SCENE CLASSIFICATION USING BAG-OF-REGIONS REPRESENTATION

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Demir Gökalp

July, 2007

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst.Prof.Dr. Selim Aksoy (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof.Dr. Özgür Ulusoy

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof.Dr. Enis Çetin

Approved for the Institute of Engineering and Science:

Prof.Dr.Mehmet B. Baray
Director of the Institute

ABSTRACT

SCENE CLASSIFICATION USING BAG-OF-REGIONS REPRESENTATION

Demir Gökalp
M.S. in Computer Engineering
Supervisor: Asst.Prof.Dr. Selim Aksoy
July, 2007

Significant growth of multimedia data creates the need for more complicated approaches in image understanding, classification and retrieval. Semantic scene classification is a popular research area which categorizes images into semantic categories for applications like content based image retrieval. In the near future, content based image retrieval will be much more important especially for the next generation internet technologies so new approaches are very welcomed in this subject. Research has showed that classifying images using components like regions, pixels or objects is a challenging work because of the ambiguity of the visual data. The main idea about image classification is to find similarities between these components to get information about the content of the image. This thesis describes our work on classification of outdoor scenes. As the first step, regions are extracted using one-class classification and patch-based clustering algorithms. The components (pixels, regions and objects) in outdoor images have particular spatial and geometric interactions so dividing images into meaningfully clustered regions has important benefits for a detailed content analysis. For region clustering, features from different levels make specific contributions but to avoid the ambiguity, we need to use low level information and more global information together for the clustering step. Also, using spatial relationships between clustered regions, we can make inference about the detailed content of outdoor images from specific to general. Therefore, after rough segmentation, scene representations are constructed with and without spatial information. At the final step Bayesian classification approach is used with the two different scene representations. The developed methods were tested on the MIT LabelMe dataset, and the results showed that using regions and their spatial relationships improved the classification accuracy.

Keywords: Semantic scene classification, region segmentation, image understanding.

ÖZET

BÖLGE GRUPLARI KULLANARAK RESİM SINIFLANDIRMA

Demir Gökalp
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Asst.Prof.Dr. Selim Aksoy
Temmuz, 2007

Görsel verinin son yıllarda her alanda kullanımının giderek artmasından dolayı, içerik tanıma ve sınıflandırma konularında daha verimli ve kullanışlı yöntemlere ihtiyaç duyulmaktadır. Resimleri içeriklerine uygun olarak gruplamayı hedefleyen anlamsal resim sınıflandırma konusu, son yıllarda üzerinde daha fazla çalışılır hale gelmiştir. Genel olarak sınıflandırma metodları; resim içindeki önemli obje ve bölgelerin tanımlanması ve daha sonra bu bilgiler kullanılarak resim içeriği hakkında genel bir yargıya varılmasına dayanır. Özellikle dış mekan resimleri içerisinde tanımlanan bu tip bölgelerin birbirlerine olan uzamsal konumları resim içeriği anlama ve sınıflandırmada çok önemli katkı yapmaktadır. Bu noktada karşılaşılan en önemli sorun görsel veride gürültü ve görüş açısı gibi nedenlerden dolayı meydana gelen tutarsızlıklardır. Bu tip sorunları önlemenin bir yolu farklı düzeyde bilgilerden yararlanmak olabilir. Bu çalışmada bu amaçla biz hem düşük seviyede hem de orta seviyede öznitelikler kullanarak, resim içinde bölge bölütlemesi yapıyoruz. Daha sonra ayrılan bu bölgeleri uzamsal konumlarına göre tek başlarına veya ikili olarak kullanarak resimleri modelliyoruz. Sahne sınıflarında baskın olan tekli veya çiftli bölgeleri belirlemek için geliştirdiğimiz bir seçme algoritmasını kullanmayı sonucu iyileştirmeye yönelik bir alternatif olarak sunuyoruz ve algoritmamızın en son aşamasında sınıflandırma tekniği olarak Bayesci sınıflandırma yöntemi kullanarak resimlerin sınıflarını belirliyoruz. Geliştirilen yöntemler MIT LabelMe veri kümesinde denenmiş ve bölgeleri kullanarak elde edilen uzamsal bilginin sınıflandırma sonuçlarını oldukça iyileştirdiği gözlemlenmiştir.

Anahtar sözcükler: Anlamsal resim sınıflandırma, bölütleme, resim anlama.

Acknowledgement

I would like to express my gratitude to Asst.Prof.Dr. Selim Aksoy, from whom I have learned a lot, due to his supervision, suggestions, and support during this research.

I also would like to thank to the members of my thesis committee; Prof.Dr. Özgür Ulusoy and Prof.Dr. Enis Çetin for reviewing my thesis.

This work was supported in part by the TUBITAK Grant 104E077 and European Commission Sixth Framework Programme Marie Curie International Reintegration Grant MIRG-CT-2005-017504.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Image understanding and scene classification, which are about interpreting the content of an image and categorizing similar ones into predefined scene classes, is one of the important problems in computer vision. According to the quality of the digital media, it is a quite problematic study especially at the feature extraction step. This process needs to extract information from different levels of data like recognizing the particular regions, objects in an image or just trying to interpret concept of the image using more global information (like general color or texture distributions). In scene classification problem, to understand the image content and classify into an abstract scene class, we believe that using some particular indicators which are extracted from low and intermediate level is more significant than using global information. These indicators can be regions, objects or patches which are related to the general concept of the image. Each specific component helps to understand the image content such as finding a table or phone object increases the possibility of the scene being an office while a sky and water region possibly means that the scene contains a beach or waterfront. At this point the main problem is the difficulty of region and object detection because of ambiguity in visual data [6]. There are several reasons for ambiguity

in visual data like illumination and camera position. Generally, different kinds of regions can be very similar at some levels such as water and sky regions can have similar color information in the RGB color space depending on the camera position or illumination. To avoid this kind of resemblance, additional information especially neighboring components in the image and their spatial relationships can be helpful. Water and sky regions can be similar on the color space but they would be separated at the texture feature space and also it is very reasonable to expect that sky must be on top of the water.

It is very well known that spatial dependencies are not random in natural images so spatial information especially the vertical one is very helpful for the analysis of natural scenes. In [19] Smith and Li demonstrated the contribution of vertical relationships especially in natural scenes by dividing images into vertical parts and extracting region strings vertically for classification. They explain the importance of vertical relations by flipping an image vertically and horizontally. When an image is flipped horizontally, spatial relations of the regions remain the same. On the other hand, if the image is flipped vertically, horizontal relations remain the same but vertical relations change. For example, in Figure 1.1 the sand region still remains near water and both are under the sky region for the horizontal flip (first row). However, for the vertical flip (second row), sky remains under the sand and water regions so the relationships are no longer intuitive. Besides that, spatial dependencies can be local and global, and both types are useful at different information levels. Local dependencies are the pixel based relations and they have interactions in the different parts of an image whereas global dependencies consist of the relations between larger regions in the image content. Modeling the content based relations in images is a problem because of the statistical variations and uncertainties in visual data but the strong side of the probabilistic methods is that they create a flexible framework by using global probability distributions defined by local constraints.

In this thesis we present a framework which combines local and global information at the clustering phase and uses probabilistic approaches for classification phase. As the first step, we developed a new approach for image segmentation. The segmentation approach includes two parts which are one-class classification

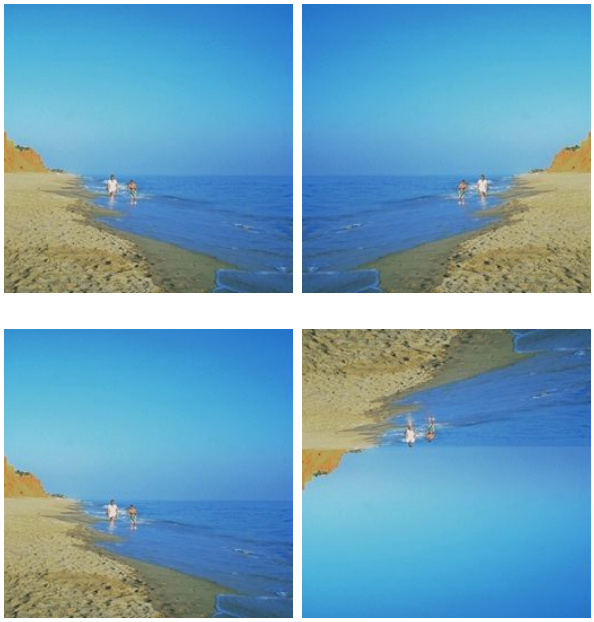

Figure 1.1: Images on the left column are original images, on the right column are flipped versions. Fist row is flipped horizontally and second row is flipped vertically.

and patch-based clustering. One-class classification detects "homogeneous" color and texture regions that mostly exist in natural images like water, grass, sky etc. Second phase of the segmentation process is the patch-based clustering which aims to find structured and more complicated regions like man-made structures where one-class classifiers are insufficient. Our segmentation process gives a roughly segmented image. At the next step, HSV histograms of the regions are extracted and regions are clustered to construct a codebook. We are expecting that similar regions belong to similar clusters. Each region cluster represents a codeword in the codebook. Next step is the representation of the scenes. We created two different scenarios. One is the representation of scenes with bag-of-regions while the other is the representation of scenes with bag-of-region pairs. For the second model, a spatial model is constructed to exploit the contribution of the spatial information. Above-below relations of the regions are modeled. We want to use region types which are dominant or characteristic for particular scene categories,

