DOE/DP/20153--T7

Rochester Institute of Technology
Center for Imaging Science

SELECTION OF OPTIMAL TEXTURAL FEATURES FOR MAXIMUM LIKELIHOOD IMAGE CLASSIFICATION

RIT/DIRS Report #89/90-66-131
January 1990

Prepared for

United States Department of Energy
Contract DE-FG01-88DP20153 (Task 1)

Prepared by

Wendy I. Rosenblum
Carl Salvaggio
John R. Schott

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DOE/DP/20153--T7

DE90 006901

ACKNOWLEDGEMENTS

The authors would like to express their deepest thanks to the research staff and graduate students of the Digital Imaging and Remote Sensing Laboratory. In particular, thanks is given to Captain Denis Robert (Royal Canadian Air Force) for the effort put forth on his Master of Science degree thesis at the Center for Imaging Science which served as a starting foundation for this work, to Eugene Kraus for his skill and patience in optimizing the computer code (written by the rest of the "hacks") needed to generate the results for this research in a timely fashion, to Sharon Cady and Ranjit Bhaskar for their assistance in production of the results and their patience with the authors, and to the sponsors for their support of our research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION AND SUMMARY

The classification or segmentation of images into land cover or object class is one of the fundamental remote sensing/image processing tasks. The classification techniques have reached a high level of maturity in the forms of statistical pattern recognition techniques applied to multispectral images. According to this approach each pixel has a spectral vector associated with it and pixels are segmented into the class they most closely resemble spectrally. While these techniques are reasonably successful they do not take advantage of the brightness patterns within a class or at object boundaries nor are they readily applicable to monochrome images or highly correlated multispectral images (e.g. true color images). To overcome these limitations several investigators have suggested the use of image derived features as additional factors for use in image classification. Robert (1989) has identified over forty textural features that have been suggested by various authors as useful for scene classification. Regrettably it is very compute intensive to generate all these features for every pixel in an image and to then use them in a classifier. Schott *et al.*(1988) developed a technique for selecting a small subset of spectral bands from a large set based on criteria intended to optimize maximum likelihood classification. Salvaggio *et al.*(1990) have implemented code to generate 46 image derived textural features and applied a two-step feature reduction and optimization process based largely on the band selection process of Schott *et al.* (1988). The current effort drew on this proof-of-concept work mentioned above and improved on several limitations noted in the earlier work. Specifically; the initial feature set reduction algorithm of Robert (1989) is shown to be deficient and an improved method implemented. The spectral band optimization routine of Schott *et al.* (1989) is applied to feature selection and refined to generate a more truly optimum feature set, the feature generation algorithms developed on the proof-of-concept effort were completely rewritten so that the feature selection and classification process could be implemented in reasonable time frames, and finely the techniques were tested for classification accuracies and consistency of features selected.

For the four scenes tested the classification accuracies on independent data sets were ~96% indicating a strong potential for the use of texture based features in classification and scene segmentation. Furthermore the optimization technique is shown to have considerable value in isolating useful features from the overwhelming number of features which could have been suggested for use in segmentation. The optimization and classification tools developed here are image and feature independent and can be applied to any classification or segmentation problem.

2.0 THEORETICAL / HISTORICAL BACKGROUND

A review of the historical development leading to this work as well as the underlying theory behind image classification is presented in this report. The review here is intended only to provide a general framework for this effort. For a more rigorous review, the reader if referred to the more specific treatments alluded to throughout this section.

2.1 Image Derived Features

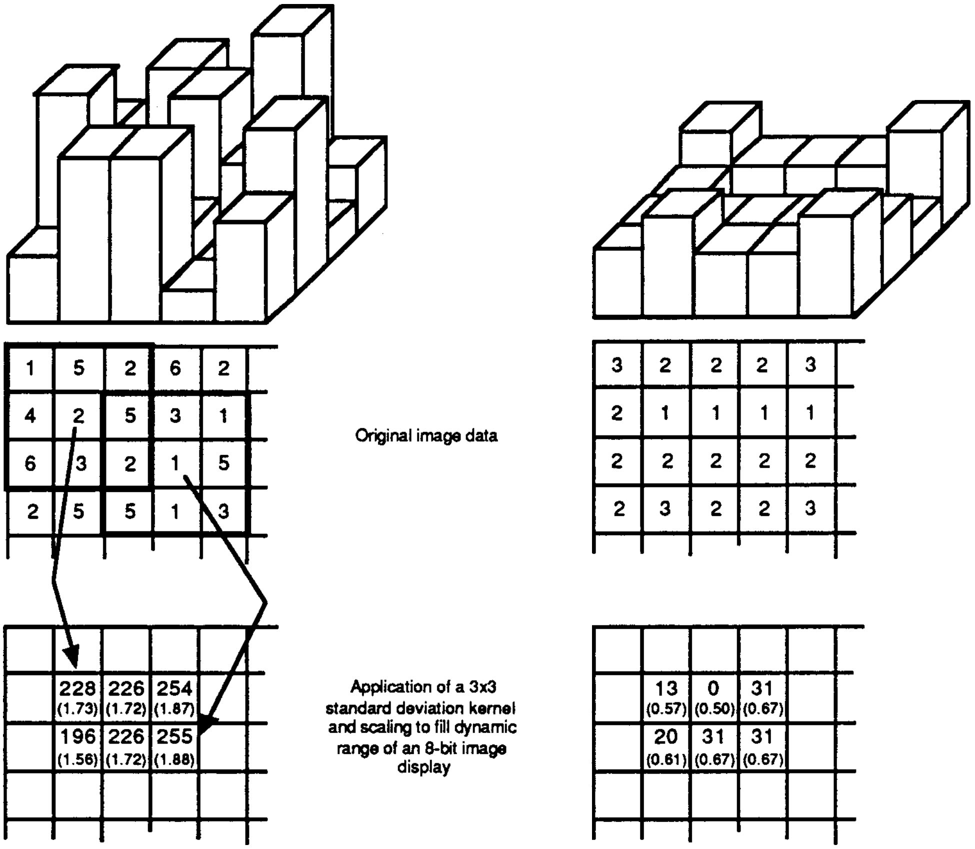
Image data is most often understood to be that collected by some sort of photographic or electrooptical device. These images contain a vast amount of information about relative brightnesses of objects within a scene, spatial proximity, size and so on. Mathematical manipulations of these images can transform much of the information which is stored in these images into a form which can aid in a specific application. This effort will look at the extraction of such information which will aid in image classification.

