Natural Scene Retrieval based on a Semantic Modeling Step

Julia Vogel and Bernt Schiele

Perceptual Computing and Computer Vision Group
ETH Zurich, Switzerland
`{vogel,schiele}@inf.ethz.ch`
`http://www.vision.ethz.ch/pccv`

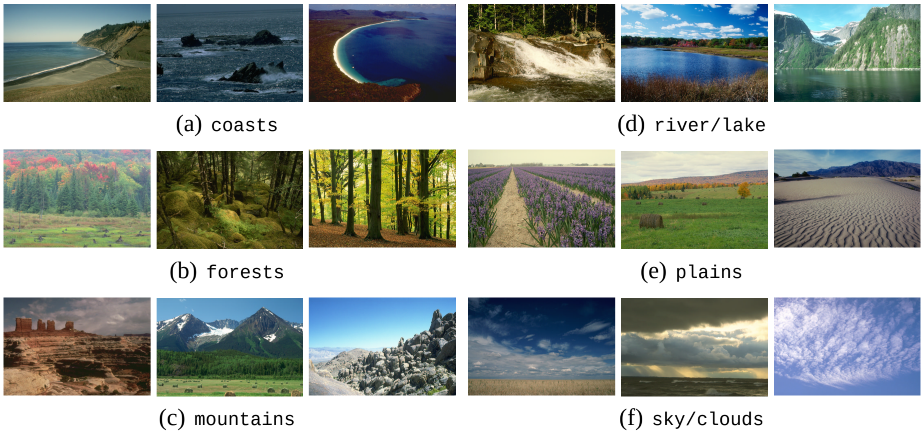
Abstract. In this paper, we present an approach for the retrieval of natural scenes based on a semantic modeling step. Semantic modeling stands for the classification of local image regions into semantic classes such as *grass*, *rocks* or *foliage* and the subsequent summary of this information in so-called concept-occurrence vectors. Using this semantic representation, images from the scene categories `coasts`, `rivers/lakes`, `forests`, `plains`, `mountains` and `sky/clouds` are retrieved. We compare two implementations of the method quantitatively on a visually diverse database of natural scenes. In addition, the semantic modeling approach is compared to retrieval based on low-level features computed directly on the image. The experiments show that semantic modeling leads in fact to better retrieval performance.

1 Introduction

Semantic understanding of images 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 [1]. 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 (for a review see [2]). Natural scene categorization is an intermediate step to close the semantic gap between the image understanding of the user and the computer. In this context, scene categorization refers to the task to group arbitrary images into semantic categories such as `mountains` or `coasts`.

First steps in scene category retrieval were made by Gorkani and Picard [3] (city vs. landscape), Szummer and Picard [4] (indoor/outdoor) and Vailaya et al. [5] (indoor/outdoor, city/landscape, sunset/mountain/forest). All these approaches have in common that they only use global information rather than local information. More recent approaches try to automatically annotate local semantic regions in images [6]-[9] but the majority does not attach a global label to the retrieved images. Oliva and Torralba find global descriptions for images based on local and global features but without an intermediate annotation step [10].

The general goal of our work is to find semantic models of outdoor scenes. In the context of image retrieval it reduces the amount of potentially relevant images. But it also allows to adaptively search for semantic image content inside a particular category (e.g. an image from the `mountains`-category, but with large forest, no rocks). Thus a



(a) `coasts`

(d) `river/lake`

(b) `forests`

(e) `plains`

(c) `mountains`

(f) `sky/clouds`

Fig. 1. Examplary images for each category.

bottom-up step, i.e.scene categorization, and a top-down step, i.e. the use of category information to model relevant images in more detail, can be combined. For that goal, we employ a semantic modeling step. Semantic modeling stands for the classification of image regions into concept classes such as *rocks*, *water* or *sand* and the scene retrieval based on this information. The advantage of an intermediate semantic modeling step is that the system can easily be extended to more categories. Also, for local semantic concepts, it is much easier to obtain ground-truth than for entire images that are often ambiguous. In this paper, we compare two implementations of the semantic modeling approach for natural scene retrieval. In addition, we evaluate how the semantic modeling approach compares with direct low-level feature extraction.

