Semantic Categorization and Retrieval of Natural Scene Images

Sharmishtha D Ronge
Sinhgad College of Engineering, Pune

Chaitali laukar
Sinhgad College of Engineering, Pune

Abstract

In this paper, we present an approach for the retrieval of natural scenes based on a semantic modelling. In this Semantic modelling stands for the classification of local image regions into semantic classes as grass, rocks or foliage and the subsequent summary in the local image regions are classified using low level features into semantic concept classes such as water, sky or sand. This paper proposes a method for semantic categorization and retrieval of natural scene images or picture with and without people. Image retrieval systems usually represent images by a collection of low-level features such as colour, texture, edge positions and spatial relationships in the image. These features are used to compute the similarity between a picture or image selected by the user and the picture or images in the database. We have attempts to provide a comprehensive survey of the latest mechanical achievements in high-level semantic-based image retrieval.

1. Introduction

In this paper, we present an innovative image representation that renders it likely to get access to natural scenes by local semantic description. Our work is inspired by the continuing effort in content-based image retrieval to extract and to form the semantic content of images. The basic concept of the semantic modelling is to classify local image districts into semantic notion classes such as water, rocks/foliage. Images are represented through the frequency of occurrence of these localized notions. Through extensive trials, we demonstrate that the likeness representation is well suited for modelling the semantic content of heterogeneous view classes, and therefore for categorization and retrieval. Humans tend to interpret images using high-level concepts; they are able to identify keywords, abstract objects or events presented in the image. However, for a computer the image content is a matrix of pixels, which can be summarized by low-level colour, texture or shape features. The lack of correlation between the high-level concepts that a user requires and the low-level features that image retrieval systems offer is the semantic gap. Content-Based Image Retrieval (CBIR) has been discussed in the technical literature as a method that may develop into an efficient image search and retrieval technique. Prior work on medical image retrieval has mainly focused on extracting low-level visual features (e.g., colour, texture, shape, spatial layout) and then using them directly to compute image similarity. Extensive experiments have shown, however, that low-level image features cannot always capture the biomedical semantic concepts in the image.1 This poses a serious shortcoming in applying CBIR to routine clinical use, where image content is defined in terms of biomedical concepts. In general, it is challenging to link high-level semantic concepts and automatically-extracted, low-level image features. After that, to support query by semantic concept, that is a compelling need for CBIR systems to provide maximum support towards bridging the „semantic gap" between the low-level visual features and the semantics in biomedical concepts.

In our work we try to reduce this semantic gap in a field of natural scene images with and without people. These sorts of pictures are common in personal family albums. Our method can help the people to search in these albums effectively. Semantic understanding of images remains an important research challenge for the image and video retrieval community. Some gain access to the content of still images. 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 context, scene categorization refers to the task to group arbitrary images into semantic categories. Scene classification and the related retrieval problems have two critical components: representing scenes and learning models for associating labels to these scenes. Given the

difficulty of image segmentation in unconstrained data sets, most Intermediate semantic models that make use of the occurrence of common concepts, such as sky, water, grass, snow, have also been shown to improve the classification of natural scenes Although, recent work was mostly limited in terms of grid based representations or the small number of classes used contextual modelling of image scenes using combinations of concepts and objects looks promising for decreasing the semantic gap.

2. Literature Survey

2.1 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 of semantic detail that can only be modelled 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.

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 modelling 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 modelling 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.

2.2 Semantic Modelling

For the scene retrieval, we selected six natural scene categories: coasts, forests, rivers or lakes, plains, mountains and sky or 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. In addition, the choice of suitable categories has been influenced by the work of Rogowitz et al.