2.1.1 What Are Image Derived Textural Features?

Image classification has historically utilized spectral image data to develop the necessary statistical pattern recognition metrics with which to assign land cover/class types to individual image elements. The human visual system utilizes "color" as a strong key to the identification and recognition of objects so this history is a well justified one. The visual system also uses "texture" of objects within a field as a key to identification. For example, if one perceives a "red" object in an isolated visual field (i.e. with no surrounding clues), one would be hard pressed to determine if this object was the side of a fire engine or the leaf of a deciduous tree during autumn based on color alone. However, if texture was considered, that is the local change in intensity of the color within some defined spatial region, then the distinction would become a trivial one since the fire engine would appear very homogeneous in its color while the surface of the leaf would have much more variation. It is this variation in color, or grey tone as it will be referred to in this report, that constitutes the monochromatic image derived features known as textures.

Textures became of interest in the field of image classification in the 1970's with the work of Haralick *et al.* (1971,1973,1979), Sutton and Hall (1972), Rosenfeld (1975), Galloway (1975), Keltig and Landgrebe (1976), Weszka *et al.* (1976), Hsu (1978), and Conners (1979). It was found that, when used in combination with the more traditional spectral data, classification accuracies could be significantly improved. It may be helpful to look at a simple example of the kind of information these derived constructs provide.

A simple class of textural features known as first-order statistics provide an intuitive feeling of the type of information provided by these metrics. Let's look at the local standard deviation of a 3x3 collection of pixels. Figure 2.1 shows two isolated areas of a digital image. The two-dimensional bar-chart plot shown provides an insight into the amount of variation in grey-tone which is occurring (a high level in the left-hand image along with a low level in the right-hand image). The relevant digital counts are then shown along with the 3x3 computational windows. The center pixel in each of the possible windows is replaced with a scaled version of the standard deviation of the digital brightness levels which occur in this window. The resulting images are shown. Notice that although the original image data had very similar "colors" (grey-tone levels) that the resulting image derived textural feature illustrate very different grey-levels. The highly variant image on the left produces a feature with very high brightness while the image segment on the right produces very low-levels of brightness. This illustrates that high texture areas produce different image information than the low texture areas. This information can be used in a statistical pattern recognition analysis to aid in material identification or classification.


Original image data
Application of a 3x3 standard deviation kernel and scaling to fill dynamic range of an 8-bit image display

1	5	2	6	2
4	2	5	3	1
6	3	2	1	5
2	5	5	1	3

3	2	2	2	3
2	1	1	1	1
2	2	2	2	2
2	3	2	2	3

	228 (1.73)	226 (1.72)	254 (1.87)	
	196 (1.56)	226 (1.72)	255 (1.88)	

	13 (0.57)	0 (0.50)	31 (0.67)	
	20 (0.61)	31 (0.67)	31 (0.67)	

Figure 2.1 Illustration of different textures within isolated sections of image data

2.1.2 Description Of Features Used In This Study

The monochromatic image derived features used in this study were taken from the work of Robert (1989). In this work Robert implemented the features described in the literature by Haralick *et al.* (1973), Weszka *et al.* (1976), Galloway (1975), Sutton and Hall (1972), Weszka *et al.* (1976), and Hsu (1978) and utilized an optimization approach in order to chose those features which best assisted classification of land cover types in a monochrome image.

The features implemented by Robert can be broken up into 6 classes. These classes include grey-level difference statistics, first-order statistics, run-length statistics, difference statistics, grey-level cooccurrence matrix and spectral features. Table 2.1 summarizes those features used. Only a brief descriptive account of the features will be presented here. For a complete mathematical description of these features, the reader is directed to the original works cited as well as a summary provided in Robert (1989).

Grey-level difference statistics as their name indicates are a class of features which derive their values from absolute differences between pairs of image elements or from some average of these elements (Weszka *et al.*, 1976). Grey-level difference statistics tend to indicate the presence and direction of "edges" within an image. The location and presence of edge information is a clue to the visual system in the recognition process and has been shown to contribute to increased classification accuracy by Sutton and Hall (1972), Rosenfeld (1975) and Rosenfeld and Thurston (1971). Similarly, first-order statistics are a class of features arising from first-order statistical metrics computed on some finite neighborhood of image brightness values (e.g. mean, standard deviation, variance).

Run-length statistics represent contiguous occurrences of identical grey-level values. Galloway (1975) has developed features which indicate the relative magnitudes and directions of these "runs" and has demonstrated increased classification accuracy with their utilization.

Grey-level cooccurrence matrix derived features tend to contain textural characteristics such as homogeneity, grey-level linear structures, contrast, number and nature of object boundaries, and image complexity. The mathematical intricacies can be found in Haralick's research (1973).

If all the features described were to be used in a classification, the amount of time necessary

Table 2.1
Listing of the image derived features used in this study

Cooccurence

Angular Second Moment	Average
	Range
Contrast	Average
	Range
Correlation	Average
	Range
Variance	Average
	Range
Inverse Difference Moment	Average
	Range
Sum Average	Average
	Range
Sum Variance	Average
	Range
Sum Entropy	Average
	Range
Entropy	Average
	Range
Difference Entropy	Average
	Range
Information Measures of Correlation A	Average
	Range
Information Measures of Correlation B	Average
	Range
Difference Variance	Average
	Range
Maximum Probability	Average
	Range

First Order Statistics

Average of Grey Tone Values
Variance of Grey Tone Values

Run-Length Statistics

Short Run Emphasis Inverse Moment	Average
	Range
Long Run Emphasis Inverse Moment	Average
	Range
Grey Level Non-Uniformity	Average
	Range
Run-Length Non-Uniformity	Average
	Range
Fraction of Image in Runs	Average
	Range

Difference Statistics

Textural Edgeness
Contrast
Entropy
Average
Gradient

Spectral Features

Red Grey Level
Green Grey Level
Blue Grey Level

for their production and implementation in a classifier is prohibitive in operation. It is not necessary to use them all and in some cases may prove detrimental to classification. This effort is therefore aimed at the selection of an optimized subset of these features.

