International Journal of Computer Vision 72(2), 133–157, 2007
© 2007 Springer Science + Business Media, LLC. Manufactured in the United States.
DOI: 10.1007/s11263-006-8614-1

Semantic Modeling of Natural Scenes for Content-Based Image Retrieval

JULIA VOGEL
Department of Computer Science, University of British Columbia, Vancouver, Canada
vogel@cs.ubc.ca

BERNT SCHIELE
Computer Science Department, Darmstadt University of Technology, Germany
schiele@mis.tu-darmstadt.de

Received April 21, 2005; Revised March 10, 2006; Accepted March 29, 2006

First online version published in July, 2006

Abstract. In this paper, we present a novel image representation that renders it possible to access natural scenes by local semantic description. Our work is motivated by the continuing effort in content-based image retrieval to extract and to model the semantic content of images. The basic idea of the semantic modeling is to classify local image regions into semantic concept classes such as water, rocks, or foliage. Images are represented through the frequency of occurrence of these local concepts. Through extensive experiments, we demonstrate that the image representation is well suited for modeling the semantic content of heterogenous scene categories, and thus for categorization and retrieval.

The image representation also allows us to rank natural scenes according to their *semantic* similarity relative to certain scene categories. Based on human ranking data, we learn a perceptually plausible distance measure that leads to a high correlation between the human and the automatically obtained typicality ranking. This result is especially valuable for content-based image retrieval where the goal is to present retrieval results in descending semantic similarity from the query.

Keywords: semantic scene understanding, content-based image retrieval, scene clasification, human scene preception, perceptually based techniques, computer vision

1. Introduction

Semantic understanding of scenes remains an important research challenge for the image and video retrieval community. Some even argue that there is an "urgent need" to gain access to the content of still images (Sebe et al., 2003). The reason is that techniques for organizing, indexing and retrieving digital image data are lagging behind the exponential growth of the amount of this data. The semantic gap between the image understanding of the user and the image representation of the computer still hampers fast progress in modeling high-level semantic content for image browsing and retrieval. Particularly, early retrieval systems have been based on the extraction of only low-level, often global pictorial features (for overviews see (Eakins and Graham, 1999; Rui et al., 1999; Smeulders et al., 2000; Veltkamp and Tanase, 2001)). Also in the early work on scene classification, semantics are often only found in the definition of the scene classes, e.g. indoor vs. outdoor, or waterfalls vs. mountains (Feng et al., 2003; Lipson et al., 1997; Maron and Ratan, 1998; Szummer and Picard, 1998; Vailaya et al., 2001).

Recently, several systems have been proposed that address a global as well as local image annotation (Barnard et al., 2002, 2003; Duygulu et al., 2002;

Feng et al., 2004; Lavrenko et al., 2003; Oliva and Torralba, 2001, 2002). In general, these approaches aim at learning the correspondence between global annotations and images or image regions, respectively, a promising trend in image understanding. Nevertheless, the fact that global annotations are more general than pure region naming, and consequently that a semantic correspondence between keywords and image regions does not necessarily exist, is often neglected. This is especially true for the correspondence between category labels and category members (Li and Wang, 2003).

Oliva et al. (1999) are among the first to bring a truly semantic component into the field of scene classification by proposing to organize images along three semantic axes. The semantic axes have been determined through psychophysical experiments. Also through a set of psychophysical experiments, Mojsilovic et al. (2004) obtain 20 semantic categories relevant for humans as well as verbal descriptions of these categories. These are extracted automatically and used for retrieval. Serrano et al. (2004) employ semantic features in addition to low-level features in order to increase indoor-outdoor classification performance. Boutell et al. (2004) propose a framework for describing natural scenes by multiple labels. The rationale is that most real-world categories are often not mutually exclusive, a point also being stressed in Section 4 of this paper. An additional way to access semantic image information is to automatically attach a set of manually selected, semantically meaningful labels to local image regions that can be searched for in a subsequent retrieval step (Kumar and Hebert, 2003; Minka and Picard, 1997; Picard and Minka, 1995; Town and Sinclair, 2000). However, the region labels are usually not combined to a global image representation that can be employed in content-based image retrieval.

In this paper, we propose a novel image representation that allows access to natural scenes by local semantic image description. Local regions descriptions are combined to a global image representation that can be used for scene categorization, retrieval, and ranking. In particular, we argue that hard-decision categorization is semantically not wise since most scenes consist of too complex semantic content for unambiguous categorization. Instead, scenes should be ranked according to their semantic similarity or typicality. This is especially true for content-based image retrieval where images are usually ranked according to their relevance for the query. We show that our semantic image representation renders it possible to rank nature scenes in a semantically meaningful way. In particular, the automatically obtained ranking correlates well with human rankings of the same scenes.

1.1. Ingredients to a Semantic Image Representation

Review of the relevant literature in the field of content-based image retrieval, image understanding, scene classification, and human visual perception suggest a set of requirements for a successful semantic image representation that are described in the following. The envisioned image representation shall be:

Semantic. The reduction of the semantic gap between the image representation of the human and the image representation of the machine is of prime importance. The ultimate goal is an image representation that is more intuitive for the user.

Descriptive. Image description is a highly intuitive means of communications for humans. Therefore, the goal is a vocabulary-supported access to images that replaces the common query-by-example paradigm with a query-by-keyword paradigm.

Region-Based. Natural scenes contain a large amount o semantic detail that can only be modeled by a region-based approach. This entails that the features are extracted from local image regions, and that the images are semantically annotated on a region level to supply the descriptive vocabulary for querying.

Segmentation-Free. Image segmentation algorithms such as the mean-shift algorithm (Comaniciu and Meer, 2002) or the NCuts algorithm (Shi and Malik, 1997) still lead to undesirable over- and undersegmentation of semantically contiguous regions. For that reason, in this paper, automatic image segmentation will be avoided and be substituted by a regular subdivision of the images.

Global from Local. In addition to the local image description, the goal is a global image representation based on local information. This global representation allows for a global, semantic comparison of scenes.

Inspired by Human Perception. The result of any image retrieval or image description system will be presented to a human user. It is therefore important to guide system design through knowledge about the human perception of natural scenes.

Evaluated Quantitatively. The proposed image representation has to be evaluated quantitatively, especially with respect to its semantic representativeness.

On the one hand, this refers to the evaluation concerning human perception as mentioned before. On the other hand, the goal is to assess the semantic applicability, the robustness, the strengths, and the weaknesses of the image representation through clearly defined and quantifiable tasks.

The last requirement is closely connected with the question of whether to employ supervised or unsupervised learning methods. The drawback of unsupervised or semi-supervised methods is that the extraction of semantics can be incidental. Also, the annotation accuracies are undesirably low as in approaches modeling word-region co-occurrences (Barnard et al., 2003; Lavrenko et al., 2003; Feng et al., 2004). For these reasons, this paper focuses on supervised learning methods and good image modeling performance in order to evaluate the proposed representation thoroughly. Certainly, the long term goal is to extent the supervised approach through semi-supervised or unsupervised learning methods.

The paper is organized as follows. In the following section, we propose an image representation, the semantic modeling, that addresses the above mentioned requirements. The semantic modeling is based on the extraction of local semantic concepts. The concept classifiers are introduced and discussed in Section 3. One field of application for the semantic modeling is scene categorization. In Section 4, several categorization approaches are compared and tested. The categorization results are analyzed with respect to their semantic applicability. Section 5 reveals that a semantic typicality ranking is preferable to a hard-decision categorization and that the proposed image representation is suitable for obtaining a semantically meaningful scene ordering. Section 6 summarizes the main results of a psychophysical study on the human ranking of our scenes. Using the human typicality rankings for evaluation purposes, we show in Section 7 that the automatic ranking based on the semantic modeling correlates well with the human ranking. In addition, we propose a psychophysically plausible distance measure that increases the correlation with the human data by another 15% relative to the correlation obtained with the Euclidean distance.

2. Semantic Modeling

As argued above, we aim for a region-based, that is local, semantic image description. For that reason, the image analysis proceeds in two stages. In the first stage, local image regions are classified by concept classifiers into semantic concept classes. In order not to be dependent on the largely varying quality of an automatic segmentation, the local image regions are extracted on a regular grid of 10 × 10 regions. Influenced by the psychophysical studies of Mojsilovic et al. (2004) and through the analysis of the semantic similarities and dissimilarities of the employed images, nine local semantic concept $s_i, i = 1 \ldots M, M = 9$ were determined as being discriminant for the desired retrieval tasks. These local semantic concepts are s = [*sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers*, *sand*]. With these nine semantic concepts, the database images can be annotated to 99.5%. On average only half an image region per image can not be assigned to one of the nine concept classes thus indirectly validating the choice of the local semantic concepts. Figure 1 depicts on the right an exemplary annotation of an image with the concepts *sky*, *water*, *rocks* and *sand*. Note that image regions that contain two semantic concepts in about equal amounts have been doubly annotated with both concepts. In Section 3, the concept annotation and classification is discussed in more detail.

In the second stage, the region-wise information of the concept classifiers is combined to a global image representation. For each local semantic concept, its frequency of occurrence is determined. This information enables us to make a global statement about the amount of a particular concept being present in the image, e.g. "This image contains 9% *sand*." In addition, the local image information is summarized in a semantics-based feature vector. The so-called concept-occurrence vector (COV) is essentially a normalized histogram of the concept occurrences in an image (see Fig. 1). The strength of the image representation using COVs is that these can also be computed on several overlapping or non-overlapping image regions thus increasing their descriptive content. This allows us to model information about which concepts appear at the top or bottom of an image. We obtain a semi-local, spatial image representation by computing and concatenating the COVs of $r = [1, 2, 3, 5, 10]$ horizontally-layered image regions resulting in a feature vector of length $N(r) = [9, 18, 27, 45, 90]$. When using $r = 2$ image areas (top-bottom), the COV of the image in Fig. 1 is $\boldsymbol{COV} = [44.5, 0, 0, 0, 0, 0, 5.5, 0, 0, 3, 23.5, 0, 0, 0, 0, 14.5, 0, 9]^T$.

The advantages of the semantic modeling are manifold. Only through the use of *named* concept classes in the first stage of the retrieval system, the semantic detail

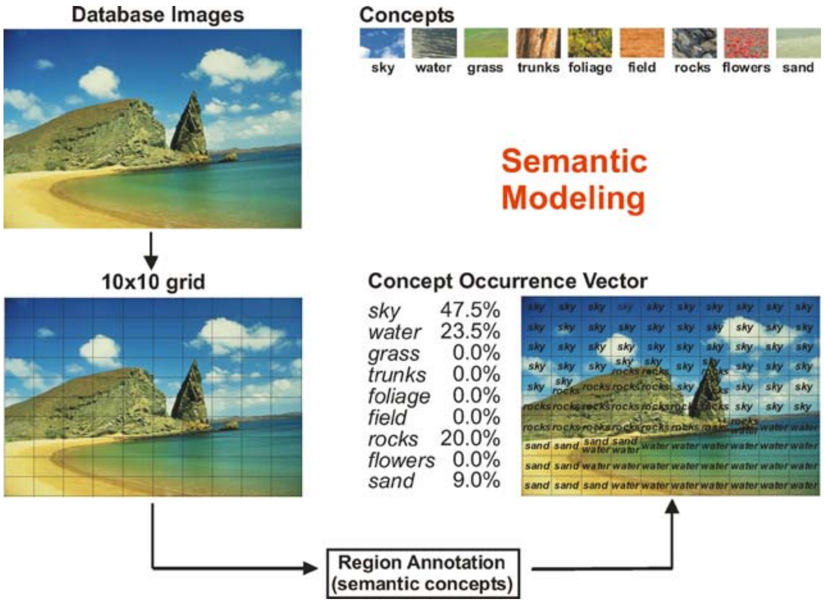

Figure 1. Image representation through semantic modeling.

of nature images can effectively be modeled and be used for description. In addition, the semantic content of the local image regions is far less complex than that of full images making the acquisition of ground-truth required for training and testing much easier. Since the local semantic concepts corresponds to "real-world" concepts, the method can also be used for descriptive image search. However, in the following, we will only relate to the global image representation through semantic modeling.

