maddox, W. T. 1990. Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance 16*, 598–612.

Ashby, F. G. and Maddox, W. T. 1992. Complex decision rules in categorization: Contrasting novice and experienced performance.*Journal of Experimental Psychology: Human Perception and Performance 18*, 50–71.

Ashby, F. G. and Maddox, W. T. 1993. Relations among prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology 37*, 372–400.

Ashby, F. G. and Townsend, J. T. 1986. Varieties of perceptual independence. *Psychological Review 93,* 154–179.

Barnard, K., Duygulu, P., De Freitas, N., and Forsyth, D. 2002. Object recognition as machine translation—Part 2: Exploiting image data-base clustering models. *European Conference on Computer Vision ECCV'02*, Copenhagen, Denmark.

Barsalou, L. W. 1987. The instability of graded structure: implications for the nature of concepts. In *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization,* U. Neisser, Ed. Cambridge University Press, Cambridge. 101–140.

Barsalou, L. W. and Sewell, D. R. 1985. Contrasting the representations of scripts and categories. *Journal of Memory and Language 24*, 646–665.

Birnbaum, M. H. 1998. *Measurement, Judgment, and Decision Making.* Academic Press, San Diego, CA.

Brooks, L. R. 1978. Nonanalytic concept formation and memory for instances. In *Cognition and Categorization,* E. Rosch and B. B. Lloyd, Eds. Lawrence Erlbaum Associates, Hillsdale, NJ.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. 1956. *A Study of Thinking.* Wiley, New York.

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

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., and Yianilos, P. N. 2000. The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing 9*, 1, 20–37.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Edelman, S. 1999. *Representation and Recognition in Vision*. MIT Press, Cambridge, MA.

Estes, W. K. 1986. Array models for category learning. *Cognitive Psychology 18,* 500–549.

Fei-Fei, L. and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. *IEEE Computational Vision and Pattern Recognition*, 524–531.

Feng, S. L., Manmatha, R., and Lavrenko, V. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Conference on Image and Video Retrieval CIVR '04*, Dublin, Ireland.

Feng, X., Fang, J., and Qiu, G. 2003. Color photo categorization using compressed histograms and support vector machines. In *International Conference on Image Processing ICIP'03,* Barcelona, Spain.

Franks, J. J. and Bransford, J. D. 1971. Abstraction of visual patterns. *Journal of Experimental Psychology 90,* 65–74.

Hampton, J. A. 1998. Similarity-based categorization and fuzziness of natural categories. *Cognition 65*, 137–165.

Hintzmann, D. L. 1986. "Schema abstraction" in a multiple-trace memory model. *Psychological Review 93,* 411–428.

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

Kalish, C. W. 2002. Essentialist to some degree: The structure of natural kind categories. *Memory & Cognition 30*, 340–352.

Kline, P. 2000. *Handbook of Psychological Testing.* London: Routledge.

Maddox, W. T. and Ashby, F. G. 1993. Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics 53*, 49–70.

Mccloskey, M. and Glucksberg, S. 1979. Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology 11*, 1–37.

Medin, D. L. and Schaffer, M. M. 1978. Context theory of classification learning. *Psychological Review 85*, 207–238.

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*, 79–107.

Mojsilovic, A., Kovacevic, J., Hu, J., Safranek, R. J., and Ganapathy, S. K. 2000a. Matching and retrieval based on the vocabulary and grammar of color patterns. *IEEE Transactions on Image Processing 9*, 38–54.

Mojsilovic, A., Kovacevic, J., Kall, D., Safranek, R. J., and Ganapathy, S. K 2000b. The vocabulary and grammar of color patterns. *IEEE Transactions on Image Processing 9*, 417–431.

Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*, 39–57.

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, 145–175.

Posner, M. I. and Keele, S. W. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology 77,* 353–363.

Posner, M. I. and Keele, S. W. 1970. Retention of abstract ideas. *Journal of Experimental Psychology 83,* 304–308.

Potter, M. C. 1976. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 5, 509–522.

Rips, L. J., Shoben, E. J., and Smith, E. E. 1973. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior 12*, 1–20.

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

Rosch, E. 1973. Natural categories. *Cognitive Psychology 4*, 3, 328–350.

Rosch, E. 1975. Cognitive reference points. *Cognitive Psychology 7*, 4, 532–547.

Rosch, E. 1977. Human categorization. In *Studies in Cross-Cultural Psychology,* N. Warren, Ed. Academic Press, London.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. 2000. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 12, 1349–1380.

Smith, E. and Medin, D. 1981. *Categories and Concepts*. Harvard University Press, Cambridge, MA.

Spearman, C. 1904. The proof and measurement of association between two things. *American Journal of Psychology 15*, 72–101.

Szummer, M. and Picard, R. W. 1998. Indoor-outdoor image classification. *Workshop on Content Based Access of Image and Video Databases,* Bombay, India.

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

Tversky, B. and Hemenway, K. 1983. Categories of environmental scenes. *Cognitive Psychology 15*, 121–149.

Ullman, S. 1996. *High-Level Vision*: *Object Recognition and Visual Cognition*. The MIT Press, Cambridge, MA.

Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H. J. 2001. Image classification for content-based indexing. *IEEE Transactions on Image Processing 10*, 1, 117–130.

VANRULLEN, R. AND THORPE, S. J. 2001. The time course of visual processing: From early perception to decision making. *Journal of Cognitive Neuroscience 24*, 454–461.

VELTKAMP, R. C. AND TANASE M. 2001. Content-based image retrieval systems: A survey. *Technical Report UU-CS-2000-34*, Department of Computer Science, Utrecht University.

VOGEL, J. 2004. *Semantic Scene Modeling and Retrieval*. Hartung-Gorre, Konstanz, Germany.

VOGEL, J. AND SCHIELE, B. 2004. Natural scene retrieval based on a semantic modeling step. *International Conference on Image and Video Retrieval CIVR 2004,* Dublin, Ireland. Springer Verlag, New York.

Received December 2005; Revised April 2006; Accepted May 2006