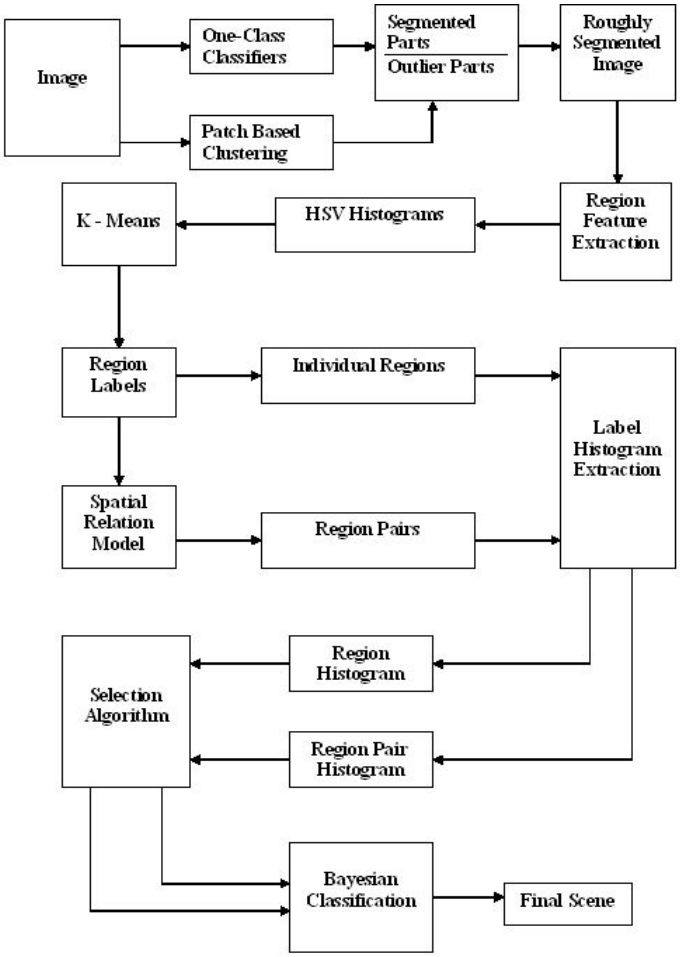

Figure 1.2: Diagram of our classification framework.

so a selection algorithm was developed. Selection algorithm determines the region types which mostly exist in particular scene categories while not often exist in other categories. At the classification process, Bayesian classifiers are used to model both individual and pairwise relations of regions. Figure 1.2 summarizes our classification framework.

1.2 Scene Classification

Scene classification is a very popular field in computer vision because of the need in management of huge amounts of visual data. Data retrieval is only one

example application of scene classification. Traditional methods for retrieval is to index images with labels according to their content. Difficulty here is to annotate each image by a person. This is very inefficient for large databases and also reliability of the system depends on the understanding of the annotater. If the visual data can be categorized by algorithms according to the content, efficiency and reliability of the applications (like data retrieval) can be increased. So far, some systems were implemented for this purpose but still there are no satisfactory results on this subject. There are approaches with and without using segmentation process. One of the main problems in computer vision is image segmentation so it is also an important problem of the image classification. Some segmentation methods were implemented like N-cut at [17] which is a well known segmentation method but like all methods it only works on specific datasets. Figure 1.3 shows an N-cut result. The problem of the N-cut algorithm is to determine the best N value which is the number of segments in an image. Determining a fixed N-value is not very reliable so N-cut gives satisfactory result only for specific images with specific N-values.

Figure 1.3: Examples for N-cut segmentation result.

Generally research is focused on the segmentation process for a detailed scene classification. The idea is that recognizing the regions in images helps to understand the whole concept of the image. When recognizing the regions in images, modeling their distributions or relations gives the general idea about the scene. The disadvantage is the difficulty in the segmentation process as mentioned. In recent years, another new approach is to find scale invariant local features. These

features can be extracted from patches around the detected keypoints. Importance of the patches is that they give intermediate-level information about the content and their distributions can be modeled in scene categories without performing any segmentation process. This kind of approaches have satisfactory results in limited scene categories like indoor-outdoor classification. Success of the method depends on the success of the feature detector or descriptors. The disadvantage of the method is that in natural images, distribution of the scale invariant descriptors can be very random. The descriptors are more reasonable on organized structures like man-made structures so randomness of the descriptors in regions like grass or sky can be problematic to represent scene categories. Furthermore, there are very old-fashion approaches which extract the color or texture distributions of the whole images and use as features. However, these approaches are very insufficient to have detailed idea about the image content and they fail in the complex datasets. These show that using only low, intermediate or high level features generally give satisfactory results only on limited datasets so the idea of using combination of the different feature types together makes more sense.

Another important subject in scene classification problem is the spatial relationships in images. Especially in outdoor images, this kind of information is very helpful. In outdoor images randomness of the spatial relations is not very possible. In an indoor image, you can expect to see any kinds of object with different spatial relation combinations. For example, a computer monitor can be under a table or on top of a table or there can be a coffee cup on a book or near the book. Extracting these kind of spatial constraints is very challenging for indoor images. However, the situation is very different for outdoor images. Intuitively, it is not possible to see a sea or grass region on top of a sky region. These kinds of constraints are very useful to construct reliable spatial models. We can assume that spatial relations in outdoor images are rational and not random. Also, their direction is very helpful for understanding the content. For example, finding a direction with sky-water-sand means a beach scene very strongly. Using this kind of information obviously increases the performance and reliability of the classification system. New research has been focusing on the use of spatial

relations more often. Commonly, used approach is to model the above-below relations because vertical relation information is more distinguishing for outdoor images.

1.3 Organization of the Thesis

The organization of the thesis is as follows. Chapter 2 summarizes the related background work about the scene classification problem. Segmentation of images into regions is described in Section 3. Chapter 4 is about the general classification process. Representation of scenes using "bag of individual regions" and "bag of region pairs" are presented in Section 4.1 while the algorithm for selection of discriminative regions is proposed in Section 4.2. Bayesian models for scene classification are described in Section 4.3. Experiments on the LabelMe data set are presented in Chapter 5 and are followed by conclusions in Chapter 6.

Chapter 2

Related Work

Human brain is the best recognition and classification system. While the most part of the brain and its working is still a big mystery, researches have showed that human brain can easily recognize the content of a complex scene. The general method is first to recognize objects in the images and then making connection with them to have the final content idea. Brain tries to make induction from local to general. This working strategy of the human brain created a very good model for some of the modern scene classification studies. There is an important amount of research about scene classification in the last decade. Earlier studies used global information of the whole images like histograms. These histograms were used as feature at the classification phase. Then, research was focused on object recognition in images and detected objects were used for scene classification. On the other hand, more recent approaches tried to model the general image concept using local descriptors instead of recognizing the objects or global features of the whole image. Distributions of these descriptors are used to represent the scene categories.

We can divide the related research into two main categories which are methods using global features and methods using local information. These two approaches have own advantages and disadvantages so we tried to combine strong sides of these approaches in our research.

2.1 Global Approaches

The most traditional scene classification approach is to use low-level global features like color and texture distributions of the pixels. These features are extracted from whole images and each image is represented by these global features. Using such features, images are classified into very limited classes like indoor/outdoor.

There are many researches using low-level features to make high level inference about scene context. Gorkani and Picard [1] studied the classification of images into city and landscape classes. Multiscale steerable pyramid is used and they try to find dominant orientations in 4 x 4 subblocks in images. The classification is simple; to classify as city, an image needs enough number of subblocks with the dominant vertical orientation. If image cannot be labeled as city, it is landscape. Yiu [2] created a similar approach to Gorkani and Picard. Vertical dominant orientation features are used with the color information. At the classification step they used the nearest neighbor and support vector machine classifiers. Their final scene classes are outdoor and indoor. Lipson [3] describes a general scene query approach; it can not be accepted as scene classifier. According to regions and their relations in images, graphical representations of the scenes are extracted. These graphical relationships contain relative color, spatial location, and highpass frequency content. The difficulty is that graphical templates must be extracted by hand for each scene class. These templates are also quite specific, which makes them suitable for limited special cases such as sky over mountain over lake but difficult for the case of capturing a broad concept like an outdoor scene. In an other research, Yu [4] learns a statistical template from examples. Vector quantized color histograms are calculated for subblocks of the image. Then training a one dimensional hidden Markov model along vertical or horizontal segments of specific scene layouts is done. The problem is that one dimensional model cannot describe spatial relationships effectively.

In a more recent research Vogel and Schiele [5] described a framework that uses color, texture and a spatial grid layout. They aimed to perform landscape scene retrieval system which is based on a two level retrieval. This two level system

enables the use of a semantic level of block classification to do retrieval that depends on the occurrence of concepts in an image. Discriminative random fields were used by Kumar and Hebert [6] to detect and localize man made structures in a scene and then representing images with these structures for classification and retrieval.

2.2 Patch Based Approaches

More recent and popular researches focused on local information on images. This local information is generally extracted by scale invariant feature detectors and images are represented by local invariant features. Scale invariant features are independent from the scale and orientation of the image so they are very useful as descriptors.

Fei-Fei and Perona at [7] extracted the patches in images to construct a codebook. Distribution of the each patch type is learned from the codeword distribution. The method depends on the probabilistic co-occurrence of the patches and it does not need any segmentation process.

Fergus et al.[15] created a joint model which includes scale invariant model and a spatial distribution model whereas Dorko and Schmid [16] used a feature selection algorithm to determine dominant local descriptors in specific object categories.

Lazebnik et al.[8] used the patch histograms and spatial information together. She created a grid from image and then calculated the histogram of patches in each grid. This caused the limited use of spatial information.

Monay et al. [11] used probabilistic aspect models on the bag-of-words representation with the assumption that there must be correlation between the aspects and semantic classes. Their models worked on classification of man-made and natural scenes. Boutell et al. [9] developed a similar approach to our approach, they used regions and their pairwise relationships as spatial information. Gemert et

al. [14] developed an alternative codebook model which used patch histograms for scene categories. The difference from the codebook method is that they used the similarities to all vocabulary elements.

