CLASSIFICATION TECHNIQUES FOR DIGITAL MAP COMPRESSION

H. Potlapalli, M.Y. Jaisimha, H. Barad, and A.B. Martinez
Dept. of Electrical Engineering, Tulane University,
New Orleans, LA 70118

M.C. Lohrenz and J. Ryan
Naval Ocean Research & Development Activity
J.C.Stennis Space Center, MS 39529

J. Pollard
Planning Systems, Inc., Slidell, LA 70458

ABSTRACT

The performance of image classification techniques as applied to color cartographic maps is compared. These color maps have a lot of graininess due to imperfections in the printing process. This graininess decreases the efficiency of compression techniques. The color maps are classified using the K-means clustering algorithm and vector quantization with neighborhood classification to improve the visual quality and compression ratio. The classification is performed in various image representation schemes. The performance of the classifier is evaluated based on the visual quality of the classified image, the time required to classify the image and compression achieved on the classified image. The compression ratio after classification was higher than before classification.

INTRODUCTION

This paper presents an evaluation of classification techniques as applied to cartographic color maps. The maps have certain specific characteristics such as large homogeneous regions and fine detail. These maps are digitized versions of printed maps. Due to the nature of the printing of the maps, the images have a grainy salt and pepper character with each pixel a different color. The printing process, which involves half-toning, also introduces some inhomogeneity in the coloring. Further, the maps have variation in coloring due to the different ages and printing techniques of the maps. The result of all this is that regions in the maps which should possess uniform coloring instead have non-uniform coloring. The maps also have fine detail such as lines of latitude and longitude and lettering. The classification techniques being considered are used to remove the color variation, to enhance the image and to increase the compression achieved over the unclassified image.

The K-means clustering algorithm and the Vector Quantizer algorithm are used to classify the image into a much smaller number of classes (8 as compared to the 256 colors present in the original images). A discussion of the approaches employed and the results obtained will be presented in the next sections.

The performance of the classifier is evaluated based on the following criteria:

1. Image quality,
2. Time required to classify the image, and,
3. Compression ratio.

The compression techniques used to determine the compression performance are run-length coding [2], [8] and Lempel-Ziv coding [10], [11]. These techniques are chosen because they are most efficient when applied to images with large homogeneous regions. The unclassified map has very few such regions due to the graininess. The performance of a classifier is measured by the extent of region smoothing obtained with that classifier. This region smoothing is reflected in the performance of the coder.

IMAGE CLASSIFICATION

Image classification, in essence, tries to allocate pixels to a fixed number of classes based on the pixel neighborhood and information drawn from the image. The color assigned to a class is the mean of the pixels assigned to that class (i.e.cluster). Classification removes color variation in a region.

Classification methods are of 4 major types:

1) methods based on the stochastic model of the scene,
2) methods based on simultaneous classification of all pixels, for example Markov random-field models,
3) relaxation methods that iteratively modify posterior probabilities using information from an increasing neighborhood,and
4) methods using non-contextual rules based on transformed data [5].

The image possesses non-uniform pixel values in regions which are to have only uniform values. A classification algorithm should be employed in order to make these non-uniform regions uniform, while at the same time preserving fine details such as lettering. K means classification was applied to a three dimensional histogram of the data.

CH2714-4/89/0000/0268$01.00 © 1989 IEEE

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 1989		**00-00-1989 to 00-00-1989**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Classification Techniques for Digital Map Compression	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Ocean Research & Development Activity,Stennis Space Center ,MS,39529**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the IEEE: 21st Southeastern Symposium on System Theory, pp. 268-272, Tallahassee, FL, 26-28 March

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**5**	

The 3-D histogram was obtained by plotting the data under three distinct representation schemes:

1) RGB

2) XYZ

where $X = R$

$Y = R - G$

$Z = R - B$

3) YIQ

While RGB is the simplest coordinate scheme(among RGB, YIQ, and XYZ), other schemes are evaluated for any improvement in cluster separation that they may offer. YIQ representation compacts most of the information along the I axis. This decorrelating effect implies that the I component alone need be used.in the calculation of Euclidian distances. Principal component analysis has the property of completely decorrelating the data. However the computational complexity involved in calculating the principal components makes this approach unattractive.

