Scene Classification in Images

B V V Sri Raj Dutt Pulkit Agrawal Sushoban Nayak

Abstract

In this paper, following the works of Oliva and Torralba, we present a procedure to classify real world scenes in eight semantic groups of coast, forest, mountain, open country, street, tall building, highway and inside city, without going through the stages of segmentation and processing of individual objects or regions. The approach essentially takes into account the diagnostic information stored in the power spectrum of each category of images and through supervised learning, separates characteristic feature vectors of each class in separate groups that helps assign a new test image to its respective group. We follow a sequential hierarchy in which images are first sorted according to their naturalness and through traversing a tree, we ultimately reach the desired node that represents the class of the image. The significance of results obtained and the ripple effect of errors have also been reflected upon.

Keywords: scene classification, Gabor filter, LDA

1 Introduction

The statement "Machines shall rule the world" is no more a mere figment of imagination, for its the promise of Machine learning. In our quest to design and manufacture robotic replicates of humans we haven’t come a long way, yet we have with us a set of techniques, algorithms and models which serve as a good starting point. The mystery of how human brain functions has evaded researchers for centuries. The aim of research has been to empower our mechanical creations to think and analyze situations in ways similar to us. In fact, modelling the human brain and realizing it physically is the greatest challenge of the century.

Artificial Intelligence combines work in fields like computer science, neuroscience, mathematics, psychology, philosophy, cognitive sciences etc. Also, often progress is made by simultaneous contributions from more than one areas. For example, advances in cognitive psychology have led us to believe that our brain processes information in a top-down” manner rather than doing it in a bottom-up” manner. It has also been proposed that the brain develops models and templates for information it encounters, i.e., it has a model for a building, for a game of cricket, for the music of guitar, and so on for almost everything which we can observe and feel. In case it encounters something new it tries to correlate to something it already knows and at times it can form a new model as well. Moreover these models and conceptions of things keep evolving with time and with our interaction with environment. Capturing this mammoth collection of models, concepts and information in a machine is a very big challenge.

One of the main features of our daily experience is the ability to distinguish between things , to identify them and to link them with our prior knowledge. This ability to recognize and interpret the environment
around us, is in principle the foundation for any higher level of processing that we do. Synonymous is the concept of clustering, segmentation and classification in artificial intelligence. Inspired by the working of human brain, the conception of learning algorithms took birth. These algorithms in essence provide us with a methodology to find parameters which would be able to identify and classify different objects in a given signal input. Researchers, in general have looked at different aspects of the brain like vision, hearing, speech etc for a better understanding of their functioning and in an attempt to model these processes.

We consider here a fundamental problem of computer vision, i.e. enabling computers to see the way we see things. We in future wish our machines would match the capabilities of human vision. It's interesting to note that, every second we receive tremendous amount of visual data and almost unconsciously we process this information very quickly. Classifying an object as table, a ball, or a scene as mountain or river is pretty trivial for us. We can in fact process amazingly more complex information. It's a well known fact that robotic vision compares miserably with our eyes. Here, we intend to make a start towards our goal by considering a very trivial problem by the standard of human vision and that is scene classification. Given an image of the scene we wish to classify it as say a mountain, forest, city, street etc.

Generally, a learning based approach is used to solve problems of this nature. A training set is initially created which would contain representative images from all categories that we need to classify. Now these images are manually labeled to the class they belong as perceived by the human. Now a learning algorithm is employed, which basically is a strategy to enable us to come up with parameters which would characterize an image for doing the classification task. Now if a random image is given as an input, on basis of parameters already identified the machine would try to classify the image. This in essence id a generic way in which learning algorithms work, i.e., by learning from a huge set of data and then using this learned information to make predictions about successive inputs.

### 2 Introduction

#### 2.1 What is a scene?

Our project mainly focuses on classifying scene and as such it’s imperative that we clarify what we mean by a scene. To distinguish scene from ‘object’ or ‘texture’, we follow the approach taken by [5] and we consider the absolute distance between observer and the fixated zone as the discriminating factor. So, an ‘object’ is something that subtends about 1 to 2 meters around the observer; but in case of a scene, the distance between the observer and the fixated point is usually more than 5 meters. To put it in common terms, object is something that is at hand distance whereas scene is mainly a place in which one can move.

