FORESTS AND PYRAMIDS: USING IMAGE HIERARCHIES TO UNDERSTAND LANDSAT IMAGES

Ezio Catanzariti* and Alan Mackworth,
Department of Computer Science,
University of British Columbia,
Vancouver, B.C., Canada V6T 1W5

ABSTRACT

Computer-based Landsat image interpretation has neglected the spatial organization of the image in favour of the spectral and temporal organization. Semantic and spatial sensitivity can be introduced by exploiting a pyramidal, hierarchical representation of the image advocated by Kelly, Tanimoto and Levine. The image pyramid is constructed bottom-up with the original image as the base. Each level is a reduced resolution version of the level below, constructed by averaging the signatures of adjacent pixels at the lower level. By classifying pixels at the higher levels one is efficiently classifying semantically uniform regions in the original image. If, however, a region's signature lies in the spectral overlap of two or more classes, its four subregions will have to be considered for classification. Several refinements of this technique, including the use of semantically-based region splitting and merging techniques at each level of the pyramid, have been developed. These techniques are used to classify forest cover types on Vancouver Island in a Landsat image. The results of several initial experiments indicate that, compared to a baseline of a traditional supervised maximum-likelihood classifier, the cost of maintaining the pyramid is balanced by a vast reduction in the number of pixel classifications. The spatial homogeneity or readability of the segmented image, as measured by the number of regions, is improved by a factor of three, while the accuracy of the classification is unaffected or slightly improved. When the region splitting and merging techniques are applied at each level of the image pyramid the accuracy and the readability of the final segmentation both increase markedly.

* Now at Instituto di Fisica Teorica,
Universita di Napoli, Naples, Italy.

Presented to the 5th Canadian Symposium on Remote Sensing, Victoria, August 1978.

RESUME

Jusqu'à présent, pour l'interprétation par ordinateur des images Landsat, on a négligé l'organisation spatiale de l'image au profit de son organisation spectrale et temporelle. On peut faire entrer dans l'interprétation une sensibilité sémantique et spatiale, en exploitant une représentation pyramidale, hiérarchique, de l'image, telle que celle que préconisent Kelly, Tanimoto et Levine. La pyramide des images est construite du bas vers le haut et l'image originale en constitue le base. Chaque étage de la pyramide est une version (vue avec un pouvoir de résolution réduit) de l'étape du dessous; cette version est construite en faisant la moyenne des signatures de plusieurs pixels contigus de l'étage du dessous. Quand on classe les pixels des étages supérieurs, on classe donc, avec un bon rendement, des régions de l'image originale uniformes quant à la sémantique, c'est-à-dire quant à la signification de l'image. Cependant, si la signature d'une région se trouve dans une zone spectrale commune à deux classes ou plus, il faut envisager le classement de ses quatre sous-régions. Plusieurs perfectionnements de cette technique ont été mis au point, entre autres le recours à des techniques de fractionnement et du fusionnement sémantique à chaque étage de la pyramide. Ce sont les techniques que nous avons utilisées pour classer les différents types de couverture forestière d'une image Landsat de l'Île Vancouver. Les résultats de plusieurs expériences initiales montrent que si l'on prend comme base de référence le coût d'un programme traditionnel de classification fondé sur la probabilité maximale, avec intervention humaine, le coût d'établissement de la pyramide est compensé par la réduction importante du nombre de classements de pixels nécessaires. L'Homogénéité spatiale, c'est-à-dire la facilité de lecture de l'image fractionnée, dont le nombre de régions donne une mesure, est améliorée (dans la proportion de trois à un), alors que la précision de la classification reste inchangée ou est légèrement améliorée. Quant on applique les techniques de fractionnement et de fusionnement des régions à chaque étage de la pyramide des images, la précision et la facilité de lecture de l'image fractionnée finale augmentent l'une et l'autre très nettement.

