An Automated Method for Digitizing Color Thematic Maps

Rick L. Lawrence, Joseph E. Means, and William J. Ripple

Abstract

There is an increasing need for methods of rapidly entering analog data into geographic information systems. Traditional methods of hand digitizing or hand tracing followed by scanning are costly and time consuming. We have developed a rapid, easy to use method for digitizing color thematic maps that makes use of standard image processing techniques. The method uses a digital camera followed by supervised spectral classification and post-classification smoothing. Although overall accuracy for an extremely challenging test map was 93 percent, our results indicate that for most applications expected accuracy is very high.

Introduction

The recent explosion in the use of geographic information systems (GIS) by a wide variety of users has created an increasing need for faster, less costly methods for entering spatial data. Many GIS users are now faced with the need to convert massive quantities of thematic data contained on analog maps into digital format.

Traditionally, analog maps have been input into GIS through manual digitizing. Although this method can be highly accurate, manual digitizing also is labor intensive and slow. The manual digitizing stage is typically the most expensive process associated with operating a GIS with large databases.

Digital scanners have been used to partially automate the conversion of analog maps. In one technique, features of interest on input maps are manually traced onto a transparent overlay. The overlay then is scanned so that the digital image is not cluttered with extraneous information such as grid lines, text, and other symbols. Digital scanners operate in a raster domain by recording a range of pixel values that correspond to the color intensity on the map document. However, color scanners of high quality are very expensive and can require highly specialized software. Even with modern scanning technology, we believe there is a need for faster, less expensive approaches for digitally capturing existing analog maps.

The objective of our study was to develop a fast method for converting thematic information on large color analog maps into digital GIS coverages using standard image processing techniques. We did not seek to convert gray-scale thematic data, or line or point information, although we believe that the technique we used may be adaptable for these purposes.

The approach we tested was to capture analog maps in three color bands using a digital camera and perform a supervised classification on the resulting digital images to produce digital thematic maps for GIS analysis. Post classification procedures were used to correct for extraneous information, such as lines, map symbols, and text, not properly accounted for by the classification.

Our analog source data used for testing the methods described in this paper was a large complex color map (partially shown on Plate 1a) representing the first detailed regional inventory of forest resources in Oregon (Andrews and Cowlin, 1940).

Methods

The object of our study was a 1933-34 map of Oregon forest vegetation (Andrews and Cowlin, 1940). This type of map has been used in our studies of historic forest patterns in western Oregon for ecosystem management purposes (Ripple, 1994). We believe that this map provides a stringent test of the method we present and is significantly more challenging than most maps that are input in GIS.

The map used is one of a set of eight, each 130 cm by 183 cm. The maps have 25 colored land-cover classes, plus colors for water and areas outside the mapped area, and cross-hatching in urban areas. Many of the colors are not solid colors, but are speckled with white or lined with other colors. The map used covers northwest Oregon. There are 36,142 polygons on this map. We believe that other approaches, such as hand digitizing or hand tracing followed by scanning, would take several weeks to capture and edit these eight maps.

To test our technique, we obtained in .tif format a 2048- by 2048-pixel, 24-bit digital image of the map containing the northwest quarter of Oregon using a Megavision T2 digital camera back mounted on a Sinar 4 by 5 view-camera body with a 90-mm Schneider super-angulon lens. Artificial lighting was used to illuminate the map.

The northwest portion of the map was subset from the scanned image to provide a specific study area (Plate 1a). The study area was 850 pixels wide by 1254 pixels long and included 18 land-use cover types and approximately 6,000 polygons. This study area on the original map is approximately 65 cm by 80 cm.

As represented on the map, each cover type is designated by a unique color or, in most cases, a color combination. Color combinations mostly consist of solid colors broken up by varying amounts of white speckling. For example, there are three "red" classes — solid red, red with a light amount of white speckling, and red with a heavy

R.L. Lawrence and W.J. Ripple are with the Environmental Remote Sensing Applications Laboratory, Department of Forest Resources, Oregon State University, Corvallis, OR 97331.

J.E. Means was with the Pacific Northwest Research Station, USDA Forest Service, 3200 Jefferson Way, Corvallis, OR 97331 and is now with the Department of Forest Science, Oregon State University, Corvallis, OR 97331.

Photogrammetric Engineering & Remote Sensing,
Vol. 62, No. 11, November 1996, pp. 1245–1248.

