# Semantic Categorization and Retrieval of Natural Scene Images

Kristína Lidayová*
*Supervised by: RNDr. Elena Šikudová, PhD.*†

Faculty of Mathematics, Physics, and Informatics
Comenius University
Bratislava / Slovakia

## Abstract

The semantic gap between the digital image representation and the user's image understanding is still a big problem. In our work we try to reduce this semantic gap in a field of natural images.

This paper proposes a method for semantic categorization and retrieval of natural scene images with and without people. These are typical holiday pictures from hiking outdoors. Our approach comprises of three stages. Preprocessing consists of image segmentation into arbitrary-shaped regions and detection of people in the image. In the next stage, local image regions are classified using low level features into semantic concept classes such as water, sky or sand. Finally the frequency of occurrence of these semantic concept classes determines the high level scene category. For the classification of local image regions the k-Nearest Neighbor and Support Vector Machine classifiers are used. The results obtained by both classifiers are validated within the paper.

**Keywords:** Semantic gap, Semantic retrieval, Content Based Image Retrieval (CBIR), Classification

## 1 Introduction

We live in a world where having a digital camera or image scanner is not a problem any more. People are used to take thousands of pictures during their vacation and they like to share them at the web galleries or social networks. Due to more and more images being generated in digital form around the world, it is important to deal with a problem how to extract the semantic content of images and then retrieve these images effectively.

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 color, 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.

Figure 1: An example of the semantic gap problem. The two images possess very similar colour and position characteristics, but differ vastly as far as the semantics are concerned.

In our work we try to reduce this semantic gap in a field of natural scene images with and without people. These sort of pictures are common in personal family albums. Our method can help the people to search in these albums effectively.

This paper is organized in the following way: The techniques in reducing the semantic gap are discussed in Section 2. In Section 3 our method for semantic categorization and retrieval is presented. In Section 4 we describe segmentation algorithm and body detection used here. Section 5 is dedicated for the low-level features and classifiers. In section 6 we deal with scene categorization and Section 7 discuss the results. Finally, Section 8 concludes the paper.

## 2 Related work

Image retrieval research is moving from text-based, through content-based, towards semantic-based image retrieval. Several systems reducing the semantic gap have been proposed.

In [6] the techniques reducing the semantic gap are divided into six categories:

**1. Object ontology** This system is using object ontology to define high-level concepts. Firstly low-level features describing the color, position and shape of each