#### 2.2 Scene Recognition

Following the approach of the experimental schools, for whom recognition of scene means providing information about the semantic category and the function of the environment, we propose a computational model of recognition of scene categories. There are fundamentally two types of scene recognition techniques found in literature. The first one is the bottom up approach of employing object recognition
to decide the category of the scene. We here follow the second approach that essentially is a top-down approach. We bypass segmentation and processing of objects and try to categorize each scene through assimilating its global information. The reason we go for the second approach is that it is supported by psychological experiments, which propose that humans accumulate enough information about the meaning of a scene in less than 200ms, which can only mean that we go for an overview rather than a detailed analysis of object and texture. Especially while dealing with environmental pictures, object information is spontaneously ignored. For example, [3] confirms that coarse blobs made of spatial frequency as low as 4 to 8 cycles per image provided enough information for instant recognition of common environments even when the shape and identity of objects could not be recovered. Some other studies have also shown that people can be totally blind to object changes, even when they are meaningful part of the scene ([5]). So effectively, at a first glance, we just assimilate the gist of an image which is built on a low resolution spatial configuration.

2.3 Defining the Problem:

Given an image as an input we wish to classify as one of the following:

1. Coast
2. Forest
3. Highway
4. Inside City
5. Mountain
6. Open Country
7. Street
8. Tall Buildings

The image classes are standard image classes which have been used extensively in research to date. We have used the Oliva Toralba dataset which is a subset of the Corel database for our project. Some representative images from the image data set have been shown in the section 4.

3 History

There are many ways in which previous works have approached this problem. [4] uses the traditional idea of bottom up approach where scene recognition is done through segmentation of image into regions instead of going for global features. [10] goes for a typicality approach. [8] eliminates the need for human-annotations of image features. [2] introduces the concept of recognizing scenes on approximate global geometric correspondence and claims to get high recognition rates on challenging data. [6] emphasizes the fact that we recognize images at different scales and not just with global features or only with local features and does a very good work of classifying images. However, given the time and resources available, and
thanks to the detailed explanation of [7] on the importance of power spectra of images in classification, we sorted out two papers that had related works, [9] and [5]. [5]’s was a pretty elaborate technique of differentiating images into 8 semantic categories where as [9] concentrated on lesser properties of images and instead of going for exclusive classification, preferred to arrange the images in semantic axes. We took into considerations both the above techniques and followed them to some extent in our work. With some experiments and new ideas, we tried to improve upon them or at least minimize the complexity of the same, the detailed description of which is given in the sections to follow. though we have overlooked [3]’s results of the importance of color in recognition by working essentially on gray scale images, we intend to account for them in future work.
4 Dataset

Some of the images from the dataset have been shown in the figure.

Some Notes:

- Open country is essentially the countryside.
- Tall Buildings includes skylines and sky scrapers whereas Inside city is a photo a characteristic city scene generally without the skyline.

5 Our work

5.1 What are we on a look out for?

We essentially need to figure out features such that $\text{Image} = f(\text{features})$. The most important characteristic of feature that we require is a very strong co-relation between the feature value and the class of the image. This makes learning an easier, faster and a less error prone job. Thus, feature identification is among the most important task. Once, the features are identified a classifier can be constructed using state of art algorithms like SVM’s, k-means, Bayes method, Linear discriminant analysis or any other suitable method. With help of this classifier, identification of an input image needs to be done.

Basic Approach towards classification: On the basis of psychological experiments, some properties capturing the most important aspects of a scene were identified ([1]). In the same work, they tried expressing as a function of these properties like naturalness, openness etc. With this work as the background, we decided to form a tree classifier of the structure of one shown in (Fig). Firstly, the image is classified as Artificial or Natural. Next an artificial image is classified into 2 groups, (street and highway) or (inside city and tall buildings) and in the next step these groups are classified into their constituent classes. This classification was done on an intuitive basis. For example, street and highway show similar characteristics, so do open country and coast (both have horizons and are open in some sense). Some different combinations of grouping were tried like clubbing street and tall buildings together (as buildings are often present on the side of the street) etc. A more rigorous approach towards formation of tree structured classifier was not used as results with some combinations we worked were not significantly different. This is discussed in more details in scope of future work.

5.2 Things we tried:

- Haar features: Motivation: Fast and is known to work well in image processing tasks such as face detection. We divided the image into blocks of sizes 64*64 and in each block applied Harr transform successively 2 times to get the LL component of each block. Then the combined set of these LL features of each block was considered in the set of feature vector corresponding to the image. While calculating the distinguishing weights for features from LDA, the matrix was nearly singular and solution of SVM did not converge and hence we had to abandon this approach. The results obtained from LDA inspite of the near singularity of matrix have been shown in the figure.
• Edges : Motivation : Edges can give a rough idea of the scene. Tall buildings for example would show many long vertical edges and possibly numerous horizontal edges, similarly a forest would show uniform distribution of dominant long vertical edges (presence of trees). Coast and open country on the other hand would give a relatively fewer edges and more of horizontal lines as prominent ones (owing to the horizon etc). Mountains on the other hand can be expected to give more of diagonal edges, due to their characteristic shape, similarly streets and highways have characteristic edges marking the outlines of the road. The results we have obtained on finding edges for different categories of images are given in Appendix 1.