2.1. Selection of Scene Categories

The selection of semantic scene categories for the categorization and ranking experiments has been strongly influenced by work in psychophysics. In psychophysics, a hierarchical structure from very general to specific, e.g. animal, mammal, dog, German shepherd, has been suggested as a particularly important way of organizing objects or scenes. The most important level of such a taxonomy of categories is the basic level. It has been found through psychophysical experiments (especially Rosch (1978) and Rosch et al. (1976)) that the basic level, a middle level of specificity, is the most natural, preferred level when for example naming particular objects. Also, basic-level categories are easier and faster to learn.

Tversky and Hemenway were the first to construct a taxonomy of abstract categories such as environments or scenes (Tversky and Hemenway, 1983). Through psychophysical experiments, they reported in their seminal work `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`. The psychophysical experiments of Rogowitz et al. (1997) revealed two main axes in which humans sort photographic images: human vs. non-human and natural vs. artificial. Note here the different impact of employing categories that require a clear decision for one of the categories and axes that connect categories but allow a—in this case semantic—transition between two or more categories. These semantic axes were further extended in Mojsilovic et al. (2004) and resulted in the 20 mentioned scene categories.

We selected the non-human/natural coordinate as superordinate for our 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`, `mountains` and `sky/clouds`. A sample of images for each category can be seen in Fig. 2. The three columns of images on the left correspond to typical examples for each category illustrating the diversity of those categories. The right column of Fig. 2 shows images

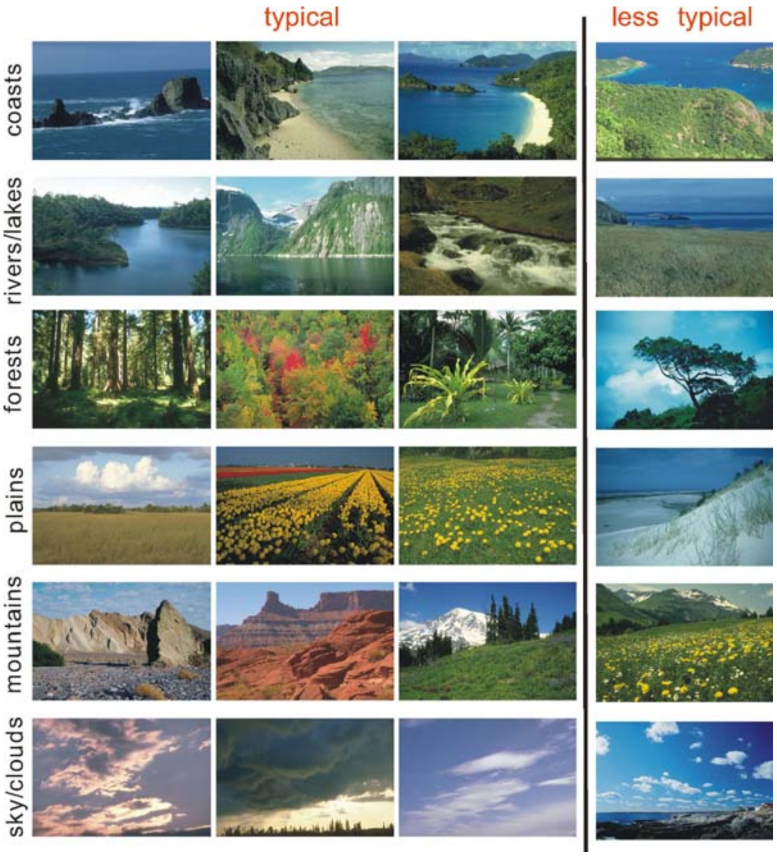

Figure 2. Exemplary images of each category. Three columns on the left: typical images. Rightmost column: less typical image.

which are far less typical but which are—arguably—still part of the respective category. Obviously, those examples are more difficult to classify and literally correspond to borderline cases.

3. Concept Classifiers

The purpose of the concept classifiers is the semantic classification of local image regions. The image regions are extracted on a regular grid of 10 × 10 regions with size 48 × 72 or 72 × 48 pixels (see Fig. 1). 700 images of nature scenes, that is 70’000 local image regions (700 images * 100 regions), have been annotated with the nine semantic concepts *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*. The visual diversity of the resulting concept classes is illustrated in Fig. 3. *Sky* is clearly not always just blue, but also overcast or partly cloudy, or *foliage* includes leaves at many scales and in many seasonal color ranges. The figure illustrates that, without any context, some of the displayed image regions are very hard to classify even for humans.

In order to be robust to “unclean” image regions that are due to the fixed grid segmentation, e.g. *water*-regions with a tiny bit of *sand*, regions containing up to 25% of a different concept were accepted both as training and testing data. Image regions that contain two semantic concepts in about equal amounts have been doubly annotated with both concepts. These regions are not used for training or testing of the concept classifiers. As a result, 60’718 singly annotated image regions are available for training and testing of the concept classifiers. However, for the categorization and ranking experiments in the following sections, all unseen image regions are classified with the trained concept classifiers. The expectation is that doubly annotated image regions are assigned to either of the two classes which is equally good.

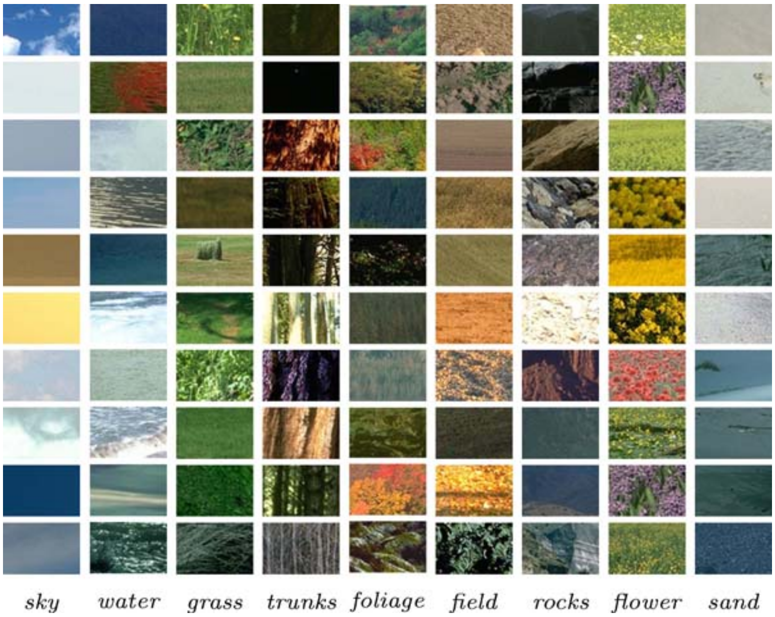

Figure 3. Semantic concept classes.

Since not all concepts are present in all images, the class sizes vary from 1’625 up to 15’296 image regions thus posing quite a challenge for the training of the concept classifiers (see Table 1). *Sky* appears in nearly every image whereas e.g. *sand* is only present in certain `coasts` or `plains` scenes.

3.1. Features

The development of the concept classifiers was not the main focus of this work. The goal is rather to evaluate the strength of the image representation based on semantic modeling. Nevertheless, several standard low-level color and texture features have been tested and evaluated. The features and the feature parameters such as the number of histogram bins have been determined in extensive pre-tests and are not further discussed here. For more details, please refer to Vogel (2004).

Table 1. Sizes of concept classes.

Concept class	# image regions	
sky	25.2%	15,296
water	12.0%	7,293
grass	5.8%	3,503
trunks	2.7%	1,625
foliage	22.5%	13,709
fields	6.9%	4,188
rocks	18.6%	11,310
flowers	3.4%	2,049
sand	2.9%	1,745
OVERALL	100%	60,718

The best concept classification results have been obtained with the concatenation of a 84-bin linear HSI color histogram (hue: 36 bins, saturation: 32 bins, intensity: 16 bins), a 72-bin edge direction histogram, and the 24 features of the gray-level co-occurrence matrix (32 gray levels): contrast, energy, entropy, homogeneity, inverse difference moment and correlation for the displacements $\overrightarrow{1,0}$, $\overrightarrow{1,1}$, $\overrightarrow{0,1}$ and $\overrightarrow{-1,1}$ (Jain et al., 1995). During concatenation, each histogram has first been normalized such that the feature types have about equal weights.

3.2. Classifier

The best classification results have been obtained with a support vector machine (SVM) classifier. For a comparison between a k-nearest neighbor classifier and a SVM classifier also refer to Vogel (2004). For the experiments, the LIBSVM package (Chang and Lin, 2001) was employed. The package offers an efficient multi-class support using internally a one-against-one approach Hsu and Lin (2002). Thus, with $M = 9$ classes, there are $\frac{M(M-1)}{2} = 36$ single classifiers. A new data point is tested by each of the 36 classifiers with the "winning" class obtaining a vote. The data point is allocated to the class that has the highest number of votes.

All experiments have been performed with 10-fold cross-validation on *image* level. That is, regions from the same image are either in the test or in the training set but never in different sets. This is important since image regions of a particular semantic concept tend to be more similar to other (for example neighboring) regions in the same image than to regions in other images.

3.3. Results and Discussion

The SVM concept classification results in an overall classification accuracy of 71.7%. Table 2 displays the corresponding confusion matrix.

The most apparent behavior of the classifier is the fact that the confusion matrix shows a strong correlation between the class size and the classification result. *Sky*, *foliage*, and *rocks* are the largest classes (Table 1), and are classified with the highest class-wise accuracies. *Sand*, *trunks*, and *flowers* are the smallest classes, and also have the smallest classification accuracies. In addition to the pure dependency on the class size, the classification confusions show that members of smaller classes are often confused with the *semantically most similar* larger class. That is, *grass* and *flowers* are almost exclusively confused with *foliage* but not vice versa. Similarly, *field* and *sand* are frequently confused with *rocks*, but also not vice versa.

Table 2. Confusion matrix of the SVM concept classification ($C = 128$, $\gamma = 0.03125$).

Overall	Classifications in %								
71.7%	*s*	*w*	*g*	*t*	*f*	*f*	*r*	*f*	*s*
True class									
sky	**95.3**	2.6	0.0	0.0	0.2	0.2	1.3	0.0	0.5
water	11.1	**66.1**	2.6	0.0	4.7	2.7	11.4	0.1	1.6
grass	0.3	6.1	**43.1**	0.7	37.5	4.9	5.7	1.2	0.5
trunks	0.6	0.6	0.6	**27.5**	38.8	3.6	27.8	0.4	0.2
foliage	0.4	1.5	2.4	1.5	**81.1**	1.0	10.9	1.1	0.0
field	0.6	6.9	6.2	1.3	17.0	**37.7**	27.6	0.3	2.3
rocks	3.1	5.1	0.3	0.9	13.6	4.2	**71.5**	0.4	0.9
flowers	0.7	1.2	2.7	1.5	43.9	4.6	2.7	**42.3**	0.5
sand	10.3	16.1	1.8	0.3	0.9	9.9	30.6	0.0	**30.2**
Precision	90.9	70.7	62.3	51.6	67.0	54.2	62.2	76.5	55.7

The fact that it is more difficult to classify small concept classes is also the main argument against using more semantic concepts. Tests with additional semantic concepts such as *snow* for snowy mountains, or *mountains* for mountains in the far background that can not be assigned to either *rocks* or *foliage* did not result in higher classification rates. On the contrary, subsequent scene categorizations based on ten or eleven instead of nine semantic concepts achieved lower accuracy. On the other hand, a reduction of the number of semantic concepts by not using small classes such as *sand* or *trunks* also resulted in a degraded categorization. These smaller classes provide the necessary detail information for discriminating between semantically similar categories.

Obviously, the obtainable classification accuracies depend strongly on the consistency and accuracy of the manual annotations. Although best care and attention was directed to that problem, annotation ambiguities are hardly avoidable. For example, the annotation of *rocks* and *foliage* is quite challenging. Imagine an image with rocky and forested hills in far distance: is it rather *rocks* or *foliage*? For that reason, it is not surprising that *rocks* or *foliage* are confused in both directions. Another major confusion appears between *trunks* and *foliage*. This results mainly from the fact that each *trunks* region contains also a fair amount of leaves whereas most *foliage* regions also include some branches or parts of trunks.