2.2 Classification Methodology

Many techniques exist for the classification of image elements into distinct land cover classes. These techniques vary in their statistical rigor and each have implied assumptions as well as advantages in particular scenarios. The techniques invoked in this study have their origin in the maximum likelihood classification schemes described in the literature (Richards, 1988 and Duda and Hart, 1973).

2.2.1 Probabilistic Description

The particular classification method used in this study is maximum likelihood classification under Bayesian assumptions. This technique as developed here will reference classes, i.e. those land cover types chosen for consideration.

If we let the M classes for an image be represented by

$$\omega_r, \quad r = 1, ..., M \qquad (1)$$

The decision to determine to which of the M classes a pixel denoted by the feature vector x belongs to is based strictly on the conditional probability,

$$p(\omega_r|\mathbf{x}), \quad r = 1, ..., M \qquad (2)$$

The vector **x** is simply a column vector of feature values associated with a pixel at position (i,j) in a digital image. The classification decision is made based on the vector **x** being assigned to the class ω_r for which the conditional probability $p(\omega_r|\mathbf{x})$ is the greatest (most probable). This is represented as

$$\mathbf{x} \in \omega_r \quad \text{if} \quad p(\omega_r|\mathbf{x}) > p(\omega_s|\mathbf{x}) \quad \text{for all } r \neq s. \qquad (3)$$

All that need be done is to determine the conditional probabilities $p(\omega_r|\mathbf{x})$. These values are unknown, however, they can be estimated provided a sufficient amount of training data can be collected from the image. If training data is collected from the feature imagery for each land cover class of interest, the multivariate probability distribution for each land cover type can be established, $p(\mathbf{x}|\omega_r)$.

The number of these conditional distributions will equal the number of land cover types. Knowing these conditional probabilities, probability values can be computed which represent the relative likelihood that the feature vector **x** at a point (i,j) in an image belongs to each class established. The Baye's theorem relates these two conditional probabilities as

$$p(\omega_r|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_r)\, p(\omega_r)}{p(\mathbf{x})} \tag{4}$$

where $p(\omega_r)$ is the probability that a class ω_r occurs in an image. These are called *a priori* probabilities. The probability $p(\mathbf{x})$ is the probability of finding a pixel from any class at location **x**. The probabilities $p(\omega_r|\mathbf{x})$ are called *posteriori* probabilities since these are the probabilities of a vector **x** belonging to class ω_r after a decision has been made. The decision rule above can be rewritten as

$$\mathbf{x} \in \omega_r \quad \text{if} \quad p(\mathbf{x}|\omega_r)\, p(\omega_r) > p(\mathbf{x}|\omega_s)\, p(\omega_s) \quad \text{for all } r \neq s \tag{5}$$

where the $p(\mathbf{x})$ is removed as a common factor. This is a more acceptable rule since the conditional probabilities can be known from training data and the analyst can make a good guess as to the values of $p(\omega_r)$ from a knowledge of the image (these *a priori* probabilities are often assumed equal). For mathematical convenience, $g_s(\mathbf{x})$ is defined as

$$g_s(\mathbf{x}) = \mathit{ln}[\, p(\mathbf{x}|\omega_s)\, p(\omega_s)\,] = \mathit{ln}[\, p(\mathbf{x}|\omega_s)\,] + \mathit{ln}[\, p(\omega_s)\,] \tag{6}$$

and the decision rule is again rewritten as

$$\mathbf{x} \in \omega_r \quad \text{if} \quad g_r(\mathbf{x}) > g_s(\mathbf{x}) \quad \text{for all } r \neq s \tag{7}$$

where $g_r(\mathbf{x})$ is known as the discriminant function. If the probability distribution of the feature vectors corresponding to each class can be assumed to be multivariate normal, the conditional probability $p(\mathbf{x}|\omega_r)$ can be computed for n features as

$$p(\mathbf{x}|\omega_r) = \frac{1}{(2\pi)^{n/2}\,|\Sigma_r|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_r)^t\, \Sigma_r^{-1}\, (\mathbf{x} - \mathbf{m}_r)) \tag{8}$$

where $\mathbf{m}_r$ and Σ_r are the mean vector and covariance matrix for the training data in class r. The discriminant function $g_r(\mathbf{x})$ can be written as

$$g_r(\mathbf{x}) = \mathit{ln}\ p(\omega_r) - \frac{1}{2} \mathit{ln}\ |\Sigma_r| - \frac{1}{2} (\mathbf{x} - \mathbf{m}_r)^t\, \Sigma_r^{-1}\, (\mathbf{x} - \mathbf{m}_r) \qquad (9)$$

If no information is known about the *a priori* probabilities this reduces to

$$g_r(\mathbf{x}) = -\mathit{ln}\ |\Sigma_r| - (\mathbf{x} - \mathbf{m}_r)^t\, \Sigma_r^{-1}\, (\mathbf{x} - \mathbf{m}_r) \qquad (10)$$

These two values of the discriminant function form the maximum likelihood classification methodology using a Bayesian decision rule depending on whether or not *a priori* probabilities are supplied by the investigator.

2.2.2 Mahalanobis Distance

Consider again the case of unknown or equal *a priori* probability demonstrated in Equation 10. If the sign of this function is reversed, the quantity can be considered a squared distance measure since this is implied by the quadratic term and the logarithmic term is a constant for class i. Rewriting Equation 10 in this form you have

$$d(\mathbf{x},\mathbf{m}_r)^2 = \mathit{ln}\ |\Sigma_r| + (\mathbf{x} - \mathbf{m}_r)^t\, \Sigma_r^{-1}\, (\mathbf{x} - \mathbf{m}_r) \qquad (11)$$

where $d(\mathbf{x},\mathbf{m}_r)$ is a measure of distance which is sensitive to direction as well as modified according to class. The classification decision is then made to assign a pixel with the descriptive vector **x** to class r if the distance $d(\mathbf{x},\mathbf{m}_r)$ is the smallest for all M classes. If we then consider the case where all class covariances are equal ($\Sigma_r = \Sigma$ for all r) then the term $\mathit{ln}\ |\Sigma_r|$ no longer contributes to the discriminating ability of the metric and can be ignored leaving

$$d(\mathbf{x},\mathbf{m}_r)^2 = (\mathbf{x} - \mathbf{m}_r)^t\, \Sigma^{-1}\, (\mathbf{x} - \mathbf{m}_r) \qquad (12)$$

the square root of which is referred to as the Mahalanobis distance. This has obvious computational advantages over maximum likelihood classification while maintaining a sensitivity based upon the covariance metric, Σ, which is often a class average metric referred to as a pooled covariance matrix. However, this approach should be used only when the covariance matrices for all classes are statistically equal.