Comparing the edge maps of various images, we found loads of noise in the edges making it to rather difficult to identify features. Thus this method was discarded without any further analysis.

• Colour Features: We asked ourselves a basic question, is it possible to classify the images only based on the colour information ?

Motivation :
We were motivated by the fact that colours provide an important clue towards discerning the image. A coast for example would have a large span of blue region, mountains, forest, open country would show a lot of green, highways and streets would contain a gray path like region in the middle of the image, tall buildings would have a skyline(blue in general) other than the colour cues from colours of buildings (which are very different from colours of natural images).

Formation of feature vector :

– Image is resized to 256*256 using bilinear interpolation.
– Image is divided into blocks of 64*64.
– The mean of 'Red', 'Green' and 'Blue' colour values are individually calculated for each block and are used as features.
– Thus we have a feature vector of size : \( \text{NumberOfBlocks} \times 3 \times 1 = 768 \) features.
– Now we find a discriminant vector for each division as we go down our classification tree using LDA.
Using these features, we do the classification as shown in the flow chart.

Conclusion: Just on the basis of colour categorization, we couldn’t separate the categories. In fact it is also important to note that colour gives lots more information in perceiving a natural scene as compared to artificial scene ([2]). This can be seen in our results as well.

6 Original Work by Oliva

The paper follows a ingenious approach in the sense that it doesn’t go for exclusive classification of scenes in categories. For large datasets, the number mis-categorizations can be whooping if we go for exclusive classification, since there is more chance of finding images belonging to ambiguous classes. Therefore, they preferred to go for arranging the figures in semantic axes. First the naturalness of the figure is decided upon. If it is natural, for example if its an image of coast or forest or mountain, we go for arranging it according to its openness. On the other hand, if we get a man-made scene, the semantic axis goes from horizontal spread at the left to vertical spread towards right. So highways are placed towards the extreme left whereas sky-scrappers find themselves at the extreme right as shown in the figure below.

So essentially they compute the structural attributes of figures and arrange them continuously along the semantic axes. The method used to compute the structural attributes are as follows:

- Important data can be arrived at from the power spectrum of an image [7] which is defined through the following equation:

  \[ \Gamma(f_x, f_y) = |FT_i(x,y)|^2 \]  

(1)
This essentially means taking the Fourier transform of the image and squaring its magnitude. They have the property of encoding energy density for each spatial frequency and orientations over the whole image.

- **Discriminant Spectral Templates:** A DST is represented by a set of low level features encoding the structure which is discriminant between the two scene categories. The structural discriminate feature $u$ is computed per image by using a DST as follows:

$$u = \int \int \Gamma(f_x, f_y)DST(f_x, f_y) df_x df_y$$

(2)

So its a weighted integral of power spectrum of the image and $DST(f_x, f_y)$ is the weighted function that describes how each spectral component contributes to the structural attributes. The DST is determined through a supervised learning stage. To represent DSTs in a low dimensional space, it is decomposed into a set of functions and in the paper, they use Gaussian envelopes corresponding to Gabor filters to this end:

$$DST(f_x, f_y) = \sum_{n=1}^{N} d_n G_n(f_x, f_y)^2$$

(3)

The coefficients $d_n$ are the weights of each of the Gabor filters and they are determined in the learning phase. From the above two equations we get,

$$g_n = \int \int \Gamma(f_x, f_y)G_n(f_x, f_y)^2 df_x df_y$$

(4)

Where $g_n$ are output energies for the N Gabor filters used as the basis of the DST. In terms of these energies, $u$ becomes:

$$u = \sum_{n=1}^{N} d_n g_n$$

(5)

The paper sampled the power spectrum with 70 Gabor filters from high spatial frequencies (1/3 cycles/image) to low spatial frequencies (1/72 cycles/image). No large difference was found when different reasonable numbers of filters were used and thus we went for 32 Gabor filters in our implementation. At this stage, each image has a unique vector representation in terms of $g_n$s. The feature vector for each image is $x = g_n$. Now, several methods can be used to find the coefficients $d_n$. The method employed here is needs to be able in finding two different sets of images that can be described with unambiguous semantic attributes. So, LDA was used to find the coefficients. It can be found trough the following sets of equations:

$$T = E[(x-m)(x-m)^T], m = E[x]; d = d_n d = T^{-1}(m_1 - m_2)$$

(6)

Here, $T = \text{the covariance matrix}$ $m_1$ and $m_2$ are the mean vectors of the feature vectors of the two classes. The only problem with this method is that when the learning set isn’t large enough, inversion of $T$ may be ill-conditioned in which case classic regularization techniques like principal components or adding a perturbation to the matrix are to be used. After the learning is done, we have all the information we need to construct the DSTs. The computational steps for obtaining the structural features are as follows: a. Pre-filtering, where the image intensity is divided at each pixel by an estimation of the local variance in order to eliminate illumination factors. B. Power spectrum computation. C. Structural feature computation. Upto this stage, our approach is almost the same. The paper and our work follow different paths hereafter. The paper uses scene discriminate filters to arrange the images along spatial axes.
• The output energy of a filter with transfer function $H(f_x, f_y)$ can be computed as:

The DST has both positive and negative values and as such, it cannot be implemented by one filter due to the square term we see in the above equation. So it’s computed as the difference between the output energies of two filters. $u$ is then $u = E_+ - E_-$ where they are the output energies of two filters $H_+$ and $H_-$. So we get:

When the output of the two filters is computed by convolution with the respective impulse responses

$$o_+(x, y) = i(x, y) * h_+(x, y) \tag{7}$$

$$o_-(x, y) = i(x, y) * h_-(x, y) \tag{8}$$

, the structural semantic features are:

The images are then projected onto this plane to be arranged along the semantic axis. This was tested on 256-by256 pixels size images from the Oliva Torralba dataset, which contains 2600 images (1500 natural scenes, 800 artificial scenes and 300 scenes which are a combination of both). The classification rate was 91% for naturalness, whereas it was 94 to 97% for ordering along open-closed axis. The results with the horizontal to vertical axes were as high as 98%.

7 Our Method

All images are converted into size 256*256 using bilinear interpolation before any further processing.

7.1 Training

• An image is picked up from the set of training images.

• The image is converted into gray scale by taking the average of Red, Green and Blue Components.

• Pre-filtering of the image is done to make it more robust to illumination changes by local illumination normalization.

We apply a pre-filtering to the input image $i(x, y)$ that reduces illumination effects and prevents some local image regions to dominate the energy spectrum. The pre-filtering consists of a local normalization of the intensity variance. The equation of pre-filtering is:

$$i'(x, y) = \frac{i(x, y) * h(x, y)}{\epsilon + \sqrt{(i(x, y) * h(x, y))^2} * g(x, y)} \tag{9}$$

$g(x, y)$ is an isotropic low-pass Gaussian spatial filter with a radial cut-off frequency at 0.015 cycles/pixel, and $h(x, y) = 1 - g(x, y)$. The numerator is a high-pass filter that cancels the mean intensity value of the image and whitens the energy spectrum at the very low spatial frequencies. The denominator acts as a local estimator of the variance of the output of the high-pass filter. $\epsilon$ is a constant that avoids noise enhancement in constant image regions. This pre-filtering stage affects only the very low spatial frequencies (below 0.015 c/p).
- Gabor functions of 4 scales and 8 orientations have been used, giving a total of 32 gabor filters. (For details refer to section...)

- These Gabor functions have been precomputed and stored.

- The image is divided into 16 windows, and in each window the filter is applied to obtain a feature vector.

- Thus we have a total of 32*16 = 512 features in our feature vector which represents the image.

- Feature Vector for each image is calculated.

- Now LDA is used to find discriminant vector for classification into natural or artificial.

- Now natural and artificial are further sub-classified as in the classification tree and discriminant vectors are obtained at each classification step.

- Thus, we have 7 discriminant vectors in total.

- At each step, projections of image classes being classified are taken on the corresponding d-vectors.

- The array of 7 discriminant vectors and the projection values constitute the training data.

**7.2 How did we decide these discriminant vectors were capable of classification**

The images in the trained data set itself were classified using the obtained d vectors. Later the results were tested out on test images. Some of the results of this process have been presented in the following graphs. Blue and red dots are used to represent the two classes. The Y co-ordinate gives the value of projection of image vector on the d vector, where as the x axis the total of number of images used for classification.

**7.3 Classification**

- The input image is converted into gray scale, resized and pre-filtered.

- The feature vector of the image is calculated.

- The feature vector is projected on the first discriminant vector (i.e. for classifying into artificial and natural)

- With the help of value of this projection value, and the training data we use a naive classifier to label this image.

- Depending on the output of the classifier we move down the classification tree and reiterate the above steps, unless we reach the bottom most node of the tree.

- The bottom most node of the tree gives the final classification result.
Figure 2: Natural v/s Artificial

Figure 3: Coast + Open Country vs Forest + Mountain
7.4 Generation of Results

- Our work allows for classification of a single image or for creation of a confusion matrix for a given set of images.

- If the training data is already available, the confusion matrix is directly calculated for the given data set.

- In case, the training data is not available a large annotated set needs to be given for analysis.

- \((9/10)\)th of the total number of images are randomly selected for training and the remaining images serve as the test set.

- A confusion matrix on this test set is generated.