*lidayova@gmail.com
†sikudova@sccg.sk

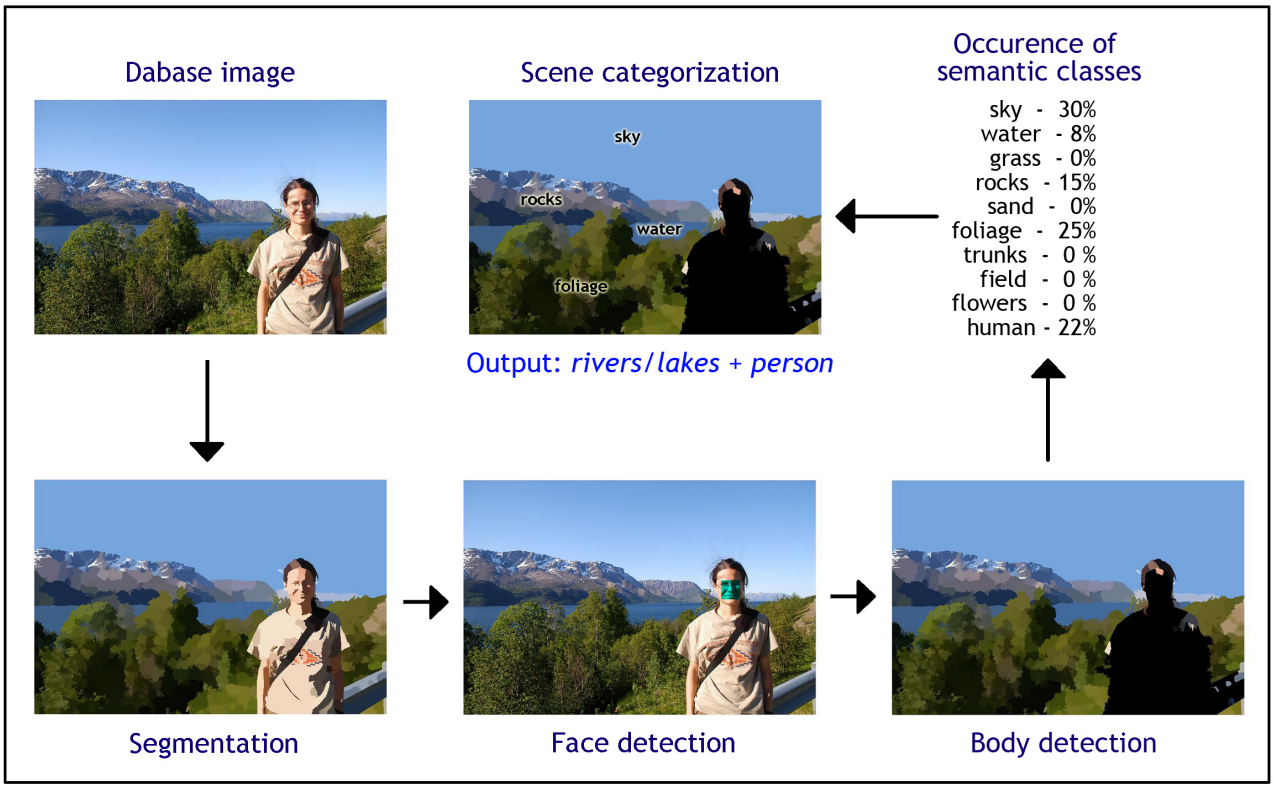


Figure 2: Overview of proposed method

region are calculated. Next different intervals for these features are defined. Each interval can be translated to an intermediate-level descriptor qualitatively describing the region attribute, that humans are more familiar with. These descriptors form a simple vocabulary, the so-called object ontology. Images can be classified into different categories by mapping such descriptors to high-level semantics based on our knowledge. For example "sky" can be defined as region of "light blue" color and "upper" spatial location. A typical example of such ontology-based system is presented in Ref. [8]

**2. Machine learning** This technique is based on using supervised or unsupervised machine learning tools to associate low-level features with query concepts. A supervised learning algorithm analyzes the training data and produces an inferred function, which should predict the correct output value for any valid input data. In unsupervised learning the goal is to describe how the unlabeled input data are organized or clustered. A novel scheme that combines semi-supervised learning, ensemble learning and active learning in a uniform framework is proposed in Ref. [13]

**3. Relevance feedback** Methods using relevance feedback technique work on-line and try to learn the user's intentions on the fly. At the beginning system provides initial retrieval results and the user marks which images he considers as "relevant" and which as "irrelevant". Machine learning algorithm learns the user's feedback and the selector retrieves another images. The process is repeated until the user is satisfied with the results. Mechanism of relevance feedback is well used in Ref. [7]

**4. Semantic template** This technique is not so widely used. Semantic templates are generating to support high-level image retrieval. Semantic template is usually defined as the "representative" feature of concept calculated from a collection of sample images. This technique is used in Ref. [14]

**5. Web image retrieval** This system has some technical difference from image retrieval in other application. Some additional information like the URL of image file or the descriptive text surrounding the image can help the semantic-based image retrieval.

**6. Frequency domain features** Image search and retrieval in this method mainly focuses on feature vectors based on the real and imaginary parts of the complex numbers of the image transformed by the Fast Fourier transform (FFT), Discrete Cosine transform (DCT) or WALSH transform. This technique was recently presented in Ref. [5]

Many systems exploit one or more of the above techniques to implement high-level semantic-based image retrieval. Our system consists of both supervised and unsupervised machine learning technique and semantic template for scene categorization.