2.3 Choosing an Optimum Set of Image Features

In the previous section a large number of feature images were discussed which can be derived from a single monochromatic image. The time required to compute all of these images and the sheer amount of data that would be present after their production is prohibitive to their effective use as classification aids. In addition, many of the features may be highly correlated such that their use in a classifier does not add significant information and may even degrade classification accuracy. It is therefore necessary to choose a subset of these feature images which serve as an optimum classification set. Many statistical techniques exist which serve as data redundancy reduction tools such as factor analysis, principal component analysis, etc., but these techniques do not serve to optimize the choice of a subset of data which will aid in a particular classification scenario. Schott *et al.* (1988) and Robert (1989) have presented a technique which attempts to produce optimized results within the context of the classification scheme used.

2.3.1 Use of Correlation as a Prescreener

Any technique which attempts to pick the best subset of size n out of a larger data set of size k very quickly becomes a large combinatoric problem if k is of any significant size. The original collection of all features used in this study amounted to 49 monochromatic image derived features (including the three spectral bands). This amount of data proved to be too large to deal with on the computational facilities available. Initial prescreening of the data needed to be performed.

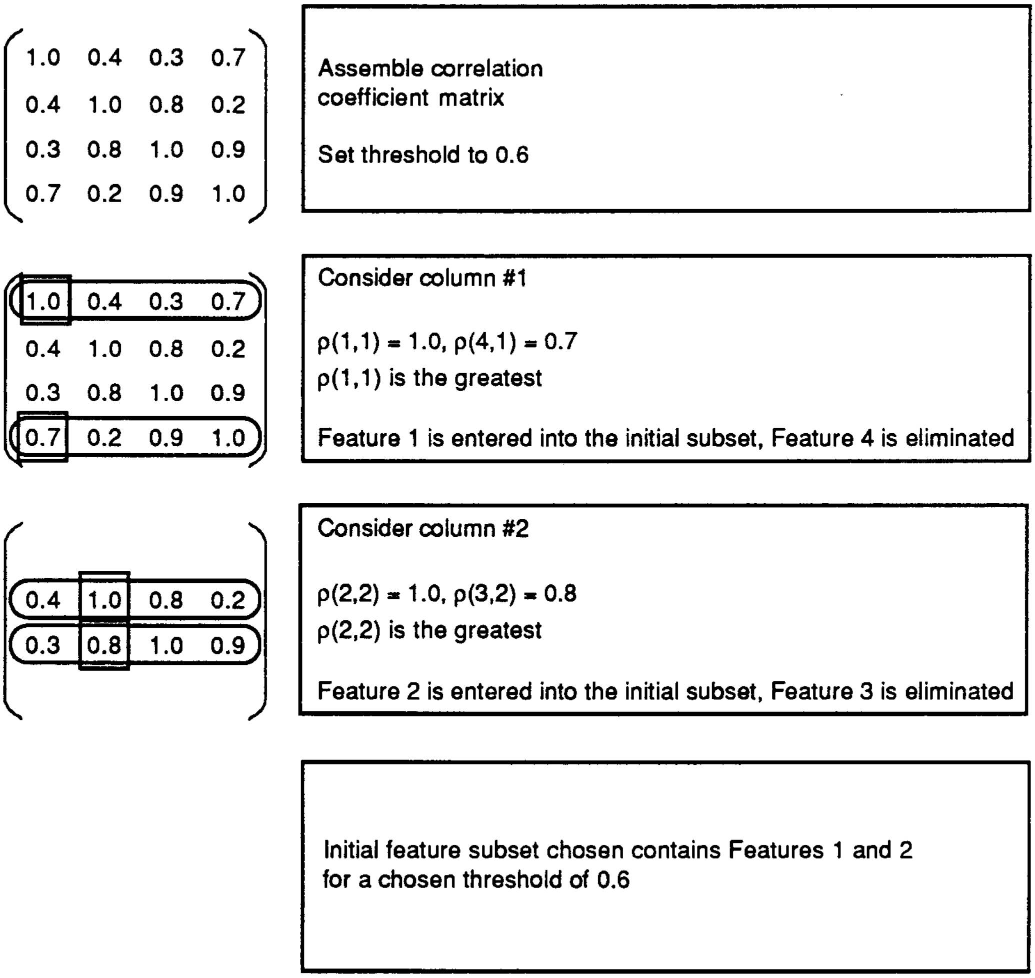
As with spectral image data, it was found that textural features derived from a single monochrome image contained a large amount of inter-feature correlation. It was deemed by Robert (1989) that highly correlated features need not all be included in the optimization analysis since the amount of unique information provided by each was minimal. The choice was made to include only one feature of each highly correlated collection.

This choice was made by the following selection criterion. It was assumed that the individual class covariances proved equal and a pooled covariance matrix defined. If the initial feature space covered k-dimensions for M defined classes then a kxk pooled covariance matrix is definable. This pooled covariance matrix in then used to define a correlation coefficient matrix of the form

$$\rho_{i,j} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,i}\,\sigma_{j,j}}} \tag{13}$$

where the σ terms indicate individual elements of the pooled covariance matrix Σ. Once this matrix is formulated, an initial subset of features is selected in the following manner.

Correlation coefficients are considered by column (i.e. column 1 of the matrix is looked at first) and all those entries whose value for correlation exceed some user defined threshold are grouped. This group is then examined and the highest correlation coefficient in this group selected. The feature corresponding to the row selected is then entered into the initial subset and all other rows are deleted from further consideration. Of the remaining rows, the second column is examined in the same manner, a grouping performed, and a selection made. This procedure is carried out until all columns are exhausted. Figure 2.2 illustrates this procedure on a 4x4 set of correlation coefficients. This initial subset of features is then entered as the selection set from which the optimum collection of features will be extracted.