Given a set of data in N-dimensional space, clustering algorithms group the data points into clusters. Typically most clustering algorithms involve the minimization or maximization of some distance criterion. Once the 3-D plot is obtained, the clustering algorithms proceed to assign each point to a particular cluster. The choice of a clustering algorithm was affected by the large number of data points (i.e.575 rows x 639 columns = 367425). Hierarchical clustering would also prove to be unwieldy because of the overhead and memory requirement that would result from having to store a tree of such a large data set. Basic agglomerative clustering involves the computation of distances between all pairs of points. This is clearly unacceptable for the large data set under consideration. The requirement to have an unsupervised classification scheme along with the need to keep computation time to a minimum led to the choice of the K-means algorithm [1,4] and the VQ algorithm [6,8].

K-MEANS CLASSIFICATION

The K-means algorithm relies on the choice of an arbitrary starting partition and moves the particles from one class to another if this will improve a certain criterion function. Since the sample set is finite, there are only a finite number of partitions of the space under consideration. Thus the clustering problem could be solved by exhaustively enumerating the partitions. Considering the large number of data, this is unfeasible. Starting from this arbitrary partition, the K-means algorithm gives results which are only locally optimal and different starting points could lead to different solutions [1].

The K-means algorithm method relies on the search method, changing partitions by moving objects from one class to another. The algorithm is a pointwise classification algorithm. The mean of each of the classes in the starting partition is computed. Since each of the particles or bins being clustered have different weights depending on the number of pixels per bin the mean of each cluster is computed in a weighted fashion.

To avoid the computational load of computing the means after each pixel reassignment, the means were recomputed after each iteration. For each point the distances to each of the cluster means is computed. The measure of distance used is Euclidian. The point is then assigned to the cluster to which it is closest.

The K-means algorithm is described below:

Let x be a point in 3-D space.

Let there be M classes each with mean m.

Let Number be the number of pixels changing class in each iteration.

Select an initial arbitrary partition of the space into M classes

compute the means for each cluster(class).

#1 : For each sample x

Compute the distance to each of the cluster means given by the following expressions,

$$\rho_j = \frac{n_j}{n_j + 1} \| x - m_j \|^2 \qquad i \neq j$$

$$\rho_j = \frac{n_i}{n_i - 1} \| x - m_i \|^2 \qquad i = j$$

Assign x to the class for which ρ is minimum.
If the new class is not the same as the old class
increment Number by one;

next x;

If no more samples remain, recompute the means m;

If the number of pixels changing classes remains the same in n attempts, stop; else go to #1.

VECTOR QUANTIZER CLASSIFICATION

Another classification algorithm that utilizes a distance criterion to form clusters of pixels is Vector Quantizer classification[6][3]. Let N gray-levels be chosen as representative gray-levels. These gray-levels are called the "codewords" and the collection of codewords is called the codebook. Then, for every input pixel, the Euclidian (RGB) distance from this pixel to every element in the codebook is computed. The RGB value of the codeword closest to the input element is the assigned to that pixel.

More rigorously, a VQ(Vector Quantizer) is defined as a mapping from a K-dimensional space to a finite subset Y of the space. The subspace is assumed to be representative of the entire space, and is so designed that any point in the space can be represented by a element in Y with a small but acceptable error.

The design of a VQ is based on the LBG algorithm [8]. N points are randomly selected from the population space. The population is passed through the VQ, and codewords are assigned to each point in the population based on the minimum distance rule. Each codeword is then replaced by the mean of all the points that were assigned to that the codeword so that the codeword is gradually moved to the center of the class it represents.. The process is repeated till the codebook does not change. The algorithm is described below in pseudocode:

Set initial codebook A[0];

Set number of iterations = N;

Set number of pixels = P

```
while(codebook[i^th iteration] not = codebook[i-1] or i
    < N)
      for( p = pixel#1 to pixel#P)
            compute distance to A[i-1];
            find codeword a[i-1] closest to p;
            assign codeword a[i-1] to p;
      compute new a[i] = centroid of all p's assigned
      to a[i-1];
end.
```

If mean squared error is used to compute the distances, the centroid of a class is the mean of the class. The most important criterion is the selection of the codewords. Pixels representative of each class, the number of classes, and the classification criteria determine the performance of the algorithm. If a large variance in the population is expected then the training set should be large so that it contains all the possible points in the space. The training set is a large sample of the population. It will be used to adjust the codewords so that they are truly representative of the class.