# 3 Proposed method

Our work is based on the work [12]. Like in their method also in our proposed method the image is segmented into local subregions. The difference is in the shape of local image subregions. While in the initial method [12] the local image subregions are extracted on a regular grid of 10x10 regions, our proposed method tries to segment the image into arbitrary-shaped subregions, which correspond to objects boundaries. This improvement reduces the misclassification of regular subregions belonging to two or even more semantic concepts.

In addition our proposed method detects presence of people in the image. This is useful because our target images are typical holiday pictures from hiking outdoors. Presence of family members on this kind of images is very common, so it is important to cover also this condition into image retrieval process. So in our system it is possible to define if the retrieval pictures should contain people or not.

Only local subregion that represent nature are further processed. Thus we identify subregions belonging to people and separate them from others. Afterwards using low-level features we classify each subregion into one of following semantic concepts: *sky, water, grass, trunks, foliage, rocks, flowers and sand*. The selection of these local semantic concepts was influenced by the psychophysical studies of Mojsilovic et al. [9] and by concepts used in Ref. [12]. For the classification of local image regions we involved the k-Nearest Neighbor and Support Vector Machine classifiers.

The last stage of this proposed method is scene categorization. In our work we have six different scene categories: *coasts, forests, rivers/lakes, sky/clouds, plains and mountains*. To each local semantic concept, its frequency of occurrence is determined. This information enables us to make a global statement about the amount of particular concept being present in the image e.g. "There is 22% of water in the image." Using this knowledge the most suitable category prototype is assigned to the image, that gives the semantic meaning of the image.

# 4 Preprocessing

Preprocessing consists of image segmentation into arbitrary-shaped regions and detection of people in the image.


Figure 3: An example of the body detection. (a) Results from the face detector with templates (b) Filtered out subregions belonging to humans bodies

## 4.1 Image segmentation

At first step of our algorithm the image is segmented into arbitrary-shaped subregions. We take advantage of Mean Shift segmentation algorithm presented in [3] based on grouping pixels, which are close in the spatial and color range domain. This algorithm iteratively detects modes in a probability density function.

It starts with a region of interest where kernel function calculates the mean shift vector. Using this vector the region is shifted to the new location. Can be shown that the mean shift vector is proportional to the normalized density gradient estimation, so the region certainly converges to a point with zero gradient. This is the mode corresponding to the initial position. Modes that are close to each other are grouped together. For segmentation purposes, each pixel is marked by color value of the corresponding mode.

## 4.2 Body detection

After the image is segmented we used algorithms for skin and face detection to identify subregions belonging to the humans body. We applied skin detection [10] which results in skin probability maps. For face detection we used the implementation of Viola/Jones Face Detector [11] found in [4]. 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 humans body is added to each detected face. Each subregion overlaid by this template is examined if the majority of subregion area lies inside the template or outside. Subregions lying for the most part within the template we consider to depict the humans body. They will not proceed to further processing and classification.

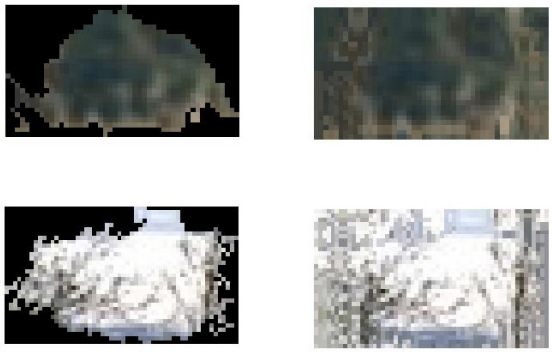

Figure 4: Example of extended subregions

# 5 Semantic Concept Classification

In the second stage three kinds of features are extracted from each subregion. Afterwards the subregions are classified by k-Nearest Neighbor and Support Vector Machine classifiers into eight semantic concept classes.