In order to improve the classification rate, a semantic hierarchical classification approach was also tested. The idea is to subsume all those classes that are often confused and to train SVM classifiers for only three or four classes resulting in higher accuracies. In a first step, the image regions were classified into the classes *sky*, *water*, *plants* and *ground*. In a second step, the *plants*-regions were further split into *foliage*, *trunks*, *flowers* and *grass* and the *ground*-regions into *sand*, *fields* and *rocks*. This two-level hierarchy did not result in a substantial classification improvement, and was for that reason not tested further.

Another idea often proposed for boosting the classification accuracy is the use of different or "better" low-level features. Putting effort into features, feature selection, and feature combination methods could improve classification. However, a better classification accuracy would not change the main conclusions of the categorization and ranking tasks, and was thus not the focus of our research.

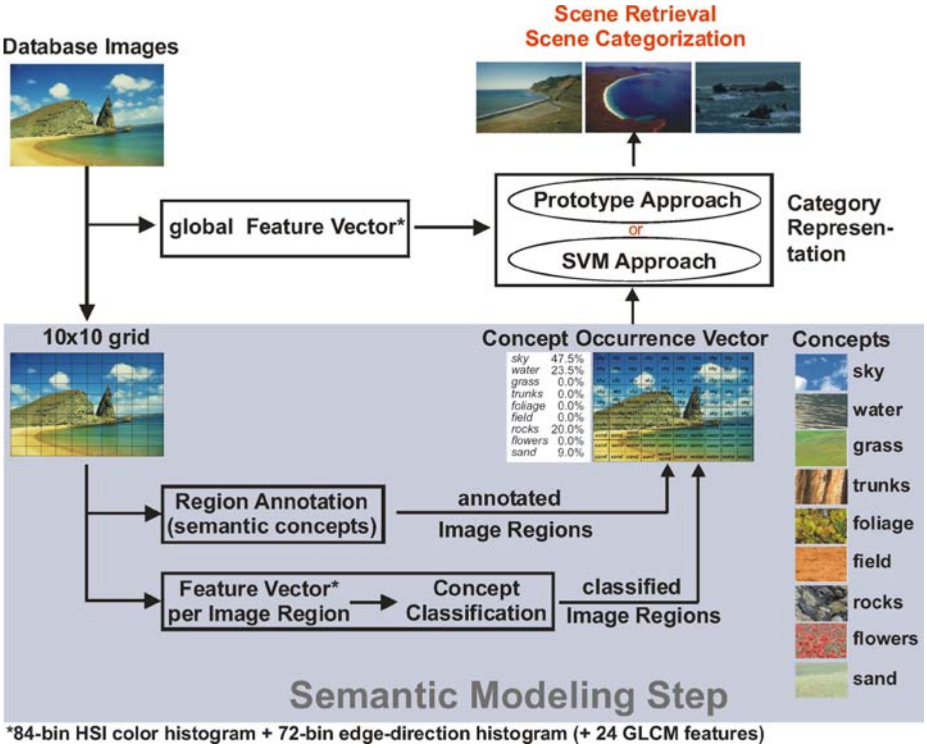

Figure 4. Overview of scene categorization approaches.

4. Scene Categorization

Scene categorization is a special case of image retrieval where the query corresponds to the scene category being searched for. Since scenes, that is full images, contain very complex semantic details, scene categorization is an appropriate task for testing the semantic representativeness of the proposed image representation. In this section, the task is hard-decision categorization whereas in the following sections, the goal is semantic scene ranking evaluated relative to human ranking data.

In the following, three pairs of conditions for the categorization task are presented and discussed. The information flow is summarized in Fig. 4. First, the semantic modeling is compared to direct feature extraction. As visualized in the lower part of Fig. 4, the concept-occurrence vectors can be obtained using annotated image regions as well as classified image regions. In the second test, the categorization performance based on annotated image regions is compared to the performance based on classified image regions. The categorization based on annotated image regions serves as benchmark: Which is the maximum performance to expect with the method? Finally, we compare two approaches to representing categories: a representative approach using category prototypes and a discriminative approach using Support Vector Machines (SVMs).

4.1. *Representative Approach: Category Prototypes*

The relevance of prototypes for categorization has been discussed in detail in the psychophysics community (Murphy, 2002). A category prototype is an example which is most typical for the respective category, even though the prototype does not necessarily have to be an existing category member. The prototype theory claims that humans represent categories by prototypes and judge the category membership of a new item by calculating the similarity to that prototype. Rosch and Mervis (1975) propose that a category prototype is not a single best example for the category but rather a summary representation. This summary representation is a list of weighted attributes through which the categorymembership can be determined. Thus, important attributes that might determine the category membership by themselves have high weights. But having various less important attributes with lower weights also renders an item a category member.

The image representation through concept-occurrence vectors provides a representation that is very close to the above mentioned attribute list. Each image is described by the frequency of occurrence of an semantic concept (see Section 2). It is thus straightforward to define the category prototype in the scene categorization task as the average COV over all members of a category.

$$\boldsymbol{p}^c = \frac{1}{N_c} \sum_{j=1}^{N_c} \boldsymbol{COV}(j) \tag{1}$$

where c refers to one of the six categories and N_c to the number of images in that category.

Figure 5 displays the category prototypes and the standard deviations for each category using one image area (for an explanation of the image areas refer to Section 2). The figure reveals which semantic concepts are especially discriminant. For example, `forests` are characterized through a large amount of *foliage* and *trunks*. In contrast, `mountains` can be differentiated when a large amount of *rocks* is detected. The attributes of the prototype hold the information about which amount of a certain concept is typically present in an image of a particular scene category. For example, a `rivers/lakes`-image usually does not contain any *sand*. Therefore, the *sand*-attribute of the `rivers/lakes`-prototype is close to zero.

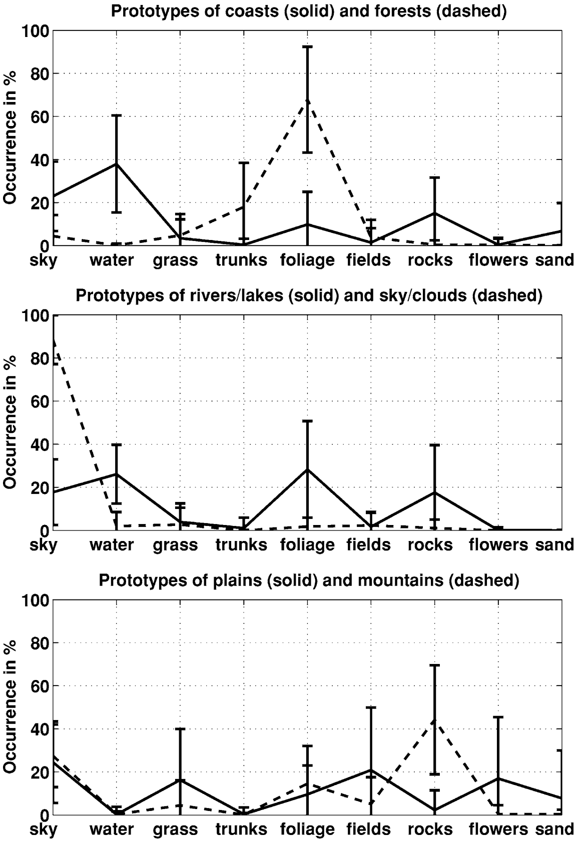

Figure 5. Prototypes and standard deviations of the six scene categories.

The distance of an image to the prototypical representation is determined by computing the sum-squared distance (SSD) between the COV of the image and the prototype. This corresponds to an unweighted attribute list. In the next section, when addressing the typicality of images, we also discuss the introduction of attribute weights. An image is assigned to the category it has the smallest distance to.

4.2. *Discriminative Approach: Multi-Class SVM*

A discriminative and thus very different approach to scene categorization is the use of SVMs. SVMs have been widely used in recent years and have been shown to be capable tools for classification and categorization (Joachims, 2002; Wang and Zhang, 2001).

For the same reasons as in Section 3, that is the efficient multi-class implementation, we employ the LIBSVM package (Chang and Lin, 2001) for the SVM-based categorization experiments. LIBSVM implements a one-against-one multi-class scheme that results in 15 two-class SVMs for the six scene categories. When determining the category of an unseen image, the COV of the image is tested by each of the 15 classifiers. Each “winning” category obtains a vote and the image is assigned to the category with the largest number of votes.

4.3. *Categorization Experiments*

The goals of the experiments were to evaluate the Prototype vs. the SVM approach, to compare the categorization performance when using annotated vs. classified image regions, and to determine if the semantic modeling step is in fact useful. The categorization performance is primarily evaluated via the overall categorization accuracy, but also via the confusion

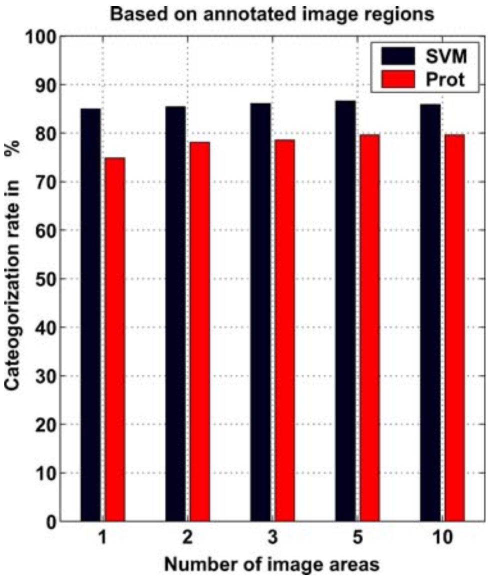

Figure 6. Categorization rates vs. Image Areas—Based on annotated image regions. The *y*-axis shows the categorization rate and the *x*-axis the number of employed image areas: 1 ↔ global image, 2 ↔ top/bottom, 3 ↔ top/middle/bottom, 5 ↔ top/upper middle/middle/lower middle/bottom and 10 ↔ ten equally sized rows.

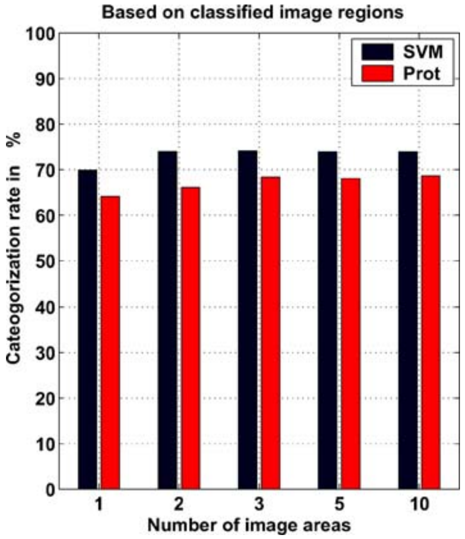

Figure 7. Categorization rates vs. Image Areas—Based on classified image regions. The *y*-axis shows the categorization rate and the *x*-axis the number of employed image areas: 1 ↔ global image, 2 ↔ top/bottom, 3 ↔ top/middle/bottom, 5 ↔ top/upper middle/middle/lower middle/bottom and 10 ↔ ten equally sized rows.

matrix and the rank statistics. All experiments are 10-fold cross-validated. Parameters were selected such that the performance on average, that is over all 10 cross-validation rounds is maximized. The images for each cross-validation round were selected randomly with the constraint that about an equal amount of images of each category is present in each training set.

Figures 6 to 8, and Tables 3 to 7 summarize the categorization results for all experiments. The following sections discuss those results in detail.

Table 3. Categorization accuracies (in %) based on annotated image regions—SVM Approach, 5 image areas.

(a) Confusion matrix

	coa	r/l	for	pla	mou	s/c
coasts	**80.3**	14.1	0.7	3.5	0.7	0.7
rivers/lakes	18.0	**73.0**	3.6	0.9	3.6	0.9
forests	0.0	1.9	**95.1**	1.9	1.0	0.0
plains	0.8	0.0	0.8	**91.6**	5.3	1.5
mountains	0.6	2.2	0.6	6.7	**89.4**	0.6
sky/clouds	0.0	0.0	0.0	5.9	0.0	**94.1**

(b) Rank statistics

	1	2	3	4	5	6
coasts	**80.3**	97.1	99.3	99.3	100.0	100.0
rivers/lakes	**73.0**	95.5	96.4	99.1	100.0	100.0
forests	**95.1**	98.1	99.0	100.0	100.0	100.0
plains	**91.6**	98.5	98.5	100.0	100.0	100.0
mountains	**89.4**	98.3	98.9	100.0	100.0	100.0
sky/clouds	**94.1**	100.0	100.0	100.0	100.0	100.0
OVERALL	**86.4**	97.7	98.6	99.7	100.0	100.0

Table 4. Categorization accuracies (in %) based on annotated image regions—Prototype Approach, 10 image areas.

