Reprinted in Canada from CANADIAN JOURNAL OF REMOTE SENSING

EXPLOITING SPECTRAL, SPATIAL AND SEMANTIC CONSTRAINTS IN THE SEGMENTATION OF LANDSAT IMAGES†

by D.W. Starr* and A.K. Mackworth**

University of British Columbia
Department of Computer Science

SUMMARY

A critique of traditional classification techniques for LANDSAT images and consideration of some scene analysis techniques, exploiting spatial organization and meaning, lead to a new approach to computer programs for LANDSAT image understanding. To justify this approach, a program that combines modified maximum likelihood techniques with interpretation-controlled region merging methods to interpret forest cover in LANDSAT images is described. For comparison purposes, a pure supervised classifier using the same data made 43% more errors and produced a segmentation twice as complex.

INTRODUCTION

The goal of interpreting LANDSAT image data automatically by computer programs has been the focus of much research effort in recent years. Achieving such a goal would have enormous economic and social benefits; however, many of the results to date are not too encouraging if judged by the accuracy of the pixel classification.

We argue here that any attempt to classify pixels which relies solely on the spectral signature of each individual pixel is, in most applications, doomed to mediocre performance. The spectral evidence alone is not enough. In suggesting an alternative approach, we propose that recent developments in an area of artificial intelligence known as computational vision offer an alternative paradigm to that of pattern recognition. The proposal is backed up by an implementation that demonstrates a substantial increase in classification accuracy for a program that combines pattern recognition and computational vision techniques.

BACKGROUND

Classification Techniques

LANDSAT classification systems in the pattern recognition paradigm can be dichotomized in a variety of ways. One such dichotomy is the distinction between sample classification systems and pixel classification systems[1, 2]. Most systems currently in use are of the pixel classification type. Another dichotomy is the distinction between non-supervised and supervised systems.

---

†Presented at the Remote Sensing Science & Technology Symposium, February 21-23, 1977 in Ottawa

*Now at Department of National Defence, Ottawa

**Assistant Professor

Non-supervised classification uses cluster analysis (or, occasionally, factor analysis) to group a set of pixels automatically into the most spectrally suitable clusters using the intensities in the four spectral bands. Such systems do not perform as well as supervised systems[3]. The supervised approach requires ground-truth data that provide for each of the required classification categories, a representative sample of pixels. For each category or class, one can then compute the mean for each spectral band and the covariance matrix showing how the hands covary with each other for that class. In a maximum-likelihood classifier based on the normal distribution, the assumption is made that the pixel's signature is normally distributed about the mean for its class. To classify an unknown pixel, one can, using Bayesian probability theory, compute the probabilities that it came from each of the classes. The pixel is assigned to the class of the highest probability hence the name, maximum-likelihood classifier[4]. Under certain simplifying assumptions, maximum likelihood reduces to minimum distance classification whereby a pixel is simply assigned to the class whose mean is closest in the spectral space to the pixel's signature. A wide variety of mathematical sophistications can be added to the scheme without changing its essence.

The attraction of the classification paradigm is its computational simplicity — it would indeed be nice if the world were so structured as to make it work; however, the world is more subtle than that. The fundamental assumption is that a pixel's interpretation depends only on its spectral attributes and not, for example, on its location in the picture or on the interpretation of neighbouring pixels. Thus, for example, Todd *et al*[5] found that most of their errors occurred near urban/rural boundary where upper income residential areas were understandably misclassified as grassy meadows. If one randomly permuted the pixels of a LANDSAT image a classifier operating in this paradigm would not even notice, let alone change its interpretation of any pixel! Clearly the paradigm is ignoring the crucial fact that all meaningful images are subject to spatial organization which must be heavily exploited in the interpretation process. Moreover, it is further assumed that there will, in fact, be good separation between the classes in the spectral space. Ideally, each class would be separated from all the rest by a set of hyperplanes in the spectral space. In reality this hoped-for separation does not occur; there is often enormous overlap of the distribution for each class. (This phenomenon will be demonstrated later for our data.) When the classifier is asked to pronounce on a pixel whose signature falls within such overlaps it can only do so by making unreliable guesses. As a consequence of the fact that the two essential assumptions do not hold, the performance of such classifiers is often poor, except, of course, in the rare cases where pixels can be classified context-free and where clear separation does occur: when distinguishing water from forest, for example.