## 5.1 Color features

The color feature is one of the most widely used visual features in the image retrieval. In our work we use linear $L^*a^*b^*$ color histogram. $L^*$ represents the lightness, $a^*$ the red-green component and $b^*$ the blue-yellow component. Colour histogram describes the distribution of color and lightness within the subregions. The histogram is invariant to rotation, translation and scaling, but does not contain semantic information.

## 5.2 Edge direction features

As the second kind of feature we use edge direction histogram. It is computed by grouping the edge pixels which fall into edge directions and counting the number of pixels in each direction. We are applying the Canny edge operator and consider 4 directional edges (horizontal, vertical and 2 diagonals) and 1 non-directional edge.

Since our subregions are arbitrary shaped we need to apply simple mirror padding to extend region to a rectangular area. Fig. 4 gives an example of extended subregions.

## 5.3 Texture features

Texture is another important property of images that helps in the image retrieval. We combine texture features with other visual attribute, because texture on its own does not have the capability of finding similar images. But it can classify textured images from non-textured ones.

In our work many subregions have same or very similar color, but they do not belong to the same semantic concept class. For example sky and water subregions have both similar shades of blue. Texture features help us classify subregion into the correct class. We applied one statistical and one transformed-based method.

### 5.3.1 Statistical method

We chose method based on co-occurrence matrices [2]. The co-occurrence matrix $C(i,j)$ shows the co-occurrence of gray-valued pixels $i$ and $j$ at a given distance $d$ and given direction $\theta$. In our case $d$ is 1 and $\theta$ takes values $0°$, $45°$, $90°$ and $135°$. Together we have four different co-occurrence matrices from where six texture features are extracted: Energy, Contrast, Correlation, Difference Moment, Entropy and Homogeneity. They are defined as follows:

$$Energy = \sum_i \sum_j C(i,j)^2$$

$$Contrast = \sum_i \sum_j (i-j)^2 C(i,j)$$

$$Correlation = \frac{\sum_i \sum_j (ij) C(i,j) - \mu_i \mu_j}{\sigma_i \sigma_j}$$

$$Difference\ Moment = \sum_i \sum_j \frac{1}{1+(i-j)^2} C(i,j)$$

$$Entropy = -\sum_i \sum_j C(i,j) log C(i,j)$$

$$Homogeneity = \sum_i \sum_j \frac{C(i,j)}{1+|i-j|}$$

where

$$\mu_i = \sum_i i \sum_j C(i,j)$$

$$\mu_j = \sum_j j \sum_i C(i,j)$$

$$\sigma_i = \sum_i (i-\mu_i)^2 \sum_j C(i,j)$$

$$\sigma_j = \sum_i (j-\mu_j)^2 \sum_j C(i,j)$$

### 5.3.2 Transformed-based method

Another texture feature in our work is Gabor Texture Feature. The two-dimensional Gabor filter is defined as

$$Gab = \frac{1}{2\pi\sigma_x\sigma_y} e^{[-\frac{1}{2}((\frac{x}{\sigma_x})^2 + (\frac{y}{\sigma_y})^2) + jW(x\cos\theta + y\sin\theta)]}$$

where $\sigma_x$ and $\sigma_y$ are scaling parameters of the filter, $W$ is the radial frequency of the sinusoid and $\theta \in [0, \pi]$ specifies the orientation of the Gabor filters. Gabor filtered output $FGab$ of the image is obtained by the convolution of the given image $F$ with Gabor function $Gab$ for each of the orientation and scale. The magnitudes of the Gabor filters responses are represented by the mean and standard deviation:

$$\mu = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} FGab$$

$$std = \sqrt{\sum_{x=1}^{X} \sum_{y=1}^{Y} \|FGab| - \mu|^2}$$

The feature vector is constructed using $\mu$ and $std$ as feature components.