(a) Confusion matrix

	coa	r/l	for	pla	mou	s/c
coasts	**64.8**	14.8	4.2	9.2	6.3	0.7
rivers/lakes	18.9	**55.9**	10.8	2.7	10.8	0.9
forests	0.0	0.0	**96.1**	2.9	0.0	1.0
plains	1.5	0.0	4.6	**89.3**	1.5	3.1
mountains	0.0	2.2	2.8	7.8	**85.5**	1.7
sky/clouds	0.0	0.0	0.0	5.9	0.0	**100.0**

(b) Rank statistics

	1	2	3	4	5	6
coasts	**64.8**	90.8	99.3	100.0	100.0	100.0
rivers/lakes	**55.9**	88.3	98.2	100.0	100.0	100.0
forests	**96.1**	99.9	99.0	99.0	99.0	100.0
plains	**89.3**	97.7	98.5	100.0	100.0	100.0
mountains	**85.5**	93.9	96.6	97.8	100.0	100.0
sky/clouds	**100.0**	100.0	100.0	100.0	100.0	100.0
OVERALL	**79.6**	94.1	98.3	99.3	99.9	100.0

Table 5. Categorization accuracies (in %) based on classified image regions—SVM Approach, 3 image areas.

(a) Confusion matrix

	coa	r/l	for	pla	mou	s/c
coasts	**71.1**	12.0	0.7	6.3	9.2	0.7
rivers/lakes	28.8	**42.3**	6.3	4.5	17.1	0.0
forests	1.0	2.9	**89.3**	3.9	2.9	0.0
mountains	4.6	0.8	5.3	**71.0**	17.6	0.8
plains	3.9	3.4	0.0	5.0	**86.6**	1.1
sky/clouds	8.8	0.0	0.0	0.0	0.0	**91.2**

(b) Rank statistics

	1	2	3	4	5	6
coasts	**71.1**	87.3	96.5	99.3	100.0	100.0
rivers/lakes	**42.3**	82.0	93.7	98.2	99.1	100.0
forests	**89.3**	95.1	96.1	99.0	100.0	100.0
mountains	**71.0**	87.8	97.7	100.0	100.0	100.0
plains	**86.6**	95.5	98.9	98.9	100.0	100.0
sky/clouds	**91.2**	97.1	100.0	100.0	100.0	100.0
OVERALL	**74.1**	90.3	97.0	99.1	99.9	100.0

Table 6. Categorization accuracies (in %) based on classified image regions—Prototype Approach, 10 image areas.

(a) Confusion matrix

	coa	r/l	for	pla	mou	s/c
coasts	**64.1**	12.7	3.5	4.2	13.4	2.1
rivers/lakes	18.0	**35.1**	10.8	2.7	29.7	3.6
forests	0.0	1.0	**94.2**	0.0	3.9	1.0
mountains	10.7	0.8	16.0	**48.1**	22.1	2.3
plains	2.8	1.1	2.8	5.0	**87.2**	1.1
sky/clouds	2.9	0.0	0.0	0.0	0.0	**97.1**

(b) Rank statistics

	1	2	3	4	5	6
coasts	**64.1**	88.7	99.3	100.0	100.0	100.0
rivers/lakes	**35.1**	78.4	96.4	99.1	100.0	100.0
forests	**94.2**	95.1	97.1	98.0	99.0	100.0
mountains	**48.1**	64.9	84.0	98.5	100.0	100.0
plains	**87.2**	91.6	94.4	98.9	100.0	100.0
sky/clouds	**97.1**	97.1	97.1	100.0	100.0	100.0
OVERALL	**68.4**	84.7	94.3	99.0	99.9	100.0

Table 7. Categorization accuracies (in %) without semantic modeling step—SVM Approach, 10 image areas.

(a) Confusion matrix

	coa	r/l	for	pla	mou	s/c
coasts	**54.2**	13.4	2.8	12.0	15.5	2.1
rivers/lakes	15.3	**45.9**	9.0	5.4	22.5	1.8
forests	1.0	4.9	**81.6**	2.9	9.7	0.0
plains	10.7	1.5	6.9	**61.1**	18.3	1.5
mountains	6.7	7.8	1.1	8.9	**74.3**	1.1
sky/clouds	8.8	2.9	0.0	0.0	0.0	**88.2**

(b) Rank statistics

	1	2	3	4	5	6
coasts	**54.2**	71.8	95.8	98.6	100.0	100.0
rivers/lakes	**45.9**	74.8	89.2	97.3	100.0	100.0
forests	**81.6**	86.4	92.2	96.1	100.0	100.0
plains	**61.1**	86.3	90.8	97.0	99.2	100.0
mountains	**74.3**	91.1	96.6	99.4	100.0	100.0
sky/clouds	**88.2**	91.2	94.1	100.0	100.0	100.0
OVERALL	**65.0**	83.0	93.4	98.0	99.9	100.0

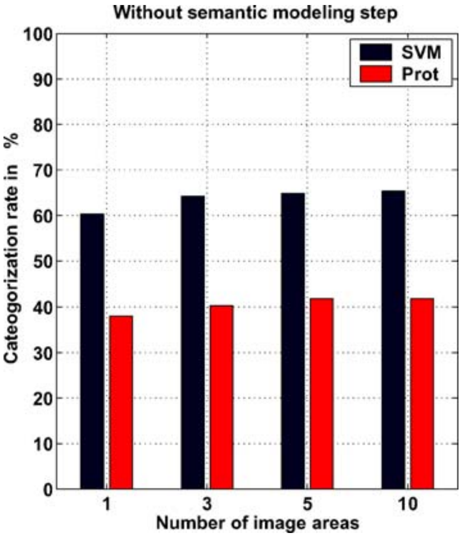

Figure 8. Categorization rates vs. Image Areas—Without semantic modeling step. The y-axis shows the categorization rate and the x-axis the number of employed image areas: 1 ↔ global image, 2 ↔ top/bottom, 3 ↔ top/middle/bottom, 5 ↔ top/upper middle/middle/lower middle/bottom and 10 ↔ ten equally sized rows.

4.3.1. Categorization Based on Annotated Image Regions. The experiments based on annotated image regions serve as benchmark: assuming that the manual annotations are consistent, the experiments reveal what is the best performance of the semantic modeling that can be expected. The result for 1, 2, 3, 5, and 10 image areas is depicted in Fig. 6. The plot suggests quite clearly that the SVM approach outperforms the Prototype approach, and that an increase in the number of image areas leads to a slight improvement in the case of the SVM approach and a larger improvement in the case of the Prototype approach.

Table 3(a) displays the confusion matrix of the best SVM categorization. The best SVM performance of 86.4% categorization rate is clearly better than the best Prototype performance of 79.6% categorization rate (see confusion matrix in Table 4(a)). Table 3(a) shows that especially the two categories `coasts` and `rivers/lakes` are frequently confused. Tables 3(b) and 4(b) display the rank statistics of the categorization for the best SVM and the best Prototype performance. The rank statistics allows us to analyze how close to each other the first and the second, third, etc. best candidate are in the feature space. The large jump from 86.4% to 97.7% in Table 3(b) shows that a large percentage of images is indeed quite close in the COV space. The same is true for the Prototype approach as displayed in Table 4(b). The finding suggests that the corresponding images are also semantically hard to categorize. This is especially true for the `coasts` and the `rivers/lakes` category. Visual inspection of the mis-categorized images shows that those images are hard to categorize unambiguously even for humans.

4.3.2. Categorization Based on Classified Image Regions. The next set of tests regard the categorization of automatically classified image regions. The image regions were classified as described in Section 3. As mentioned in the previous section, the concept classifiers were only trained and tested on singly annotated image regions. However, for the categorization experiments, all 100 image regions had to be classified. In the case of doubly annotated image regions, a classification into either of the two classes is acceptable. The experiments showed that about 65% of the 12.6% doubly annotated regions were classified into one of the two classes. This classification rate is thus in the same range as for the singly annotated regions, and acceptable as input for the categorization.

Figure 7 displays the categorization results based on classified image regions. A higher number of image areas tends to result in a higher categorization rate, although 3, 5, and 10 image areas often perform very similarly. As with the annotated image regions, the SVM approach clearly outperforms the Prototype approach. The gain in categorization accuracy relative to the Prototype approach is up to 7%. The highest categorization rate (74.1%) is obtained with the SVM approach using 3 image areas. The corresponding confusion matrix and rank statistics are displayed in Table 5. The best Prototype categorization results in 68.7% categorization accuracy as shown in Table 6. The confusion matrix and the rank statistics show that also based on classified image regions the `coasts` and the `rivers/lakes`-category are often confused, and that the performance jump between the first and the second best is for those two categories among the highest. In addition, images of the `plains`-category are often confused with `mountains`.

4.3.3. Categorization Without Semantic Modeling. The final set of experiments was conducted in order to determine whether the data reduction from low-level image features to more semantic concept-occurrence vectors is in fact a wise step or whether it harms the achieved categorization accuracies. For that reason, the same low-level feature vectors that have been used for the concept classification were extracted directly from the image. The results of these experiments are depicted in Fig. 8.

Besides the observations that have been made before (SVM approach better than prototype approach, and more image areas better tha fewer image areas), the diagram clearly shows that the categorization without semantic modeling performs worse than with the use of concept-occurrence vectors. Even the best result without semantic modeling (65.0% with the SVM approach, 10 image areas) is below the worst result in the previous section (70.1%, 1 image area). The confusion matrix and the rank statistics of the best categorization are printed in Table 7. When not employing semantic modeling, there are confusions between all classes. Before, confusions appeared mainly between semantically similar classes. The performance jump between the first and the second best is still large, but does not reach the same level as with the semantic modeling.

These results clearly show that the semantic modeling step leads to a meaningful image representation. Although the dimensionality is reduced by a factor of 20, the categorization performance is better by nearly 10% (compare Tables 5 and 7).

4.4. Categorization Experiments—Discussion

One of the goals in the previous section was to evaluate the performance of the discriminative, that is the SVM Approach, vs. the representative, or the Prototype Approach. The experiments give a nearly unanimous answer to that question: The SVM Approach outperforms

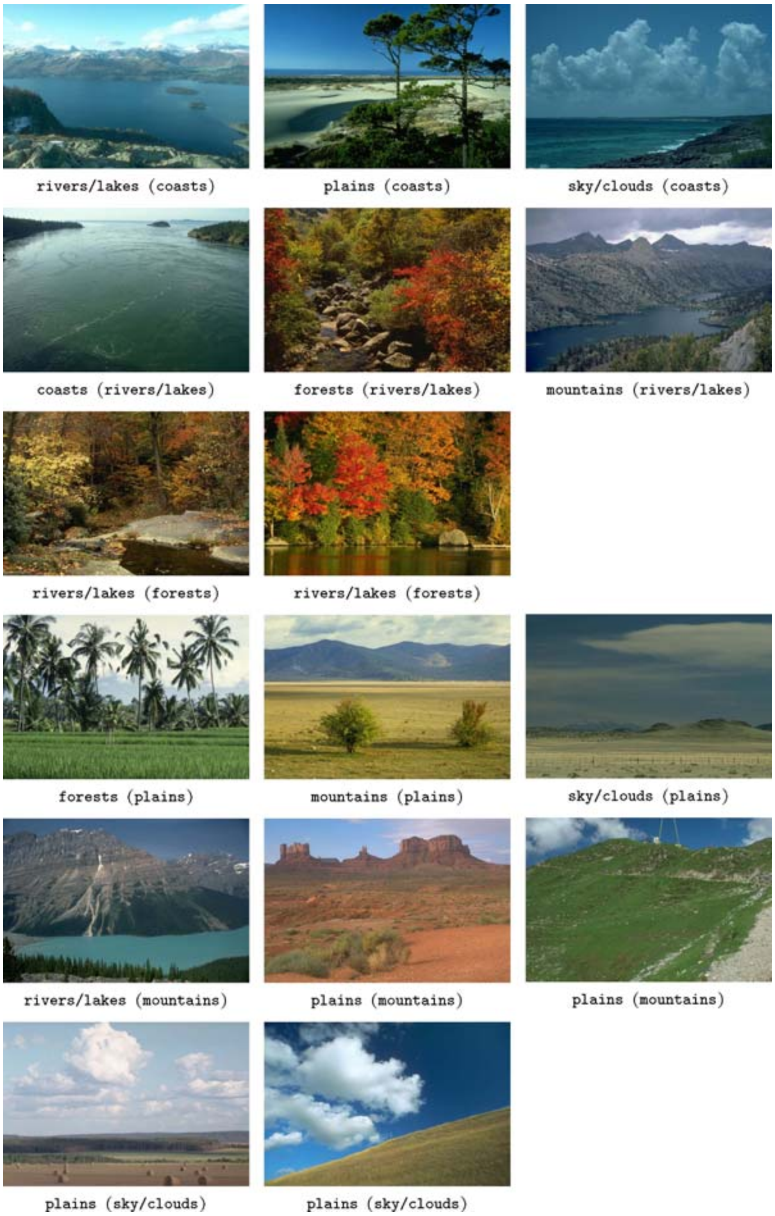

Figure 9. Examples of mis-categorized scenes in each category, "correct" category in parentheses (SVM approach on annotated image regions).

the Prototype approach. Based on annotated image regions, the performance increases by 6.8% when using SVMs, based on classified image regions, it increases by 5.4%, and without semantic modeling step even by 10.9% compared to using prototypes.