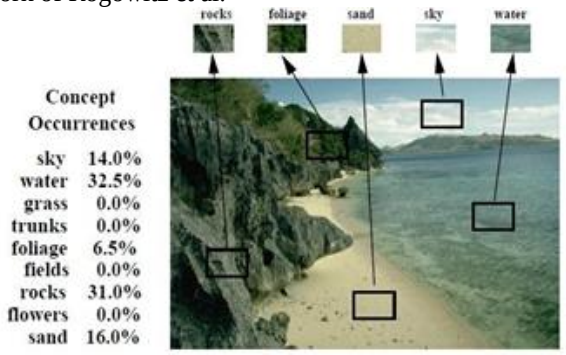


Figure 1. Semantic Modelling

Obviously, these scene categories are visually very diverse. Even for humans the labelling task is the nontrivial. 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.

2.3 Concept Occurrence Vectors

By analysing the local similarities and dissimilarities of the scene categories, we identified nine discriminate local semantic concept are sky, water, grass, trunks,

foliage, field, rocks, flowers and sand in order of avoid a potentially faulty segmentation step, the scene images were divided into an even grid of 10x10 local regions, each comprising of the image area. Through so-called concept classifiers, his 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 in the 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.

2.4 Database

Our database consists of natural scenes: 143 coasts, 114 rivers/lakes, 103 forests, 128 plains, 178 mountains and 34 sky/clouds. Images are present both in and scape and portrait format. In that order to obtain ground-truth for the concept classifications, all 70‟000 local regions (700 images * 100 sub regions) have been annotated manually with the above mentioned semantic concepts. Again, a realistic setting was the 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 sub region (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 has been annotated doubly.

3. Preprocessing

3.1Segmentation using color information

At first step of our algorithm the image is segmented into arbitrary-shaped sub regions. We take advantage of Mean Shift segmentation algorithm presented in based on group of pixels, which are close in the spatial and color range domain. In this algorithm iteratively detects modes in a probability density function. Image segmentation is still an unsolved problem in computer vision. Although numerous algorithms have been shown to work well for images with only a few objects and a simple background, it seems impossible to find a fixed set of parameters that produces reasonable results in a large unconstrained data set. In this paper, we assume that a very precise segmentation of an image is not required for the scene classification and retrieval problem. Therefore, our goal is to obtain a rough estimate of important regions using unsupervised classification of color features and grouping of line segments. For segmentation using color, we use the combined classifier approach in [6] because it fuses color and spatial information, and does not require the number of regions as an input parameter. First, an initial labeling of an image is done using k-means clustering of only the HSV values of pixels. Next, these pixel labels are used to train a multi-class nearest mean classifier on the HSV color features and a Parzen window classifier with a Gaussian kernel using the pixel positions as spatial features. The nearest mean classifier is selected for its simplicity and the Parzen window classifier is selected for its nonparametric nature for modeling a distribution with an indefinite number of modes (each mode corresponds to a segment). Then, the posterior probability outputs of each classifier for each pixel are combined using the product rule, and the pixels are assigned to the class with the largest probability. A new pair of classifiers is trained using these new pixel labels and the iterations continue until the pixel labels stabilize. Note that the number of clusters in the initial k-means clustering does not directly correspond to the number of segments, and can be empirically estimated using the number of dominant colors that can be found in the images in the selected data set. Figure 2 shows example segmentations. The regions that are smaller than an area threshold are removed from the final segmentation where the results contain only contiguous sets of pixels that have a relatively uniform color distribution and are large enough

Figure 2 Segmentation using color

3.2 Body detection

After the image is segmented we used algorithms for skin and face detection to identify sub regions belonging to the humans body. We applied skin detection. Which results in skin probability maps? For face detection we used the implementation of

Viola/Jones Face Detector found in this face detector is applied only in the regions, where skin was detected. This combination with skin detector produces more precise results, because occurrence of false faces in treetops and rocks was eliminated. As next step a template in the shape of human's body is added to each detected face. Each sub region overlaid by this template is examined if the majority of sub region area lies inside the template or outside. Sub regions lying for the most part within the template we consider to depict the human's body. They will not proceed to further processing and classification.