One very simple way to determine the number of classes (of colors) is by visual inspection. Another is to assume that all images of one kind, say maps, have only so many principal colors (8 in the case of maps) and fix these as the number and types of classes. Once the representative colors have been selected, the image can be classified pixel by pixel, by measuring the distances to each codeword and assigning the color of the closest codeword to the pixel. This is called supervised VQ classification and is a very unsatisfactory approach. The method requires a lot of human interaction new classes and new representative pixels have to be determined for every image that needs to be classified. The classes selected are very subjective. Furthermore, the classifier needs to be redesigned for every image.

Unsupervised VQ classification involves very little human interaction. The subjectiveness in class-selection (common to supervised classification) is avoided. A small number of points in the image representation space, say RGB, are picked randomly such that they are distinct and a minimum distance(in the RGB co-ordinate system) apart.. These RGB values are coded into the codebook.

Typical distances are 70 units to separate codewords and typical widths of each class are 20 units. This means that every codeword is a minimum of 70 units distant from every other codeword and if a pixel is assigned a certain codeword then the original gray-level of that pixel was a maximum of 10 units from that codeword.

The training set has to be large, but for a small vector size(of 1 in this case) and small number of classes, the training set is approximated by the image itself. Once the optimal codebook has been designed, the image can be classified using this codebook and the minimum distance rule. If the image can be considered representative of a whole class of images then, the codebook can be hard-coded and used for other images too. This allows for considerable savings in computation time. If the codebook design algorithm is run with more than one image then it is possible that the codebook designed will indeed be the optimal universal codebook.

Due to the arbitrary choice of the starting partition(clustering approach) or codewords(Vector Quantizer approach) several pixels in the images were misclassified. A neighborhood classification scheme is used to eliminate these misclassified pixels. The main consideration in the formulation of a neighborhood classification algorithm is that the misclassified pixels were the same color as the lettering. The improved classification scheme has heuristics to determine whether a pixel is misclassified or merely belongs to some line or lettering on the image.

NEIGHBORHOOD CLASSIFICATION ALGORITHM

Once the map has been classified using either the K-means approach or the VQ approach, neighborhood based classification is performed to reclassify misclassified pixels using neighborhood information. The algorithm for neighborhood classification is described below in pseudocode.

For each pixel check the classes of its eight neighbors.

If (the class of each of the eight neighbors is different from that of the pixel and if each of the eight neighbors has the same class)

then

reclassify the pixel to the class of the eight neighbors.

If (some of the eight neighbors belong to the same class as the pixel under consideration)

then

check classes of four neighbors of the pixel

If(the four neighbors of the pixel have the same class as the pixel OR the four neighbors in either the horizontal or vertical directions are the same class)

then

reclassify the pixel to this class.

If (some of the pixels in the eight neighborhood are of the same class as the pixel AND the four neighbors of the pixel are not of the same class)

then

If(the diagonally placed eight neighbors are of the same class)

then

do not reclassify the pixel.

RESULTS

The number of classes was chosen as to be eight since this is the number of distinct colors required in the classified map. In the K-means classification approach, the number of pixels changing classes becomes constant and the loop terminates after about 20 iterations. The number of bins was reduced by a factor of 4 along each axis to 64 x 64 x 64. The output from this algorithm consisted of a three dimensional array of classes.

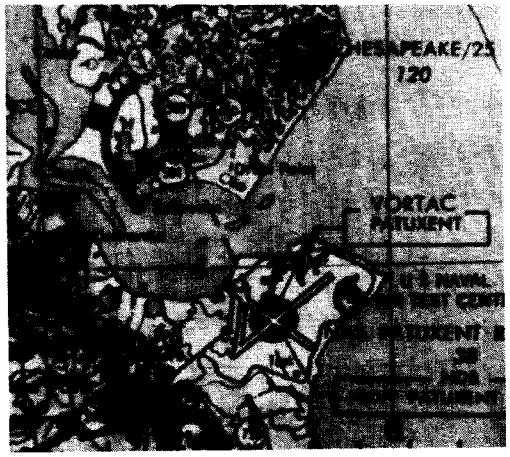

Fig. 1 Original unclassified map (green band)