Furthermore, the results show that the semantic modeling approach leads to higher categorization rates. The best categorization based on classified image regions has an accuracy of 74.1%, whereas the best categorization without semantic modeling reaches only 65.0% categorization rate. Thus, besides being a means to semantically describe natural scenes, and besides leading to a dimensionality reduction by a factor of 20, the semantic modeling also achieves higher categorization rates.

Obviously, the categorization performance based on classified image regions is lower than the benchmark based on annotated image regions. A closer analysis of the confusion matrix in Table 5 and the confusion matrix of the concept classification in Table 2 shows that the categorization performance is strongly correlated with the performance of the concept classifier that is most discriminant for the particular categories. Three of the six categories have been categorized with high accuracy: `forests`, `mountains` and `sky/clouds`. The main reason is that *sky*, *foliage* and *rocks* have been classified with high accuracy and thus lead to a good categorization of `forests`, `mountains` and `sky/clouds`. Critical for the categorization especially of the category `plains` is the classification of *fields*. Since *fields* is frequently confused with either *foliage* or *rocks*, `plains` is sometimes mis-categorized as `forests` or `mountains`. Another semantic concept that is critical for the categorization is *water*. If too much *water* is misclassified, `rivers/lakes` images are confused with `forests` or `mountains` depending on the amount of *foliage* and *rocks* in the image. If too much *water* has incorrectly been detected in `rivers/lakes` images, they are confused with `coasts`.

Experiments using more semantic concepts than in the current system decrease the categorization performance. As mentioned in Section 3, the use of more semantic concepts leads to even smaller classes and, due to the size of the classes, a low classification accuracy. A low classification accuracy in turn leads to a lower categorization accuracy. We experimented with ten and eleven instead of nine local semantic concepts and did not observe a categorization improvement. However, also the decrease to seven or eight semantic concepts did not improve categorization. The smaller semantic concept classes are necessary for a final discrimination between visually similar categories such as *sand* for the differentiation between `rivers/lakes` and `coasts`.

In addition to the categorization rate, we also analyzed the rank statistics. With the SVM approach, an image is allocated to the category with the maximum number of votes and with the Prototype approach, an image is allocated to the category with the smallest distance to the prototype. In Tables 3(b), 5(b), and 7(b), the categorization rates are displayed when using the best as well as the second best, third best, etc. for categorization. The result is surprising: When accepting also the second best for the categorization, the overall categorization rate jumps on average by 16.8%. When analyzing the number of votes or distances, respectively, of the second best vs. the best image, it appears that these values are actually quite close. Does this mean that the images are also semantically very close to both, that is the best and the second best category? Fig. 9 shows for each category exemplary images where the second best is actually the "correct" category. The "correct" category is written in parentheses. One might argue that the person that did the annotation of the images did a good job or not. But the goal is not to model the opinion of one single annotating person. In fact, the images show that there is sometimes no "correct" or "incorrect" answer when categorizing images. How much *foliage* makes a `rivers/lakes`-image to a `forests`-image and vice versa? How far must a mountain be so that the image moves from the `mountains`-category to the `plains`-category.

The conclusion we draw from these observations is that pure hard-decision categorization should not be the goal. Rather, some sort of semantic typicality ranking as discussed in psychophysics (Murphy, 2002) and aimed for in content-based image retrieval should be performed.

5. Semantic Typicality Transition Between Categories

With the goal to obtain a semantic ordering of scenes, we performed an additional experiment. Using the Prototype approach with annotated image regions, three pairs of prototypes were selected: `rivers/lakes` and `forests`, `forests` and `mountains`, `mountains` and `rivers/lakes`. The sum-squared distance of all database images to the selected prototypes was computed, and those images were chosen that lie inside a constrained region between two prototypes. With the intention to obtain a normalized distance between two prototypes, the concept-occurrence vector of the selected images was projected onto the connecting line between two prototypes. Thus, the normalized distance relative to prototype 1 is $D_{\text{prototype1}} = \frac{d_1}{d}$ (see Fig. 10). Using this normalized distance measure, the database images were sorted semantically "between" two prototypes.

Figures 11 to 13 show the exemplary sorting result for the three prototype pairs. In each figure, the reference prototype is on the left and the normalized distance D is displayed below the images. The figures illustrate that with the concept-occurrence vectors of the semantic modeling, a *semantic* ordering of natural scenes can

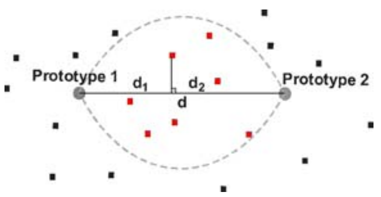

Figure 10. Normalized distance computation for scene ordering.

be obtained. From left to right in Fig. 11, the concepts that are typical for a `rivers/lakes`, that is mainly water, decrease whereas the typical `forest` items, that is foliage, greenery, or trunks increase. The same happens in Fig. 12, where the scenes change from a `forests` via forested mountains to rocky mountains. In Fig. 13, the transition goes back from `mountains` to `rivers/lakes`. Note here the difference in the transition with respect to Fig. 11. In Fig. 13, the intermediate scenes are waterscapes with a mountain in the background whereas in Fig. 11, the border of the lake or river is forest. This illustrates that we are indeed able to separate these scenes semantically.

6. Human Ranking of Natural Scenes

In general, the sorting of the images in Figures 11 to 13 seems to make sense semantically. But do humans on average really agree with the ranking? In order to study the human ranking of nature scenes, two psychophysical experiments were conducted employing a subset (250 images) of the images used in the previous sections. The psychophysical experiments were carried out in collaboration with Dr. Adrian Schwaninger and Dr. Franziska Hofer from the Visual Cognition Group at the Psychology Department, University of Zurich, Zurich, Switzerland.

The goal of the experiments was to gain insights into the human perception of nature scenes. In the first experiment, subjects were asked to categorize the displayed scenes into one of the five scene categories `coasts`, `rivers/lakes`, `forests`, `plains`, and `mountains`. The experiment thus provided ground-truth for the succeeding experiments: Each scene was assigned to the category the majority of subjects selected. In the second experiment, a different set of subjects was asked to rate the typicality of the displayed scenes

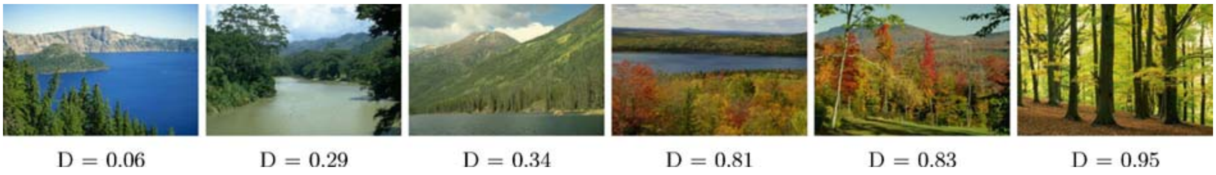

Figure 11. Transition from `rivers/lakes` to `forests` with normalized typicality value.

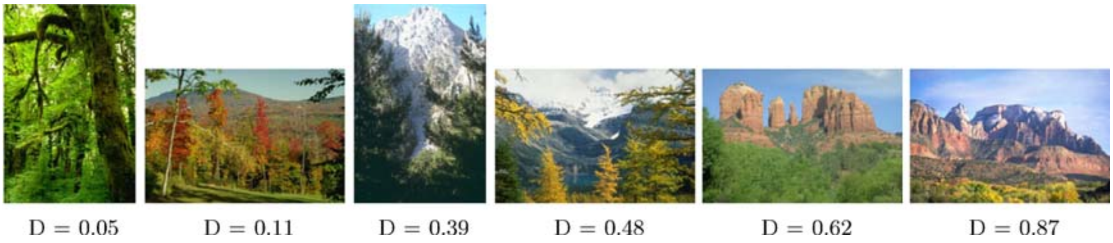

Figure 12. Transition from `forests` to `mountain` with normalized typicality value.

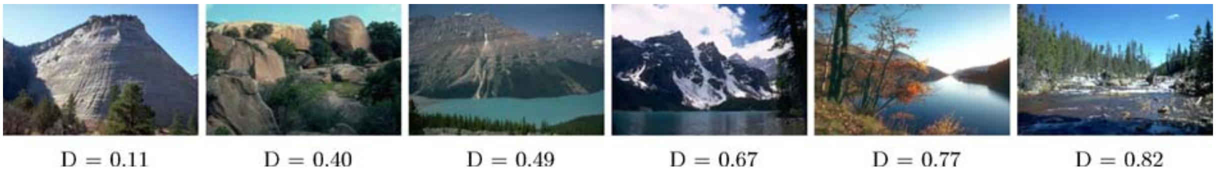

Figure 13. Transition from `mountain` to `rivers/lakes` with normalized typicality value.

relative to each of the five categories using bar sliders from 1 (very atypical) to 50 (very typical).

A main finding of the experiments was that participants were very consistent in their responses. The consistency of the typicality judgments can be measured using the averaged Spearman's rank correlation between participants. Since the absolute ranking score is meaningless for the task, Spearman's rank correlation has to be employed instead of the correlation coefficient (Bortz, 1999). Spearman's rank correlation is equivalent to the correlation coefficient when variables are in ordinal scale as is the case with rank orders. Special care has to be taken when ties are present, i.e. when multiple items in ranking A have the same rank in ranking B, or vice versa. Spearman's rank correlation with ties is computed using the following equation ($M =$ number of data points, $d_j =$ rank difference between the two compared rankings, $k(A); k(B) =$ number of tied rank groups in ranking A or B, $t_l =$ number of ties in each rank group of ranking A, $u_l =$ number of ties in each rank group of ranking (B):

$$r_s = \frac{2 \cdot \left(\frac{M^3-M}{12}\right) - T - U - \sum_{j=1}^{M} d_j^2}{2 \cdot \sqrt{\left(\frac{M^3-M}{12} - T\right) \cdot \left(\frac{M^3-M}{12} - U\right)}}, \qquad (2)$$

where

$$T = \frac{1}{12} \sum_{l=1}^{k(A)} \left(t_l^3 - t_l\right), \qquad (3)$$

$$U = \frac{1}{12} \sum_{l=1}^{k(B)} \left(u_l^3 - u_l\right). \qquad (4)$$

In our case, Spearman's rank correlation of two typicality rankings has been computed for each combination of two participants and averaged. Table 8 shows the average inter-rater correlation of the second experiment. The averaged rank correlation between participants between 0.65 (`mountains`) and 0.81 (`forests`) is very high for these kind of psychophysical experiments (see e.g. Kline (2000)). In addition, all results are significant. This indicates that there is a large agreement between participants concerning the typicality ranking of the scenes used in the experiment. This result is especially interesting as the first experiment confirmed that the database also contains ambiguous images that cannot be undoubtedly assigned to only one category. These images are at medium typicality for several categories and, on average, the subject agreed about their ratings.