It is very obvious that patches are very helpful about the content analysis but the disadvantage is to have only local information. To improve the contribution of the local patches, using spatial content where both local and global information are used together creates more reliable results.

Chapter 3

Finding Local Regions

We believe that low level local information has important contributions to the overall classification method so creating a well performed segmentation algorithm is the first step of our study. First of all, we aimed to get roughly segmented images in which important and dominant regions are segmented, because this kinds of regions gives detail about the general image content. Trying to recognize all regions is not necessary and costly so we are just trying to extract large and characteristic areas in the images. We do not need to segment all regions detailed with clear borders because to find the important part of a region is enough to extract features at next steps. On the other hand, all kinds of segmentation is a very though and complex process in computer vision. There are many kinds of methods like color based methods, geometric methods, texture based methods, etc., but all methods generally give satisfactory results only on the specific and restricted datasets. For reliable segmentation results, simple images with clear backgrounds and compact structures should be used whereas more complicated images fail at the segmentation step because of the physical conditions and intrinsic nature of the data.

In this thesis, we are using our segmentation approach which is the combination of two different methods. These two methods have weak and strong sides so we aim to combine strong parts in a reasonable framework. We decided that

defining some common region types for natural scenes will be very helpful for segmentation. Regions like water, sky and green areas exist in most of the outdoor images so predefining these kinds of common regions make the segmentation process easy. Predefining all region types is impossible and inefficient, especially for the regions belonging to structured areas but to find some very common regions at the first step can be very helpful. For predefined semantic classes, we used color and texture based classification while patch based local approach is applied for more complicated regions (like man-made structures). For the natural regions like sky and water, using low-level information like color and texture is very reasonable but it does not work on complicated regions like cars and buildings. Complicated components need more local and complicated information. This is why we are using a two level approach for segmentation. The steps of this two level segmentation algorithm is as follows. First, we defined most possibly existing semantic classes for outdoor images like sky, water, sand etc. These classes have different color and texture distributions. Classes like sky and water can have similar color information but they will be separated on the texture space. Combination of the color and texture information is used with one-class classification algorithm and these classes were determined firstly. Secondly, outlier regions where predefined semantic classes were not found are given as input to the patch based clustering. Aim of the patch based approach is to use information which is not random on the images like man-made structures. The notion of the segmentation at this stage is to have roughly segmented images with reasonable regions. As a summary, our proposed algorithm is; first, one-class classifiers segment the regions that consist of pixels with relatively uniform color and texture properties (Section 3.1), then, finding keypoints on the outlier regions, extracting patches and clustering of patches to detect more complicated regions in the remaining image (Section 3.2). We are not labeling regions in these steps, we are just extracting regions and to prevent noise and computational load, we are eliminating regions according to their sizes.

3.1 One Class Segmentation

One class classification is an approach to the classification problem developed by David M. J. Tax during his Ph.D. thesis studies [13]. One class classification, also called as data description is a special classification problem. In traditional multi-class classification, classifiers are trained using training data samples from each class in order to estimate a model that will provide the decision boundaries. We hope that these decision boundaries will separate observations of different classes in the feature space into regions that do not overlap much. From this aspect, traditional multi-class classification can be said to define differences among observations of different classes.

On the other hand, in one class classification, we try to define properties of only one class. The aim is to discriminate observations of the target class from all other observations called the outlier class [13]. So, a testing sample is either detected as belonging to the target class or it is rejected.

In real word classification problems, sampling a sufficient number of training data from each of the classes is not always possible. In some cases sampling one of the classes might be very difficult, and even in some cases it might be impossible. One-class classification is a method that can effectively be used in such cases since it only needs data from the target class to be trained. Using this data, a one class classifier divides the feature space into two regions one of which covers the region that best defines the target class. Since only samples from a single class is assumed to be available, in order to train a one-class classifier a high number of training samples from the target class is needed. Although it is not necessary, the performance of some one class classifiers can further be improved by supplying them with some samples from the outlier class [13].

All one class classifiers take a parameter that is called the rejection threshold. This rejection threshold is used when describing the region for the target class in the feature space. For example, setting a rejection threshold of 0.1 means that only %10 of the training samples are allowed to reside out of the decision region of the target class. Since the rejection threshold is an upper bound on the error

rate of the classification of training observations, the error rate for the testing observations is independent from it. In our method, we tried some different threshold values and selected 0.05 as the best threshold value.

One class classification aims to minimize the classification error for both target and outlier classes. In an one-class classification problem, the total error is the sum of samples belonging to the target class labeled as outlier and samples belonging to the outlier class labeled as target. In the absence of training samples from the outlier class, one method of minimizing the total error is to minimize the decision region for the target class. We have seven different classifiers for our predefined region classes which are sky, water, road, sand, rock and vegetation. Application of these classifiers is as follows. Each classifier is applied to each pixel in test images. If all classifiers give the outlier label, the pixel is outlier otherwise it belongs to the target class whose probability is the highest. Figures 3.1 and 3.2 show some one class classifier results in our dataset. To create the one-class classifiers, we used Gaussian mixture models. Its parameters can be calculated by the Expectation-Maximization (EM) algorithm. The probability density function of a mixture model with k components for the feature vector x is defined as

$$p(x) = \sum_{j=1}^{k} \alpha_j p(x|j) \tag{3.1}$$

where α_j is the mixture weight and $p(x|j)$ is the Gaussian density model for the j'th component.

$$p(x|j) = \frac{1}{(2\pi)^{\frac{d}{2}} |\sum_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_j)^T \sum_j^{-1}(x-\mu_j)} \tag{3.2}$$

where μ_j is the mean vector and $\sum_j$ is the covariance matrix for the j'th component and d is the dimension of the feature space. As the feature vector, RGB, HSV and Gabor texture values of the pixels are used.

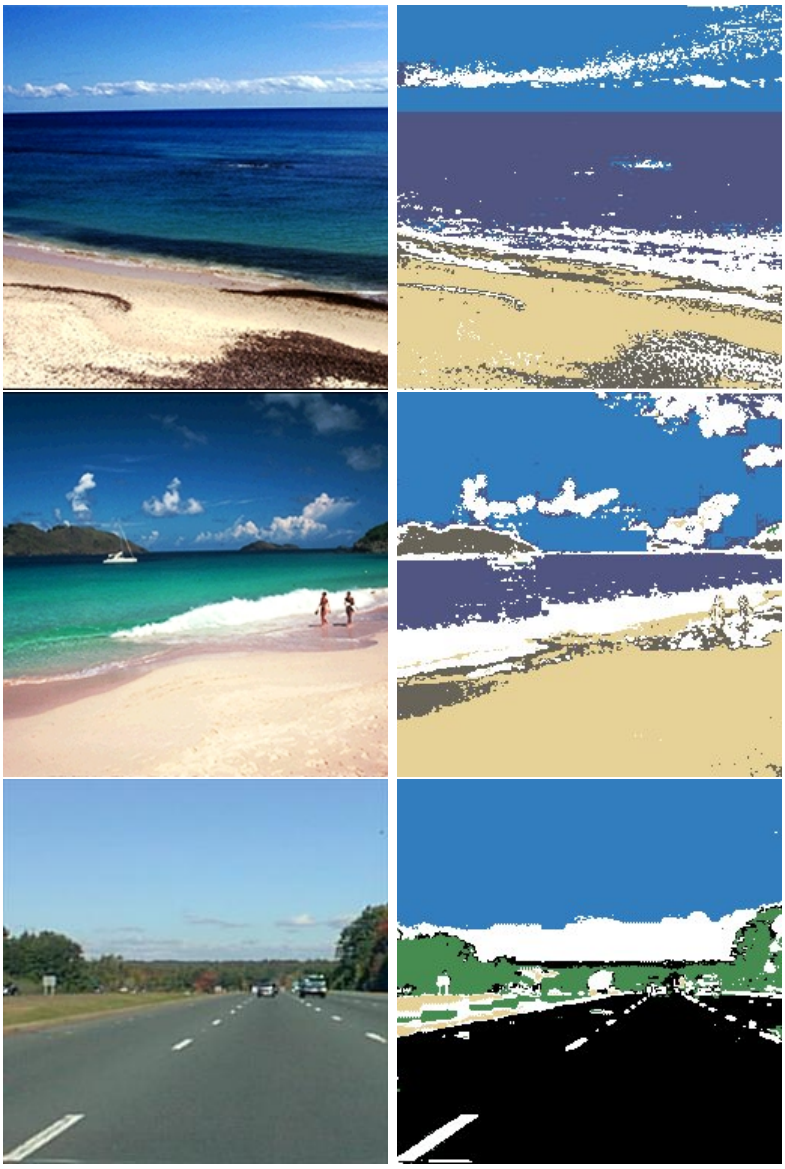

Figure 3.1: Examples for one-class classifier-based segmentation. White pixels are outlier class, blue pixels are sky, purple pixels are water, gray pixels are rock and beige pixels are sand.