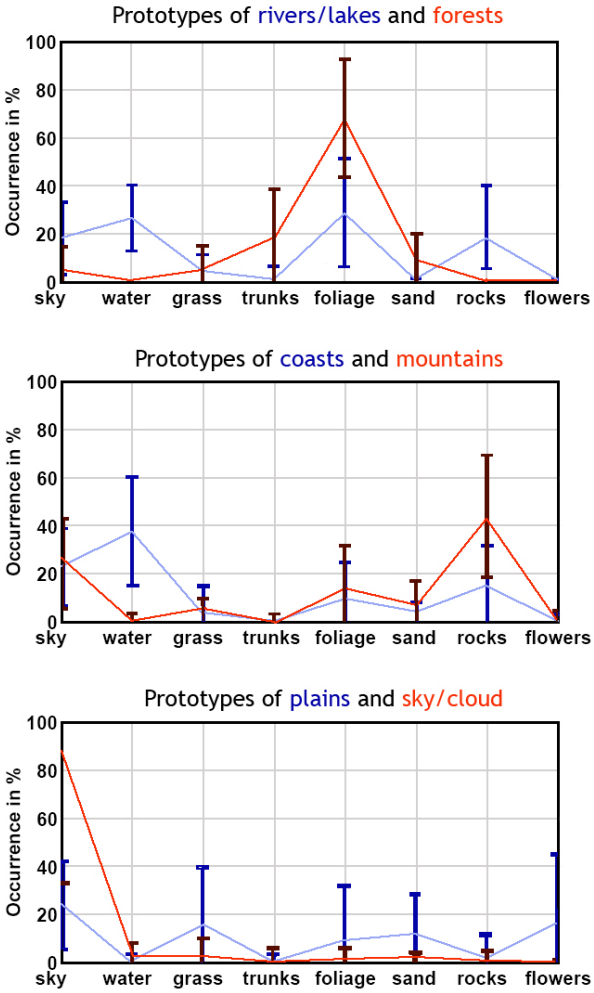


Figure 5: Prototypes of the six scene categories

## 5.4 Classification

We used two methods for the classification of the local semantic concepts, k-Nearest Neighbor and Support Vector Machine classifiers. Same classification methods were used in the initial method [12].

### 5.4.1 k-Nearest Neighbor classifier

The k-Nearest-Neighbor (kNN) classification is one of the most fundamental and simple non-parametric classification methods. For k-nearest neighbors, the predicted class of test sample x is set equal to the most frequent true class among k nearest training samples.

In our work we used the matlab implementation of kNN classifier. We tested several values of $k$. Best results were obtained by $k = 10$.

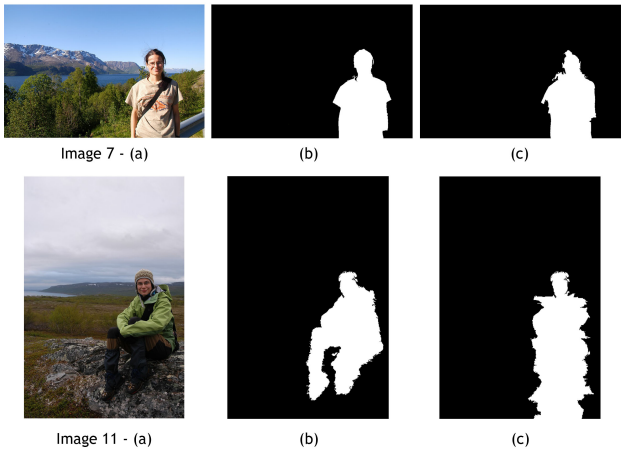


Figure 6: Human body detection (a) Original Image. (b) Manual detection. (c) Obtained result.

### 5.4.2 Support Vector Machine classifier

Support Vector Machines (SVM) are based on the concept of decision hyperplane. The SVM finds a linear separating hyperplane with a maximal margin in the higher dimensional space.

