SCENE CLASSIFICATION

DIP Project for Group-I at IIIT Sri City

Group Members

Chris Andrew - IS201401015
Santhoshini Reddy - IS201401040
Nikath Yasmeen - IS201401042
Sai Hima - IS201401038
Sriya Ragini - IS201411009

Introduction

The problem of scene classification is a vital problem in the field of computer vision that models the human perception of visual information and maps it to a mathematical problem. The problem consists of being able to classify a given scene into one or more of specified or unspecified categories of scenes based on the content of an image that given as input. Usual approaches to scene classification include extracting a set of features from the image and using a classifier to discriminate between images based on these features.

With advances in the field of machine learning, people are now able to classify images without providing either categories or features using Deep Neural Networks and Convolutional Neural Nets. These machine learning algorithms can not only discriminate between different scenes in an image but can also provide a description of it's contents.

In our project, we have limited our approach to the feature based classification method using two categories of scenes. We differentiate between **Natural** and **Artificial** scenes. To do so we provide images of coasts, forests, mountains and open fields which are to be classified as **Natural** scenes and images of highways, streets, cities and building which are to be classified as **Artificial** scenes.

Survey

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, [2] 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 ([3]). So effectively, at a first glance, we just assimilate the gist of an image which is built on a low resolution spatial configuration. Torralbla et. al [4] proposed three semantic axes along which real world images are organized.. The semantic axes suggested by them were: degree of naturalness of a scene, degree of openness of a scene and the degree of verticalness of a scene. They used linear discriminative filters to calculate the degree for each scene. Whereas, Lazebnik et.al [5] achieved high performance results by generating a spatial pyramid containing the image information. Spatial pyramid was formed by dividing the image into fine sub-regions and computing histograms of the local features for the sub-regions. [6] proposes a segmentation based approach and a classification scheme in the descriptor space to solve the problem by using kernel-based methods and feature selection. [7] proposes a typicality measure for scene classification problem, wherein ambiguous images can be marked as being less typical for a particular image category, or the transition between two categories can be determined. [8 showed that multiple-instance learning can be used to classify images for natural scenes.

We found that Olivia and Torralba's [1] work in this area to be of considerable significance and we used their research as a reference for our project.

Hypothesis

Gabor based features have been proven to be good descriptors in many cases for facial recognition, document analysis, etc. The use of Gabor features in scene classification is warranted by the fact that images of man-made or artificial structures are more regular than natural patterns. This can be observed in the images of a building and a tree. When converted to grayscale both have a similar structure to them, the difference lies in the regularity in pattern of the two images. Using this as observation we researched a paper that utilised Gabor based features for scene classification.

The Power Spectrum of the image is the square of the absolute values of the Fourier Transform of an image.

$$\Gamma(f_x, f_y) = |FTi(x, y)|^2 \quad (1)$$

The Gabor filtering of an image is obtained by convolving a rotated Gaussian filter or the Gabor kernel over an image, the resultant image is said to be Gabor Transformed.

$$G(x, y) = i(x, y) * g(x, y) \quad (2)$$

The Gabor transform is done using the following equation, where f is the frequency and T is the orientation.

$$G_x(t, f) = \int_{-\infty}^{\infty} e^{-\pi(\tau - t)^2} e^{-j2\pi f\tau} x(\tau)\, d\tau$$

Gabor filters of inverse symmetry can be obtained by separating the sin and the cosine terms from the above equation.

The features described in the work of [2] used Discriminant Spectral Templates(DST) features for classification of images which are based on Gabor Features of the image. The DST is a weighted summation of the Gabor convolved images using different filters.

$$DST(f_x, f_y) = \sum_{n=1}^{N} d_n G_n(f_x, f_y)^2 \quad (3)$$

The final feature described is the combination of both the Power Spectrum and the Discriminant Spectral feature.

$$u = \int \int \Gamma(f_x, f_y) DST(f_x, f_y) df_x df_y \quad (4)$$

The value of d_n is currently unknown for the given set set of images and must be calculated during the training phase. We do this using Fischer's Linear Discriminant Analysis. The set of features from each filter is computed and converted to a feature vector. The vector is then fed into the LDA where the features are linearly combined to get the best representation using a supervised learning algorithm. We therefore do not worry about the values of d_n at this stage.