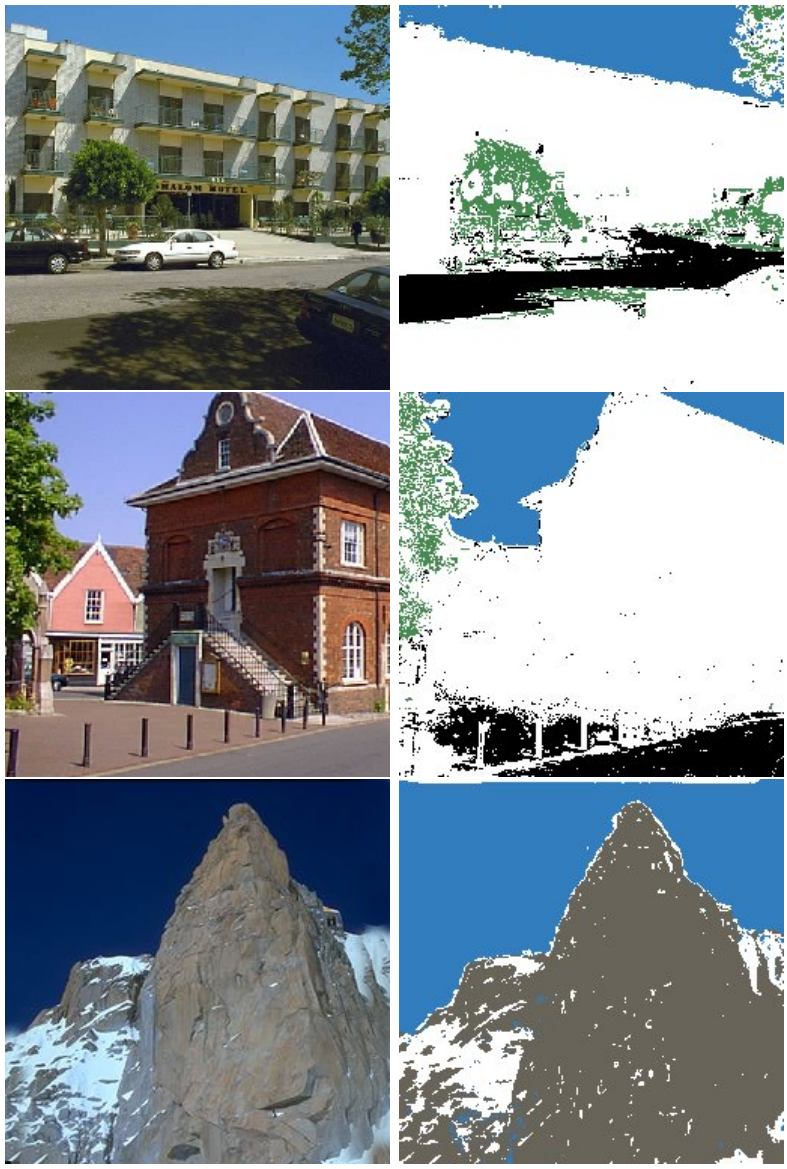

Figure 3.2: Examples for one-class classifier-based segmentation. White pixels are outlier class, blue pixels are sky, purple pixels are water, gray pixels are rock and beige pixels are sand.

3.2 Patch Based Segmentation

Patch based image segmentation or object detection and recognition algorithms are newly popular and very effective approaches. The main idea is to extract important patches and using these patches as features at the classification step. There are many kinds of different patch extraction and classification methods in the literature. General approach is to find important key points in the image and then crop patches around these points. In this research, we are finding key points or interest points in the images using two different keypoint detectors; one is David Lowe's SIFT (Scale Invariant Feature Transformation) algorithm [10] and the other is the Kadir-Brady detector [18]. Our main algorithm is implemented using the SIFT approach but we also used the Kadir-Brady detector as an alternative approach. When we compared the results of these two detectors on the same images, we decided that SIFT gives more reliable keypoint results so our final experiments are made using the SIFT approach.

Method of extracting distinctive features is presented by David Lowe in [10]. These features are generally used for matching and tracking but in recent years, they have been also used for object recognition. These are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise [10]. In addition, the features are distinctive, which allows a single feature to be correctly matched, providing the basis for object and scene recognition. Extraction of the features has some steps;

1. Scale space extrema detection: Searches on all scales and locations using the difference of Gaussian function to identify potential interest points which are invariant to scale and orientation.

2. Keypoint localization: Model is fit to all candidate locations for finding location and scale.

3. Orientation assignment: One or more orientations are assigned to each

keypoint location based on local gradient directions.

4. Keypoint descriptor: Local gradients are measured at the selected scale in the region around each keypoint.

This is the approach of scale invariant feature transform (SIFT) which transforms image data into scale invariant coordinates [10]. Our aim for using patch

Figure 3.3: Results of keypoint detector.

based approach is to narrow the boundaries of outlier class. The keypoints that are outside the regions labeled as outlier are ignored because we already segmented those regions with one-class classifiers. Actually, we are not doing a recognition process; we are just saying that the probability of being a man made regular structure (buildings, cars, bridges etc.) is very high on those regions. Figure 3.3 shows all keypoints found on the example images.

The main problem at this step is the polysemy like having similar keypoint

descriptors from different structures. To solve this problem we used color information in patches. When keypoint detector finds the keypoints in the outlier part of the image where one class classifiers labeled as outlier, 16x16 patches are extracted around these keypoints. The reason to apply keypoint detector on the outlier part is to get rid of unnecessary keypoints coming from possible natural regions. The priority is on the one class classifiers and keypoints are detected on the outlier parts. After extracted patches, they are divided into four 4x4 quadrants. RGB and HSV averages of the pixels in each quadrant are calculated and a 24 feature descriptor is created for each patch; texture information is not preferred because of the insufficient size of quadrants.

The basic idea and assumption is that the keypoints with similar features must belong to the same structure so we clustered keypoints using k-means with k = 25. Value of the k is selected after different runs. An important step in the clustering process is to preserve rotation invariance because keypoints from different parts of the same structure can have similar features but in different quadrants. An example is shown in Figure 3.5.

This invariance is achieved by considering four possible rotations of the quadrants in the computation of the Euclidean distance between the descriptors of two keypoints, and taking the rotation corresponding to the smallest distance as the degree of similarity between these keypoints. After the keypoints are clustered, final step is the dilation of the keypoints to form regions. We used morphological dilation and each keypoint region grows iteratively on the outlier area. This process merges the little regions with same labels and growing process is done only on the outlier parts. Figure 3.4 shows patch based segmentation results and Figure 3.6 shows the all levels of our segmentation approach.

Another alternative approach for keypoint detection is the Kadir and Brady detector which uses Affine Invariant Scale Saliency method [18]. The main idea of the Kadir and Brady approach is that salient image regions exhibit unpredictability in their local attributes and over spatial scale. There are three main parts of the method: First is the calculation of Shannon entropy of local image attributes over a range of scales. Second is the selection of the scales at which the

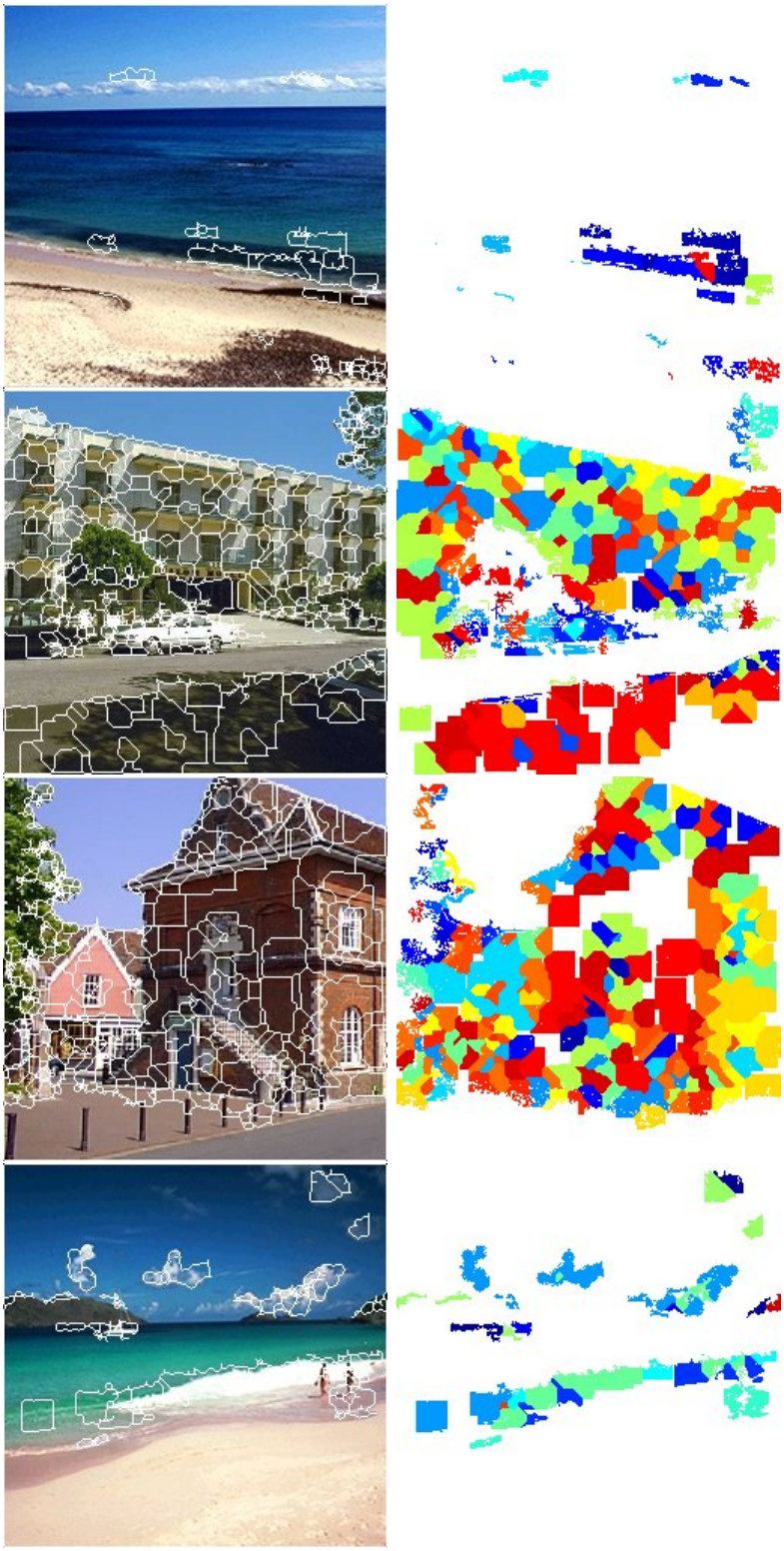

Figure 3.4: Examples for patch based clustering results in outlier part. First column shows the overlayed regions on the original images. Second column shows the patch based clustering results in false color.