0099-1112/96/6211-1245$3.00/0

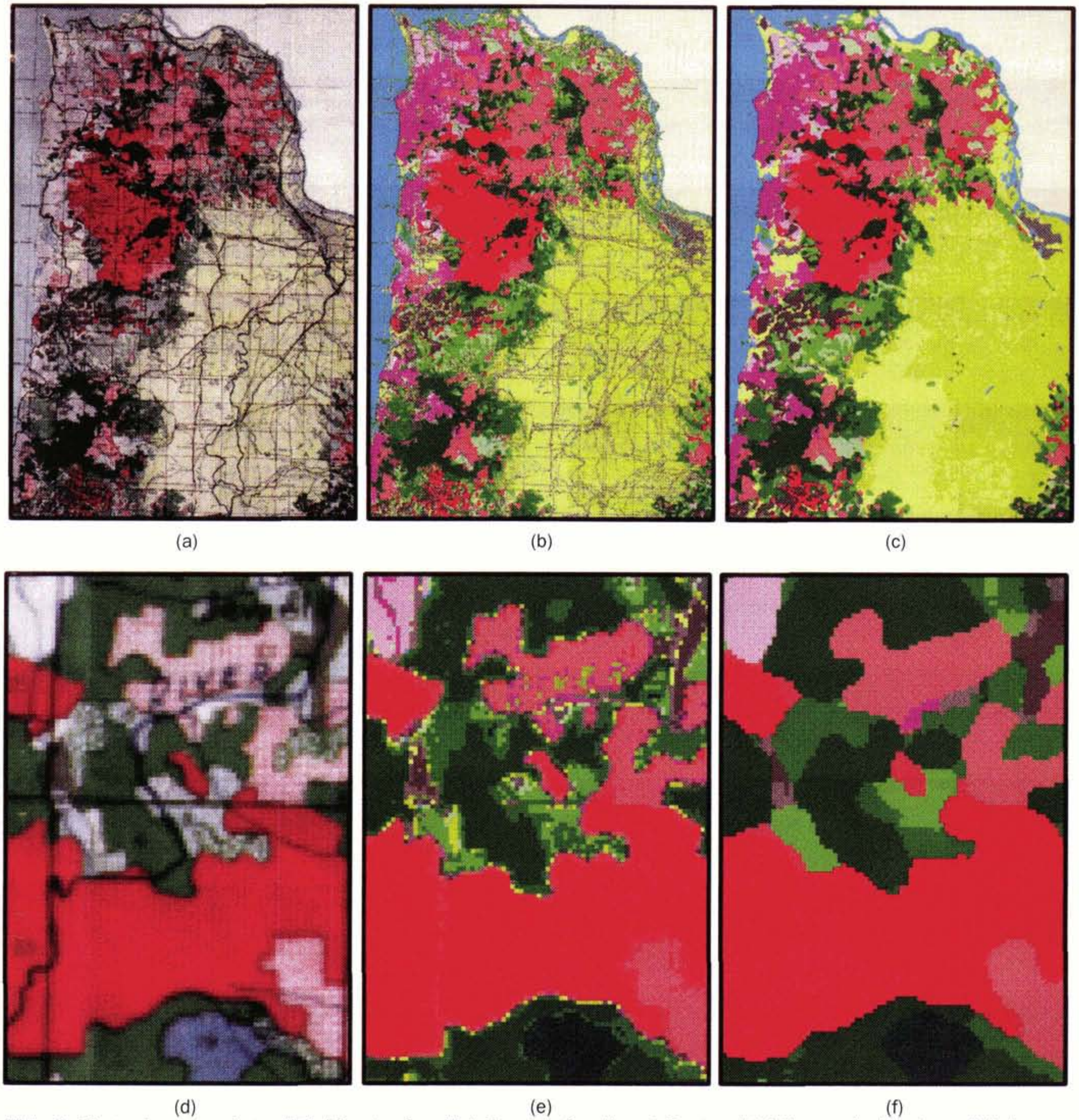

Plate 1. Stages in processing a digital image of an historic map of northwest Oregon. (a) A three color band raw digital image, showing roads, grid lines, and text. The original map represented by this image is 65 cm by 80 cm. (b) The results of the supervised classification of the image. (c) The results of smoothing the classified image to reduce misclassified pixels. (d), (e), and (f) Approximately full size reproductions of a small portion of (a), (b), and (c), respectively.

amount of white speckling. In addition to these solid and white speckled classes, there are two yellow classes — pale yellow and pale yellow with yellow-green dashes.

A close examination of the image of the map reveals important information about the imaging process (Plate 1d). The camera lacked both the resolution and the acuity to differentiate the white speckling from the surrounding color. Instead, the speckling blurred with the colors. Thus, the solid red class still appears solid, dark red, the lightly speckled red class appears as a generally medium red, and the heavily speckled red class appears as light red or pink. Similar effects are seen in other classes.