Concerning the database, we paid special attention to using highly varying scenes. The database contains hardly two visually similar images. All experiments have been fully cross-validated in order to average out the fact that in such diverse databases certain test sets perform better than others. The goal is to find out how much profit semantic modeling brings in a realistic setting.

The paper is organized as follows. In the next section, our scene database and the image representation are discussed in detail. Section 3 explains the interplay between the semantic modeling step and the retrieval stage. Finally, Section 4 is devoted to several experiments that compare two different implementations of the system and quantify the performance of the semantic modeling approach vs. a low-level feature-based approach.

2 Natural Scene Categories

For the scene retrieval, we selected six natural scene categories: `coasts`, `forests`, `rivers/lakes`, `plains`, `mountains` and `sky/clouds`. Exemplary images for each category are displayed in Figure 1. The selected categories are an extension of the natural basic level categories of Tversky and Hemenway [11]. In addition, the choice of suitable categories has been influenced by the work of Rogowitz et al. [12].

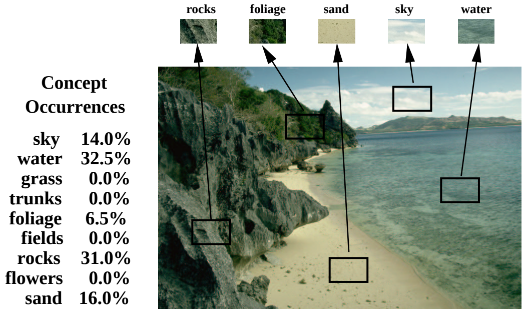


Fig. 2. Semantic modeling

Obviously, these scene categories are visually very diverse. Even for humans the labeling task is non-trivial. Nonetheless, pictures of the same category share common local content, such as for example the local semantic concepts *rocks* or *foliage*. For example, pictures in the `plains`-category contain mainly *grass*, *field* and *flowers*, whereas `mountains`-pictures contain much *foliage* and *rocks*, but also *grass*. Based on this observation, our approach to scene retrieval is to use this *local semantic* information.

2.1 Concept Occurrence Vectors

By analyzing the local similarities and dissimilarities of the scene categories, we identified nine discriminant local semantic concepts: *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand*. In order to avoid a potentially faulty segmentation step, the scene images were divided into an even grid of 10x10 local regions, each comprising 1% of the image area. Through so-called concept classifiers, the local regions are classified into one of the nine concept classes. Each image is represented by a concept occurrence vector (COV) which tabulates the frequency of occurrence of each local semantic concept (see Figure 2). A more detailed image representation can be achieved if multiple COVs are determined on non-overlapping image areas (e.g. top/middle/bottom) and concatenated.

2.2 Database

Our database consists of 700 natural scenes: 143 `coasts`, 114 `rivers/lakes`, 103 `forests`, 128 `plains`, 178 `mountains` and 34 `sky/clouds`. Images are present both in landscape and in portrait format. In order to obtain ground-truth for the concept classifications, all 70’000 local regions (700 images * 100 subregions) have been annotated manually with the above mentioned semantic concepts. Again, a realistic setting was of prime interest. For that reason, each annotated local region was allowed to contain a small amount (at maximum 25%) of a second concept. Imagine a branch that looms into the sky, but does not fill a full subregion (*sky* with some *trunks*) or a lake that borders on the forest (*water* with *foliage*). Due to these quantization issues, only 59’582 out of the 70’000 original annotated regions can be used for the concept classifier training since only those contain the particular concept with at least 75%. The rest

Table 1. Confusion matrix of the local concept classification (k-NN classifier)