7.5 Analysis of Results

<table>
<thead>
<tr>
<th></th>
<th>Coast</th>
<th>Forest</th>
<th>Highway</th>
<th>Inside City</th>
<th>Open country</th>
<th>Mountain</th>
<th>Street</th>
<th>Tall Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>23</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Forest</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Highway</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Inside City</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Open country</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mountain</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Street</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>Tall Building</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>18</td>
</tr>
</tbody>
</table>

The confusion matrix gives us the information in the following format:
The value of Confusion(i,j) gives us the number of images which belong to category j but are classified as category i.

Thus, the diagonal represent the number of correctly classified matches.

The off diagonal elements give us the false positives and the false negatives.

Table shows a general confusion matrix that we obtain. Note that the elements in the diagonal are large as compared to off diagonal elements. There are 161 correctly classified images out of 267. This gives accuracy of $161/267 = 60.29\%$ correct classification. The trends in false positives and false negatives are well captured by this confusion matrix. On observation we find that the trends in false positives and false negatives are very similar. We get general trends in conflict which are discussed below.

### 7.6 General trends in Errors

We’ll look at each column one by one, and analyze where we are getting not so good results.

#### 7.6.1 Wrong Classifications and error rates

- Coast images are often wrongly labelled as Open Country. (8/36 times)
  Accuracy for coast image identification = 64%

- Forest images are often wrongly labelled as Mountain. (4/32 times)
  Accuracy for forest image identification = 63%

- Highway images are often wrongly labelled as Inside City. (6/25 times)
  Accuracy for Highway image identification = 60%

- Inside City images are often wrongly labelled as Highway. (7/35 times)
  Accuracy for Inside City image identification = 54%

- Mountain images are often wrongly labelled as Forest. (10/42 times)
  Accuracy for Mountain image identification = 64%

- Open Country images are often wrongly labelled as Coast. (12/35 times)
  Accuracy for Open Country image identification = 57%

- Street images are often wrongly labelled as Highways. (5/27 times)
  Accuracy for street image identification = 71%

- Tall Buildings images are often wrongly labelled as Street. (9/35 times)
  Accuracy for Tall Building image identification = 49%

#### 7.6.2 Insight into the errors

- Semantically Mountains and forest rate similarly on the openness level.
• Open country and Coast both comprise of a vast open space, and in this sense they have a similar scene gist.

• Street and highways have a similar structure semantically, yet the classification is pretty good.

• Often tall buildings are present on the side of the streets, hence we get such large errors in this classification.

7.7 Analysis of results at various levels of the tree

We used 2528 images for learning and 160 previously unseen images by our classifier for performing the test. The test images consisted of 20 images from each class. The results have been detailed below.

7.7.1 At tree level 1

• 75 images were correctly classified as Natural. This accounts for 94% accuracy in natural image identification.

• 67 images were correctly classified as Artificial. This accounts for 84% accuracy in artificial image identification.

• Thus a total of 142 images out of 160 images were correctly classified at the first level. This accounts for 89% accuracy at the topmost level.
7.7.2 At tree level 2

We formed 2 subgroups of the test images, namely artificial and natural (80 images each). Now on the subset of test images belonging to natural class ran the classifier for distinguishing between (Coast and Open Country) Vs (Mountain and Forest) macro classes. Similar procedure was followed for the artificial image test set as well. Results are given below.

- 38 images were correctly classified to belong to Coast & Open Country macro class. This accounts for 95% accuracy.
- 33 images were correctly classified to belong to Mountain & Forest macro class. This accounts for 83% accuracy.
- Thus a total of 71 images out of 80 images were correctly classified at this level. This accounts for 89% accuracy at the second level of natural images.

7.7.3 At tree level 3

Now each macro class was broken into constituent classes and we ran our code on the test data for each pair. (example: Coast Vs Open Country). Some results are presented here.

- 15 images were correctly classified to belong to Coast class. This accounts for 75% accuracy.
- 13 images were correctly classified to belong to Open Country class. This accounts for 65% accuracy.
- Thus a total of 28 images out of 40 images were correctly classified at this level. This accounts for 70% accuracy.
15 images were correctly classified to belong to Forest. This accounts for 75% accuracy.

17 images were correctly classified to belong to Mountain class. This accounts for 85% accuracy.

Thus a total of 28 images out of 40 images were correctly classified at this level. This accounts for 80% accuracy

This the total accuracy for classification at this level = 75% for natural subcategory. Similar accuracy was obtained for classes under artificial sub category as well.

7.8 Errors due to Cascading

We realized that due to tree classification structure we obtain a ripple effect in errors. Errors made at one level in the tree are propagated down. For example, if a artificial image has been wrongly classified as natural image then further classifications into subcategories of natural is irrelevant and is sure to produce a wrong classification. The accuracy of our method can be increased by providing a reliability measure of the result along with a way to prevent this rippling of errors. This has been discussed in detail in scope of future work(??).