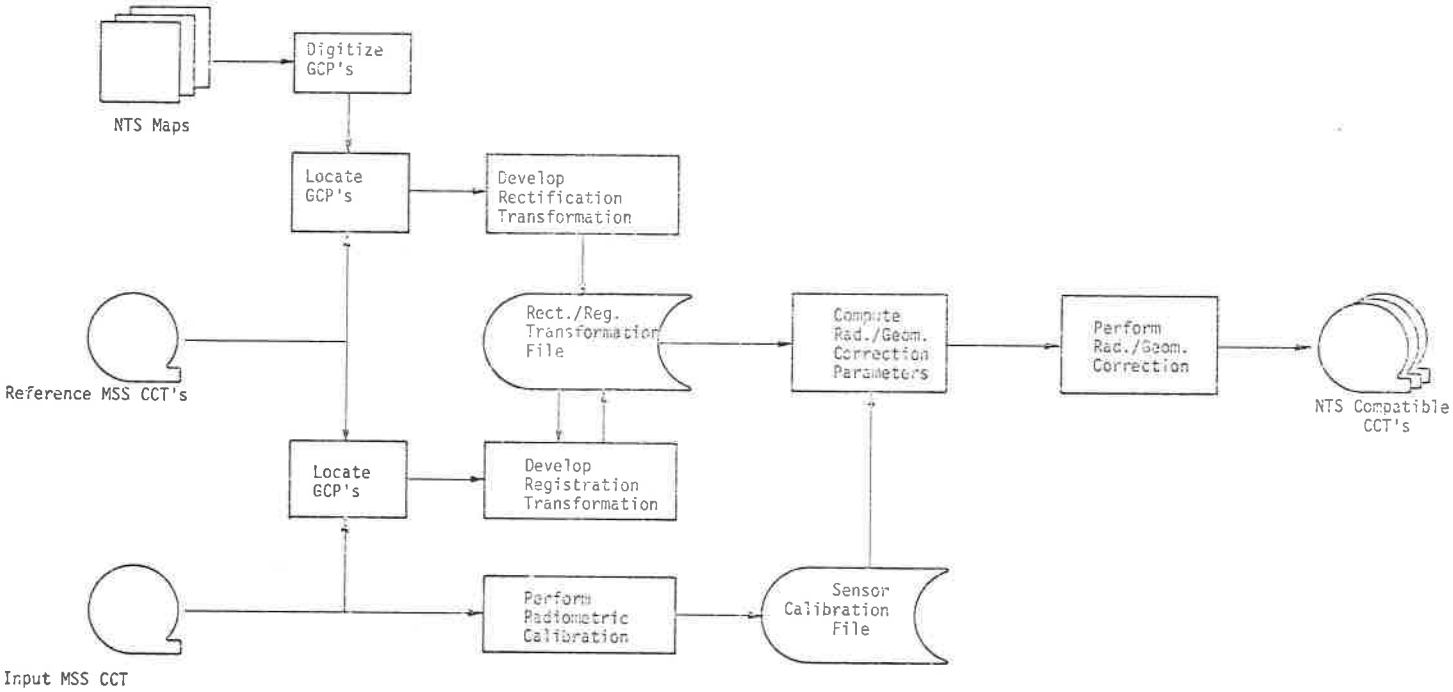

Figure 7. DISCS Operations and Data Flow

1. BACKGROUND

1.1 Classification Techniques

All remotely sensed images are characterized by the fact that the information about the scene is conveyed to the sensors through the spectral, spatial and temporal variations of the electromagnetic field. Landsat image processing systems rely most heavily on multispectral and multitemporal techniques. To a great extent they neglect spatial variation. This is due to the high cost of the few successful attempts to exploit the spatial context of the scene, together with the lack of a substantial theory on which to base suitable techniques. Most systems rely on the well-known multispectral Pattern Recognition paradigm for digital image interpretation. One way of stating this approach is in decision-theoretic terms: the objective of the classification is to arrive at an assignment of every pixel to one of a number of classes in the way that best fulfills a certain decision criterion. Supervised or unsupervised techniques can be used, depending on the availability of ground truth data for the classes of interest. A wide variety of mathematical sophistications can be added to the classification model without changing its essence.