Further, various black features in the original map are not represented as pure black in the digitized image. These features include grid lines, roads, and text. Because of the blurring, a road passing through a red class, for example, appears not as black, but as very dark red.

The basic approach we used to prepare a land-cover type raster GIS layer from the map was supervised spectral classification. We used a Sun Sparc 10 workstation with Imagine 8.2 digital image processing software for all analytical procedures.

The .tif image was imported into Imagine .img format, and the study area was subset. If the image was going to be used for GIS analysis, it should be georeferenced at this point. For each class, a training site was selected from one of the larger patches within the class. Each training site was hand digitized on the computer screen using the cursor. Each training site included a portion of a road, grid line, or text (collectively, "black noise"). A maximum-likelihood classifier was then used to classify all of the pixels. The total time for training and classification was approximately one hour.

As a result of the inclusion of black noise in the training sites, in many cases the resulting spectral signature for a class was broad enough to include such extraneous data. Thus, much of the black noise was properly classified into the class through which it passed. This reduced the amount of post-classification correction that was necessary.

In spite of the apparent success of the supervised classification, the classified image showed considerable misclassification corresponding to black noise (Plates 1b and 1e). This was expected, and was more often the case in lighter colored classes. For example, generally a road passing through a solid red patch would appear as dark red and be properly classified. However, often a road passing through a light red patch would also appear dark red and be misclassified with the solid red class.

To reduce the effects of misclassified pixels, we performed a post-classification smoothing. The smoothing was conducted by passing a square moving window across the image and reassigning to the pixel in the center of the square the class value of the majority of the pixels in the square. The appropriate size of the moving window was chosen by trial and error, followed by visual examination of the results. The best results were achieved with a seven by seven window. This may be explained by much of the black noise (roads and grid lines) generally appearing three pixels wide. A three by three or five by five window is not large enough to smooth this size feature.

The initial smoothing worked well for most of the image. However, the large yellow classes in the center of the image retained a large percentage of misclassified pixels. This area is the lightest area in the original map and is heavily laced with roads and grid lines. The black roads and grid lines passing through the yellow classes were often classified as brown classes. Therefore, a second, selective smoothing was performed to remove only the two brown classes from within the yellow classes. The total time for training, classification, and smoothing was less than two hours, although this might vary depending on how many smoothing options are compared.

Two other techniques were explored, but did not provide acceptable results. Unsupervised classification was attempted, where a specified number of classes are designated based on spectral similarity of pixels. Because of the spectral similarity of some classes, such as the green and blue-green classes, nearly 100 classes were required to adequately segregate all classes. This required extensive post-classification work to recombine classes to the desired 18 cover types.

A Fourier transformation prior to classification was also attempted to remove the grid lines and reduce the need for post-classification smoothing. This technique is commonly used in digital image processing to replace bad scan lines in satellite imagery. The Fourier transformation successfully removed most of the grid lines. However, the resulting image had degraded color purity, and visual analysis of the subsequent classification revealed very low accuracy.

Results

The final classification was first evaluated visually (Plate 1c). The full image appeared to successfully capture the broad-scale patterns present in the original map. At finer scales (Plate 1f), differences between the original map and the final classified image are apparent.

Accuracy of the classification was quantified using standard accuracy assessment analysis (Lillesand and Kiefer, 1994). For this purpose, 255 random pixels were selected. To ensure that all classes were represented, we used a stratified selection procedure with a minimum of five pixels per class. Selected pixels were constrained to be at least one pixel from cluster edges, so that differences between original map resolution and digital map resolution would not be considered a source of inaccuracy. Class 17, which represented areas outside the study area, was excluded from the assessment.

Overall accuracy, which is the percentage of correctly classified pixels, was 93 percent. The Kappa statistic, which represents the additional level of accuracy achieved versus a random classification and is therefore lower than overall accuracy, was 0.92. Table 1 presents the error matrix for the accuracy assessment. The columns represent reference values for each pixel (what the pixel's class should have been), while the rows represent the results of the classification (how the pixel was classified). Thus, in column 5, row 4, we see that one pixel that should have been classified as class 5 was misclassified as class 4. The diagonal of the matrix represents correctly classified pixels. All other values represent misclassified pixels.

Examining the error matrix reveals that almost all misclassification resulted from two errors. First, seven pixels that should have been classified as class 2 (light yellow with yellow-green dashes) were misclassified as class 1 (light yellow). The difference between these classes is very subtle and the classifier was not fully capable of distinguishing them. If we remove these pixels from the calculation of accuracy statistics, overall accuracy is 95 percent and the Kappa statistic is 0.95.