		Classifications in %									#regions
		sky	*water*	*grass*	*trunks*	*foliage*	*field*	*rocks*	*flowers*	*sand*	
True class	*sky*	**91.8**	5.7	0.0	0.1	0.5	0.2	1.6	0.0	0.2	15360
	water	9.5	**68.1**	2.4	0.0	6.0	3.8	9.0	0.1	1.2	7309
	grass	0.9	6.4	**34.4**	0.5	43.1	9.0	4.5	0.9	0.5	3541
	trunks	0.8	0.8	1.5	**28.0**	45.6	5.9	16.3	1.1	0.0	1516
	foliage	0.5	1.0	2.5	1.0	**85.1**	1.2	7.3	1.4	0.0	13470
	field	1.2	7.4	6.4	1.3	18.8	**34.8**	27.4	1.8	0.9	4070
	rocks	1.7	3.5	0.7	1.0	24.6	6.6	**61.0**	0.4	0.6	10567
	flowers	0.9	0.7	2.2	0.3	53.0	2.4	4.7	**35.5**	0.4	2051
	sand	6.3	19.7	6.3	0.4	2.2	16.5	32.6	0.3	**16.8**	1773

has been annotated doubly. As some concepts exist in nearly all images and some only in a few images, the size of the nine classes varies between 1'516 (*trunks*) and 15'405 (*sky*) regions.

3 Two-Stage Scene Retrieval

In order to implement the semantic modeling step, the natural scene retrieval proceeds in two stages. In the first stage, the local image regions are classified into one of the nine concept classes. In the second stage, the concept occurrence vector is determined and the images are retrieved based on that concept occurrence vector. The following describes those two stages in more detail.

3.1 Stage I: Concept Classification

The local image regions are represented by a combination of a color and a texture feature. The color feature is a 84-bin HSI color histogram (H=36 bins, S=32 bins, I=16 bins), and the texture feature is a 72-bin edge-direction histogram. Tests with other features, such as RGB color histograms, texture features of the gray-level co-occurrence matrix, or FFT texture features, resulted in lower classification performance. The classification has been tested with both a k-Nearest-Neighbor ($k = 30$) (k-NN) and a Support Vector Machine ($C = 8, \gamma = 0.5$) (SVM) [13] concept classifier.

With 68.9% classification rate, the k-NN concept classifier showed a slightly inferior performance than the SVM concept classifier with 69.9% classification rate. Nevertheless, its resulting classifications perform better in the subsequent retrieval stage and will therefore be employed in all following experiments. The reason for this behavior is that the global classification rate usually improves to the benefit of the large classes (*sky*, *foliage*) and at the expense of the smaller classes (*field*,*flowers*,*sand*). Since these smaller classes are essential for scene retrieval, the overall classification accuracy on the first stage is not the most important performance measure.

The 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 from the same semantic concept tend to be more similar to other (for example neighboring) regions in the same image than to regions in other images. The confusion matrix of the experiments with the k-NN concept classifier is depicted in Table 1. The confusion matrix shows a strong correlation between class size and classification result. In addition, we observe confusions between similar semantic classes, such as *grass* and *foliage*, *trunks* and *foliage*, or *field* and *rocks*.

The trained concept classifier is used to classify all regions of an image into one of the semantic classes. The experience showed that doubly annotated regions (e.g. with *sky* and *rocks* at the border between the sky and a mountain) were usually classified as one of those two semantic concepts.

3.2 Stage II: Scene Retrieval based on Concept Occurrence Vectors

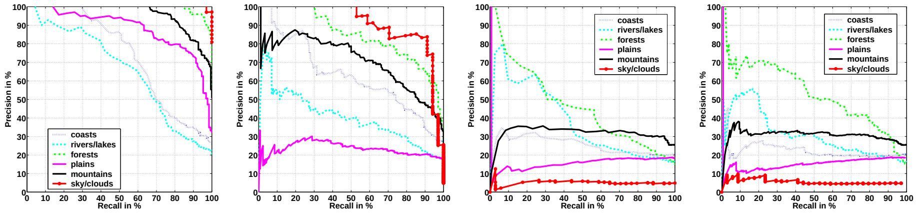
The output of the first stage is localized semantic information about the image. It specifies where in the image there are e.g. *sky* or *foliage*-regions and how much of the image is covered with e.g. *water*. From that semantic information, the concept occurrence vectors are determined. Experiments have shown that the retrieval performance improves if multiple concept occurrence vectors are computed either on three (top/middle/bottom) or five image areas. This leads to a resulting concept occurrence vectors of either length=27 or length=45.