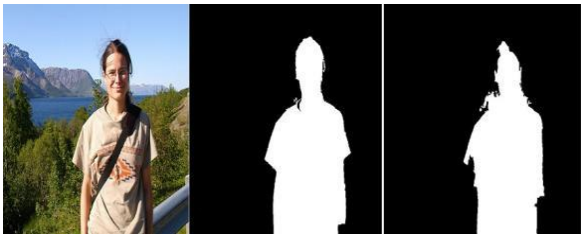


Figure 3. Human body detection (a) Original Image (b) Manual detection (c) Obtained result

3.3 Scene Categorization

This stage of our method is scene categorization. Scene categorization refers to the task of grouping images or scenes into a set of given categories. In our work we have six different categories: coasts, forests, rivers/lakes, sky/clouds, plains and mountains. We define for each of these categories a category prototype. This is most typical for the respective category in these category prototypes and the standard deviations for each category. Using the frequency of occurrence of eight semantic concept classes in the image the most similar category prototype is defined and that determines the high level scene category.

4. Experimental Analysis

4.1 Scene Retrieval: Experiments

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 is retrieved). The precision-recall curves of the experiments are depicted in Fig. 3 for the Prototype approach and shows in 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 training set is used to train the concept classifier, the SVM and the prototypes. Classification and retrieval are evaluated on the corresponding test set.

4.2 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 the 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 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. By varying the accepted distance to the prototype, precision and recall for the retrieval of a particular scene category can be influenced

4.3 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 is e.g. sky or foliage-regions and how much of the image is covered by the 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.

4.4 Retrieval without semantic modeling

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 an 84-bin linear HSI color histogram and a 72-bin edge direction histogram. These features were once computed directly on that image and resulting in a global feature vector of length 156, and once on three image areas (top/middle/bottom) and 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.

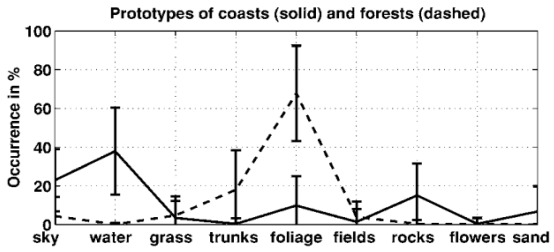
4.5 Scene Categorization

Scene categorization is a special case of image retrieval where the query corresponds to the scene category being searched for. Since scenes, that are 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.

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 category membership 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 a semantic concept

$$(\mathrm{P})^{n} = \frac{1}{N_c}\sum_{j=1}^{Nc} \mathrm{COV}\ (\mathrm{j})$$

Where c refers to one of the six categories and Nc to the number of images in that category. Figure 4 displays the category prototypes and the standard deviations for each category



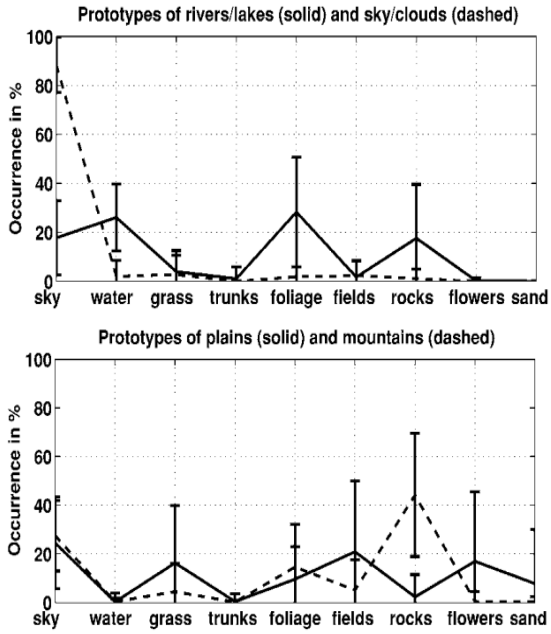


Figure 4. Prototypes and standard deviations of the six scene categories.

The figure reveals which semantic concepts are especially discriminate. 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.