Figure 3.5: Examples for different patches from the same structure.

entropy over scale function exhibits a peak and the third step is calculation of the magnitude change of the PDF as a function of scale at each peak. The final saliency is the product of entropy and magnitude change at each peak. Generally, entropy is measured for the gray level image intensity but other attributes, e.g. color or orientation, may be used instead.

The entropy of local attributes measures the predictability of a region with respect to an assumed model of simplicity. In the case of entropy of pixel intensities, the model of simplicity corresponds to a piecewise constant region [18].

In the second step, scales are selected where the entropy is peaked. Through searching for such extreme, the feature-space saliency is locally optimized. Moreover, since entropy is maximized when the PDF is flat, i.e. all present attribute values are in equal proportion, such peaks typically occur at scales where the statistics of two (or more) different pixel populations contribute equally to the PDF estimate [18].

For example, in a noise image the pixel values are highly unpredictable at any one scale but over scale the statistics are stationary. However, a noise patch against a plain background would be salient due to the change in statistics. We can interpret the role of the inter-scale unpredictability measure as a weighting of the entropy value such that some pixel permutations are preferred over others. It is defined as the magnitude change of the PDF as a function of scale. This detection approach is used at the patch based segmentation step to find keypoints

Figure 3.6: Final segmentation results. First column shows the original images, second column shows the results of our segmentation process in false color.

as an alternative to the SIFT. Visual interpretation of the results showed that SIFT gives more reliable keypoints so we integrate the SIFT as primarily keypoint detector. Figure 3.7 shows some results of SIFT and Kadir&Brady detectors.

Figure 3.7: Results of two keypoint detectors. First column shows Kadir-Brady keypoint results with scale circles. Second column shows the Kadir-Brady keypoint coordinates while the third column shows SIFT detector keypoint coordinates.

The main problem for the segmentation process is the over segmentation. At the patch based approach, huge number of regions are existing. We just want to use big and compact regions from patch based approach so we are using a growing algorithm to our patch regions to get regions as compact as possible. Thresholding is the other preprocessing step to get rid of small regions. Still contribution of the patch based regions is fuzzy in the framework so far. To understand the contribution, at the experiment stage, we run our framework

with and without patch based segmentation. Results showed that unless it seems over segmented, patch based regions have contribution on the scene classes like street and inside city. These kinds of regions create the contribution in scene categories where predefined classes in one-class classification are not dominant. Detailed results are given in Chapter 5.

Chapter 4

Bayesian Scene Classification

4.1 Scene Representation

After the segmentation step, scenes can be represented by regions and their features. In this research we used color information and spatial relationships of the regions for scene representation.

4.1.1 Region Features

Features of the regions are the multivariate HSV histograms. A histogram is constructed by 8 bins for the H channel, 3 bins for the S channel and 3 bins for the V channel. Total feature dimension of a region is 72.

4.1.2 Region Vocabulary Construction

To construct the region vocabulary, regions are clustered by the k-means algorithm. This process assigns a discrete region type (codeword) to a region. The whole codeword set creates our vocabulary or codebook. Note that at the segmentation step which used one-class and patch based methods, there was

no recognition, i.e. that process just created the connected components in the images.

When features are extracted from each region, the clustering algorithm is applied to the regions. We preferred to use k-means algorithm with $k = 50$. At this point k represents the number of codewords in our codebook. For the k-means algorithm, determining the k value is the main problem and the common solution is to determine the k-value empirically. For this research, we preferred to use value of k as 50 after different runs with different k values. We tried 30, 50, 75 values and decided to use 50. There is no an important difference on the overall results while k is equal to 30, 50 and 75. More detailed results are shown in Chapter 5. Figure 4.1 shows some regions which are in the same cluster after k-means.

Figure 4.1: Region clustering results. Rows represent the clusters.

4.1.3 Spatial Modeling of Regions

For content analysis, spatial relationships of the components in the image have very significant benefits. In this research we worked on outdoor images and spatial information in outdoor images is more significant. Because of the

spatial interactions between regions like sky, water, sand and road, modeling spatial relations contributes to the understanding of image content very much. We are considering only vertical spatial relations because especially in outdoor images, vertical locations of the regions are more distinctive. For example, we expect to find a water region under the sky region or more obviously we can find a sand region under the sky region. Horizontal relations are more vague so we

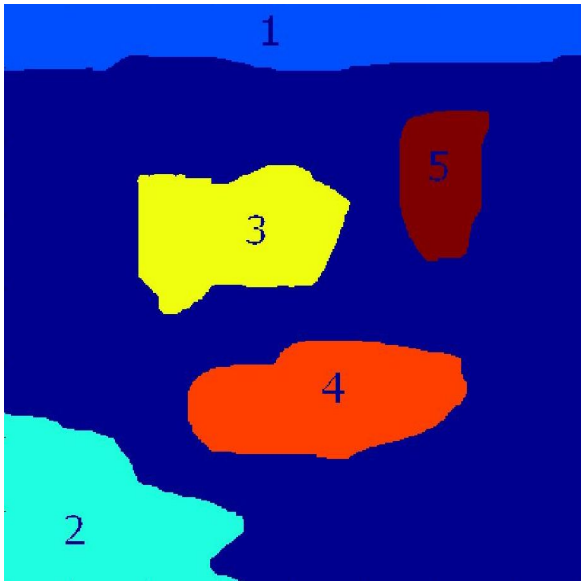

Figure 4.2: Example of the spatial model. The list on the right shows the region pairs extracted from the image.

assumed that using vertical information makes more contribution to the content of the image. Difficulty of the spatial models is the relativity of the position. Everyone can perceive some positions differently. To prevent the relativity, we are dealing with clear relations and also, some constraints are defined to restrict definitions. We constructed a simple spatial model which finds the region pairs that have vertical spatial relationships. First of all, projection of two regions on the x-axis is calculated. If there is no overlap, algorithm finds "no relation". If there is an overlap on the x-axis, it is possible to have an above-below relation so projection on the y-axis is calculated. If there is no overlap on the y-axis, these two regions have a vertical spatial relation. The one whose centroid is above is over the other region. If there is overlap on the y-axis, the overlap area is calculated. If overlap area is larger than %50 of the smaller region's area, there

is no above-below relation otherwise, the one whose centroid is above, is over the other region. Figure 4.2 shows an example of our spatial model.

4.1.4 Scene Features

Segmentation of the regions, clustering of the regions and extracting vertical spatial relations of the regions provide the necessary information to extract scene features. Each scene can be represented according to its feature content. In this research, scene features are the regions and region pairs so we created two different scenarios to represent the scenes. First is using regions individually and constructing "bag of individual regions" while second is using region pairs and building "bag of region pairs" to use the spatial information in images. Scene classification process uses these two assumptions separately. We consider two settings for this bag of regions representation:

1. each region is regarded separately and and a "bag of individual regions" representation is generated.

2. regions that satisfy the above-below relationship are grouped together and a "bag of pair regions" representation is constructed.

4.2 Region Selection

Determining the feature space is a significant problem in computer vision. In this research we represented scenes with regions and region pairs. In our representation, an image can have at most k different types of regions or k^2 different types of region pairs. Our aim is to find the characteristic region or region pair types in the scene categories because using all region-pair types cannot be contributing to all scene categories. Some region and pair types represent a particular scene category better than the others. We want to use these kinds of region and pairs in the representation to have more reliable results. Decreasing the size of our codebook will contribute to the results because if the selected

region and pair types will be the more characteristic ones for all scene classes. There are two important criteria for this process; one is that region types that are frequently found in particular class of scenes but rarely exist in other classes. Second is that region types that consistently occur together in the same class of scenes.

We formulate the feature selection process as a multisubset search problem as in [12] with the major difference being the definition of the optimization criteria for finding the best set of subsets of regions (because the motivations, requirements and inputs for keyword selection are different from our region selection setting). Let $\mathcal{T}$ be the set of region types in the codebook ($|\mathcal{T}| = k$) and

$$\boldsymbol{\mathcal{X}}_d = \{\mathcal{S}_j | j = 1, \ldots, c\} \tag{4.1}$$

be a set of subsets, also called a multi-subset, of these types where

$$\mathcal{S}_j = \{t_n^j | n = 1, \ldots, c_j; t_n^j \in \mathcal{T}\} \tag{4.2}$$

is the subset for the j'th class, c is the number of classes, and c_j is the size of the j'th subset. The size of the multi-subset $\boldsymbol{\mathcal{X}}_d$ is $|\boldsymbol{\mathcal{X}}_d| = \sum_{j=1}^{c} |\mathcal{S}_j| = d$.

Given a criterion $J(\cdot)$ that describes the quality of a multi-subset, the goal is to find such subset $\boldsymbol{\mathcal{X}}_d$ for which the criterion is maximum. A suboptimal solution to this problem can be found using the sequential forward selection algorithm that starts with an empty set $\boldsymbol{\mathcal{X}}_0$ and iteratively finds a new set $\boldsymbol{\mathcal{X}}_{i+1}$ by adding a new feature to the set $\boldsymbol{\mathcal{X}}_i$ such that

$$J(\boldsymbol{\mathcal{X}}_{i+1}) = \max_{\substack{j=1,\ldots,c \\ t \in \mathcal{T} \setminus \mathcal{S}_j}} J(\{\mathcal{S}_1, \ldots, \mathcal{S}_{j-1}, \mathcal{S}_j \cup \{t\}, \mathcal{S}_{j+1}, \ldots, \mathcal{S}_c\}) \tag{4.3}$$

until the multi-subset $\boldsymbol{\mathcal{X}}_d$ with the required size is obtained.