Second, class 12 was not successfully classified. This class is brown with white speckling. It appears in only one small patch in the entire study area with less than 50 pixels. We believe that the small size of the patch prevented obtaining a quality spectral signature and resulted in confusion with other pixels. This is most notable in class 18, which is light yellow with thin black lines. If we remove the pixels resulting from both of these errors from the calculation of accuracy statistics, overall accuracy is 98 percent and the Kappa statistic is 0.98.

Discussion

Acceptable Levels of Accuracy

The utility of the method presented in this paper is dependent on the user's acceptable level of accuracy in the final digitized GIS layer. Although the level of accuracy that can be expected from this method will probably always be less than hand digitizing, the actual level of accuracy for any single application will depend on several factors.

The first important factor affecting accuracy is the spectral distinctiveness of the polygons being digitized. Based on our results, we believe that this method is capable of distinguishing any solid colors likely to be used by a cartographer. Greens and similar blue-greens and yellow-greens that were hard to distinguish with unsupervised classification were accurately distinguished with the supervised classification. In addition, most non-solid colors were well distinguished. The only difficulty in separating colors occurred with the extremely similar yellow classes and the speckled brown class,

Table 1. Error Matrix for Determining Accuracy of the Image Processing.

		Reference Classes																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	1	22	7																
	2		42																
	3			18															
	4				17	1													
Pixel Classifications	5					17													
	6					1	9												
	7							14	1										
	8								8										
	9									5									
	10										5								
	11											10							
	12	1							1				0				1		4
	13													19					
	14														7	2			
	15															15			
	16																22		
	17																	0	
	18																		6

for which an adequate signature was unobtainable because of the size of the class.

The second important factor affecting accuracy is the amount of noise present in the original map. Classification errors result from this noise. Post-classification smoothing increases accuracy by removing most of the effects of these misclassifications. When such lines are one, two, or occasionally three pixels wide, smoothing can generally remove problems they create.

The third important factor affecting accuracy is size of smaller features relative to size chosen for the smoothing window. While smoothing reduces noise, the procedure also decreases accuracy along borders between class types and will often reduce or eliminate small features such as peninsulas, bays, or polygons. Although few of the errors detected in the accuracy assessment were the result of these errors, a detailed comparison on the map and the final GIS layer reveals many differences, especially along borders (Plates 1d and 1f). Because the vast majority of the area of most maps is occupied by areas within large polygons, these errors will usually not have a significant effect on overall accuracy, but do affect accuracy at fine scales. If the size of the smoothing window is small relative to the sizes of smaller features, this problem will be minimized; otherwise, it can be an important source of error.

We suggest that people who use this technique experiment with pixel size (image resolution) and sizes of smoothing windows. Coarser resolutions will blur unwanted features such as roads and streams but will reduce accuracy at relatively fine scales. Larger smoothing windows will greatly facilitate reduction of misclassified pixels caused by such unwanted features, but will progressively degrade accuracy at relatively sharply curved edges and may eliminate relatively small features. A digital camera allows experimental variation in resolution relative to the analog map by either using a different lens or changing the camera to map distance.

Potential Applications, Advantages, and Disadvantages

We believe that there are many potential applications for this and similar techniques that others may develop. Libraries contain many paper maps with historical information that would be much more usable in digital form. Such maps could be used in studies of landscape development by landscape ecologists, resource managers, and land-use planners.

The primary advantages of this technique over hand digitizing or tracing are speed and resultant low labor cost. The image of the paper map was obtained with the digital camera in less than an hour; and most of the time was used in setting up. The cost for the digital camera for one hour was $90.00, and we could have obtained four to eight images in that time.

The main disadvantage of this technique is its reduced accuracy relative to hand digitizing approaches, where highly accurate data are required. Also, some analysts may not have access to a digital camera or other rasterizing device.

In short, this technique offers a speedy, low-cost approach to digital conversion of color analog maps where equipment is available and the high accuracy of traditional methods is not required.

Acknowledgments

Partial funding for this project was provided by the U.S. Fish and Wildlife Service (grant #144500091577, RWO #7) through the Oregon Wildlife Cooperative Research Unit, and the National Science Foundation (grant #GER-9452810) under the auspices of the NSF Graduate Research Fellowship in Landscape Studies.

References

Andrews, H.J., and R.W. Cowlin, 1940. *Forest Resources of the Douglas-fir Region*, U.S. Government Printing Office, Washington, D.C., 169 p.

Lillesand, T.M., and R.W. Kiefer, 1994. *Remote Sensing and Image Interpretation*, 3rd ed., John Wiley & Sons, Inc., New York.

Ripple, W.J., 1994. Historic spatial patterns of old forests in western Oregon, *Journal of Forestry*, 92:45-49.