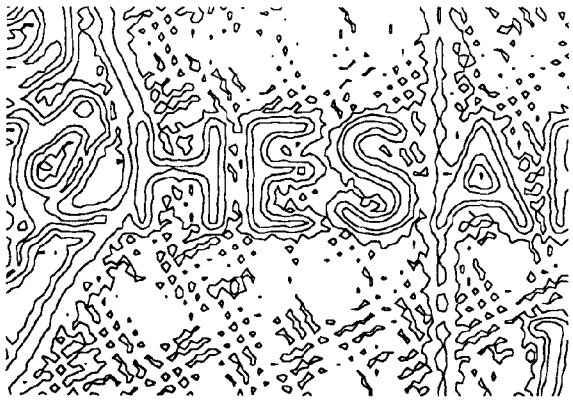

Fig. 2 Contours for unclassified map

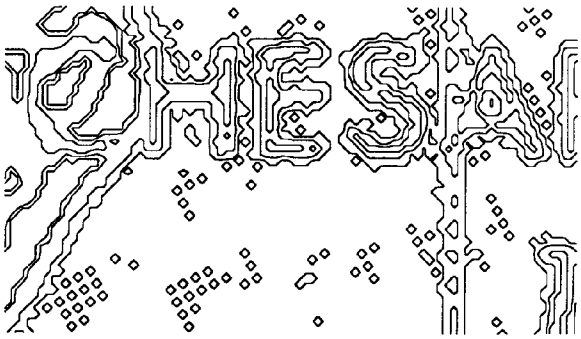

Fig. 3 Contours for VQ classified map

Fig. 4 Contours for K-means classified map

This array was passed to a another program and used as a lookup table to create a classified image.

The VQ classification terminates in less than 10 iterations on the average. The feature space was not partitioned unlike K-means classification. Codebook generation took most of the time. The codebook was not hardcoded; therefore the time required to classify an image using an existing codebook was not investigated. After the classification, some standard file compression methods, such as Lempel-Ziv and run-length encoding were applied to the classified image. The map size was 367425 bytes. The best result from Lempel-Ziv could compress the unclassified map to only 266 KB.

The K-means classified image gave a compression ratio of 7.34:1 while the VQ classified map gave a compression ratio of 7.9:1 using Lempel-Ziv coding. Neighborhood classification improved the performance of the K-means classification algorithm resulting in a compression of 8.06:1 with Lempel-Ziv coding. Run-length coding[2],[9] was performed on the classified images. The run-length coded VQ classified image occupied 0.167 MB of memory versus 0.914 MB for the unclassified image. This gives a compression ratio of 2.2:1 The XYZ and YIQ feature space representations did not improve the appearance of the image. The compression ratios achieved were also lesser than those obtained under RGB representation. The map classified under XYZ representation gave a compression ratio of 5.55:1 with Lempel-Ziv coding. YIQ representation gave a compression of 6.93:1 with Lempel-Ziv coding.

K-means classification was completed in under 480 seconds. VQ classification required 1600 seconds with XYZ, 1300 with RGB and 780 seconds with YIQ representations respectively.

On visual inspection the image did not possess the grainy nature of the original image, but still had some isolated pixels which were misclassified. The photograph of the map is shown in Fig.1. Contours were for a small region taken from the top right hand portion of the map. Figs. 2 to 5 show these contours

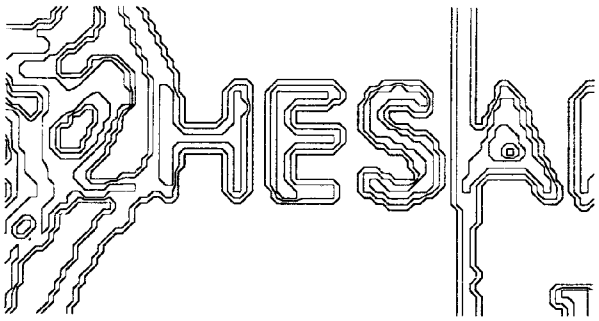

Fig. 5 Contours for K-means neighborhood classified map

for the classified maps. The extent of smoothing achieved through classification is a good indicator of the advantage of classification.The histogram of the gray-levels of the unclassified map, and classified map(Fig.6) shows the extent of color equalization.