The fundamental assumption of this paradigm is that a pixel's interpretation depends only on its spectral attributes, not, for example, on its location in the picture or on the interpretation of neighbouring pixels. An essential requirement of such approach is that the classes be spectrally separable, and, moreover, that the class statistics are stationary over the image. Neither assumption holds for real images. When the classifier is asked to pronounce on a pixel whose signature falls within an overlapping area of two given classes, it can only do so by making unreliable guesses. For example, in an experiment performed at LARS[1], the classification performance on a set of Landsat data was compared with the performance on a data set collected by an airborne multispectral scanner system with more wavelength bands over a wider region of the spectrum. The interesting result was that the overall performance for the data set was nearly identical. Classification performance of any technique based on the classification model depends largely on the degree of spectral separability of the classes of interest. If the classes of interest are spectrally similar then, using this approach, one cannot discriminate among them, regardless of the amount of training data used or the number of spectral bands available.

1.2 The Use of Spatial Information

There have been several attempts to use spatial information or, more generally, to introduce context-sensitivity into the interpretation of remotely sensed images. The majority of them are formulated within the classification paradigm. Usually some spatial features, such as texture, are computed for pixels or groups of pixels. These features are used as additional inputs to a point by point classifier[1,2].

Several systems successfully exploit the fact that spatially adjacent pixels are more likely to belong to the same class than distant pixels. Therefore the image is partitioned into "homogeneous" groups of pixels which are considered as single entities and, as such, classified by a traditional classifier using spectral and spatial characteristics[3,4]. Obviously techniques like these considerably reduce the number of needed classifications, but the total time required to process the entire image is often considerably augmented by the preprocessing procedures. Robertson, for example, reports an increased accuracy over the point by point classification by 2.5%, while the computing time increased by a factor of 10.

Others have used the same idea in devising techniques to postprocess the image after it has been first segmented by a point by point classifier. Goldberg *et al.*[5] give a technique for relabelling each pixel using the spatial information contained in the surrounding three-by-three region. Kan[6] and Davis and Peet[7] give procedures which eliminate small regions in the scene. Postprocessing techniques such as the above are especially effective in improving the readability of a classified image.

1.3 Scene Analysis and Image Understanding

Starr and Mackworth[8] used an approach that differed from most other Landsat systems. They identified different types of forest cover in a Landsat image. Traditional classification methods were used to obtain an initial segmentation of the image into atomic regions. Then they used Artificial Intelligence region merging techniques to merge regions with similar intensities. The region merging process goes hand-in-hand with the interpretation process: regions with unambiguous interpretation are allowed

to sequentially influence the interpretation of ambiguous regions. Context sensitivity is thus introduced from the beginning into the interpretation process. Using this technique, they showed a 9% improvement in classification accuracy over the point by point classifier, at a cost of increasing the computing time by a factor of 3.5.

This program is an application of an alternative approach that has recently emerged for machine interpretation of visual data within the field of Artificial Intelligence. The conventional approach to machine vision, on which the decision theoretic classification model for Pattern Recognition is based, sees the interpretation phase of the whole recognition process as sequentially following the segmentation phase. On the other hand, the Artificial Intelligence approach (the cycle of perception paradigm for scene analysis[9]) states that segmentation requires semantic information, interpretation, to be performed sensibly. Segmentation is interpretation and vice versa. To be meaningful segmentation needs to be driven by a real-world model, but such a model cannot be invoked without having first partially segmented the picture. Following this paradigm the whole vision process becomes a cycle, alternating segmentation and interpretation.

1.4 Hierarchical Image Structures

In the simplest application of Pattern Recognition to image understanding, each pixel in the image is uniformly processed in the course of interpretation. This approach is inefficient, at best, because not all the available detail is always necessary to interpret the image. The amount of detail that is needed strongly depends on the purpose of the study.