In the following we propose two different implementations to semantically categorize images based on the concept occurrence vectors, namely a Prototype approach and an SVM approach. In the experiments those two implementations are compared and analyzed.

Prototype approach to scene retrieval. The prototype for a category is the mean over all concept occurrence vectors of the respective category members. Thus, the prototype can be seen as the most typical image representation for a particular scene category where the respective image does not necessarily exist. The bins or attributes of the prototype hold the information which amount of a certain concept an image of a particular scene category typically contains. For example, a `forest`-image usually does not contain any *sand*. Therefore, "*sand*-bin" of the `forest`-prototype is close to zero.

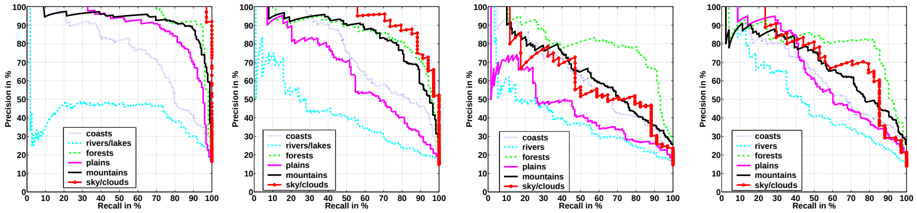
When determining the category of an unseen image, the Euclidean or the Mahalanobis distance between the image's concept occurrence vector and the prototype is computed. The smaller the distance, the more likely it is that the image belongs to the respective category. By varying the accepted distance to the prototype, precision and recall for the retrieval of a particular scene category can be influenced.

SVM approach to scene retrieval. For the SVM-based retrieval of natural scenes we employ the LIBSVM package of Chen and Lin [13]. A Support Vector Machine is trained for each scene category. The input to the SVM are the concept occurrence vectors of the relevant images. The margin, that is the distance of an unseen concept occurrence vector to the separating hyperplane, is a measure of confidence for the category membership of the respective image. By varying the acceptance threshold for the margin, precision and recall of the scene categories can be controlled.



(a) COV based on annotated image regions (b) COV based on classified image regions (c) Based on low-level 156-bin feature vector (d) Based on low-level 468-bin feature vector

Fig. 3. Scene retrieval with Prototype approach



(a) COV based on annotated image regions (b) COV based on classified image regions (c) Based on low-level 156-bin feature vector (d) Based on low-level 468-bin feature vector

Fig. 4. Scene retrieval with SVM approach

4 Scene Retrieval: Experiments

Using the database described in Section 2.2, we conducted a set of experiments in order to compare the performance of the two retrieval implementations. In addition, it is evaluated whether the semantic modeling approach is superior to using low-level features of the images directly for retrieval. Performance measures are precision (percentage of retrieved images that are also relevant) and recall (percentage of relevant images that are retrieved). The precision-recall curves of the experiments are depicted in Fig. 3 for the Prototype approach and Fig. 4 for the SVM approach. Tables 2 and 3 summarize the Equal Error Rates (EER) of the experiments. Both concept classification and scene retrieval experiments are 10-fold cross-validated on the same ten test and trainings sets. That is, a particular trainings set is used to train the concept classifier, the SVM and the prototypes. Classification and retrieval are evaluated on the corresponding test set.

Retrieval based on annotated image regions. In the first experiment, we compared the performance of the Prototype vs. the SVM approach based on annotated patches. The goal of the experiment is to evaluate if the semantic modeling approach is effective given perfect data.