Table 8. Spearman's rank correlation r_s for evaluating the inter-rater reliabilities in the typicality rating experiment.

	r_s
`coasts`	0.69
`rivers/lakes`	0.78
`forests`	0.81
`plains`	0.68
`mountains`	0.65

The second experiment provided us with a mean typicality rating between 1 (very atypical) and 50 (very typical) for each of the 250 images with respect to each of the five categories. This typicality rating allows us to rank the scenes in descending typicality relative to each of the five categories. This information will be used in the following section to evaluate the automatically obtained typicality rankings based on semantic modeling. For further information concerning the setup, the evaluation, and the discussion of the psychophysical experiments please refer to Schwaninger et al.

7. Perceptually Plausible Ranking of Scenes

The topic of this section is the automatic ranking of natural scenes relative to five of our scene categories (`coasts`, `rivers/lakes`, `forests`, `plains`, and `mountains`). The automatically obtained ranking is compared to human rankings using the results of our psychophysical experiment 2.

A prototype approach is most appropriate when aiming for a similarity ranking. For that reason, we employ the Prototype approach of Section 4.1 with two different distance measures for obtaining a semantic typicality ranking. In the next section, the sum-squared distance (SSD) is used for ranking scenes. In Section 7.2, a psychophysically plausible distance measure is learned and analyzed.

7.1. Typicality Ranking using Prototypes and the SSD

As described in Section 4.1, each scene category may be represented by the mean over the concept-occurrence vectors $\boldsymbol{COV}$ of all images belonging to the respective category. This leads to a prototypical representation $\boldsymbol{p}^c$ of the scene categories where the semantic concepts s_i act as attributes and their occurrences as attribute scores. The typicality of a scene relative to a category c is computed by the sum-squared distance

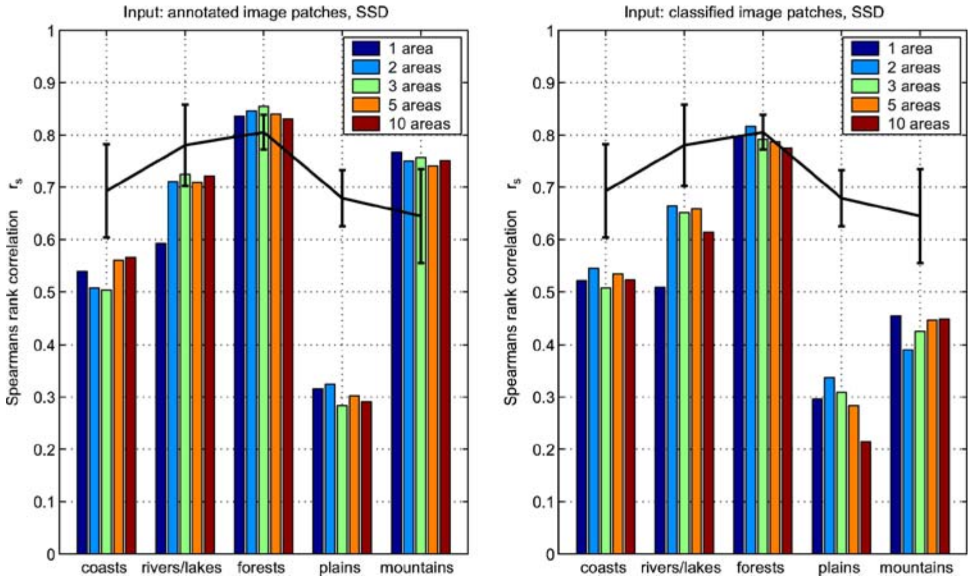

Figure 14. Spearman's rank correlation: Prototype Approach with sum-squared distance. Black line: Average human inter-rater correlation.

Table 9. Best correlation r_s in each category using the sum-squared distance.

	Regions	
	Annotated	Classified
coasts	0.57	0.54
rivers/lakes	0.72	0.66
forests	0.86	0.82
plains	0.32	0.34
mountains	0.77	0.46

Table 10. Best correlation r_s in each category using the perceptually-plausible distance.

	Regions	
	Annotated	Classified
coasts	0.64	0.59
rivers/lakes	0.80	0.75
forests	0.87	0.77
plains	0.72	0.58
mountains	0.74	0.60

(SSD) between the scene representation ***COV*** and the prototype $\boldsymbol{p^c}$ of the respective category:

$$d^c_{SSD}(r) = \sum_{j=1}^{N(r)} \left(COV_j(r) - p^c_j(r)\right)^2 \tag{5}$$

In this equation, $r = [1, 2, 3, 5, 10]$ refers to the number of image areas and $N(r) = [9, 18, 27, 45, 90]$ to the corresponding length of the concept-occurrence vector (compare Section 2).

In the experiments, ability of the semantic image retrieval system to rank natural scenes similarly to humans is evaluated. All experiments have been 5-fold cross-validated. In each round, $\frac{4}{5}$th of each category have been used as training set for the computation of the prototype. The images of the test set were ranked according to $d^c_{SSD}(r)$, and correlated with the corresponding human typicality ranks. The reported Spearman's rank correlation is the average over cross-validation rounds.

As before, the tests with the manually annotated image regions serve as benchmark for the maximum correlation performance that can be expected with our computational model. In a second experiment, the correlation performance is determined when using classified image regions as input. In both cases, the prototypes are learned based on annotated image regions.

Figure 14 left shows the obtained correlation between the human typicality ranking and the machine typicality ranking using annotated image regions. Each group of bars belongs to the category noted below the plot. In each category, $r = [1, 2, 3, 5, 10]$ horizontally-layered image areas have been tested. The black line with the error bars displays the average inter-individual rank correlation from the second psychophysical experiment and its standard deviation for each category. The

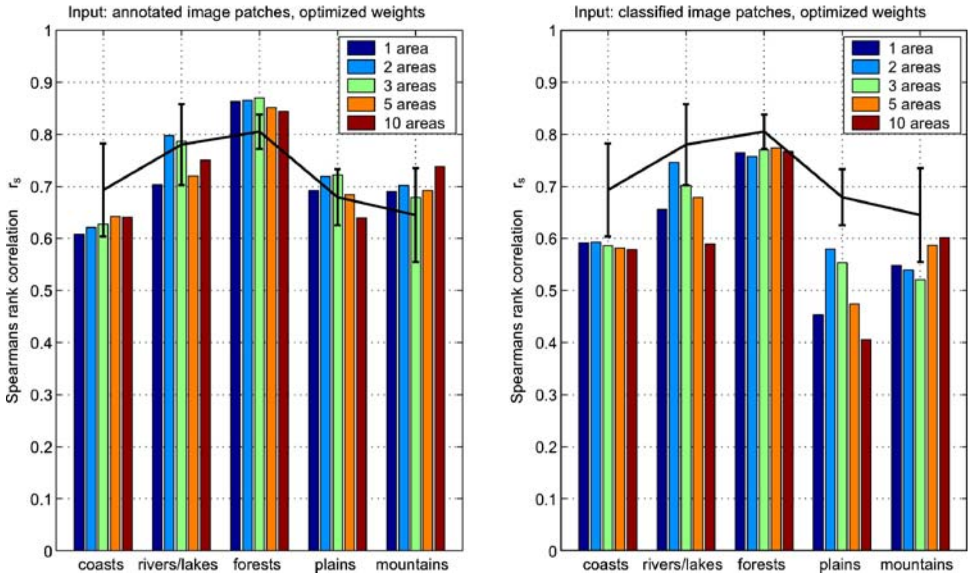

Figure 15. Spearman's rank correlation: Prototype approach with perceptually-plausible distance. Black line: Average human inter-rater correlation.

results are partially promising. The machine ranking using prototypes and the SSD performs for `forests` and `mountains` at least as good as the average inter-individual correlation, and `rivers/lakes` lies inside the 1σ-interval. But the ranking of `coasts` and `plains` correlates poorly with that of humans.

The results of the experiment using classified image regions are shown in Fig. 14 right. In addition, Table 9 displays for each category separately the best correlation over all image areas for the experiments with annotated as well with classified image regions. Although the classification rate of the concept classifier is with 71.7% not extremely high, `coasts`, `forests` and `plains` seem to be quite stable with respect to the misclassifications of image regions. The maximum correlation reached in these categories is less than 0.03 below that of the experiment with the annotated image regions. Also the ranking of `rivers/lakes` looses only little correlation whereas the ranking of `mountains` breaks down with a correlation difference of 0.3.

In summary, the performance of the computational model employing prototypes with SSD is encouraging but not fantastic. With annotated image regions, we obtain for `rivers/lakes`, `forests`, and `mountains` a correlation that is in the same range as the inter-rater correlation. That means, the system behaves similar to the average of a group of humans. But with the more realistic case of classified image regions, most correlations lie quite below the average inter-rater correlations.

7.2. *Typicality Ranking using Prototypes and a Perceptually Plausible Distance Measure*

The results of the previous section suggest that the semantic concepts in each category should be weighted differently. In the previous section, the typicality ranking was performed using SSD as distance measure. When introducing a set of weights $w^c(r)$ in Eq. (5)

$$d^c_{PPD}(r) = \sum_{j=1}^{N(r)} w^c_j(r)(COV(r)_j - p^c(r)_j)^2, \qquad (6)$$

then SSD corresponds to the case where the weight vector is composed of only ones. The weights $w^c(r)$ model in fact the relative importance of the local semantic concepts in each category. In the case of the SSD, all local semantic concepts in all categories are given the same weight. This is not necessarily the best method. *Flowers* for example are very discriminative for the `plains`-category, but hardly appear in any other category (see Fig. 5). For that reason, the relative importance of the concept *flowers* should be increased for the `plains`-category and decreased for all other categories. The goal must thus be to adapt the concept weights $w^c(r)$ depending on the category.

The typicality scores of the human participants in the second psychophysical experiment provide us with powerful information to learn these concept weights. For each number of image areas r and each category c, the weights $w^c(r)$ are optimized through gradient ascent such that the correlation between the machine typicality ranking and the average human typicality ranking in the training set is maximized. This procedure optimizes the concept weights for maximum typicality ranking performance. For the optimization, a constrained minimization problem is solved where the weights $w^c(r)$ are adapted to find the minimum of $1 - r_s(typ_{human}, typ_{machine})$ under the constraint that the weights are bounded: $0.0001 < w_j^c(r) < 10000$. The `fmincon` function of Matlab's Optimization Toolbox was used for the optimization. Through the optimization, the distance $d^c_{PPD}(r)$ becomes perceptually and psychophysically meaningful because the weights are learned from the average human typicality score. $d^c_{PPD}(r)$ is abbreviated in the following PPD for **P**erceptually **P**lausible **D**istance. The images of the test set in each cross-validation round are ranked using the distance with the weights optimized on the training set. Subsequently, Spearman's rank correlation relative to the human ranking is computed. As before, the reported correlation is the average correlation over all cross-validation rounds.

The results of the experiments with the PPD and annotated image regions are displayed in Fig. 15 left. Using classified image regions, the results are plotted in Fig. 15 right. In addition, Table 10 shows the best correlations for each category. In comparison with Table 9, the correlation performances of all categories but `mountains` make a large jump. Notably, the correlation of `plains` increases by 0.4. Except for `coasts`, the correlation between the machine and the human ranking exceeds in all categories the inter-rater reliability of the second psychophysical experiment (black line with error bars), and all categories lie inside the 1σ-interval. Thus, the typicality judgements of the participants in the psychophysical experiment have been modeled by our system. The human-machine correlations also follow the varying inter-individual correlations between categories. That is, `forest`-scenes are ranked more consistently by humans and also by the machine than `mountains`-scenes that exhibit a lower average correlation and higher variance. The reason for this behavior is that the `mountains`-category consists of more visually ambiguous scenes.