This idea underlies some multistage techniques which make sequential use of pictures at different scales in the analysis of remotely sensed data. Nichols et al.[10], for example, described a three-stage sampling method using satellite and aircraft imagery and ground sampling. Information gathered at any stage is used to direct the selection of samples at the successive lower stage. They used this technique to estimate the timber volume in a forest inventory application. At each stage timber volume estimates are made from sampling units whose probabilities of selection in the sample are biased by a factor proportional to the corresponding predicted volumes, as interpreted from the previous smaller scale imagery. This technique is shown to be more efficient, in terms of cost, than purely random sampling because fewer ground samples need to be taken.

This same idea has been clearly stated and successfully exploited in some recent scene analysis work.

Kelly[11] described an approach motivated by the idea of selective attention. He considers an image of a human face and extracts a smaller picture from it. The idea is that this reduced, lower resolution picture exhibits only the gross features of the face without the surrounding noise of the fine features. These features are therefore detected and used as a plan to find all the features in the original picture.

Several authors have extended Kelly's idea on planning. Tanimoto and Pavidlis[12] described a pyramid structure capable of handling image processing at different levels. The structure consists in a sequence of matrices where every matrix is a digitization of the picture at lower resolution than the previous matrix in the sequence.

This pyramidal data structure was successfully exploited by Levine[13] to segment an outdoor scene.

These programs, exploiting planning in the image interpretation process, suggested to us that elaborations of this technique could give us the spatial and meaning sensitivity of the interpretation-guided region merging techniques of our earlier Landsat work[8] without the associated high cost in CPU time.

2. USING IMAGE HIERARCHIES IN REGION FINDING

A segmentation obtained exploiting the pyramidal approach is in the spirit of the cycle of perception paradigm. The segmentation at every level gives a context to, and drives, the segmentation at the level below. This observation, together with the previously mentioned aspect of computational efficiency supporting the planning idea, suggested that, although current image and scene domains in Artificial Intelligence are typically much simpler than those for a Landsat image, we could combine the best features of the Pattern Recognition approach with these scene analysis techniques to interpret a Landsat scene.

To test these ideas, we used a Landsat image of a forested area of Vancouver Island taken

on August 12, 1973, covering a ground area of 3.65 x 5.12 km. The same ground truth map and the same modified maximum likelihood classifier described by Starr and Mackworth were used. The objective of the classification was to identify regions of old growth (class 1), second growth (class 2), recent logging (class 3) and water (class 4).

The original image is stored as a 64 x 64 array. This is level 1 in the data structure. The spectral signature of a pixel at level L + 1 in the pyramid is constructed by successively averaging the signatures of a square cell of four adjacent pixels at level L. As the program works up the pyramid pixels in the middle of regions are averaged with pixels belonging to the same class, while pixels on region boundaries are mixed together with pixels belonging to different classes. In statistical terms, areas composed of pixels that have equal probabilities of belonging to two or more classes expand as a result of the averaging process, while clusters of pixels that have a high probability of belonging to a single class shrink.

One could build the pyramid until eventually getting to the highest level, which consists of one pixel with a value for the feature vector equal to the average for the whole image. But as one goes up in the pyramid, the gross features (central areas) of the small regions start disappearing until the noise of the boundaries eventually covers the whole image. An optimal level at which to stop building the pyramid can be evaluated by having an estimate of the average sizes of the classes of interest. In this case, the small lakes in the scene (average size 15 pixels) suggest using level 4 ($2^4 = 16$) as the top level. Incidentally, notice that this is the only operation that could not be automated. In fact, we need a priori information (an interpretation!) to perform it.

A labelling procedure is started with the application of the maximum likelihood classifier to the compressed picture at the top level of the pyramid. A pixel that gives a 'high enough' probability of belonging to one of the four classes of interest is labelled as strong; otherwise it is labelled as ambiguous.

The segmentation proceeds down the pyramid from this level a level at a time until level 1 is reached. The strong pixels at level L are the starting points of the region growing process. They are simply expanded into groups of four pixels at level L-1 retaining the same label, while ambiguous pixels sent to level L-1 will be re-classified by the maximum likelihood classifier. Since there is a compression factor of four between two successive levels of the pyramid, every time a pixel is labelled as strong at level L, the total number of pixels classified is reduced by $4^{(L-1)}-1$ over the point by point classifier.