The results of the experiment are depicted in Fig. 3 (a) and Fig. 4 (a). The SVM approach outperforms the Prototype approach in 4 of 6 cases (Tables 2 and 3). Obviously, `coasts` and `rivers/lakes` are the most difficult categories. In fact, the detailed analysis of the retrieval results of those two categories shows that they are frequently

Table 2. Equal Error Rates for Prototype approach

Retrieval based on	coasts	rivers lakes	forests	plains	moun-tains	sky clouds
annotated regions	64.3%	61.9%	95.1%	79.7%	86.0%	97.1%
classified regions	57.3%	43.0%	74.8%	28.9%	66.9%	85.2%
156-bin feature vec.	29.4%	41.8%	45.6%	11.6%	33.8%	2.9%
468-bin feature vec.	25.7%	32.2%	50.5%	11.2%	32.5%	8.8%

Table 3. Equal Error Rates for SVM approach

Retrieval based on	coasts	rivers lakes	forests	plains	moun-tains	sky clouds
annotated regions	70.5%	47.4%	91.4%	81.3%	89.3%	97.2%
classified regions	61.0%	42.1%	80.6%	54.7%	78.1%	85.3%
156-bin feature vec.	56.6%	40.3%	77.6%	46.1%	59.0%	52.9%
468-bin feature vec.	57.3%	47.3%	81.4%	54.6%	63.8%	70.5%

confused. The main reason is that these two categories are in fact quite ambiguous. Even for the human annotator it is not clear into which category to sort a certain image that contains some water. It is especially those ambiguous images that are also wrongly retrieved by the retrieval system.

The SVM implementation has difficulties in modeling the `rivers/lakes`-category for small recall values since this category is not compact in the COV space. All other categories, that is `plains`, `mountains`, `forests` and `sky/clouds`, are retrieved with good to very good accuracy. Again the analysis of the retrieval results show that wrongly retrieved images are often semantically closer to the category that has been requested than to the “correct” category.

Retrieval based on classified image regions. In the next experiment, images with automatically classified local regions were considered. The concept classifier described in Section 3.1 and Table 1 was employed for the Stage I classification. Based on these classifications, the concept occurrence vector is determined. The retrieval result is depicted in Fig. 3 (b) and Fig. 4 (b). Here, the SVM approach again outperforms the Prototype approach in 5 of 6 cases (Tables 2 and 3). `sky/clouds`, `mountains` and `forests` have been retrieved especially well by the SVM. The loss compared to the annotated scenes is quite low. Compared to the retrieval in the annotated case, `coasts` are retrieved reasonably well.

The Prototype approach fails completely to retrieve `plains`, whereas the SVM is able to achieve an EER of 54.7%. The reasons for the general worse performance in the `plains`-category are the confusions of the concept classification stage. The `plains`-category can be discriminated by the detection of *field*, *grass* and *flowers*. These three concepts are confused to a large percentage with *rocks* and *foliage* (refer to Table 1). These strong mis-classifications lead to the observed low retrieval performance.

Retrieval without semantic modeling step. The last two experiments were carried out in order to find out whether the semantic modeling step is in fact beneficial for the retrieval task. Therefore this section will describe an experiment where we compare the retrieval results based on the concept occurrence vector vs. the performance using the low-level features directly as image representation. The same features as for the concept classification were used for the image representation: a concatenation of a 84-bin linear HSI color histogram and a 72-bin edge direction histogram. These features were once computed directly on the image, resulting in a global feature vector of length 156, and once on three image areas (top/middle/bottom), resulting in a feature vector of length 3*156=468. The "Prototype" approach now refers to the learning of a mean vector per category and the computation of the Euclidean distance between the mean vector and the feature vector of a new image. The results of these experiments are depicted in Fig. 3 (c)-(d) for the "Prototype" approach and Fig. 4 (c)-(d) for the SVM method.

Both the figures and the EERs in Table 2 clearly show that the "Prototype" approach based on low-level features fails compared to the semantic modeling based approach both for one image area and for three image areas. Probably the feature space is too high-dimensional and too sparse. For that reason also the introduction of more localized information through the use of three image areas does not bring any improvement compared to one image area.