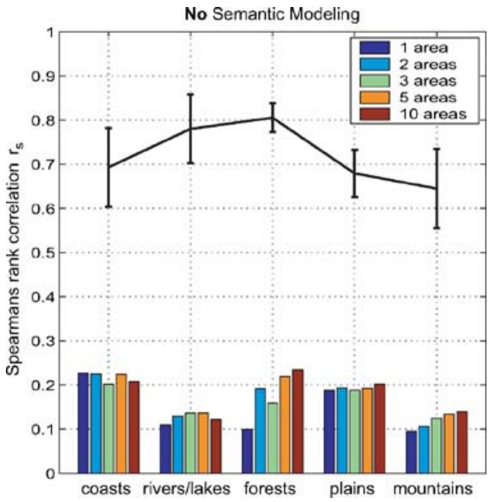

Figure 16. Ranking *without* semantic modeling, direct features extraction. Black line: Average human inter-rater correlation.

Table 11. Best correlation r_s in each category *without* semantic modeling.

	Direct feature extraction
`coasts`	0.22
`rivers/lakes`	0.13
`forests`	0.23
`plains`	0.20
`mountains`	0.14

The weights were optimized on the training set using annotated image regions as input. Figure 15 right shows the results when ranking images with classified image regions using these weights. The correlations based on classified image regions lie on average less than 0.1 below the correlations based on annotated image regions. Spearman's rank correlation is close to the 1σ-interval of the inter-individual correlation in all categories. As before, the correlations follow the variations of the inter-individual correlations between categories. This results suggest that the system consisting of full image analysis, typicality ranking, and computational model is indeed perceptually plausible.

The averaged variance of Spearman's rank correlation over the cross-validation rounds is small with 0.0026 for the annotated images regions and 0.0076 for the classified image regions. The variances correspond to 0.34% of the average rank correlation for the annotated image regions and 1.2% of the average rank correlation for the classified image regions. This

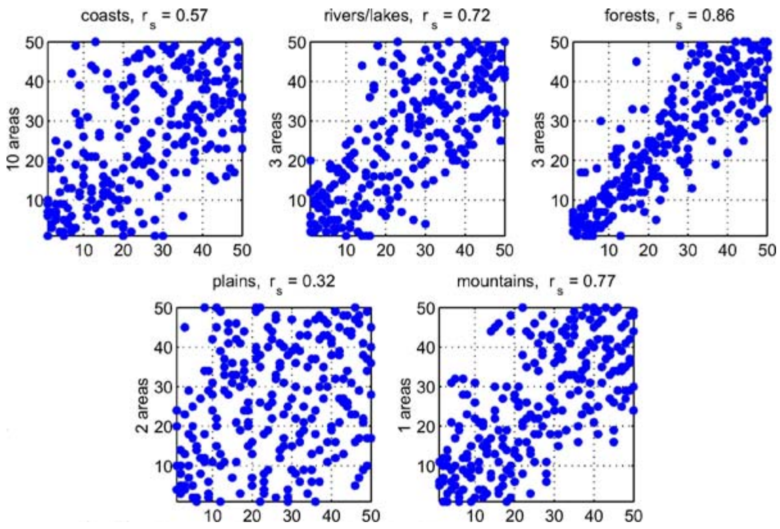

Figure 17. Scatter plots of all categories: Machine ranking vs. human ranking. Prototype approach with SSD. Annotated image regions.

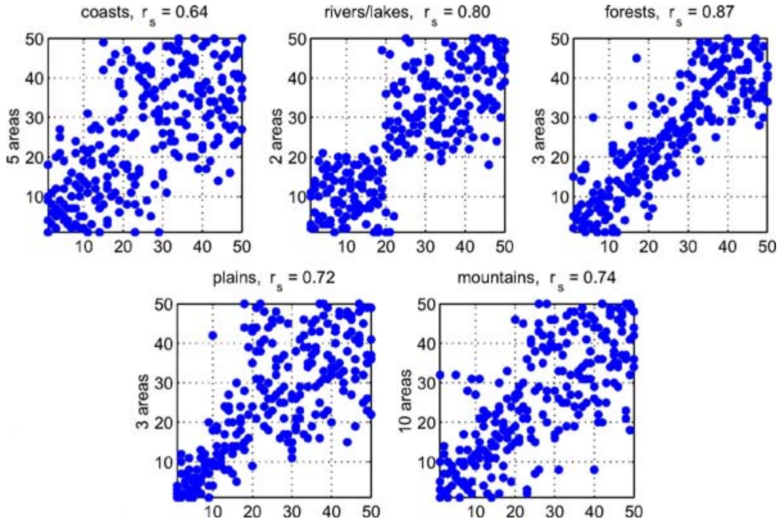

Figure 18. Scatter plots of all categories: Machine ranking vs. human ranking. Prototype approach with PPD. Annotated image regions.

suggests that the results are stable, that the training sets represent the data well and that the method generalizes.

7.3. *Ranking without Semantic Modeling*

In Section 4, it was already shown that the categorization performance is better when employing the semantic modeling instead of directly extracting low-level features. Here, the questions remains whether this is also true for the ranking performance. As before, we extractedthe low-level features mentioned in Section 3 directly from the images, and based the ranking, i.e. the computation of the category prototypes, on those feature vectors. The ranking performance based on these low-level features and using the SSD is displayed in Fig. 16 and in Table 11. The experiment shows that the ranking obtained without semantic modeling hardly correlates at all with the human ranking. In fact, the results indicate quite clearly that the extraction of the nine local semantic concepts and the image representation based on concept-occurrence vectors model the important semantic image details that are needed for a meaningful typicality ranking.

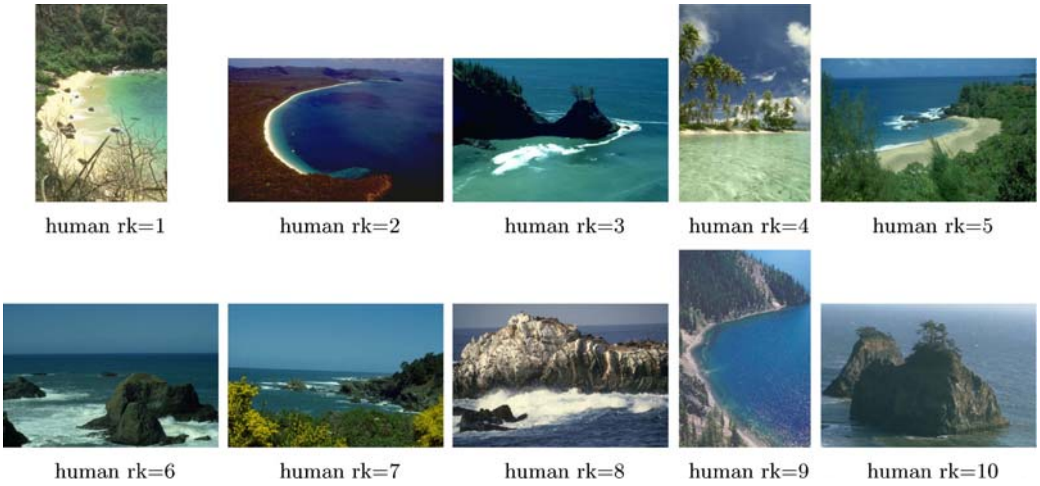

Figure 19. Human rankings: Top 10 ranked `coasts`-images.

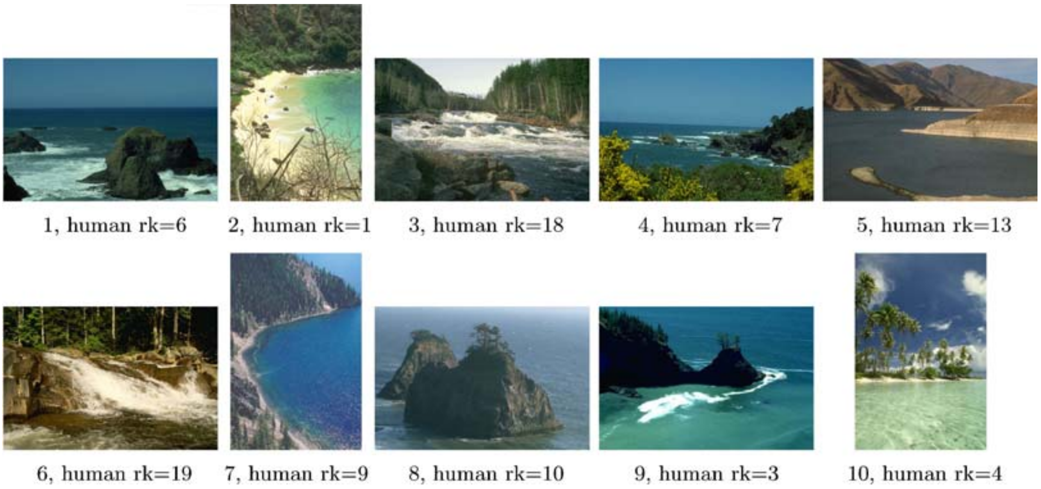

Figure 20. Prototype apporach with PPD, annotated image regions: Top 10 ranked `coasts`-images.

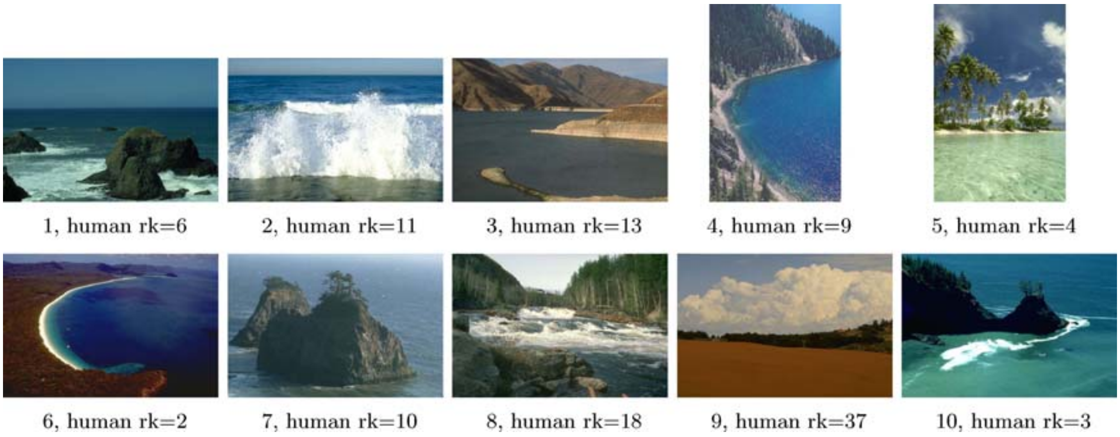

Figure 21. Prototype apporach with PPD, classified image regions: Top 10 ranked `coasts`-images.

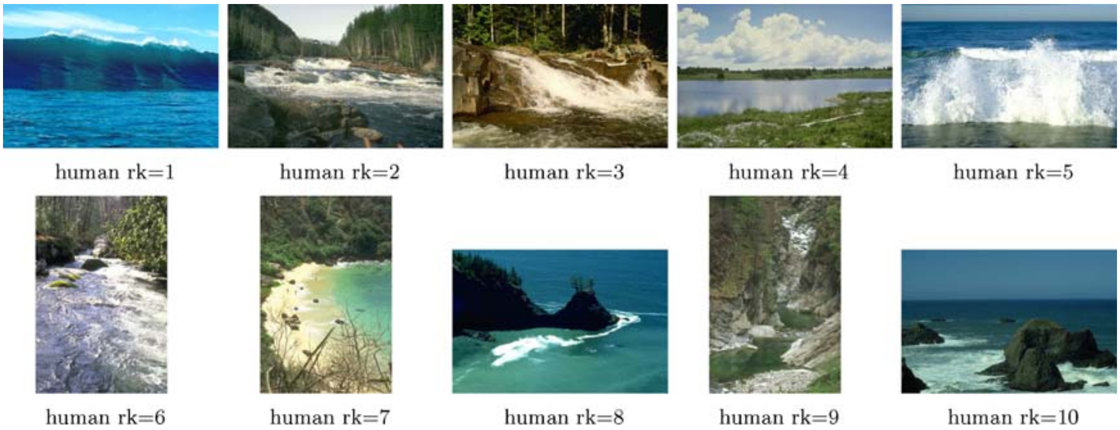

Figure 22. Human rankings: Top 10 ranked `rivers/lakes`-images.