The first set of experiments with the pyramidal structure is performed by going up to some level and then successively segmenting down to the highest resolution picture at level 1.

The simple averaging operation going up the pyramid proves to be appropriate for homogeneous areas. But when the spatial transition from one region to another is very abrupt, that is, when neighbouring pixels give a high probability of membership to different classes, then the averaging operation sometimes gives a resulting pixel whose value would lie in an unambiguous, but incorrect, area of the feature space. For example, pixels resulting from the boundary between water (class 4) and forest (class 1 or 2) were often classified as recent logging (class 3) in some higher level of the pyramid. To overcome this difficulty, the value of the dispersion vector of each four pixel cluster is tested against a threshold vector before the averaging operator is applied as one goes from one level up to the next higher level. When abrupt changes along the boundaries are detected by the test, holes are created that are propagated upwards in the pyramid. On the way down, the classification of these areas is delayed by the labelling algorithm until the level at which the hole was created is again reached. Optimal global values for the dispersion thresholds are automatically computed by the classifier.

The classifier so far described gives some improvement over the point by point classifier both in classification accuracy and in readability; that is, it gives a smaller final number of regions. This improvement is obtained without any increase in processing time because the overhead used for building and maintaining the pyramid is balanced by the smaller number of calls to the maximum likelihood classifier.

Further context-sensitivity is introduced in the segmentation process by applying some region merging and splitting techniques at any level in the pyramid. In considering these techniques, remember that a pixel at

level L actually represents a square region of side length $2^{(L-1)}$ of the original picture.

The immediate neighbourhood of each pixel in the image is scanned. A pixel already labelled as strong that does not have a sufficient number of neighbours belonging to its own class is considered to be possibly misclassified and therefore is sent to the next lower level to be classified again. Effectively, the square region in the original image corresponding to that pixel has been split into its four quadrants. On the other hand, a pixel whose probability of membership for any class is not high enough for it to be classified as strong, but which has a large number of strong neighbours all belonging to the same class, has its classification influenced by theirs and is therefore merged into that class.

At the lowest level (highest resolution) of the pyramid, a similar region merging clean-up procedure can be applied to clean up the final segmented image, eliminating the "salt and pepper" noise caused by the small and isolated regions. In the case of the pyramidal classifier this effect, which is otherwise almost completely eliminated as a side effect of the pyramidal structure, is partly reintroduced by the application of the test that delays the classification of many pixels until the lowest level.

3. RESULTS

The pyramidal classifier was implemented in ALGOL W on an IBM 370/168 running under the Michigan Terminal System.

There are 36 regions in the group truth map. The stability of the maximum likelihood classifier was verified with different sets of training data. The size of the set varied from 5 to 20% of the total number of pixels for each class. The correctness of the point by point classification varies from 73 to 75%; as expected, it slightly improves with an increase in the size of the training set. The experiments with the pyramidal structure should be compared with a point by point correctness of about 74% and 220 regions, achieved in 8 seconds of CPU time.

At any level in the pyramid, a pixel is labelled as strong if pmax > Ki(p1+p2+p3 +p4)/100 where pmax = max(p1,p2,p3,p4) is the maximum value for the probability density function for the four classes and Ki is a threshold at level i. If a pixel is labelled as strong at a level, it is classified at that level; otherwise it is sent down to the next lower level to be classified.

Many experiments were performed. They were intended to test the performance of the pyramid classifier in three crucial dimensions, mainly: efficiency, as expressed by the computing time; accuracy, as expressed by the fraction of pixels correctly classified, and readability, as expressed by the number of regions remaining in the final output. For a complete presentation and discussion of the following results see Catanzariti[14].

Up to four levels were used. The first experiments were performed using the straight pyramidal structure; that is, without using the homogeneity test going up, and without doing any region merging. As expected, the thresholds Ki play an important role in the pyramidal classifier. The results showed that the lower (less conservative) the values for the thresholds, the worse the correctness of the classification; on the other hand, efficiency and readability improve (Table 1).

