DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

final report for

Applications of High Performance Statistical Modeling to Scenic Beauty Estimation

submitted to fulfill the semester project requirement for

EE 8993 Pattern Recognition

May 5, 1999

submitted to:

Dr. N. H. Younan

Department of Electrical and Computer Engineering
225 Simrall, Hardy Rd.
Mississippi State University
Box 9571
MS State, MS 39762

submitted by:

Zhiling Long and Xinping Zhang

The Institute for Signal and Information Processing

Mississippi State University
Box 9671
Mississippi State, Mississippi 39762
Tel: 601-325-8335
Fax: 601-325-8192
email: {long, zhang}@isip.msstate.edu

1 ABSTRACT

The objective in the Scenic Beauty Estimation (SBE) problem is to develop an automatic classification algorithm that matches human subjective ratings. Previously, various algorithms have been applied to this problem, with limited success. In this project, we developed a new statistical modeling technique. For this scheme, we divided each image into three regions according to their nature, extracted features from each region separately and then applied the algorithm of Principal Components Analysis (PCA) respectively to each region to do image classification. Compared with the previous PCA experiments carried out on the whole images, this new classification method generated better results on some aspects.

2 INTRODUCTION

The United States Forest Service (USFS) has a long-term interest in the development of automatic methods [1] for managing forest resources. These methods use a database of forestry images to determine the utility of a plot of forest land both in terms of timber use and scenic quality. Traditional methods used to determine the scenic quality are very tedious and involve a large group of people manually rating each of the images. Obviously, an automatic method has advantages both in consistency and efficiency.

To facilitate this research, an extensive database of forestry images [3] has been developed, which contains 637 images taken from Winona National Forest under all kinds of controlled conditions. And a wide array of features were extracted from these images. Currently a total of 65 features are available, which include histograms for red, green, blue, yellow and brown, count of long lines and short lines, entropy, sharpness, standard deviation, compression ratio and fractal dimensions. Also various pattern recognition algorithms, such as Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Support Vector Machines (SVMs) and Decision Trees (DT), have been investigated on this problem [1][3][4]. Their performance is shown in the table below.

Table 1: Performance of different algorithms

Algorithm	Average error rate
PCA	37.7%
LDA	35.2%
ICA	33.4%
SVMs	32.2%
DT	43%

As we can see, the classification results are not very satisfactory in general. Can we do better? Noting the fact that a typical forestry image usually can be divided roughly into three strips, with the upper strip standing for the sky, the middle one standing for trees and foliage, and the bottom strip standing for grass and ground, and the fact that the significant features for those different regions may not be the same due to their various nature, we believe that if we split the images into regions and do classification on those regions respectively, we may get some better results. Based on this, we investigated PCA on split images in this project. We segmented images evenly into three horizontal strips, extracted features from each region, and applied PCA on those regions respectively. Our experiments generated encouraging results and indicated that this segmentation scheme is very promising in the scenic beauty estimation problem.

3 THEORY OF PRINCIPAL COMPONENTS ANALYSIS

Here we will describe briefly the theory of Principal Components Analysis [5].

PCA is a kind of linear classification algorithm. In PCA we use the distances from the test vector to the sample means to do classification. But we don't directly use the Euclidean distances in the feature space. Instead, we use distances weighted by the covariance matrices as follows,

$$d = \sqrt{(\vec{x} - \vec{x}_{\mu})^{T} C_{x}^{-1} (\vec{x} - \vec{x}_{\mu})}$$

In PCA, it is assumed that the first principal component of a sample vector lies parallel to the direction along which there is the largest variance over all samples. This direction corresponds to the eigenvector associated with the largest eigenvalue. The k^{th} principal component is chosen to be the linear combination of the input features that has the largest variance, under the constraint that it is also uncorrelated to the previous *k-1* principal components. PCA works very well when the direction along which there is maximum variation also contains the information about the class discrimination.

4 EXPERIMENTS AND ANALYSIS

First, we did experiments on all three regions, using the feature vector of blue+brown. The results were compared with those generated by PCA algorithm on the non-split images, as shown in the following table.

Table 2: Results using feature vector of blue+brown

Region	Data set 1	Data set 2	Data set 3	Data set 4	Average
top	44.0%	38.0%	43.8%	45.0%	42.7%
middle	37.7%	46.1%	71.3%	60.0%	52.5%
bottom	50.3%	35.4%	53.1%	42.5%	45.3%
complete	31.4%	50.6%	35.0%	56.9%	43.5%

From these results, we found some encouraging information. First, for different regions, the performance of PCA varied significantly on the same feature vector. Secondly, we did get some improvement by this splitting method. Classification error rate based on the complete images is 43.5%, while we achieved 42.7% on the top region. These preliminary experiments confirmed our assumption that it should make sense to try classification algorithms on different regions respectively and we could get better performance by this splitting method.