For our experiments, the LIBSVM package [1] with the radial basis function (RBF) kernel was employed. LIBSVM implements the "one-against-one" approach for multi-class classification. For $n = 8$ classes there are $\frac{n(n-1)}{2} = 28$ single classifiers and each one trains data from two classes. Each binary classification is considered to be a voting, where a new data point is allocated to the class with the highest number of votes.

# 6 Scene Categorization

The last 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 (Figure 8 shows an example for each category). We define for each of these categories a category prototype. It is an example which is most typical for the respective category. Figure 5 displays 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.

# 7 Results

This section summarizes the results of the proposed approach.

| Overall **67,8%** | w | g | r | s | s | f | f | t |
|---|---|---|---|---|---|---|---|---|
| water | **71,8** | 1,4 | 7,0 | 1,4 | 14,1 | 4,2 | 0,0 | 0,0 |
| grass | 9,2 | **40,0** | 7,5 | 0,8 | 0,0 | 18,3 | 23,3 | 0,8 |
| rock | 7,4 | 0,5 | **77,5** | 6,9 | 0,5 | 2,0 | 2,5 | 2,9 |
| sand | 0,0 | 6,7 | 23,3 | **53,3** | 10,0 | 3,3 | 0,0 | 3,3 |
| sky | 8,0 | 0,0 | 0,9 | 1,8 | **82,1** | 2,7 | 4,5 | 0,0 |
| foliage | 3,1 | 14,8 | 4,8 | 0,0 | 0,0 | **71,6** | 3,9 | 1,7 |
| flowers | 0,0 | 11,6 | 2,0 | 1,5 | 0,0 | 17,2 | **67,7** | 0,0 |
| trunks | 1,6 | 3,1 | 23,4 | 7,8 | 0,0 | 9,4 | 1,6 | **53,1** |
| Precision | 54,26 | 43,24 | 75,24 | 38,10 | 86,79 | 69,2 | 73,63 | 73,91 |

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

| | |
|---|---|
| Color | 52,3% |
| Co-ocurance matrix | 41,2% |
| Gabor feature | 43,4% |
| Edge direction | 25,3% |
| Color+Co-ocurance matrix | 59,8% |
| Color+Gabor feature | 62,5% |
| Color+Edge direction | 56,7% |
| All features | 67,8% |

Table 2: Low level feature relevance

We measured the quality of human body detection by comparing the obtained results with manual detection. We calculated the overlap and left-out feature. Overlap feature determines what percentage of the manual detection*(MD)* is covered by the obtained result*(OR)*.

$$Overlap = \frac{area(OR \cap MD)}{area(MD)}$$

Left-out feature determines what percentage of the obtained result is not covered by the manual detection.

$$Left-out = \frac{area(OR-MD)}{area(OR)}$$

Our method for human body detection was tested on 15 images and we achieved average Overlap $92,06\%$ and average Left-out $15,42\%$. The method works well if the person is standing straight. It is a typical pose on holiday pictures. If person is sitting or lying some errors may occur. (See Figure 6)

As a next step we tested which low level features are most relevant in classification process. Results obtained using SVM classifier can be find in Table 2. It is obvious that color feature give a good result, but its combination with texture feature leads in even better accuracy.

The ground truth for subregion membership to one of the eight semantic concepts was annotated manually. Together we annotated 1028 subregions. The class sizes vary from 54 (trunks) up to 192 (sky), because sky appears more often in the images than trunks. The classifiers are challenged with the inequality in the class sizes and the visual similarity of image regions that belong to different classes.

| | Class size | Classification accuracy kNN | Classification accuracy SVM |
|---|---|---|---|
| sky | 192 | 77,2% | 82,1% |
| water | 139 | 53,4% | 71,8% |
| grass | 111 | 20,7% | 40,0% |
| trunks | 54 | 43,8% | 53,1% |
| foliage | 166 | 66,7% | 71,6% |
| sand | 103 | 47,6% | 53,3% |
| rocks | 171 | 66,0% | 77,5% |
| flowers | 94 | 57,7% | 67,7% |

Table 3: kNN and SVM classification accuracies

The Table 3 shows that the SVM classification performs better than the kNN classification. We can see a correlation between the class size and the classification result. Sky, foliage, and rocks are the largest classes and they are also classified with the highest accuracy. In Table 1 is displayed confusion matrix of the SVM concept classification.