Notice that in the third of the cases shown in Table 1 the pixels are classified only once, at the fourth (top) level of the pyramid. In other words, only 64 pixels (1.8% of the total number of pixels) are classified in this case. Table 1 shows also some results obtained using three and two levels of the pyramid. A drastic improvement in readability is obtained while at the same time also improving slightly efficiency and accuracy. As already pointed out, 3 and 2 levels are more appropriate than 4 levels for the particular scene under study. It should be noted that the execution time is relatively independent of the number of levels used. The increased number of pixels to be classified when using fewer levels is balanced by the smaller overhead required for building and maintaining the pyramid.

The next set of experiments was intended to test the performance of the classifier, including the test for homogeneity in the creation of the pyramid. Table 2 shows some results obtained using four levels.

Global values for the dispersion thresholds are automatically computed by the classifier. The test, although intentionally rough so as to be inexpensive, is very effective in stabilizing the small regions in the image. Detail is not lost. The detail is 'saved' while going up and is 'recaptured' on the way down the pyramid. This time, even the

straight classification in the four main classes at the fourth level outperforms the point by point classifier (first case of Table 2). As expected, the test improves the classification accuracy but, as a side effect, reintroduces some salt-and-pepper noise in the final output. In this case, the clean-up final procedure becomes particularly effective.

In the last set of experiments, the performance of the local region merging and splitting procedures was tested at all levels in the pyramid (Table 3).

These results are not strongly dependent on the number of levels used. This is probably because some of the semantics used by these procedures are already implied in the use of the pyramidal structure. An improvement up to 6% in classification accuracy is obtained over the point by point classification. The final number of regions left is close to the number of regions in the ground truth map. The overall computing time is rarely more than twice the time taken by the point by point classifier.

4. CONCLUSIONS

The efficiency and feasibility of a pyramidal structure have been extensively tested on a typical Landsat image. In evaluating the results presented, a few main points have to be taken into consideration.

Often the results show small improvements in execution time or, in other words, in the total number of pixels classified. The scene under study was 64 x 64 pixels in size, while a full size picture has at least 2048 x 2048 pixels. The difference of a few seconds of CPU time involved in the experiments herein described may become several hours when one or more full size pictures have to be classified. It takes about 8 hours to classify an entire Landsat scene on the IMAGE 100[15].

An important part of the execution time of the whole classification procedure in the pyramid structure consists on the time necessary to build the pyramid (from two to three seconds in the 64 x 64 pixel image). This time could be significantly reduced by making use of special purpose parallel hardware in the pyramid representation. Computer systems connecting parallel hardwired arrays of processors to a serial computer system have been developed over the past few years. Some of these devices are already at the experimental and marketing stage[16]. A software system which makes use of large numbers of regular iterative parallel-serial operations, as is the case of pyramidal classifier, can take enormous advantages of parallel-serial architecture and open totally new perspectives to Landsat data classification.

With regard to the main problem raised in this work, the correctness/readability/efficiency tradeoff, attempts to balance these different factors can be seen as attempts to balance different, often conflicting, points of view.

From the Artificial Intelligence point of view, as long as there is enough memory to contain all the information needed by the program, and as long as the execution time is kept to a reasonable level (less than 24 hours, say), efficiency is not the main concern. Rather, the main concern is the performance of the program in the given domain. The Artificial Intelligence researcher seeks a procedure that can correctly and adequately recognize a given scene and be general enough to be used on other scenes.

From the Remote Sensing point of view, on the other hand, computational efficiency is the first requirement for a classification program that will probably be used in a production environment.

At other times, for example, when producing forest inventory maps, readability becomes a prime factor, for a classified map with a salt-and-pepper appearance, even if correctly classified, is not too meaningful to the user.

The pyramidal classifier here described can quickly classify a scene, giving a very clean and readable output with a correctness comparable to or markedly better than the correctness of the point by point classifier. But also any of the previously described points of view can be stressed by simply changing some parameters in the structure. One might either have a fast rough glance at the scene; one might efficiently classify the image to meet production requirements better than a point by point classifier does, or one might use the pyramid as a fast segmentation component of a more intelligent image understanding system.