7.8.1 Illustrating Cascading

We can get an estimate of overall accuracy by multiplying fractional accuracies acheived at each level. For example consider the given image is coast.

- Probability that image is classified as natural = 0.94.
- Probability that natural image is now classified into Coast & OpenCountry = 0.95.
- Probability that this image is further classified as Coast = 0.75.
- Thus the overall accuracy = 0.94*0.95*0.75 = 0.67

This is also the achieved acuracy on the average case for the coast. Other results can also be similarly explained. In natural category we find at level 3 we get minimum accuracy whereas for artificial classification we get lowest accuracy at level 2.
8 Comparisons

8.1 Classification using Power Spectrum templates

[3] The paper essentially describes the basis of using power spectra as the scheme for classifying images.

Power spectral forms of the prototypical images are shown in figure Power Spectral Forms. From these we can clearly see the basis for power spectra based classification, which we have used with a little modification. An examination shows that beach scenes have power spectrum strongly organized in the horizontal direction, i.e. power spectra displays dominance of energy on the \( f_y \) axis, mostly at low \( f_y \) spatial frequencies. Cities’, on the other hand, are structured along vertical and horizontal directions. Spectra of forests are mainly isotropic from low to high spatial freq. On the whole, all the categories of spectra can be divided into 5 families, to wit, horizontal shape, cross shape, vertical shape, oblique shape and circular shape. Horizontal spectrum exhibits an horizontal dominant line on the \( f_x \) axis from low to high spatial frequencies and representative of this class is a city with tall buildings. Cross shape has equal representation from vertical and horizontal direction and a typical scene is an indoor image. Power spectrum of vertical scene shows a vertically dominant line along \( f_y \), cue to the fact that the scene has horizontal structure ans as such mainly representative of coast and fields. Oriented at about 45 degrees to the axes, oblique shapes represent mountains. Circular shapes on the other hand represent forests, fields etc. and have an isotropic power spectrum.

From the above discussion, its clear that cross shape is of not much interest to us as we are essentially classifying outdoor signs. And they also inspire us to classify the images based on naturalness first. Since cross and horizontal shape proclaim that the image is most probably artificial and the other three point towards natural images. Furthermore, due to representative nature of the above shapes, we are best served by this method if we go for the 8-category classification we have adhered to.

Now that the scheme is decided, they go for preprocessing of images, which is needed for two reasons:
1. reduce effect of large shadows;
2. minimize effect of high contrast objects with the potential to disturb power spectrum of background.
The process is simple. First they apply logarithmic functions to intensity distribution. Then they attenuate very low spatial freq by high pass filtering. After that adjustment of local standard deviation at each pixel is done. These steps are carried out in our work too. But then, the prime component, that of choosing the features, is slightly different here.

Instead of using a single Gabor at each stage, they use a combination of two filters and finally take a normalized version as their feature vector and this brings about some difference in performance. Then instead of using all the feature vectors as we have done, they use PCA to find 8 dominant features and project the image onto this plane. Then onwards they follow the same approach of Discriminant Analysis to place them on a semantic axis. The results are impressive. They got 90% success rate in classifying images according to their naturalness. Openness and expandedness brought about 88% and 8% success respectively. So, even though their results are more impressive than ours, it falls short of [4], the original paper our work is based upon.

8.2 Semantic typicality measure approach

[5] The paper goes for a different approach than what we have dealt with so far. Oliva and previous works focused on global information rather than localized ones. Oliva also attached global labels to images based on local and global features. However, even he didn’t use intermediate semantic annotations. The present paper emphasizes the importance of typicality. They argue that typical items serve as better cognitive reference points. In fact, learning of category representations is faster if subjects are taught on mostly typical items, since typical items have common attributes. Therefore, they propose the use of ‘local semantic categories’ as scene category attributes. They found out 9 local semantic concepts that are most discriminant: sky, water, grass, trunks, foliage, field, rocks, flowers and sand. Once they were decided, they were extracted on arbitrary 10 by 10 grids of the image. Their frequency of occurrence was determined and the ‘concept occurrence vector’ calculated. For each group of classification, prototypes of concept vectors were found on the basis of statistics of occurrence of each semantic concept.

The new images to be classified were tested with these prototypes to decide which class they belong to. The results they got are as follows: with annotated concept regions, the classification rate was 89.3%. However, when semantic concept classifier was used, the rate dropped to 67.2%. On the other hand, if second best match was also considered, the success rate was a whooping 98% and 83.1% respectively which might be interpreted as that the wrong classifications were actually border cases.

Though the performance was good, even with the second best test, they could not match the performance of the original paper. At the same time, the method is heavily dependent upon the performance of concept classifiers. For example, sky, foliage, rocks are easily and accurately classified and thus categories of forest, mountain and sky/clouds gave better results. On the other hand, field is confused with foliage or rock, so plains were mis-categorized.