In contrast, the low-level feature-based SVM approach performs surprisingly well compared to the SVM based on the semantic modeling step. The introduction of localized information by using three image areas also leads to a performance increase. The variation of the EER in the three-area feature-based approach is smaller than the approach based on the COV. Categories such as `sky/clouds` or `mountains` are not retrieved as good as with the semantic modeling approach and categories such as `rivers/lakes` are retrieved better than with the semantic modeling approach. But in summary, the performance increase in the `rivers/lakes`-category does not counterbalance the performance decrease in the `sky/clouds`- and `mountains`-category.

Discussion. Summarizing, we can draw two conclusions from the experiments. Firstly, the SVM implementation of the retrieval system outperforms the Prototype approach. Only single categories are retrieved better when using prototypes. Here, a combination of both methods might be advantageous.

Secondly, the semantic modeling step and an approach such as the concept occurrence vectors is beneficial for the retrieval of natural scene categories considered in this paper. For most categories, the EER obtained with the semantic modeling step is equal to or better than without the semantic modeling. Many of the wrongly retrieved images are in fact content-wise on the borderline between two categories. For that reason quantitative retrieval performance should not be the only performance measure for the semantic retrieval task. Still, the performance of the problematic categories `rivers/lakes` and `plains` can be improved by better concept classifiers in order to retrieve discriminant concepts with high confidence or better category models. One might, for example, employ different numbers of discriminant concepts and/or image areas per category in order to differentiate between `rivers/lakes` and `coasts`.

5 Conclusion

In this work, we presented an approach to natural scene retrieval that is based on a semantic modeling step. This step generates a so-called concept-occurrence vector that models the distribution of local semantic concepts in the image. Based on this representation, scene categories are retrieved. We have shown quantitatively that Support Vector Machines in most cases perform better than the retrieval based on category prototypes. We have also demonstrated that the semantic modeling step is superior to retrieval based on low-level features computed directly on the image. In addition, since ground-truth is more easily available for local semantic concepts than for full images, the system based on semantic modeling is more easily extendable to more scene categories and also to more local concepts. Further advantages of the semantic modeling are the data reduction due to the use of concept occurrence vectors and the fact that the local semantic concepts can be used as descriptive vocabulary in a subsequent relevence feedback step.

Acknowledgments This work is part of the CogVis project, funded by the Comission of the European Union (IST-2000-29375) and the Swiss Federal Office for Education and Science (BBW 00.0617).

References

1. Sebe, N., Lew, M., Zhou, X., Huang, T., Bakker, E.: The state of the art in image and video retrieval. In: Conf. Image and Video Retrieval CIVR, Urbana-Champaign, IL, USA (2003)
2. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. on PAMI **22** (2000)
3. Gorkani, M., Picard, R.: Texture orientation for sorting photos at a glance. In: Int. Conf. on Pattern Recognition ICPR, Jerusalem, Israel (1994)
4. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: IEEE Int. Workshop on Content-based Access of Image and Video Databases, Bombay, India (1998)
5. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.: Image classification for content-based indexing. IEEE Trans. on Image Processing **10** (2001)
6. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: Int. Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (1998)
7. Town, C., Sinclair, D.: Content based image retrieval using semantic visual categories. Technical Report 2000.14, AT&T Laboratories Cambridge (2000)
8. Naphade, M., Huang, T.: A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans. on Multimedia **3** (2001)
9. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D.: Object recognition as machine translation - part 2: Exploiting image data-base clustering models. In: European Conf. on Computer Vision, Copenhagen, Denmark (2002)
10. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. Journal of Computer Vision **42** (2001)
11. Tversky, B., Hemenway, K.: Categories of environmental scenes. Cogn. Psychology **15** (1983)
12. Rogowitz, B., Frese, T., Smith, J., Bouman, C., Kalin, E.: Perceptual image similarity experiments. In: SPIE Conf. Human Vision and Electronic Imaging, San Jose, California (1998)
13. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.