Assemble correlation coefficient matrix

Set threshold to 0.6

Consider column #1

$\rho(1,1) = 1.0$, $\rho(4,1) = 0.7$
$\rho(1,1)$ is the greatest

Feature 1 is entered into the initial subset, Feature 4 is eliminated

Consider column #2

$\rho(2,2) = 1.0$, $\rho(3,2) = 0.8$
$\rho(2,2)$ is the greatest

Feature 2 is entered into the initial subset, Feature 3 is eliminated

Initial feature subset chosen contains Features 1 and 2
for a chosen threshold of 0.6

Figure 2.2 Example of the use of correlation criteria as a prescreener to the optimization code

This prescreening method has a weakness in it's choice of the best feature from the group assembled from each column. For each column, more than one feature may produce a correlation value of equal magnitude at which point this selection method depends on the order of the original data. A method needs to be developed which will not be affected by such factors. For this study a prescreening method incorporating an eigenvalue/eigenvector transformation of the data was used. This method is discussed in Section 3.1.

2.3.2 Pure Mahalanobis Distance Approach

Since the classification scheme chosen for this study was a Mahalanobis distance based implementation, an optimization approach which enhanced separability of classes in the context of this classifier is most appropriate. Schott *et al.* (1988) proposed a method by which a class separation metric Z was determined. This metric had the form

$$Z = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n} d_{i,j}^{2}} \tag{14}$$

where $d_{i,j}^2$ are the Mahalanobis distances computed between all class means. This metric was computed for all combinations of n features chosen from the original set of size k. The subset producing the maximum value of Z was then deemed the optimal set of n features (i.e. that set which yields the largest class separability).

A significant problem exists with this approach. The value of Z can be affected very heavily by the position of one class mean. Figure 2.3 shows two situations in a two-dimensional feature space for simplicity. The first situation shows the relative positions of three class mean values arranged in an equilateral-triangular orientation. The second situation shows the same three class mean values with a different orientation. This second situation obviously produces a larger metric Z however the class separability is not as well established for classes 1 and 2 and is over-established for class 3. A modification to this approach is implemented in this research as detailed in Section 3.2.

2.3.3 Mahalanobis Distance Approach With Class Weighting Factors

A further refinement to the feature optimization approach was established by Schott *et al.* (1988) which allowed the user to specify relative importances on particular sets of class separabilities. For example, it may be important to tell class 1 apart from classes 2 and 3 however it may seem relatively unimportant to tell classes 2 and 3 apart. A scenario of this

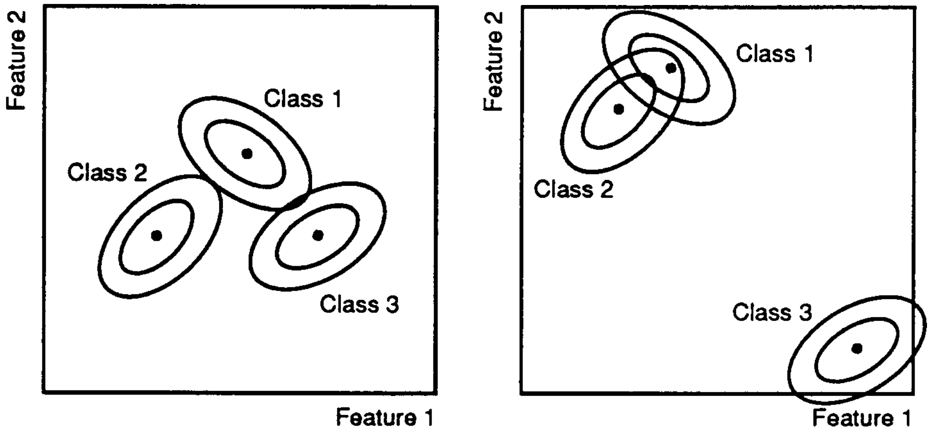


Figure 2.3 Illustration of problems which may occur with the separability metric Z of Schott *et al.* (1989)

sort may exist in isolating a ship on the water apart from the surrounding water and landmasses. If the scenario is to establish which pixels in a scene are boat pixels, calling all the others background, then this would be the case. This relative weighting of the importance of particular separabilities warrants further attention.

The class separation metric defined in Equation 10 can be rewritten as

$$Z = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{n} w_{i,j} \, d_{i,j}^{2}} \qquad (15)$$

where the term $w_{i,j}$ is an element of a weighting factor matrix W. This matrix W contains the relative importances of being able to separate class i from class j with statistical directionality implied. For example, if you had the 3-class problem referred to above, the matrix W may look like

$$W = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

where the only important separability to establish is between class 1 and classes 2 and 3, with the separation between classes 2 and 3 carrying no importance at all.

The weights do not need to be 1's or 0's however this is typically the case in target/background scenarios.

The methods of Schott *et al.* (1988) and Robert (1989) which are drawn on here each have some intrinsic limitations. Attempts are made in this implementation to rectify these problems.

3.0 EXPERIMENTAL APPROACH

There are approximately two dozen cooccurrence features which are defined along with many run-length features. Most of these are dependent on the angle for which they are computed (this is a four time increase for 0°, 45°, 90° and 135° orientation angles). The number of features possible for use in classification can rapidly get out of hand (Haralick, 1979). The main goal of feature selection is to select a subset of n features out of N features (n<<N) without significantly degrading the classification accuracy obtainable. The number of features used in the classification should still give a minimal probability of misclassification (Fu, 1976).

Choosing a subset of features is very important when textural features are being used to classify an image. The generation of most textural features is computationally intensive, so it is advantageous to produce as few feature images as possible. To perform an image classification with textural features, full-resolution images of each textural feature must be calculated, making a reduction in the entire feature set necessary. Data reduction with spectral images using techniques such as eigenvector transformations may be used since a linear combination of the spectral bands is sometimes better for classification than the individual bands. With textural features, calculating all of the feature images in order to form a linear combination would be very slow, defeating the purpose of choosing a subset.

Feature selection algorithms are characterized by a search procedure, a selection criterion, and possibly a stopping criterion (Queriros and Gelsema, 1984). In the search procedure, combinations of subsets of n features out of N are tested and some measure of their classification abilities made. In the selection criterion, the subset with the most potential for correct classification is chosen. For the stopping criterion, the effect of including additional features would be measured and the addition to the accuracy of the classification is weighted against the extra computation. These three steps should lead to the selection of the best subset for classification.