There have been several attempts to use spatial information in the interpretation of remotely sensed images. Robertson[5], for example, classified blocks of adjacent pixels using texture measures and other spatial characteristics. This increased the accuracy over the pixel by pixel classification by 2.5% (with an average accuracy of 82%). Following a similar approach, Kettig and Landgrebe[6] reported a method that partitioned the image into blocks of statistically similar pixels before classifying them. Initially, 2x2 cells of pixels are created provided that the four pixels are sufficiently homogeneous. Similar adjacent cells are merged to form blocks which are classified using a sample classifier. These techniques considerably reduce the number of classifications required to segment the entire image.

Gupta *et al*[7] used an edge detector to find roughly homogeneous regions in airphotos. Since their domain of interpretation was agricultural fields with long, straight boundaries, one would expect this approach to be useful. In their case the original pixel classification was so good, at about 95% correctness, that they only demonstrated a marginal improvement.

An interesting program that exploits both the pattern recognition paradigm and the computational vision paradigm, described in the next section, is one written by Bajscy and Tavakoli[8]. They identified bridges, islands, rivers and lakes in LANDSAT images. The initial scene segmentation used classification techniques to separate water from non-water. They then used a world model describing the structure of the objects they were interested in to make sense of the image. Using these techniques they found, for example, bridges whose width was less than the width of a single pixel!

Computational Vision

Over the last decade or so there has developed, within artificial intelligence, an alternative approach to the machine interpretation of visual data. Usually known as scene analysis, or, more broadly, computational vision, this approach is characterized by computer programs that

are expert in interpreting pictures of a particular domain. A prerequisite to writing such programs is a careful analysis of the structure of the knowledge that we have about the picture in the domain and the knowledge we have of the objects depicted, the scene[9]. Although current image and scene domains are, typically, much simpler than those for a LANDSAT image, there is now emerging a paradigm for perceptual processes captured by the cycle of perception[10]. Cues in the picture can invoke models of the scene which have to be tested. If the models are established then certain consequences follow including the instigation of a search for new cues in the picture which then invoke new models . . .

The implications of this paradigm for the LANDSAT task are that we should attempt to find cues that can be sensibly interpreted. These can then suggest interpretations for new areas of the picture. To be specific, we can introduce the spatial sensitivity we are after by dealing with regions of connected pixels in the picture, not individual, isolated pixels. Region merging techniques have been developed in scene analysis[11] that show, starting with atomic regions of pixels with identical intensities, how one can merge regions with similar intensities to produce a region segmentation for subsequent interpretation. Moreover, we can go further and allow the interpretation of the initial regions to control the segmentation or merging process itself[12, 13].

The strategy is an emerging principle in artificial intelligence. It is coming to be known as best-first or island-driving: in this case, use the interpretations of those segmented partial regions we are most confident of as a guide and context for further region segmentation and interpretation.

RESOLVING AMBIGUITY IN CONTEXT

In adapting these ideas to this task we can combine the best features of the pattern recognition approach with these scene analysis techniques. As a front-end we propose a supervised maximum likelihood classifier that is traditional in all but one important respect: if a pixel is highly ambiguous, that is, if no one class is overwhelmingly likely, the classifier is simply to report the two most likely classes. Pixels are either unambiguously put in a class or ambiguously related to two classes. The subsequent region merging process uses this information to form regions of connected pixels for each of these unambiguous (strong) and ambiguous categories. It then should sensibly merge regions: the strong regions grow by devouring, in amoeba-like fashion, the ambiguous regions according to a systematic set of rules. As we shall see the interpretation of the spatial context of a pixel can then occasionally suggest that its most likely interpretation should be discarded in favour of its second most likely interpretation.