At the end we discuss results obtained by our proposed method and those obtained by the initial method [12]. Because of using a regular grid in the initial method, rectangular subregions 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 subregions.

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%. An example can be find in Figure 7.

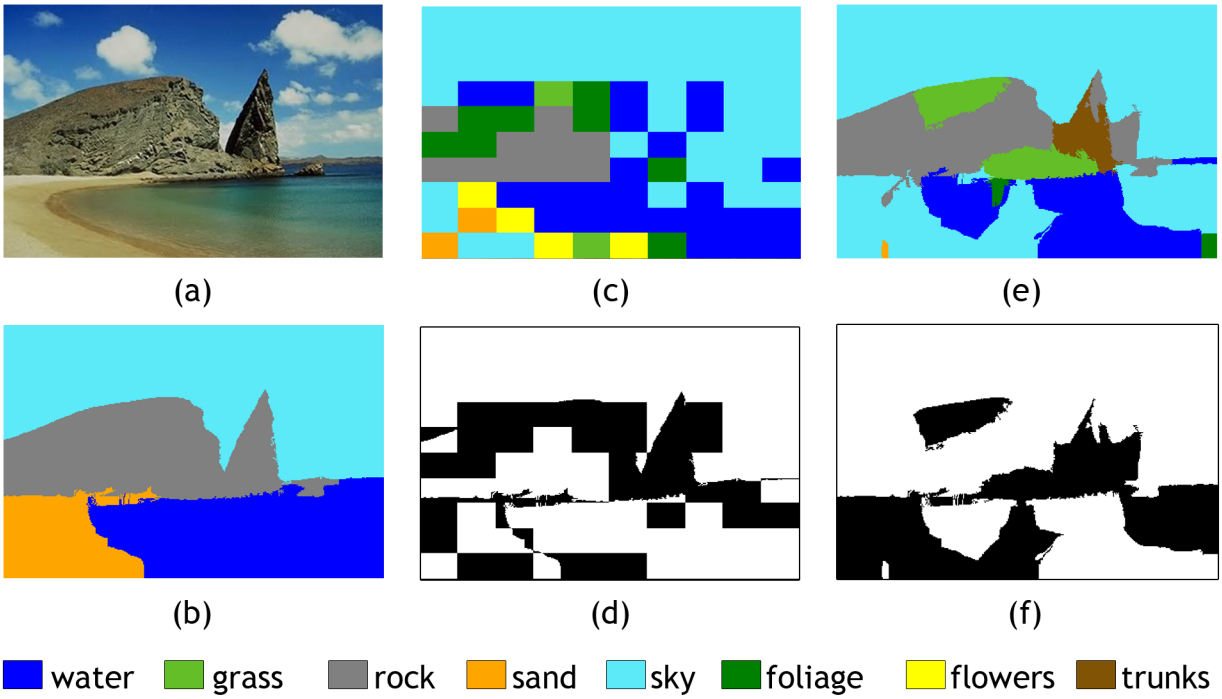


Figure 7: Local semantic concept classification (a) Original image. (b) Ground truth. (c+d) Result of initial method and equality map (e+f) Result of our proposed method and equality map