3.1 Prescreening Approach (Eigenvector Criteria)

The first step in the optimal selection of features for use in image classification is a prescreening of the features. This step serves to eliminate highly correlated features, those features which do not contain unique information and bring the number of features being compared down to a size which can be quickly analyzed by a more rigorous optimization process.

As mentioned before, the prescreening by analysis of correlation has several problems. By examining the correlation between features in the first column and working across to the last column, one set of uncorrelated features is chosen. If one were to start in the last column and work towards the first, a different set of features would be chosen. Both orders of operation would produce sets of uncorrelated features, but a more robust selection method is required. A second problem with feature selection by correlation is that it only examines combinations of features based on their inter-correlations with no regard for information content.

A second prescreening method was developed based on eigenvector transformations which considers both the information content of the feature and its correlation to the other features. A covariance matrix is calculated from the training data representing the covariance between all calculated features for each class in the image. These covariance matrices for the individual classes are then pooled and the eigenvectors/eigenvalues calculated. The eigenvectors represent linear combinations of the features which explain the amount of the information in the original matrix proportional to the corresponding eigenvalue. The eigenvectors are ranked according to their eigenvalues so the first eigenvector represents the most information and so on. The eigenvectors have the additional property that they are orthogonal, i.e. each is independent of all others.

To preselect the best k features from the entire set of N, the preselection method examines the first k eigenvectors and chooses the features which contributes the most to that eigenvector. In this manner, the features which contribute to the most significant eigenvectors will be chosen, and the features chosen should be minimally correlated since they come from orthogonal eigenvectors. This process provides a reduced set of features with reduced correlation and high information content. It is recognized that this is not an optimal selection technique, but it should provide an adequate set of initial features from which to choose an optimal set.

3.2 Optimization Process

The most common approach for a selection criterion is to define a distance or separability measure between the probability distributions corresponding to the classes under investigation (Fu, 1976). The separability measure which should best represent distance between classes is a function of the Mahalanobis distance. For each subset of features, the distances between all possible combinations of two classes can be measured according to

$$\mathbf{d(m_r,m_s)^2 = (m_r - m_s)^t \Sigma^{-1} (m_r - m_s)} \qquad (16)$$

where $\mathbf{m}_r$ and $\mathbf{m}_s$ are the mean vectors for classes r and s, respectively, and Σ is the pooled covariance matrix between all features. If equality of class covariance matrices can not be assumed, the distance measure becomes

$$d(\mathbf{m}_r,\mathbf{m}_s)^2 = \mathit{ln}\ |\Sigma_r| + (\mathbf{m}_r - \mathbf{m}_s)^t\, \Sigma_r^{-1}\, (\mathbf{m}_r - \mathbf{m}_s) \qquad (17)$$

where Σ_i is the individual class covariance matrix. This defines the statistical distance of the mean of class s from the mean of class r in class r's probability space. Because the distances are measured from class r to all other classes in r's probability space, the distance from class r to class s will be different than the distance from class s to class r in s's probability space. Figure 3.1 illustrates a possible example.

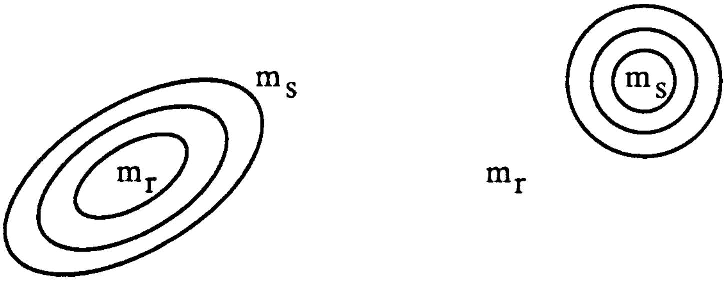


Figure 3.1 Illustration of the discrepancy between statistical distance measured between points in different 2-dimensional probability spaces

Once the interclass distances are defined the subset providing the largest sum of distances between all combinations of classes would seem to provide the greatest separation between classes and therefore the best classification results. This is not always the case.

Using the distance measure between two classes (Equation 17), a standard distance was defined to improve the separability of classes. Measuring in class r's probability space (using r's covariance matrix), the distance between the means of classes r and s needed to make the probability of misclassification small was calculated as follows. Given that the probability of finding the mean of class s in a sample from class r is defined as

$$p(\mathbf{m}_s|\omega_r) = \frac{1}{(2\pi)^{n/2}\,|\Sigma_r|^{1/2}}\ \exp(\,-\frac{1}{2}\,(\mathbf{m}_r - \mathbf{m}_s)^t\, \Sigma_r^{-1}\, (\mathbf{m}_r - \mathbf{m}_s)\,) \qquad (18)$$

where $p(\mathbf{m}_s|\omega_r)$ is the probability of misclassification. If this probability is fixed by the analyst at some maximum acceptable misclassification probability P and the quantity

$(\mathbf{m}_r - \mathbf{m}_s)^t \Sigma_r^{-1} (\mathbf{m}_r - \mathbf{m}_s)$ replaced by the class dependent constant D_r, Equation 18 can be simplified as

$$P = \frac{1}{(2\pi)^{n/2} |\Sigma_r|^{1/2}} \exp(- \frac{1}{2} D_r) \qquad (19)$$

The quantity $D_r + ln\ |\Sigma_r|$, the distance between the means of the two classes modified for the use of individual covariance matrices can be solved for as

$$D_r + ln|\Sigma_r| = -2ln(P) - n\,ln(2\pi) \qquad (20)$$

where n is the number of optimal features chosen from the initial subset of size k. Now $D_r + ln\ |\Sigma_r|$ is the distance modified for the use of individual covariances which the mean of class s must be from the mean of class r such that they have the probability P of misclassification. The probability of misclassification, P, can be set so that its value is known and the distance necessary for that probability, defined as the threshold distance, can be solved for between the means of classes r and s. This measurement will be used as a threshold for the actual distances between classes calculated during the optimal feature selection, the steps of which are illustrated in Figure 3.2.