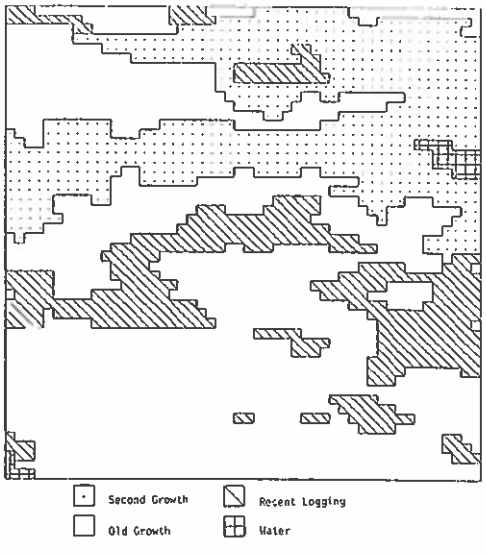

Figure 1
Ground truth data

The Data

To test this method we used ground truth data for a forested area of Vancouver Island and a LANDSAT image of the same area both obtained in August, 1974. The task is to classify the ground cover into regions of old growth, second growth, recent logging and water. The ground truth data for a 50 x 50 pixel area (2.85 km x 4 km) are shown in Figure 1. There are 24 regions in that data.

Baseline Maximum Likelihood Classifier

As a baseline, we implemented a traditional supervised maximum-likelihood classifier. Using a subset of about 10% of the ground truth data we obtain the mean and covariance statistics for the four classes of interest. Shown in Figure 2 are the ellipses of concentration in spectral bands 5 and 7 (cross sections of equiprobable surfaces) for the four classes. These two bands give the best class separation of any pair of bands. Notice that

water is easily distinguishable from the other three classes, but old growth and new growth are most confusing and new growth and recent logging only slightly less so. The maximum likelihood classification shown in Figure 3 when compared to the ground truth has, not surprisingly, 30% of the pixels misclassified. Note also the usual annoying "salt and pepper" appearance of the classification with many small regions. There are 150 regions in that classification.

THE ALGORITHMS

The system is comprised of four main routines which are invoked in sequence:

1) Training
2) Classification with ambiguity
3) Initial region finding
4) Interpretation-guided region merging

Training

The training data and routines are exactly the same as those for the baseline classifier, with the same results as shown in Figure 3.

Classification With Ambiguity

In this phase, for each pixel the probabilities of its membership in each of the four classes, p_1, p_2, p_3 and p_4 are computed. If the largest of those, p_i, is greater than some fixed threshold, k, $(0.25 < k < 1)$ then the pixel is labelled as class i. If, however, the largest is not greater than k then the pixel is labelled as ambiguous: it is either in class i or class j corresponding to the second largest probability, p_j. In theory, the number of classes is increased from the basic four to ten, viz class 1, class 2, class 3, class 4, class 1 or 2, class 1 or 3, etc. Or, in general, from n classes to $n(n+1)/2$. In fact, only those pairs of classes which do have significant overlap in the spectral space will occur. For our data there are no ambiguous classes that include water as one of the pair but the classes old growth, new growth and recent logging are all often, pairwise, ambiguous.