Then we did experiments on each region separately. For each region, we chose some feature vectors supposed to be related closely to that specific region. Also these features have been proved to be significant features by previous PCA experiments on the complete images. By these experiments, we expected to hit the significant feature vectors and achieve some better classification error rate. The features we used were: blue, green, brown, ylw+brn, blue+brn, rgb, rgb+ent. Here, ylw+brn stands for yellow+brown; blue+brn stands for blue+brown; rgb means red+green+blue; and rgb+ent is red+green+blue+entropy. The results are shown in the following tables.

Table 3: Classification results on the top region

Features	Data set 1	Data set 2	Data set 3	Data set 4	Average
blue	48.4%	31.6%	43.1%	50.0%	43.3%
rgb	51.6%	30.4%	66.9%	36.9%	46.5%
rgb+ent	54.1%	27.8%	30.6%	35.6%	37.0%

Table 4: Classification results on the middle region

Features	Data set 1	Data set 2	Data set 3	Data set 4	Average
green	49.7%	58.9%	65.0%	60.6%	58.5%
brown	44.7%	57.0%	66.9%	51.9%	55.1%
rgb	43.4%	40.5%	39.4%	41.9%	41.3%
rgb+ent	42.8%	40.5%	31.3%	40.6%	38.8%

Table 5: Classification results on the bottom region

Features	Data set 1	Data set 2	Data set 3	Data set 4	Average
brown	37.7%	65.8%	46.3%	41.3%	47.8%
ylw+brn	44.0%	52.5%	45.0%	41.3%	45.7%
rgb	45.3%	32.9%	41.3%	33.1%	38.2%
rgb+ent	38.4%	32.3%	43.8%	36.9%	37.9%

As we can easily understand, the performance of PCA on the same region by different feature vectors varied. We chose the best classification results for each region and compared them with the best average misclassification rate we got previously for the complete images.

Table 6: Comparison with PCA on complete images

Region	Features	Error rate
top	rgb+ent	37.0%
middle	rgb+ent	38.8%
bottom	rgb+ent	37.9%
complete	rgb+ent	37.7%

Obviously, the best results of this splitting scheme are comparable to those of the non-splitting version. On top region, we even achieved a better result, though it is not a very significant improvement. Recall we divided the images evenly into three regions. That certainly was a very rough segmentation. If we can segment the images more accurately, we are sure we will get much better results. With larger size of database, more extensive exploration on feature combinations, and more accurate segmentation, we believe this segmentation scheme will undoubtedly enhance the performance of those advanced pattern recognition algorithms on this scenic beauty estimation problem.

5 CONCLUSION

Based on the experiment results and the analysis shown above, we can arrive at the following conclusions. First, features play a very important role in the SBE problem. Secondly, it is reasonable to split forestry images into regions according to their different nature. And significant features for one specific region are not necessarily good features for other regions. Finally, because we achieved 37.0% error rate by this evenly-splitting scheme, which is better than the previous best result of 37.7% for the non-splitting version, we believe that a more accurate segmentation scenario will be very promising.

6 REFERENCES

[1] N. Kalidindi, A. Le, L. Zheng, H. Yaqin, V. Rudis and J. Picone, "Scenic Beauty Estimate of Forestry Images," *Proceedings of the IEEE Southeastcon*, pp. 337-339, Blacksburg, Virginia, USA, April 1997.

[2] N. Kalidindi, A. Le, J. Picone and V. A. Rudis, "The USFS Scenic Beauty Estimation Database," *USFS Southern Research Station*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, December 1996.

[3] N. Kalidindi, "Scenic Beauty Estimation of Forestry Images," http://www.isip.msstate.edu/publications/seminars/masters_oral/1997/usfs_imaging/index.html, Master of Science Special Project Presentation, Mississippi State University, MS State, MS, USA, November 1997.

[4] X. Zhang, V. Ramani, Z. Long, Y. Zeng, A. Ganapathiraju and J. Picone, "Scenic Beauty Estimation Using Independent Component Analysis and Support Vector Machines," *Proceedings of IEEE Southeastcon*, pp. 274-277, Lexington, Kentucky, USA, March 1999.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, California, USA, 1990.

7 ATTACHMENT

c++ source code for this project is of very large size, and will be provided on requirement.