ACKNOWLEDGEMENT

This work was supported by grants from the National Research Council and the University of British Columbia.

REFERENCES

1. Landgrebe, D.A. (1975) NASA contract NA-S9-1416 final report, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

2. Haralick, R.M., Shanmugam, K., and Dinstein, I. (1973) Textural features for image classification, IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-3, pp. 610-621.

3. Robertson, T.V. (1973) Extraction and classification of objects in multi-spectral images, Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

4. 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 University, West Lafayette, Indiana.

5. Goldberg, M., Goodenough, D. and Shlien, S. (1975) Classification methods and error estimation for multispectral scanner data, Third Canadian Symposium on Remote Sensing, Edmonton, Alberta, pp. 125-131.

6. Kan, E.P. (1976) A new computer approach to map mixed forest features and post-process multispectral data, Fall Convention of the American Society of Photogrammetry, Seattle, Washington, pp. 386-401.

7. Davis, W.A. and Peet, F.G. (1977) A method of smoothing digital thematic maps, Remote Sensing of the Environment, 6, pp. 45-49.

8. Starr, D. and Mackworth, A.K. (1978) Exploiting spectral, spatial and semantic constraints in the segmentation of Landsat images, Technical Report 78-1, Department of Computer Science, University of British Columbia, Vancouver, B.C. and Can. J. of Remote Sensing 4, 2, 101-107.

9. Mackworth, A.K. (1978) Vision research strategy: black magic, metaphors, mechanisms, miniworlds and maps, Computer Vision Systems, (eds.) Riseman, E. and Hanson, A. Academic Press, (to appear).

10. Nichols J.D., Gialdini, M. and Jaakkola, S. (1973) A timber inventory based upon manual and automated analysis of ERTS-1 and supporting aircraft data using multi-stage probability sampling, Third Earth Resources Tech. Satellite Symp., NASA SP-351 Goodard Space Flight Center, Washington, D.C.

11. Kelly, M.D. (1971) Edge detection in pictures by computer using planning, Machine Intelligence 6, pp. 397-409.

12. Tanimoto, S.L. and Pavidlis, T. (1975) A hierarchical data structure for picture processing, Computer Graphics and Image Processing 4, No. 2, pp. 104-119.

13. Levine, M.D. and Leemet, J. (1976) A method for non-purposive picture segmentation, Third Intern. Joint Conference on Pattern Recognition, Coronado, California pp. 494-498.

14. Catanzariti, E. (1977) Using image hierarchies to interpret Landsat data, M.Sc. Thesis, Department of Computer Science, University of British Columbia, Vancouver, B.C.

15. Strome, W.M. (1975) Remote Sensing: the future, Third Canadian Symposium on Remote Sensing, Edmonton, Alberta pp. 7-25.

16. Uhr, L. (1977) 'Recognition Cones' and some test results; The imminent arrival of well-structured parallel-serial computers; Positions and positions on positions, Computer Vision Systems, (eds.) Riseman, E. and Hanson, A. Academic Press, (to appear).

TABLE 1: Results from varying the number of levels and the thresholds at each level

No. levels	K4	K3	K2	K1	Correctness %	N. regions	CPU-time
4	85	80	75	0	72	73	8.8
4	80	75	75	0	71	60	8.4
4	0	0	0	0	64	8	7.5
3	-	90	85	0	75	140	9.5
3	-	80	75	0	74	99	8.8
3	-	0	0	0	71	29	7.7
2	-	-	0	0	76	70	7.5

TABLE 2: Results using the homogeneity test going up

K4	K3	K2	K1	Correctness %	No. regions	CPU-time
0	0	0	0	75	124	8.3
90	80	0	0	76	50	8.2
90	90	0	0	76.5	60	8.5

TABLE 3: Results using region splitting and merging at each level

Correctness %	No. regions	CPU-time
77	100	9.3
79	60	14.4
79.5	31	21