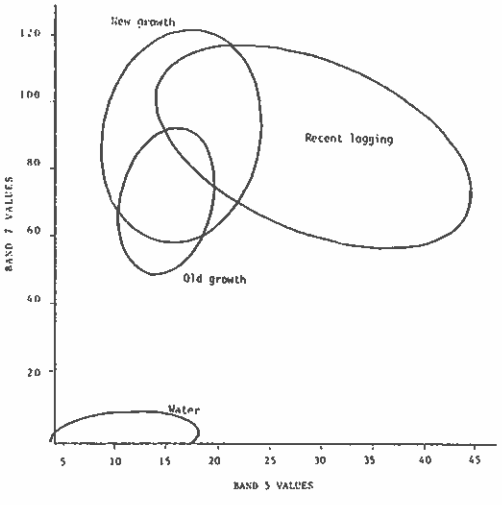

Figure 2
The intensity distributions in bands 5 and 7

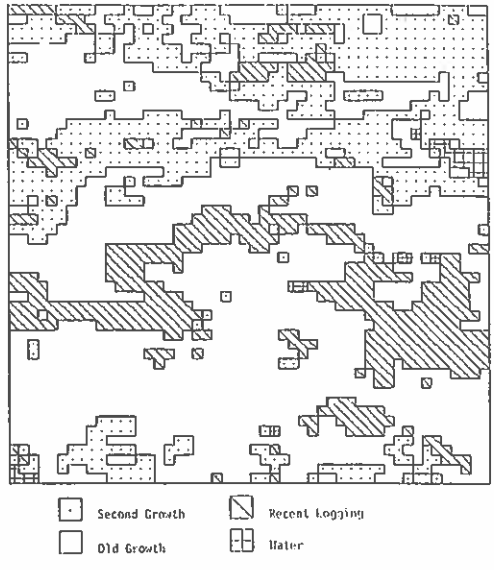

Figure 3
Maximum likelihood classification on individual pixel basis

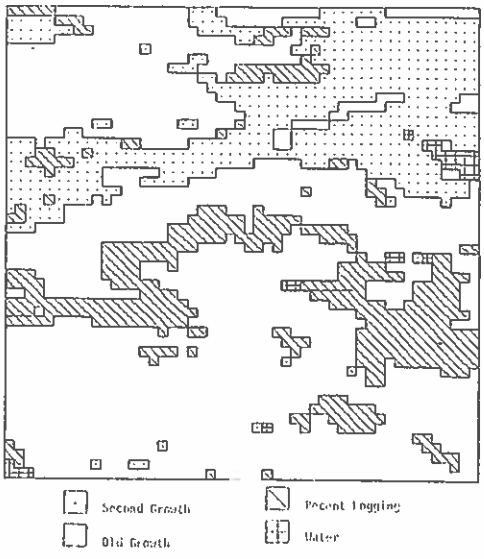

Figure 4
Initial segmentation with ambiguous regions

The interpretation of this procedure in the spectral space is that we have not exhaustively partitioned the space into the 4 classes, rather we have put a much smaller, more conservative boundary around the mean of the class and said anything inside is indeed in that class. Outside those conservative bounds, we have, for now, refused to jump to a conclusion: the pixel is genuinely ambiguous. The spatial context for such a pixel will later have to decide its interpretation.

Initial Region Finding

This phase simply collects together into regions all pixels that are adjacent to one another and have the same interpretation or, in the case of the ambiguous pixels, the same two possible interpretations. The details of the region finding algorithm need not concern us — a straightforward recursive algorithm is sufficient. The regions are represented by lists of their boundary points. The regions found from the classification with ambiguity with the level of confidence for an unambiguous pixel, k, set at 0.7 are shown in Figure 4.

Interpretation-guided Region Merging

The aim of this phase is to eliminate the ambiguous regions by absorbing them all into unambiguous (strong) regions. This is a very context-sensitive process that must be done carefully. An ambiguous region should assume one of its interpretations if there is local strong evidence in favour of that class existing there. If there is no evidence for either class then it may as well be taken to be its most likely class. This is implemented as a 3 pass relaxation-like algorithm as follows:

Pass 1

For each ambiguous region, A, merge it into an adjacent strong region, S, if

i) The most probable class for A is that of S and
ii) The new region that would be formed, A S, is also strong and
iii) S and A have in common at least some minimum fraction of their total boundary and
iv) The fraction of the common boundary that is weak is above a certain threshold

Iterate this process until no more ambiguous regions can be merged into strong regions. A region is said to be strong or ambiguous according to whether its average signature is strong or ambiguous. A boundary element between two regions is weak if the pixels on one side of the boundary do not have a significantly different signature from the pixel on the other side.

Pass 2

As for pass 1 but now relax requirement i to

i') The most or second-most probable class of A is that of S

Requirements ii, iii and iv are unchanged.