Our definition of $J(\cdot)$ combines two components (as in [12]):

$$J(\boldsymbol{\mathcal{X}}_d) = \sum_{\substack{j=1,\ldots,c \\ n=1,\ldots,c_j}} A^j(t_n^j) \left(\sum_{\substack{i=1,\ldots,c \\ i \neq j \\ m=1,\ldots,c_i}} E^{j,i}(t_n^j, t_m^i) \right) \tag{4.4}$$

where $A^j(t)$ describes the intra-subset importance of region type t within $\mathcal{S}_j$ and $E^{j,i}(t, \bar{t})$ describes the inter-subset relation between region types $t \in \mathcal{S}_j$ and $\bar{t} \in \mathcal{S}_i$.

Given $\mathcal{I}$ as the whole set of training images, $\mathcal{I}_j$ as the set of training images for the j'th class, and $H_l(t)$ as the frequency of the t'th region type in the l'th image, we define these components as follows (different from [12]):

$$A^j(t) = \frac{\left(\sum_{l\in\mathcal{I}_j} H_l(t)\right)\left(1+\sum_{l\in\mathcal{I}_j}\sum_{\bar{t}\in\mathcal{S}_j\setminus\{t\}} \min\{H_l(t), H_l(\bar{t})\}\right)}{\left(1+\sum_{\substack{i=1,\ldots,c\\ i\neq j}}\sum_{l\in\mathcal{I}_i} H_l(t)\right)} \tag{4.5}$$

promotes region types that are frequently found in examples for a particular class (first term in the numerator) and consistently occur together with other region types selected for the same class in the same examples (second term in the numerator) while demoting types that are also similarly frequent in examples for other classes (term in the denominator), whereas

$$E^{j,i}(t,\bar{t}) = \left(\sum_{l\in\mathcal{I}_j} \max\{H_l(t) - H_l(\bar{t}), 0\}\right) \times \left(\sum_{l\in\mathcal{I}_i} \max\{H_l(\bar{t}) - H_l(t), 0\}\right) \tag{4.6}$$

promotes pairs of region types of which each one is frequent in examples of one class but is rarely found in examples of the other class. This setting does not depend on a specific classifier unlike most of the traditional feature selection algorithms because it performs selection only based on the frequencies of region types in example images for different classes.

Using this algorithm, we reduced our codebook size from 50 to 20. In this case, 20 is obtained from empirical runs of the algorithm. We used different values like 10, 20 and 30 for the subset size. Detailed results are discussed in Chapter 5.

4.3 Scene Classification

As the classification step, we used Bayesian decision theory with posterior probabilities. A scene containing regions $\{x_1, \ldots, x_m\}$ will be assigned to the class with the largest posterior probability:

$$w_j^* = \arg\max_{j=1,\ldots,c} p(w_j|x_1, \ldots, x_m) \tag{4.7}$$

In this equation, w_j represents the j'th class, c is the number of classes, and m is the number of regions in the scene. Posterior probabilities are calculated as follows

$$p(w_j|x_1,\ldots,x_m) = \frac{p(x_1,\ldots,x_m|w_j)p(w_j)}{p(x_1,\ldots,x_m)}. \tag{4.8}$$

All classes can be assumed to have equal priors so we only consider $p(x_1,\ldots,x_m|w_j)$. This is equal to class conditional probabilities. We have two classification model; individual and pairwise.

4.3.1 Individual Classification Model

At the final Bayesian classification step, firstly we assumed that scenes are represented with the bag of individual regions. Each region is independent from others in the same class. Therefore, the class conditional probability reduces to

$$p(x_1,\ldots,x_m|w_j) = \Pi_i^m p(x_i|w_j). \tag{4.9}$$

Using the multinomial model, probabilities are calculated by maximum likelihood estimation:

$$p(x_i = u|w_j) = P_{ju}, \qquad j = 1,\ldots,c \qquad u \in \{1,\ldots,k\} \tag{4.10}$$

From maximum likelihood estimation, $P_{ju} = \frac{n_{ju}}{n_j}$ where n_{ju} is the number of regions with label u in the images for class j and n_j is the total number of regions in class j. This gives a total of k parameters for each class.

The possible problem here is the sparsity. Some scene categories may not contain regions from some clusters so P_{ju} values will be zero for those cases. When the P_{ju} values are zero, $p(x_1,\ldots,x_m|w_j)$ values will also be zero and this is an unwanted situation. To avoid this problem, while calculating the P_{ju}, we add 1 to the n_{ju} and k (number of clusters) to the n_j. Instead of having zero probability, it assigns a probability value very close to zero. To investigate the sparsity problem and visualize the distribution of the region types in different scene categories, we plotted the n_{ju} values for different scene classes in Figure 4.3 and 4.4. Similarly, Figure shows the probability values estimated using Equation (4.10).

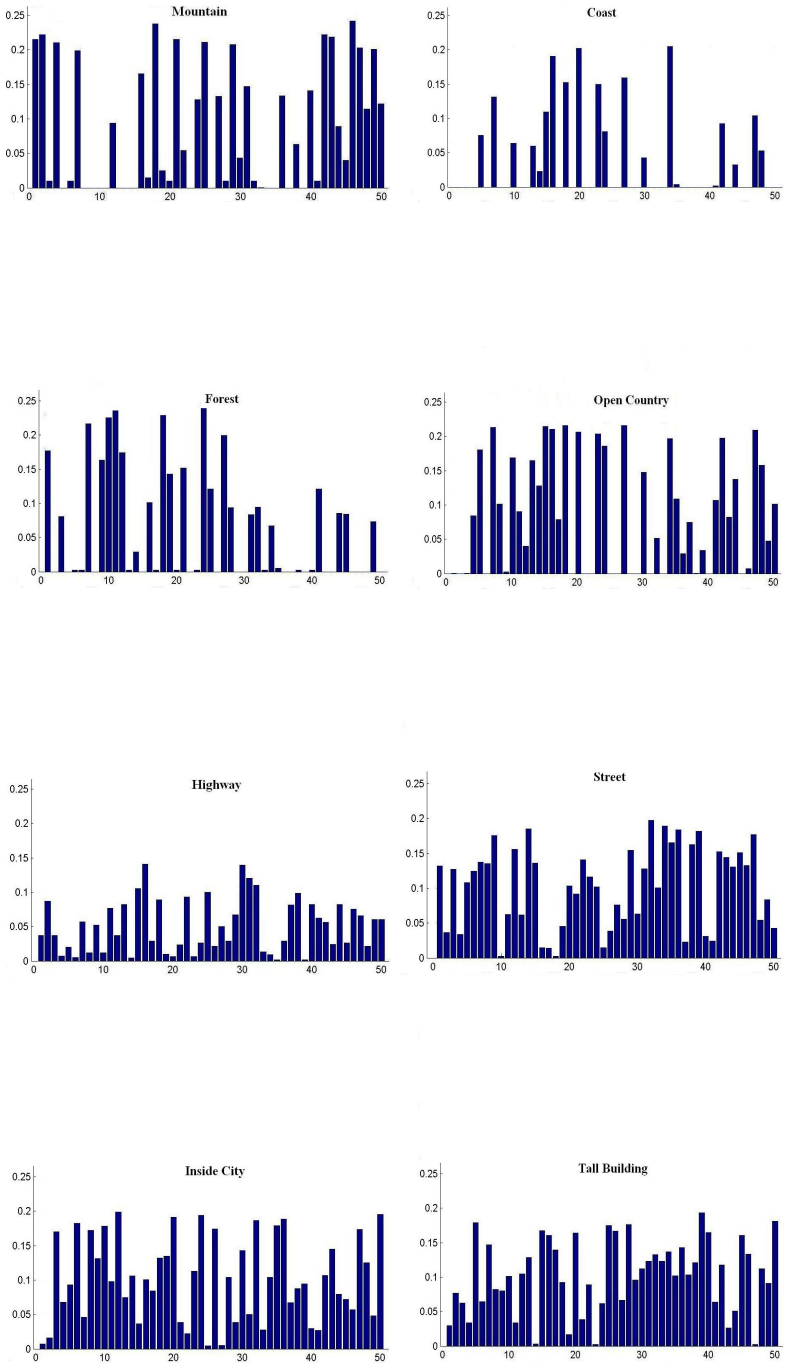

Figure 4.3: Region type histograms of the scene categories. x-axis shows the 50 region types

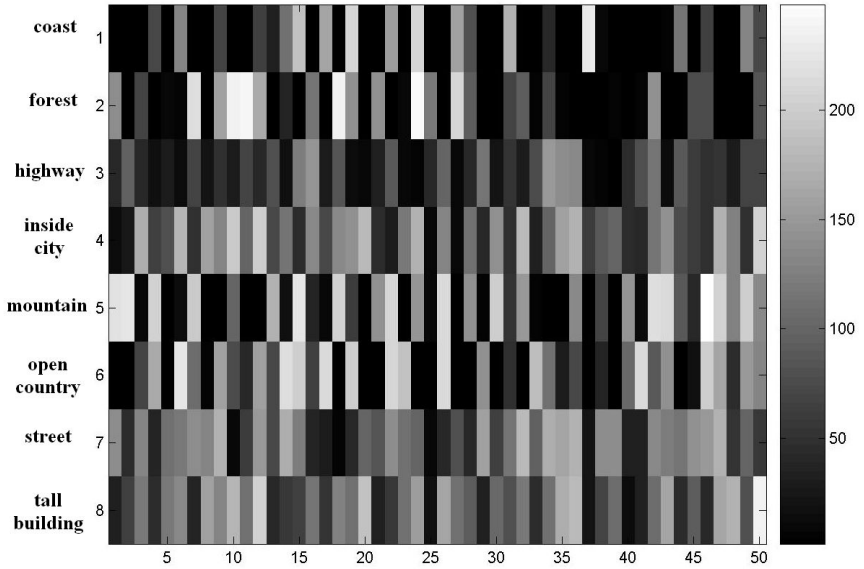

Figure 4.4: Region type histograms of the scene categories where brighter colors represent larger values.