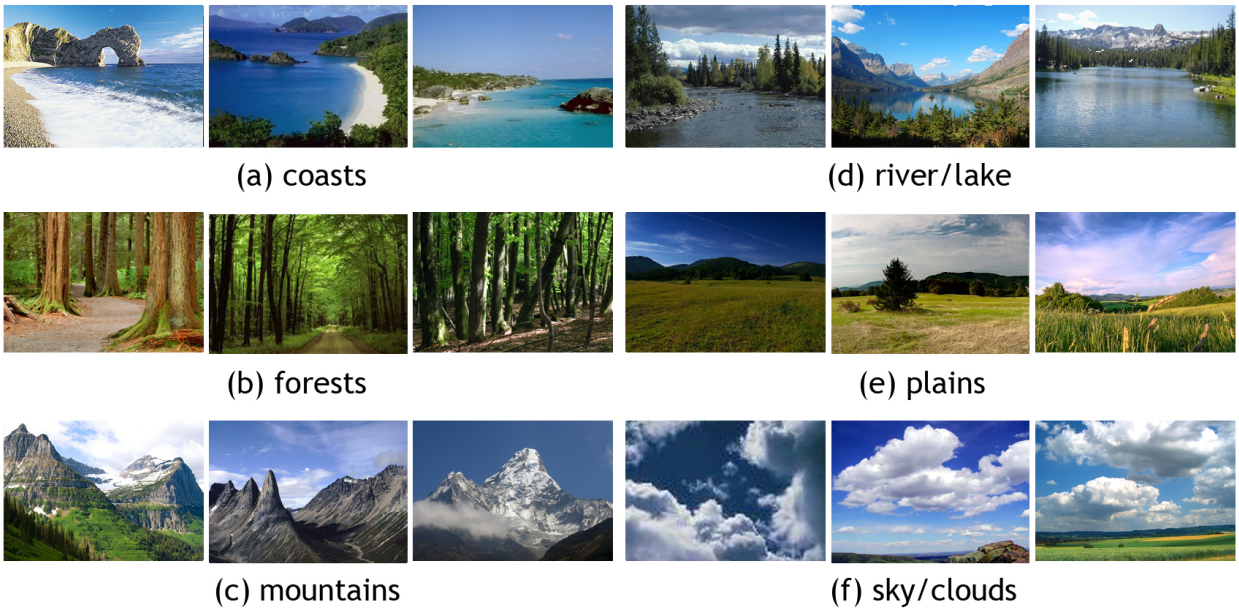


Figure 8: Exemplary images for each category

## 8 Conclusion

We implemented method for semantic categorization and retrieval of natural scene images presented in [12]. Since segmentation into 100 rectangular subregions used in this method can cause mistakes in semantic concepts classification, we modified this method by using segmentation into arbitrary shaped regions.

Our target images were typical holiday pictures from hiking outdoors. Due to frequent presence of family members in these pictures we enhance this method with automatic detection of people in the image.

## 9 Acknowledgment

The author wish to thank Elena Šikudová, PhD. for her support and the excellent leadership in this project.

## References

[1] Ch. Chang and Ch. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] 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.

[3] 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.

[4] V. Kazemi. Face detector (boosting haar features), 2010. `http://www.mathworks.com/matlabcentral/fileexchange/27150\-face-detector-boostinghaar-features`.

[5] H. B. Kekre and D. Mishra. Cbir using upper six fft sectors of color images for feature vector generation. *International Journal of Engineering and Technology*, 2:49–54, 2010.

[6] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262 – 282, 2007.

[7] Jianjiang Lu, Zhenghui Xie, Ran Li, Yafei Zhang, and Jiabao Wang. A framework of cbir system based on relevance feedback. In *Proceedings of the 3rd international conference on Intelligent information technology application*, IITA'09, pages 175–178, Piscataway, NJ, USA, 2009. IEEE Press.

[8] 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.

[9] A. Mojsilovic and B. Gomes, J.and Rogowitz. Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues. *International Journal of Computer Vision*, 56:79–107, 2004.

[10] E. Šikudová. *On some possibilities of automatic image data classification*. PhD thesis, Comenius University, Bratislava, Slovakia, March 2006.

[11] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154, May 2004.

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

[13] J. Wu, Z. Lin, and M. Lu. Asymmetric semi-supervised boosting for svm active learning in cbir. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 182–188, New York, NY, USA, 2010. ACM.

[14] 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.