8.3 Bayesian Hierarchical Model

[6] The paper claims its different in the sense that it eliminates the process of human-annotations of required features. To make it clear, we might note that [1] asked people to annotate pictures on the basis of openness, roughness etc. and then went for machine learning. In [4], the local semantic concepts were experimentally determined through human participation (the subjects decided upon the 9 concepts of rock,
sky etc.). Here they claim to learn such things directly from the picture without involving humans. Its true that the images are differentiated into the prime categories of coast, mountain etc. through human effort, but no effort is spent in classifying underlying concepts and features once the category of the image is decided upon.

First the images are classified into 13 sections, the 8 already mentioned and some more indoor categories. Thereafter, local features are extracted to form a `codebook` that essentially is a set of factors favouring the particular image type, learned without human supervision. Each image is represented through code words and the machine learns Bayesian hierarchical models for each class.

Now that learning stage is over, when we need to classify a new image, the unknown image goes through local feature extraction and from the previously learned code book its model is formed and compared with the previous Bayesian models to get the best possible match. Local regions are extracted through various methods like evenly sampled grids of 10-by-10, random sampling (500 per image), Brady Saliency Detector etc.

The performance was not that good as compared to the previous papers. The average score was 64%. If second best option was also taken, the performance became 82.3%. Low efficiency was essentially attributed to the difficulty in classifying 4 indoor types(kitchen, bedroom, office, living room) of 13 classes used. For the core group of forests, mountains, coasts etc., the efficiency was 74%, which combined with no human annotations and the fact that the learning sample was small, is respectable.

But still, our paper achieved a better efficiency using simpler methods.

### 8.4 Spatial Pyramid Matching

[7] The paper introduces the concept of recognizing scenes on approximate global geometric correspondence. They say that global characteristics only help in getting `gist` of the image. With this method, they claim to be able to get high recognition rates on challenging data. They define `locally orderless images` as something that returns the histogram of image features aggregated over a defined fixed hierarchy of rectangular windows. They play very imp role in visual perception. The method involves sub-dividing and dis-ordering.

The old method of pyramid matching is as follows: A sequence of increasingly coarser grids are placed over feature space and the weighted sum of the number of matches that occur at each level of resolution are taken. If matches are found at finer resolution, more weightage is given. Here they used spatial pyramid matching. This essentially involves performing pyramid match in 2D image space and using traditional clustering techniques in feature space, i.e quantizing all feature vectors into discrete types, and assuming only feature of the same type can be matched to one another. After that feature extraction is done with respect to oriented edge points, i.e. points with gradient magnitude in a given direction exceeding a minimum threshold. Also, SIFT descriptors of 16-by-16 pixel patches computed over a grid with spacing of 8 pixels are used. In the training phase, multi-class classification is done through SVM which is trained using one-vs-all rule.

This method provides an efficiency of 74.7% on the 13 classes. Main problem is as before, in distinguishing indoor categories. place the confusion matrix of the paper here So, on the whole, the results were not far better from Oliva.
8.5 Colours in Recognition

[2] This paper review was essentially to get an idea of impact of color on recognition. The previous papers discussed here, along with our own method, either operate on gray scale images or the convert colored images to gray scale before going through the process. But the paper relates three experiments that point towards the importance of color in scene recognition. What they have found is that color influences recognition when it’s diagnostic in nature, i.e. it’s representative and helpful in classifying the class. For example, color is considered diagnostic in classifying beach, desert, forest, field etc. and non-diagnostic in identifying kitchen, diner etc. Our project, concerned with the diagnostic group, is in need of this implementation. The paper also points to the fact that color helps in recognition at coarse spatial scale. So there is a possibility of coarse organization of diagonistically colored blobs to effectively support the categorization of complex visual scenes, that might be further looked at.

8.6 Object based recognition

[8] This paper presents a good idea of the bottom up method, i.e. scene recognition through segmentation of images into regions instead of going for the global picture. They propose a new image classifier. the feature extraction goes through the following steps: a. Collecting region types-A training set of generic images is prepared, each segmented into regions using the mean-shift algorithm. The regions are then pulled together and indexed using standard feature, and clustered using Fuzzy C-means algorithm. Each category then represents a region type. b. The presence vector of the image is formed by assigning 1 or 0 to each region type based upon its presence or absence in the image. Thus, we get a vector representation of the image. c. Feature selection- Through entropy considerations on the training sample a few of the image regions are selected as features d. After that Kernel-Adatron classifier is used to classify the images. The statistics are like this: Train error was around 4% over 4 categories of snowy, countryside, people, streets. Test error was about 8% over all the categories they tested. As we see, they gave pretty good results. So we might further look at this and integrate this bottom up with our approach to try to get a better result.