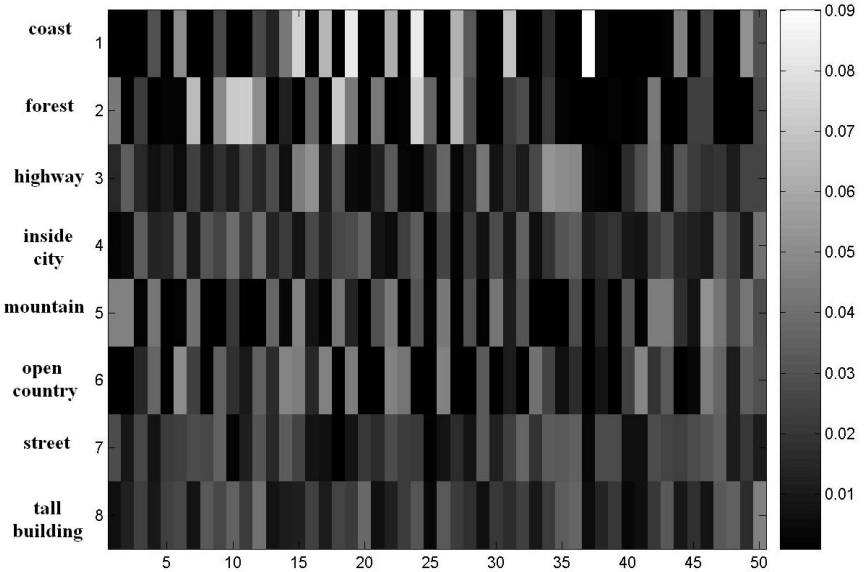

Figure 4.5: Region type probability of the scene categories where brighter colors represent larger values.

4.3.2 Pairwise Classification Model

We assumed that regions have pairwise spatial dependencies which are defined according to the above - below relation. In our second scenario, scenes are represented with bag of region pairs. Regions have pairwise dependencies but the pairs are independent from others in the same class. Therefore, the class conditional probability reduces to

$$p(x_1, \ldots, x_m|w_j) = \Pi_{(r,s)\in R} p(x_r, x_s|w_j) \tag{4.11}$$

where R is the set of region pairs that satisfy the spatial relationship where x_r is above x_s. Probabilities calculated by maximum likelihood estimation consist of k^2 parameters for the pairwise model.

Calculation of the pair probabilities is very similar to the individual one. Instead of regions, region pairs are taken in each scene category and proportion to the total region pairs in that category gives us the pair probabilities for each scene category. Same sparsity problem is possible for calculating region-pair probabilities and same solution is valid for this case, too.

Chapter 5

Experiments

In this research we used a subset of the MIT LabelMe data set that contains 8 scene categories: coast, forest, highway, inside city, mountain, open country, street and tall buildings [24]. Total number of images is 2696 with size 256 x 256 pixels and in "jpg" format. Numbers of images according to scene categories are shown in Table 5.1. This dataset was divided into training and testing sets

Table 5.1: Number of images in each scene categories.

	Train	Test
coast	150	211
forest	150	179
highway	150	111
inside city	150	159
mountain	150	225
open country	150	261
street	150	143
tall building	150	207
Total	1200	1496

randomly. Our train data contains 1200 images which are randomly selected 150 images from each category. The remaining 1496 images were used as test data. At the training process our segmentation framework was applied to each image and all data were segmented into regions. At the scene representation step, two types of representation model which are "bag-of-regions" and "bag-of-region

pairs" were applied. Before running the experiment cases, we needed to decide some important parameters.

For region clustering, value of k is needed to be determined for the k-means clustering algorithm so we tried different k values to find the most suitable one. To determine the number of region types, we tried k=30, 50 and 75. Using k very big or small probably would reduced overall success rate so we tried more reasonable and common values. Results for the bag of regions model is shown in Table 5.2. There are no significant differences in the results so we can say that

Table 5.2: Success rates depending on the number of region types.

	k=30	k=50	k=75
Coast	%72	%75	%76
Forest	%70	%69	%65
Highway	%48	%51	%53
I.C.	%60	%57	%45
Mountain	%60	%61	%56
O.C.	%37	%41	%32
Street	%40	%42	%37
T.B.	%53	%51	%48
Average	%55.00	%55.88	%51.55

the interval between 30 and 75 is quite reasonable. We decided to use 50 as the k value in the rest of the experiments.

Another critical parameter is the size of subset for the selection algorithm. It was decided by empirical trials with different values. The value must be lower than the cluster number so we used 10, 20 and 30 for subset size. Results for the bag of regions representation are shown in the Table 5.3. According to the results, contribution of the selection algorithm is limited so change for subset value in a close range did not created a significant difference in the overall results. For this reason, a value between 10 and 30 can be used for subset size. We preferred to use 20 for subset size in the rest of the experiments.

After the initial experiments to set the parameters, number of clusters will be 50 for the rest of the experiment runs whereas number of cluster will be reduced to 20 for the cases where the selection algorithm is used.

Table 5.3: Success rates depending on subset-size used for selection algorithm.

	Subset Size=10	Subset Size=20	Subset Size=30
Coast	%77	%78	%76
Forest	%65	%68	%71
Highway	%63	%60	%60
I.C.	%51	%49	%47
Mountain	%65	%62	%63
O.C.	%31	%34	%30
Street	%57	%55	%54
T.B.	%51	%54	%50
Average	%57.55	%57.50	%56.38

At the experiment process, we created four different cases which are:

- bag of individual regions representation without selection.
- bag of region pairs representation without selection.
- bag of individual regions representation with selection.
- bag of region pairs representation with selection.

5.1 Individual Classification Model

"Bag of individual regions" is our first experiment case. We tested this representation with and without region selection algorithm which is mentioned in section 4.2. Results of the individual representation without region selection are in Table 5.4.

Overall success rate is %55.88 and the best success percentage belongs to the "coast" scene category whereas worst percentage is belongs to the "open country" category. The important point is that the misclassification occurs between the similar scene categories. For example most of the misclassification is shared by the scene categories "coast", "mountain" and "forest" while misclassification in "open country" is very low for classes like "street", "tall building" and "inside city". It is very obvious that most of the misclassifications occur between

Table 5.4: Confusion matrix for the bag of individual regions representation without region selection algorithm.(Rows are true labels, columns are assigned labels.)

	Coast	Forest	Highway	I.C.	Mountain	O.C.	Street	T.B.	%
Coast	158	12	4	2	11	23	1	0	75
Forest	16	123	5	2	7	25	0	1	69
Highway	7	8	57	2	7	24	5	1	51
I.C.	2	3	10	91	3	1	38	11	57
Mountain	10	28	8	10	137	22	6	4	61
O.C.	25	34	35	12	36	107	4	8	41
Street	1	0	13	50	5	4	60	10	42
T.B.	3	4	12	41	3	2	36	106	51

the scene categories which have similar contents. Considering the difficulties in segmentation process, using regions without selection and spatial information, overall result is acceptable.

Same experiment is repeated with the our selection algorithm where we reduced the codebook size from 50 to 20. Selection algorithm just determined the 20 most representative region clusters for the whole scene categories and the rest of the processes were done with the new codebook. The new results for individual regions are in Table 5.5. Average success rate of the scene classes is 57.50. Selection increased the overall rate and still important amount of the misclassifications occur between similar scene categories.

Table 5.5: Confusion matrix for the bag of individual regions representation with selection algorithm where k reduced from 50 to 20.(Rows are true labels, columns are assigned labels.)

	Coast	Forest	Highway	I.C.	Mountain	O.C.	Street	T.B.	%
Coast	169	11	2	1	5	22	1	0	78
Forest	8	118	8	4	21	12	5	3	68
Highway	2	1	82	3	2	2	19	0	60
I.C.	0	1	20	89	5	9	25	10	49
Mountain	10	7	3	2	173	25	1	4	62
O.C.	36	38	22	5	28	125	2	5	34
Street	1	2	8	36	2	2	77	15	55
T.B.	4	5	16	33	4	1	30	114	54

5.2 Pairwise Classification Model

Our second image representation is the "bag-of-region pairs". In this case region pairs were extracted by our spatial model which selects the region pairs based on the above-below relationship. Results of the region pairs representation are in Table 5.6. Average success rate of the classes is %62.75 which is a much better result than individual region presentation. As expected, we can say that using spatial information created very important contribution to the overall results.

Table 5.6: Confusion matrix for the bag of region pairs representation without selection algorithm.(Rows are true labels, columns are assigned labels.)

	Coast	Forest	Highway	I.C.	Mountain	O.C.	Street	T.B.	%
Coast	173	7	3	3	5	18	2	0	80
Forest	8	141	4	2	12	6	1	1	66
Highway	0	0	78	6	4	2	13	8	74
I.C.	2	3	15	81	1	4	25	28	48
Mountain	3	12	8	3	175	22	0	2	77
O.C.	37	33	10	5	35	136	2	3	48
Street	4	4	15	36	2	0	57	25	54
T.B.	2	3	4	25	5	1	49	118	55

The fourth experiment is to use region-pairs and selection algorithm as mentioned in section 5.1. Results of the region pairs representation with selection algorithm are shown in Table 5.7. Selection algorithm improved most of the scene categories as expected. There is a small difference in the overall result which is %63.63. Using the region pair representation and selection algorithm created the best results as we expected. Considering the average results, we can say that contribution of the spatial information is much more than the selection algorithm. However, we can not ignore the improvement caused by selection algorithm to the results. Success rates of all cases are shown in Table 5.8. As the best case, "bag-of-region pairs with selection" representation results are shown in Figure 5.1 with good results and Figure 5.2 with bad results (each row is a scene category).