5 Results

Photo Gallery has a large amount of images of various contents and is often used to evaluate the performance of image retrieval systems. It is assumed that each image has at least one dominant region that expresses its semantics. For most categories in our database, there is one dominant region in the relevant images. At the end we discuss results obtained by our proposed method and those obtained by the initial method. Because of using a regular grid in the initial method, rectangular sub regions belonging to two semantic concepts can be classified inaccurately. This is successfully improved by proposed method. On the other hand, in proposed method some problems occur in classification of small sub regions.

Table 1: Confusion matrix of the SVM concept classification (C=8, g=0.125). Classification is in %

		Classifications in %									
		sky	*water*	*grass*	*trunks*	*foliage*	*field*	*rocks*	*flowers*	*sand*	#regions
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

Table 2: Low level feature relevance

Color	52.3 %
Co occurrence matrix	41.2 %
Edge direction	43.4 %
Gabor features	25.3 %
Color+ Co occurrence matrix	59.8 %
Color+ Edge direction	62.5 %
Color+ Gabor features	56.7 %
All features	67.8 %

For comparison, both methods were tested pixel-by pixel with the manually annotated original image. An array of same size as original image was obtained, where logical 1 (white color) mean that pixels represented the same semantic category and logical 0 (black color) when different category. Initial method matched the ground truth in 68,05% compared to proposed method which reached 70,59%. Shows in figure 5

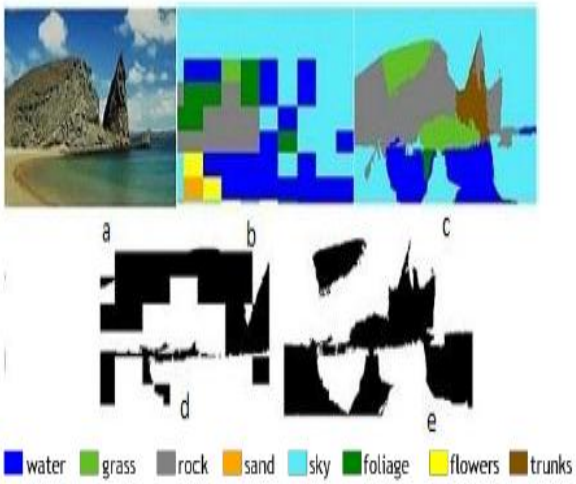


Figure 5: Local semantic concept classification (a) Original image. (b) Ground truth. (c+d) Result of initial method and equality map (e) Result of our proposed method and equality map

Figure 6. Typical images for each category

6. Conclusion

In this work, we presented an approach to natural scene retrieval that is based on a semantic. This is

generates a concept-occurrence vector that models the distribution of local semantic concepts in the image. Based on this representation, scene categories are retrieved. 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 Categorization 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 modelling 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.

7. References

[1] Sebe, N., Lew, M., Zhou, X., Huang, T., Bakker, E.: The state of the art in image and video retrieval. In: onf. 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] 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)

[6] Tversky, B., Hemenway, K.: Categories of environmental scenes. Cogn. Psychology 15 (1983)

[7] R. S. Choras. Image feature extraction techniques and their application for cbir and biometrics systems International Journal of Biology and Biomedical Engineering,1:6–16, 2007.

[8] D. Comaniciu, P. Meer, and Senior Member. Mean shift: A robust approach toward feature space analysis.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 24:603 – 619, 2002.

[9] Y. Zhuang, X. Liu, and Y. Pan. Apply semantic template to support content-based image retrieval. In Proceeding of IST and SPIE Storage and Retrieval for Media Databases, pages 23–28, 2000.

[10] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. Int. J. Comput. Vision, 72:133–157, April 2007.

[11] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. An ontology approach to object-based image retrieval In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, volume 2, pages II – 511–14 vol.3, sept. 2003.

[12] H. Feng, T.-S. Chua, A boostrapping approach to annotating large image collection, Workshop on Multimedia Information Retrieval in ACM Multimedia, November 2003, pp. 55–62.

[13] 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.