The standardized distance is used to assure adequate separation between two classes. If that distance is much greater than adequate (i.e. with a ratio greater than 1.0) the value of the ratio is truncated to a maximum of 1.0. By truncating the ratio, a very large separation between two classes would not inflate the value of the summed entries of the divergence matrix. With this method all classes must be well separated in order for the sum from the divergence matrix to be high and the subset of features which produces the highest summation is chosen.

3.3 Choosing Training Samples

The training samples were chosen with their textures in mind. Classes must be chosen carefully so that the elements may be recognized by a textural feature. The textural features derived in this study were calculated over a 5x5 pixel window. Since this was the case, each training sample had to be at least 5 pixels square. Problems arose with such classes as railroad tracks which are only 5 to 7 pixels wide in the imagery used. At the same time, the texture in the training sample must remain fairly constant from one 5x5 pixel block to the

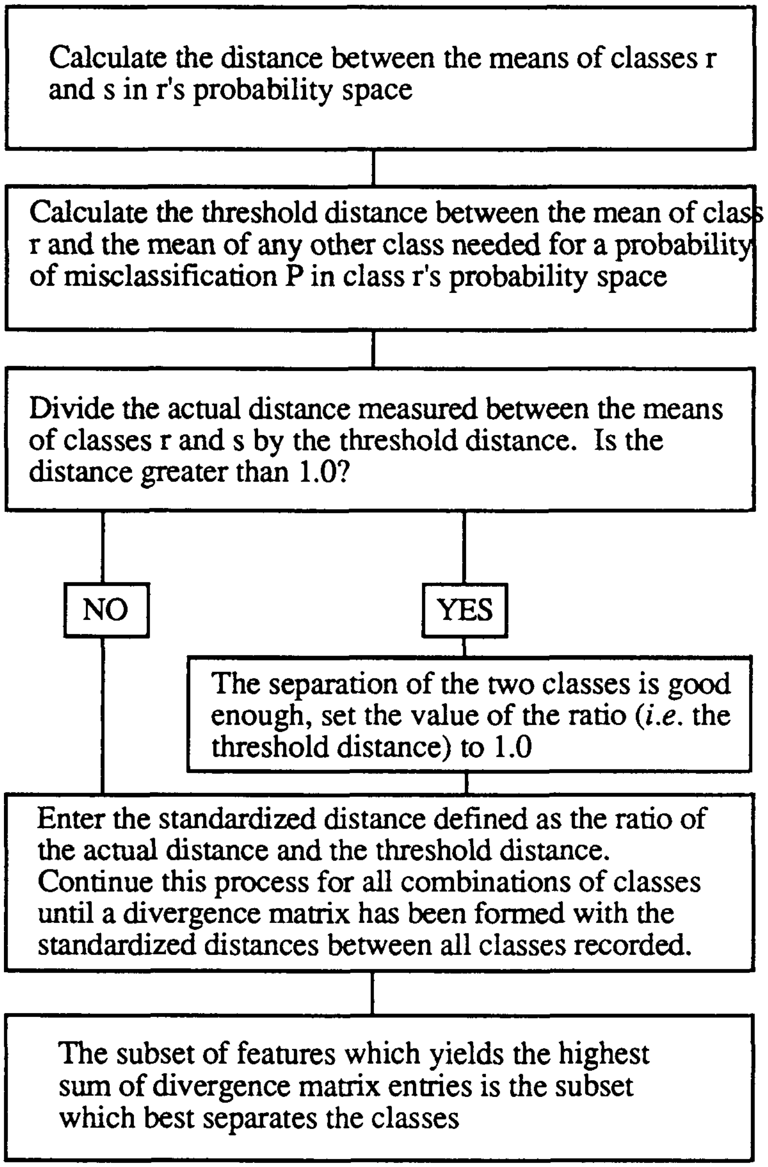

Calculate the distance between the means of classes r and s in r's probability space
Calculate the threshold distance between the mean of class r and the mean of any other class needed for a probability of misclassification P in class r's probability space
Divide the actual distance measured between the means of classes r and s by the threshold distance. Is the distance greater than 1.0?
NO
YES
The separation of the two classes is good enough, set the value of the ratio (i.e. the threshold distance) to 1.0
Enter the standardized distance defined as the ratio of the actual distance and the threshold distance. Continue this process for all combinations of classes until a divergence matrix has been formed with the standardized distances between all classes recorded.
The subset of features which yields the highest sum of divergence matrix entries is the subset which best separates the classes

Figure 3.2 Flowchart demonstrating the procedure carried out to choose the optimum set of n from k features

next within the training sample. This posed a great problem when training on houses. One 7x7 pixel area would consist of the house, while there would be large blocks of pixels around the house consisting of trees or roads. If training was done for one large area over the entire neighborhood, the values calculated for the textural features would vary widely in that class since it would have composite values for trees, houses and streets. To overcome this problem, training samples were made up of samples of individual houses with some surrounding trees. The same was done for a class of suburban roads. By breaking the training samples up into sizes and patterns that were recognizable by the textural feature, it was possible to greatly increase the overall accuracy of the classifier.

3.4 Determination of the Number of Features to Use

In order to determine the number of features to select, sets of two, three, four, ... optimal features were chosen and the classification accuracy for a dependent set of pixels calculated. The overall classification accuracy for each set of features was calculated by weighting the accuracy of each class by the number of pixels in that class. Plots of the classification accuracy vs. number of features chosen are shown in Figures 3.3 and 3.4.