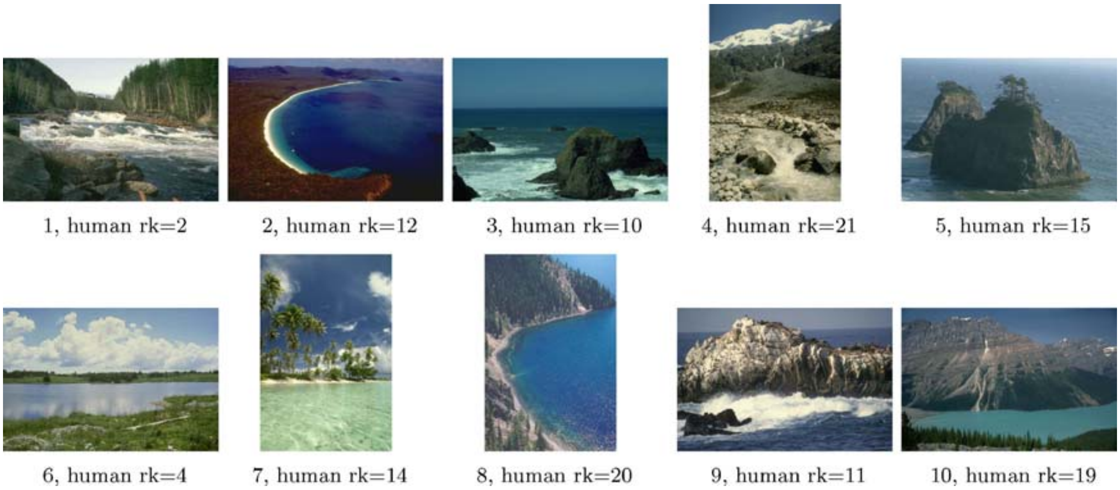

Figure 23. Prototype apporach with PPD, annotated image regions: Top 10 ranked `rivers/lakes`-images.

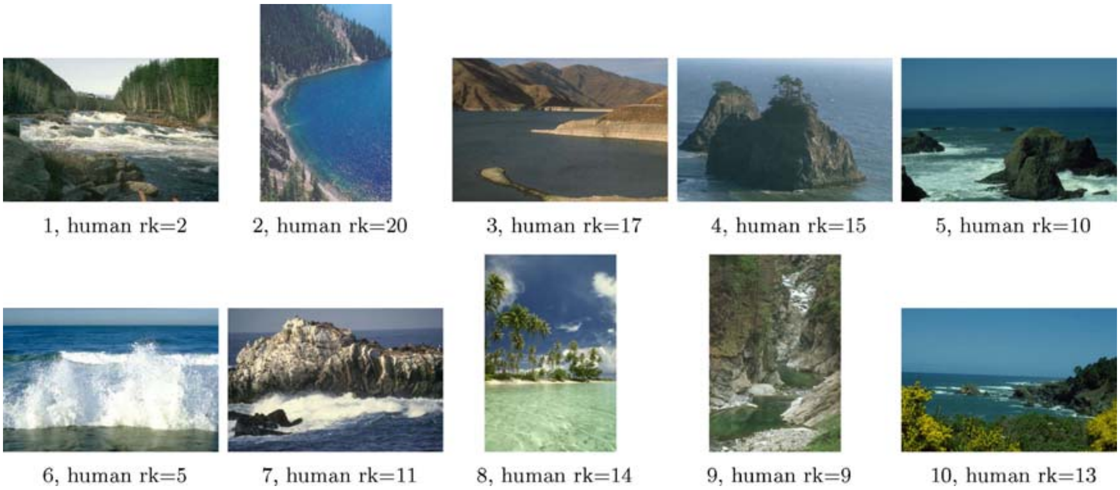

Figure 24. Prototype apporach with PPD, classified image regions: Top 10 ranked `rivers/lakes`-images.

7.4. Discussion

The experiments of the previous sections clearly show the superiority of the semantic modeling over direct feature extraction and the superiority of the psychophysically-plausible distance (PPD) over the sum-squared distance (SSD). With annotated image regions as input, the correlation performance of the PPD is up to 0.4 higher than with the SSD. Only for mountains, the correlation decreases by 0.02. But it is worthwhile noting that in all cases the rank correlation is clearly inside the 1σ-interval of the inter-individual correlation.

The performance increase when using classified image regions as input is of higher importance since it suggests a high robustness of our approach. The correlations achieved with the PPD are up to 0.25 (plains) higher than with the SSD. The mountains-category shows an especially interesting behavior. Based on annotated image regions the ranking performance is better when using the SSD (SSD: 0.77, PPD: 0.74), but based on classified image regions the PPD outperforms the SSD (SSD: 0.46, PPD: 0.60).

Figures 17 and 18 show the correlation increase from the SSD to the PPD visually. The figures display for each scene category the scatter plot of the machine ranking vs. the human ranking. The scatter plots achieved with the PPD are clearly closer to a line than those achieved with the SSD. The point clusters at low and high ranks especially for coasts, rivers/lakes and plains result from the bimodal typicality score distribution in Experiment 2.

The final figures of this paper (Figures 19–24) visualize the ranking performance of our system by plotting the top 10 ranked images for different situations. For two categories, that is coasts and rivers/lakes, exemplary images are displayed in Figs. 19 to 24. The top figure on each page shows the average ranking obtained from the human participants, the second figure the ranking obtained with the Prototype Approach and the PPD when using annotated image regions as input, and in the third figure when using classified image regions as input. The third figure on each page thus displays the ranking result obtained from a fully automatic retrieval system. Below each image in the second and third figure, its corresponding rank as well as the human rank are printed. The images in the Fig. 19 and Fig. 22 show that even for humans the ranking of the images is a non-trivial task. Some of the images appear both in the Top 10 of the coasts- and in the Top 10 of the rivers/lakes-category. Figures 20 and 23, respectively, illustrate three things: Firstly, many of the images are the same as in the Top 10 human-ranked images. Secondly, the ranking of the images corresponds closely to the ranking in the first set of images. And thirdly, the "new" images, that is images that do not appear in Figs. 19 and 22, respectively, are semantically hardly distinguishable. The observations for Figs. 21 and 24, respectively, are very similar. In these figures, images with automatically classified image regions were ranked. We detect only one "real" mistake in the ranking: rank 9 in Fig. 21.

8. Conclusion

In this paper, we presented a computational image representation that reduces the semantic gap between the image understanding of humans and the computer. This semantic modeling of natural scenes is based on the classification of local semantic concepts. Image regions are classified into one of nine concept classes that subsume the main semantic content of the database images. Images are represented through the frequency of occurrence of the semantic concepts. The semantic modeling constitutes a compact, semantic image representation that allows us to describe specific image content and to model the semantic content of natural scene categories.

The semantic modeling has been intensively studied for the categorization of natural scenes. Depending on the classification method and on the quality of the concept classification, good to very good categorization performance has been obtained. In particular, we showed that the semantic modeling leads to considerably better categorization performance compared to directly employing low-level features. Nevertheless, the analysis of the mis-categorized scenes reveals that the regular semantic ambiguity of the database images demands rather for a typicality ranking than for hard-decision categorization. This is in accordance with the goals of content-based image retrieval system where images have to be ranked according to their similarity to the query.

In the last part of this paper, it is shown visually and quantitatively that the proposed semantic modeling is also well-suited for the semantic ranking of images. In particular, the typicality transitions between two scene categories can be modeled. In addition, we introduced a perceptually plausible distance measure that allows us to rank natural scenes semantically. The typicality ranking obtained with this distance measure correlates highly with human ranking of the same images.

Acknowledgments

This work was in part supported by the CogVis project, funded by the Commission of the European Union (IST-2000-29375) and the Swiss Federal Office for Education and Science (BBW 00.0617). We thank Adrian Schwaninger and Franziska Hofer from the Visual Cognition Group at the Psychology Department, University of Zurich for the realization of the psychophysical experiments.

References

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

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., and Jordan, M.I. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Bortz, J. 1999. *Statistik für Sozialwissenschaftler*, 5th edition. Springer.

Boutell, M.R., Luo, J., Shen, X., and Brown. C.M. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.

Chang, C.-C. and Lin, C.-J. 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at: http://www.csie.ntu.edu.tw.

Comaniciu, D. and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5).

Duygulu, P., Barnard, K., de Freitas, J.F.D., and Forsyth, D.A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision ECCV'02*, Copenhagen, Denmark.

Eakins, J.P. and Graham, M.E. 1999. Content-based image retrieval, a report to the JISC Technology Applications programme. Technical report, Institute for Image Data Research, University of Northumbria at Newcastle.

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.

Hsu, C.-W. and Lin, C.-J. 2002. A comparison of methods for multiclass support vector machines. In *IEEE Transactions on Neural Networks*, 13(2):415–425.

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

Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines—Methods, Theory, and Algorithms*. Kluwer Academic Publishers.

Kline, P. 2000. *Handbook of Psychological Testing*, 2nd edition. Routledge.

Kumar, S. and Hebert, M. 2003. Man-made structure detection in natural images using a causal multiscale random field. In *Conference on Computer Vision and Pattern Recognition CVPR'03*, Madison, Wisconsin, pp. 119–126.

Lavrenko, V., Manmatha, R., and Jeon, J. 2003. A model for learning the semantics of pictures. In *17th Annual Conference on Neural Information Processing Systems NIPS'03*, Vancouver, Canada.

Li, J. and Wang, J.Z. 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088.

Lipson, P., Grimson, E., and Sinha. P. 1997. Configuration based scene classification and image indexing. In *Conference on Computer Vision and Pattern Recognition CVPR'97*, Puerto Rico, pp. 1007–1011.

Maron, O. and Ratan, A.L. 1998. Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning ICML'98*, Morgan Kaufmann, San Francisco, CA, pp. 341–349.

Minka, T.P. and Picard, R.W. 1997. Interactive learning using a society of models. In *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 30(4).

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):79–107.

Murphy, G.L. 2002. *The Big Book of Concepts*. MIT Press.

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.

Oliva, A. and Torralba, A. 2002. Scene-centered description from spatial envelope properties. In *Second Workshop on Biologically Motivated Computer Vision BMCV'02*, Tübingen, Germany.

Oliva, A., Torralba, A., Guerin-Dugue, A., and Herault, J. 1999. Global semantic classification of scenes using power spectrum templates. In *Challenge of Image Retrieval CIR*, Newcastle, UK.

Picard, R.W. and Minka, T.P. 1995. Vision texture for annotation. *ACM Journal of Multimedia Systems*.

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

Rosch, E. 1978. Principles of categorization. In E. Rosch, and B.B. Lloyd, (Eds), *Cognition and Categorization*, Erlbaum.

Rosch, E. and Mervis, C.B. 1975. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

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

Rui, Y., Huang, T.S., and Chang, S. 1999. Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10:39–62.

Schwaninger, A., Vogel, J., Hofer, F., and Schiele, B. A psychophysically plausible model for typicality ranking of natural scenes. Submitted to ACM Transactions on Applied Perception.

Sebe, N., Lew, M.S., Zhou, X., Huang, Th.S., and Bakker. E.M. 2003. The state of the art in image and video retrieval. In *Conf. Image and Video Retrieval CIVR*, Urbana-Champaign, IL, USA, pp. 1–8.

Serrano, N., Savakis, A.E., and Luo, J. 2004. Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition*, 37(9):1773–1784.

Shi, J. and Malik, J. 1997. Normalised cuts and image segmentation. In *Conference on Computer Vision and Pattern Recognition CVPR'97*, Puerto Rico.

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.

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

Town, C.P. and Sinclair, D. 2000. Content based image retrieval using semantic visual categories. Technical Report 2000.14, AT&T Laboratories Cambridge.

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

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

Veltkamp, R.C. and Tanase, M. 2001. Content-based image retrieval systems: A survey. Technical report, Department of Computer Science, Utrecht University.

Vogel, J. 2004. *Semantic Scene Modeling and Retrieval*. Number 33 in Selected Readings in Vision and Graphics. Hartung-Gorre, Verlag Konstanz.

Wang, Y. and Zhang, H. 2001. Content-based image orientation detection with support vector machines. In *Workshop on Content-Based Access of Image and Video Libraries CBAIVL'01*, Kauai, Hawaii, USA.