Table 5.7: Confusion matrix for the bag of region pairs representation with region-pair selection while k is reduced from 50 to 20.(Rows are true labels, columns are assigned labels.)

	Coast	Forest	Highway	I.C.	Mountain	O.C.	Street	T.B.	%
Coast	165	16	5	4	4	15	1	1	82
Forest	10	122	3	1	12	27	2	2	79
Highway	1	2	67	5	3	4	15	14	70
I.C.	6	4	7	78	5	3	31	25	51
Mountain	12	35	10	7	140	8	5	8	78
O.C.	42	30	34	8	40	89	12	6	52
Street	5	9	10	22	1	2	79	15	40
T.B.	4	3	6	35	12	7	28	112	57

Table 5.8: Summary of success rates for all cases.

	Individual	Pairwise	Individual with Selection	Pairwise with Selection
Coast	%75	%80	%78	%82
Forest	%69	%66	%68	%79
Highway	%51	%74	%60	%70
I.C.	%57	%48	%49	%51
Mountain	%61	%77	%62	%78
O.C.	%41	%48	%34	%52
Street	%42	%54	%55	%40
T.B.	%51	%55	%54	%57
Average	%55.88	%62.75	%57.50	%63.63

Separately from all experiments, we made a test run to show the contribution of the patch-based segmentation which is mentioned in Section 3.2. For this run, we used only one-class segmentation at the segmentation step of our framework. The aim of the test is to observe the difference in results with and without patch-based segmentation. Success rates of this experiment is shown in Table 5.9. Results showed that scene categories which have limited predefined

Table 5.9: Success rates for all cases without patch-based segmentation.

	Individual	Pairwise	Individual with Selection	Pairwise with Selection
Coast	%77	%75	%81	%83
Forest	%71	%56	%73	%75
Highway	%57	%71	%68	%67
I.C.	%26	%8	%21	%22
Mountain	%63	%74	%68	%77
O.C.	%44	%47	%41	%53
Street	%7	%5	%10	%3
T.B.	%15	%11	%9	%12
Average	%45.00	%43.37	%46.39	%49.00

class regions like street, inside city and tall building have very poor classification results. When compared to the results in Table 5.8, it is very obvious that patch-based segmentation approach has contribution to the scene categories which have complicated man-made structures.

5.3 Comparisons

There are two comparison cases. First one is the comparison with the global histogram method and the other is the comparison with the bag-of-words method.

First of all, we compared our results for the 4 cases with the most traditional classification method which uses global histograms of the images. We extracted the HSV histograms (with $8 \times 3 \times 3$ bins) of the images and classified with 5 different statistical classifiers. The classifiers are "Linear Gaussian classifier", "Quadratic Gaussian classifier", "k-nearest neighbor classifier ", "Parzen window classifier" and "Support vector machines". Results of these classifiers with global histogram results are shown in Table 5.10. It is very obvious that all of the cases

have much better results than the global histogram method and reasons can be summarized as using region segmentation(local information), spatial relation information and region selection.

Table 5.10: Classification rates for global histograms method.

Linear Gaussian classifier	36.21%
Quadratic Gaussian classifier	31.15%
k-nearest neighbor classifier	35.07%
Parzen window classifier	44.65%
Support vector machines	43.81%

Second and the more important comparison is with the bag-of-words method [7] which uses methods such as probabalistic latent semantic analysis (pLSA)[21] and latent dirichlet allocation (LDA) [22] . The matlab code of the method was taken from [20]. We randomly selected 200 images for each scene category from our dataset. Algorithm randomly used half of each category as train data and the other half as test data. Totaly, we gave 1600 images of our 2696 image dataset and the reason is that we want each category to have equal number of images. Confusion matrix of the bag-of-words method is shown in Table 5.11. Average

Table 5.11: Confusion matrix for the bag-of-words method.

	Coast	Forest	Highway	I.C.	Mountain	O.C.	Street	T.B.	%
Coast	67	1	7	0	10	12	3	0	67
Forest	1	88	0	0	6	4	0	1	88
Highway	6	11	54	4	9	2	3	1	54
I.C.	0	1	2	76	2	0	9	11	76
Mountain	0	10	7	0	75	3	0	5	75
O.C.	10	14	9	0	16	51	0	0	51
Street	0	0	0	16	0	0	80	4	80
T.B.	0	5	4	23	10	2	20	36	36

success rate of the method is %67.13. It is better than all of our cases but it is very close to our 2nd and 4th cases. Our best case (the 4th case) is slightly worse than the bag-of-words method but as summary we can say that results are close. If we compare the distribution of the misclassification, consistency can be seen between the similar scene categories like in our cases. We can say that our approach is weaker in scene categories like street and inside city because of the

patch-based features but in other categories like coast, forest, mountain we have equal or better results. To summarize, both methods have close results and both methods have very consistent result for individual scene categories.

Figure 5.1: Correct classification results. Each row shows a scene category.

Figure 5.2: Images wrongly classified with the best case. Each row shows images wrongly assigned to a particular category.

Chapter 6

Conclusion and Future Work

As a summary, in this research we aim to construct a reliable outdoor scene classification framework. As the first step, we created a two level segmentation process to find regions in images. One-class classifiers that are learned from different semantic classes are applied to images. These classifiers divide an image into predefined semantic classes and an outlier class. Next step is the segmentation in outlier regions with patch-based processing. Patches are extracted around the interest points found within outlier regions. After that, regions in roughly segmented images are clustered with k-means algorithm. Clustered region types created our codebook and images are represented as "bag-of-regions" and "bag-of-region pairs". We developed a spatial model which extracts above-below relations for 'bag-of-region pairs" representation. Region selection algorithm is applied to the codebooks to find the characteristic region types. As final step, Bayesian classification approach is used for final classification.

For the future work, better interest point detectors can be used to extract more reliable patches. We saw that our patch-based clustering method is a little weak in some categories. Also to improve the use of spatial relation information, different and more complicated spatial models can be established. We believe that better spatial relations improve the final results in a good way. Another important point is the selection algorithm; different algorithms can improve the results and different codebook techniques can also be beneficial.

Bibliography

[1] M. Gorkani and R.W. Picard. Texture orientation for sorting photos at a glance. In Proc. ICPR, volume I, pages 459–464, Jerusalem, Israel, Oct. 1994.

[2] E.C. Yiu. Image classification using color cues and texture orientation. Master's thesis, MIT, Dept EECS, 1996.

[3] P.R. Lipson. Context and Configuration Based Scene Classification. PhD thesis, MIT, 1996

[4] H.H. Yu, and W. Wolf. Scenic classification methods for image and video databases In Proc. SPIE, Digital Image Storage and Archiving systems, pages 363–371, 1995.

[5] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In Proc. CIVR, March 2004.

[6] S.Kumar and M.Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In Proc. ICCV, Oct.2003.

[7] L.Fei-Fei and P.Perona. A Bayesian hierarchical model for learning natural scene categories. In Proc. CVPR, June, 2005.

[8] S. Lazebnik and C.Schmid and J. Ponce. Beyond the bag of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. CVPR, June, 2006.

[9] M. R. Boutell, J. Luo, and C. M. Brown. Factor graphs for region-based whole-scene classification. In CVPR, Semantic Learning Workshop, New York, NY, June 17-22, 2006.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, November 2004.

[11] F. Monay, P. Quelhas, J.-M. Odobez, and D. Gatica-Perez. Integrating co-occurrence and spatial contexts on patchbased scene segmentation. In CVPR, Beyond Patches Workshop, New York, NY, June 17-22, 2006.

[12] P. Somol and P. Pudil. Multi-subset selection for keyword extraction and other prototype search tasks using feature selection algorithms. In ICPR, volume 2, pages 736-739, Hong Kong, August 20-24, 2006.

[13] D. M. J. Tax. One-Class Classification. PhD thesis, Delft University of Technology, Delft, The Netherlands, 2001.

[14] J. C. van Gemert, J. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in context. In CVPR, Semantic Learning Workshop, New York, NY, June 17-22, 2006.

[15] R. Fergus and P.Perona and A. Zisserman. Object class recognition by unsupervised scla-invariant learning. In Proc. CVPR, June, 2003.

[16] G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In Proc. ICCV, Oct, 2003.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. In Proc. CVPR, pages 731–737, 1997.

[18] Kadir, T. , Zisserman, A. and Brady, M. An affine invariant salient region detector. In Proc. 8th European Conference on Computer Vision, Prague, Czech Republic, 2004.

[19] Smith, J.R. , Li, C.S. Image Classification and Querying Using Composite Region Templates. In Proc. CVPR, 1998.

[20] ICCV 2005 Course on Recognizing and Learning Object Categories, http://people.csail.mit.edu/fergus/iccv2005/bagwords.html

[21] Hofmann, T. Probabalistic Latent Semantic Analysis. In Proc. UAI, 1999.

[22] Blei, D. and Jordan, M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022, January 2003.

[23] Sivic, J. and Russell, B. and Efros, A. and Zisserman, A. and Freeman, W Discovering object categories in image collections. Proc. ICCV, Beijing, 2005.

[24] MIT Computational Visual Cognition Laboratory, LabelMe Dataset, http://cvcl.mit.edu/database.htm