8.7 ARTSCENE

[?] This one is a pretty recent paper involving scene recognition. The basis of the paper is that neither local nor global information is more predictive than the other at all times (humans did better a job at categorizing rivers, mountains etc when presented images were globally blurred than locally scrambled; converse was true in case of coast, forest, plain). So brain uses scenic info from multiple scales.

It assumes that we first take the gist of the image at a first glance, before we go for detailed fixation. So, they extracted gist as a 304-D feature vector. Gist is first learned by spatial layout of colors and orientation. Spatial attention is then sequentially drawn to principal textures. The database used is the same as ours.

During training, images are broken down into gist, 1st principal texture, 2nd principal texture and 3rd principal texture. Then they are categorized in similar boxes for each scene class.

The test image is then divided into the same 4 parts of gist, 1st, 2nd and 3rd principal textures and each one compared with the trained boxes. Then evidence is accumulated and their weighted sum is taken to decide the class to which the image belongs. A detailed discussion of the gist feature vector is due here.
Gist is found using dominant orientation (ex: horizontal in coast, vertical in forest etc.). Basically, the figure is split into 16 patches, each patch characterized by the average values of 4 orientation contrasts at four different scales. So vector G is 304D orientation and color vector for each of 16 patches (19*16=304).

Again, usually 92.7% of total area of landscape of image is covered by 3 principal textures. to decide these textures, common object recognition approach is taken and the textures are sorted as 1st to 3rd by their relative size in the visual field.

Using only gist, they got 81% results on an average and the maximum score was 90.22%. However, when they used the three features along with the gist, the respective scores were 83% and 91.85%. The results are pretty good and comparable to our original paper. The confusion in classification was mainly between coast and countryside, forest and mountain, and mountain and countryside. Examination showed, forest-mountain misclassification was due to co-occurrence of trees and mountains. Country side is loosely defined; hence its conflict with others. It argues that even humans face the same conflicts during classification.

8.8 Spatial Envelope Approach

[1] This one is a pretty detailed paper on scene classification based upon global features, and actually an extension of the paper our present work is based on. The concept of spatial envelopes is introduced here. The spatial envelopes have the following properties:

- degree of naturalness: structure of scene differs in man-made and natural environments. Straight lines of horizontal and vertical nature dominate man-made scenes, where as natural images show textured zones and undulating contours.

- degree of openness: Openness gives a sense of enclosure. A scene can be enclosed by visual references like in the case of forest, mountains etc or they can be vast like the coastal area.

- degree of roughness: roughness in a scene primarily refers to the size of its major components. It depends on the size of elements at each spatial scale, their abilities to build complex elements and their relations between elements that are also assembled to build other structures.

- degree of expansion: especially the concept of parallel lines giving the perception of depth. so a flat view of a building has low degree of expansion where as a street with long vanishing lines has a high degree of expansion.

- degree of ruggedness: it refers to the deviation of the ground with respect to the horizon. A rugged environment produces oblique contours in the picture and hides the horizon.

The images are classified on this basis. As in the original paper, DSTs are calculated (here in addition to the DST we have encountered before, a windowed DST, WDST, is also used to compare the results). Karhunen-Loeve (KL) decomposition of energy spectrum and localized energy spectrum (found by using WFT on 8-by-8 windows) provides 2 sets of features. So, we now have some new features and a new WDST. Use of WDST shows radical changes in results. It improves the 86% performance of DST to 92%. Along with this, the paper also proposes to go for indoor vs outdoor classification and also new factors that can be taken into account like depth perception and environmental scenes vs object discrimination.
The approach taken here is a bit more complicated than what we have implemented, but it gives very good results.

9 Conclusion and Future Work

Tasks such as scene classification pose a tough problem, specially due to large variations in everyday images. It is rather tough to capture this tremendous amount of variation. Selection of appropriate features poses a tough challenge. The selection procedure seems rather non intuitive and maybe modelling the human vision requires much greater insight into the understanding the way our eyes processes information. We tried experimenting with various features but they didn’t work out quite well. Taking cues from the past work we took the approach of finding the gist of the scene using gabor functions. We deviated from the past and a used a tree classification approach. A tree classification approach very high accuracy at each and every step, else cumulation of errors at each levels and their further percolation can affect the overall accuracy of the classifier badly. Our classifier too faced the same problem. We tried limiting the number of false detections with help of first and second order statistics of projections of image vectors on classification vectors, but due to large variability in data these methods proved to be inefficient. We had tried bounding the projection values in some range like \( \pm k \ast S.D. \) and tried to optimize k, but the results were rather poor. In further work, we can include the color and texture information in the feature vector considered and try out different permutations through learning to get the best combination of different levels in the tree classification hierarchy to improve the results. We also intend to provide a reliability measure in the future.

References


10 Appendices

10.1 Appendix 1

(a) Forest
(b) Mountain

(c) Street

(d) Highway