Again iterate until no more merges are possible.

If there are any conflicts in pass 1 or pass 2, that is, if more than one strong region, S, satisfies the requirements needed to gobble A, then resolve them by the extent to which criterion iv is satisfied.

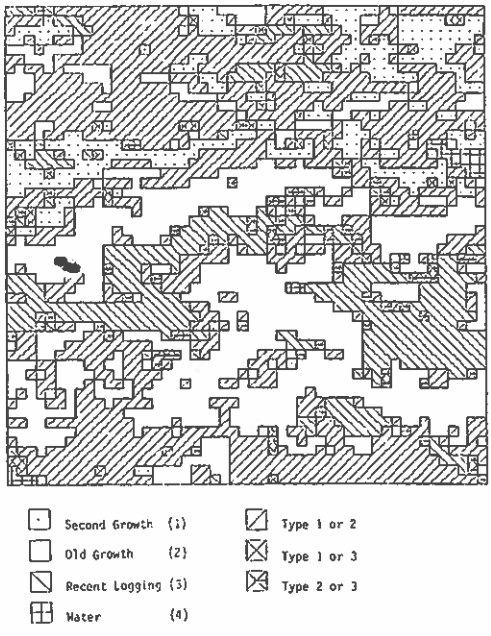

Figure 5
Final segmentation after semantic region merging

Pass 3

If there are any ambiguous regions left, interpret them as their most likely class.

The essence of this algorithm is that region nuclei of pixels most confidently assigned to a class are formed initially. These nuclei grow, selectively absorbing the ambiguously classified regions, guided by the interpretations and the local image structure. Crucially, in pass 2, an ambiguous region may be merged into an unambiguous region that has only the second most likely interpretation of the ambiguous region. When this happens, the spatial organization and meaning are partially overriding the spectral evidence.

RESULTS

The final merged regions are shown in Figure 5. The fraction of misclassified pixels compared with the ground truth is 21%. This contrasts with the 30% error rate (43% higher) of the maximum-lieklihood classifier using exactly the same training data.

Moreover, the classification with ambiguity produced over 350 regions (Figure 5) but the merging process has reduced that to 75 regions (compared with 150 regions for the pure classifier). This reduction by half of the number of regions is of considerable importance for human use of these results. This approach should not be confused with non-semantic, post-processing elimination of small regions[14, 15, 16].

It should be noted that the pass 1, 2 and 3 region-merging algorithms used are order-dependent. The final segmentation will depend somewhat on the order in which regions are considered for merging. This undesirable property is tolerated because the efficiency of the algorithms, implemented on a serial machine, appears to require it. Experiments with reordering the regions show that the order is not a major influence on the structure of the final image.

The results, while they support very strongly the thesis of this paper, must be carefully interpreted. They depend, for example, on the settings of a number of parameters or thresholds. These parameters have been set by hand for the current program. The results are not finely-tuned with respect to the parameters but further work must be done on setting them automatically using the same ground-truth data used to tune the classification parameters.

The system is written almost entirely in ALGOL W. The baseline maximum likelihood classifier for 2500 pixels used 10 seconds of CPU time on an IBM 370/168 while the merging program used about 35 seconds in total (including the classification with ambiguity).

CONCLUSIONS

The techniques of maximum likelihood classification based on the pixel signature alone are context-free. They can be used as a starting point which is modified by the interpretation-driven region merging techniques which are sensitive to the spatial and interpretation context of each pixel. Further progress in this area will require the development of computational mechanisms whereby we can express the wide variety of knowledge that a skilled human photo-interpreter brings to this task. We need to develop theories and programs that can express and use, in addition to the knowledge about colour, adjacency and meaning that we use here, knowledge about shape[7], texture[5], lighting, geography[8, 17], *a priori* information about an area, the process of forest clear cutting . . . The current program demonstrates vividly that a substantial increase in the accuracy and simplicity of the classification can be achieved by exploiting a combination of constraints from the spectral, spatial and semantic domains in the segmentation process.