CONCLUSIONS

In terms of computation times, K-means has a clear lead over the VQ classification scheme. However, the VQ classifier converges in fewer iterations than the K-means algorithm. The speed of the VQ classifier is decided by the number of classes and the minimum class separation desired. If the universal codebook is developed, then the classification time is small and the algorithm can be tailored to suit specific applications by adjusting the minimum-distance partition that identifies new codewords.

The best representation scheme in terms of compression ratio achieved, is RGB. The XYZ space does not decorrelate the pixel values along each axis and therefore gave poor results. This is useful only for reducing the number of distance computations since the computations are made only along one axis. The YIQ representation scheme was better than the XYZ scheme in terms of compression ratios. Computation time for classification using the YIQ feature space representation scheme is much lower than for both RGB and XYZ.

The algorithms eliminated almost all misclassified pixels that were present in the image. The K-means algorithm with neighborhood classification however resulted in the filling in of one of the letters as also in a deterioration in the quality of the lines.. An improved neighborhood classification algorithm with heuristics capable of recognizing lettering and lines is under development.

ACKNOWLEDGEMENT

This work was funded under Aircraft Procurement, Navy H1CC Subhead AV - 8B Harrier, Program Elements 940101(64262N) and 980101 (APN), and NORDA contract N00014-88-K-6006. Approved for public release; distribution is unlimited. NORDA contribution number PR 89:013:351.

REFERENCES

[1] Richard O. Duda and Peter E. Hart,*Pattern Classification and Scene Analysis*, John Wiley and Sons Inc., New York, 1973.

[2]. R.C.Gonzalez and Paul Wintz, *Digital Image Processing*, 2nd ed., 1987, Addison Wesley Inc.,Reading, Ma.

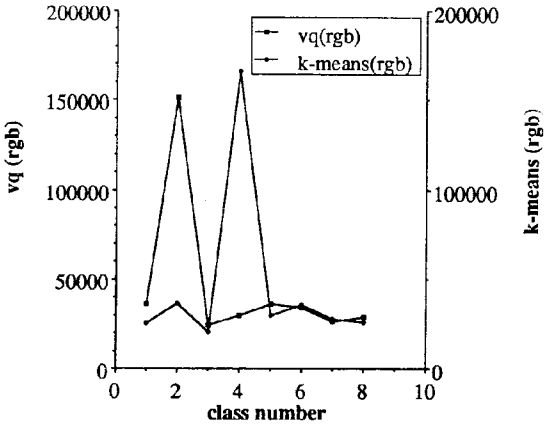

Fig. 6 Histogram of gray levels

[3]. R.M.Gray., "Vector Quantization" , IEEE ASSP magazine, vol.1, number2, pp. 4 - 29, April 1984.

[4] John A. Hartigan, *Clustering Algorithms*, John Wiley and Sons Inc., New York 1975.

[5]. R.P.Heydorn and H.C.Takacs, "On the design of classifiers for crop inventories", IEEE Trans. on Remote Sensing, Jan 1986.

[6]. N.S.Jayant and P.Noll, *Digital Coding of Waveforms*, Prentice Hall Inc., Englewood Cliffs, NJ.

[7] Donald E. Knuth, *The Art of Computer Programming Vol.2, Fundamental Algorithms*, McGraw-Hill, 1978.

[8]. Y.Linde, A.Buzo, and R.M.Gray, "An algorithm for vector quantizer design", IEEE Trans on Comm.,vol COM-28, pp 84-95, Jan.1980.

[9] A.Rosenfeld and Avinash C. Kak, *Digital Picture Processing*, 2nd ed. Academic Press Inc. New York 1982.

[10] Terry A. Welsh,"A Technique for High Data Compression, IEEE Computer", vol. 17, no.6, pp 8 - 19, June 1984.

[11] J.Ziv and A.Lempel,"A Universal Algorithm for Sequential Data Compression," IEEE Trans. on Information Theory, vol IT-23, no.3, pp. 337-343, May 1978.