We also introduced certain modifications to the method proposed [1] by modifying the features to compute the mean and the standard deviations of the filtered images rather than the integral. We found that this approach performed better than the integral by increasing our accuracy by more than 9%.

Procedure

The procedure or algorithm that was followed for the feature extraction process is as follows:

Each image in the dataset is first resized into a 256x256 image. The image is further divided into 16 regions of 64x64 each.

On each of the regions, we calculate the Power spectrum of the region as described in (1). We also compute the DST feature using

a specified Gabor Filter as described in (3). We do this using filters of opposite symmetry and compute the difference between the two. We later multiply this difference with the Power Spectrum.

This procedure is repeated for the same region for frequencies of 4, 8, 16, 32 cycles and orientations of 0, 45, 90, 135, 180, 225, 270, 315 degrees. This gives us a total of 32 filters which are applied to each region.

The mean and the standard deviation of each filtered region is computed and stored in a feature vector. The feature vector for one image therefore will have 16x32 = 512 features.

The feature vector for all the images in the dataset are computed and labeled accordingly as **Natural** or **Artificial**. This set of features are then given to the LDA module to learn and reduce. We obtained a one dimensional feature after reduction using our 512 dimensional feature set. This feature is used as our discriminating feature.

We classify our images using the obtained features with the help of the Nearest Neighbour Classifier.

Implementation

An implementation of the above defined procedure is attached along with this report. The "Readme" file present is a helpful documentation whose aim is to assist in the execution of the implementation. Please refer to the "Readme" file for instructions about the execution process.

Data

The data that we used to evaluate our method is MIT's Urban and Natural Scene Categories dataset. The set consists of 2688 images. The images consist of scenes of Coast/Beach, Open Country, Forest, Mountain which are classified as **Natural** and scenes of Highways, Streets, Cities and Building which are classified as **Artificial.** The data can be downloaded from http://cvcl.mit.edu/database.htm

Results

Features	Correctly Classified	Incorrectly Classified	Percentage Accuracy
Integral Feature	2257	431	83.96%
Mean and deviation	2367	321	88.058

We were able to obtain considerable accuracy using the Mean and Deviation features derived from our procedure. The final feature vector is a single value that is able to accurately classify between **Natural** and **Artificial** scene images.

Conclusion

We classified images according to global features into two basic categories, Artificial and Natural. Features were extracted and combined using a supervised learning algorithm. The Nearest Neighbour Classifier is employed for classification.

REFERENCES

[1] Oliva, Aude, and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope." *International journal of computer vision* 42.3 (2001): 145-175.

[2] Oliva A., Torralba A., Guerin-Dugue A., and Herault J., "*Global semantic classification of scenes using power spectrum templates*", CIR99, Elect. work in ComputingSeries, Springer-Verlag, Newcastle.1999, 1999.

[3] Julia Vogel and Bernt Schiele, "*A semantic typicality measure for natural scene categorization*", Pattern Recognition Symposium DAGM, 2004.

[4] Torralba A. and Oliva A., "Semantic organisation of scenes using discriminat structural templates," Proceedings of International Conference on Computer Vision ICCV99 Korfu Greece, pp. 1253–1258, 1999.

[5] Lazebnik S., Schmid C., and Ponce J., "Beyond bags of features: Spatial pyramid matching for recognising natural scene categories," CVR-TR-2005-04, 2004.

[6] Le Saux B. and Amato G., "Image classifier for scene analysis," Computer Vision and Graphics International Conference, ICCVG 2004, Warsaw, Poland, September 2004.

[7] Julia Vogel , Bernt Schiele , 2004, 'A semantic typicality measure for natural scene categorization', Pattern Recognition Symposium, DAGM

[8] Oded Maron, Aparna Lakshmi Ratan, 1998, 'Multiple-Instance Learning for Natural Scene Classification', Proceedings of the Fifteenth International Conference on Machine Learning, 341 - 349