ACKNOWLEDGEMENTS

We are grateful to Dr. P. Murtha of the Faculty of Forestry at UBC for providing the ground-truth data for this study. This research was supported by grants from the National Research Council of Canada.

REFERENCES

(1) **Wacker, A.G. and Landgrebe, D.A. (1972)**— *Minimum Distance Classification in Remote Sensing*. **Proc. of 1st Cdn. Symposium on Remote Sensing, Ottawa, pp. 577-592.**

(2) **Peet, F., Mack, A. and Crosson, L.S. (1974)** — *Developments in Methods for the Automatic Estimation of Crop Production from ERTS and Supporting Data*. **Proc. of 2nd Cdn. Symposium on Remote Sensing, Guelph, pp. 565-570.**

(3) **Todd, W., Mausel, P. and Baumgardner, M.F. (1973)** — *An Analysis of Milwaukee County and Use by Machine-Processing of ERTS Data*. **Laboratory for Applications of Remote Sensing (LARS) Information Note 022773, Purdue Univ.**

(4) **Landgrebe, D.A. (1975) — NASA contract NAS9-1416 final report, Laboratory for Applications of Remote Sensing, Purdue Univ.**

(5) **Robertson, T.V. (1973)** — *Extraction and Classification of Objects in Multispectral Images*. **Machine Processing of Remotely Sensed Data, Purdue University, 3b27-3b34.**

(6) **Kettig, R.L. and Landgrebe, D.A. (1975)** — *Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects*. **Symposium on Machine Processing of Remotely Sensed Data, Purdue Univ.**

(7) **Gupta, J.N., Kettig, R.L., Landgrebe, D.A., and Wintz, P.A. (1973)** — *Machine Boundary Finding and Sample Classification of Remotely Sensed Agricultural Data*. **Machine Processing of Remotely Sensed Data, Purdue Univ., 4B25-4B35.**

(8) **Bajcsy, R., and Tavakoli, M. (1973)** — *A Computer Recognition of Bridges, Islands, Rivers and Lakes from Satellite Pictures*. **Machine Processing of Remotely Sensed Data, Purdue University. 2A54-2A68.**

(9) **Mackworth, A.K. (1977)** — *How to See a Simple World*. **Machine Intelligence 8, E.W. Elcock and D. Michie (eds.), Wiley, pp. 510-537.**

(10) **Mackworth, A.K. (1977)** — *Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds and Maps*. **Proc. Workshop on Computer Vision Systems, Univ. Mass., Amherst, Mass.**

(11) **Brice, C.R., and Fenema, C.L. (1970)** — *Scene Analysis Using Regions*. **Artificial Intelligence 1, 3, 205-226.**

(12) **Feldman, J.A., and Yakimovsky, Y. (1974)** — *Decision Theory and Artificial Intelligence: I. A Semantics-Based Region Analyzer*. **Artificial Intelligence 5, 349-371.**

(13) **Tenenbaum, M. and Barrow, H.G. (1976)** — *IGS: A Paradigm for Integrating Image Segmentation and Interpretation*. **In Pattern Recognition and Artificial Intelligence, Academic Press.**

(14) **Kan, E.P. (1976)** — *A New Computer Approach to Map Mixed Forest Features and Process Multispectral Data*. **Proc. American Society of Photogrammetry Fall Conv., Seattle, pp. 386-401.**

(15) **Goldberg, M., Goodenough, D. and Shlien, S. (1975)** — *Classification Methods and Error Estimation for Multispectral Scanner Data*. **Proc. of 3rd Cdn. Symposium on Remote Sensing, Edmonton, pp. 125-143.**

(16) **Davis, W.A., and Peet, F.G. (1977)** — *A Method of Smoothing Digital Thematic Maps*. **Remote Sensing of the Environment, 6, 45-49.**

(17) **Mackworth, A.K. (1977)** — *On Reading Sketch Maps*. **Proc. 5th Int. Joint Conf. on Artificial Intelligence, M.I.T., Cambridge, Mass., pp. 598-606.**