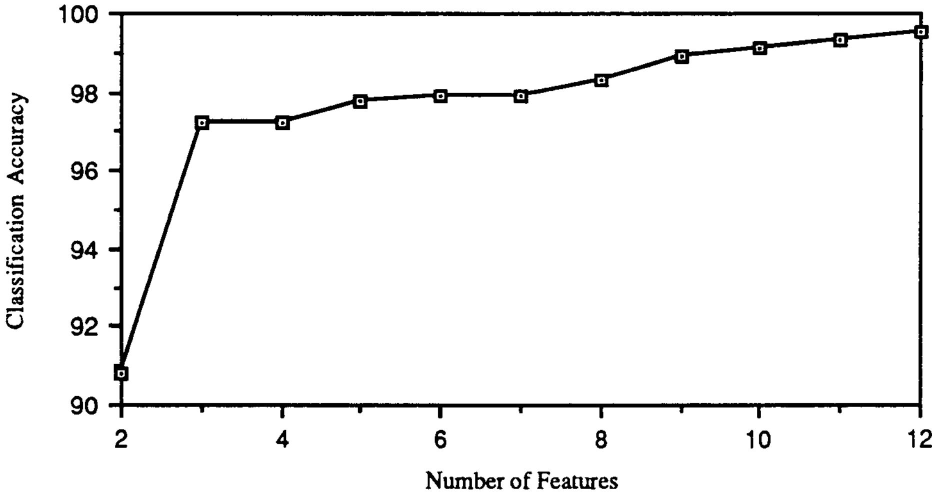


Figure 3.3 Classification accuracy as a function of number of optimal features chosen for image RR1

The plots shown illustrate that the classification accuracy levels off after six features were included in an optimal set. This value was used throughout the rest of this effort although it is clear from Figure 3.4 that a smaller number of features should work quite well for some scenes.

3.5 Determination of Classification Accuracy

The procedure of determining how good a classifier is performing can only be definitively

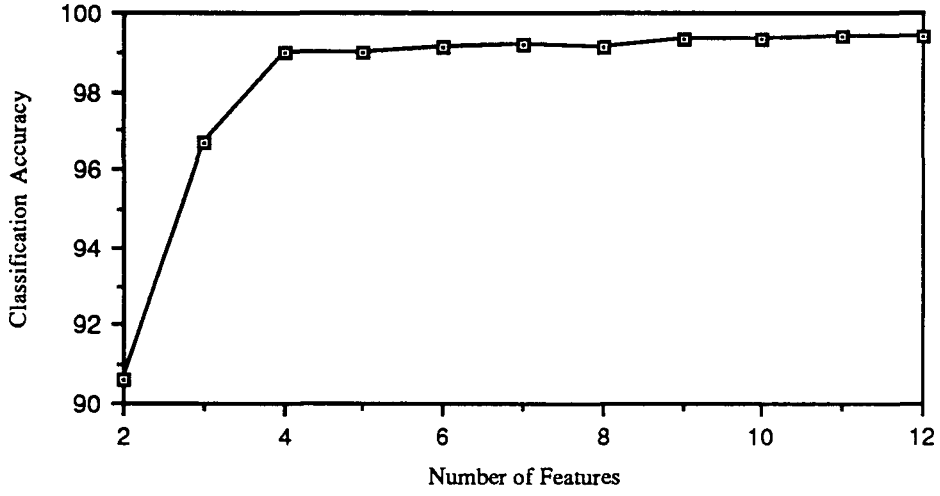


Figure 3.4 Classification accuracy as a function of number of optimal features chosen for image ROME2

answered by comparing the land cover map produced with ground control data for every point composing the imaged area. This process is not a viable solution to the question due to the size of the scenes involved. Also, if you knew the answers before you performed the classification, there would be no need to carry it out. Sampling techniques are therefore implemented for evaluation of classification accuracy.

The most simple method for evaluating the success of a classifier is defined as a dependent/ independent analysis. The training of a classifier involves the collection of "representative" pixels which describe classes. One can choose to utilize a certain percentage of the collected pixels to compute the necessary statistics for the development of the classifier. These data are termed the dependent data since the derived statistical method relies on their values. The remaining percentage of trained pixels while still representing unique classes within the scene do not have a direct influence on the classifier. Therefore it is possible to call these independent data and use them to evaluate the effectiveness of the classification scheme.

To accomplish this task a percentage split of 80% to 20% of the trained pixels was employed representing the dependent and independent data, respectively. The classifier was developed based on the 80% of the training pixels and the theory presented earlier. A classification of the remaining 20% of the data was then perform and the percentage correct classification for each class recorded. This provides a useful "quick-and-dirty" evaluation of how well your training data represented the actual phenomena unique to each class. This same procedure can also be applied to the dependent data. Repeated refinement of the training data can be carried out until the results of classification accuracy for the dependent and independent data coincide within reasonable limits.

While seemingly adequate to many researchers as a method for accessing class accuracy, the method described has some weaknesses. First, the training samples represent data from localized proximity within the scene. That is the data collected for a grass field represents a particular kind of grass (e.g. be it in type or in biological condition). Testing accuracy in the manner presented is a sort of self-fulfilling prophecy since the classifier may be able to do very well only on data similar to that with which it was trained. Therefore this method lacks robustness. Second, a statistical test of this sort requires that the analysis be random, where in fact the method presented is very structured in it's implementation.

A method to address both of these problem was used in this study. The procedure uses training data to develop the classifier just as the previous implementation did, however, all the data are used. The classification accuracy is then evaluated by a random selection of pixels from within each class determined by the classifier (a set number in all classes) and presentation of these pixels to the analyst for ground truth identification. For example, if the classifier determined a pixel to be grass, the user is blindly asked to supply the class to which ground truth indicates a particular pixel belongs. A tally is kept of the correct as well as incorrect classifications according to the supplied ground truth and classification accuracy determined.

This method is completely random and robust, i.e. it selects pixels from outside as well as inside the training population. The accuracy determined therefore represents an uncorrelated estimate of the actual accuracy expected from the classifier. This method can be time consuming but the confidence in the estimates provided can be much higher. The analysis carried out in this endeavor randomly choose 50 points from each class and presented them to the user for comparison with ground truth data (which in this case was a high-resolution color aerial photograph). Classification accuracies defined by this procedure invariably yield lower, more conservative results than other methods. These results should, however, more closely resemble actual results on whole image classification. Included in this measure of accuracy is error due to mixed pixel misclassification as well as misclassification due to failure to include significant classes (or sub-classes) during the training process. Since other investigators commonly report only dependent and independent classification accuracies for their classifiers, these values are also reported here along with the results from the random sampling procedure.

4.0 RESULTS

The approach presented was applied to two different scene types, the first containing railroad systems and the second containing camouflaged targets. Two images with approximately 1 meter per pixel spot size were analyzed for each scene type. Figures 4.1 through 4.4 show the four scenes used in this study. For the railroad system images classes of railroad tracks, grass, trees, houses, roads and highway were chosen as cover types while for the camouflaged target scenes cover types of grass, trees, camouflage and two types of roadways were chosen.

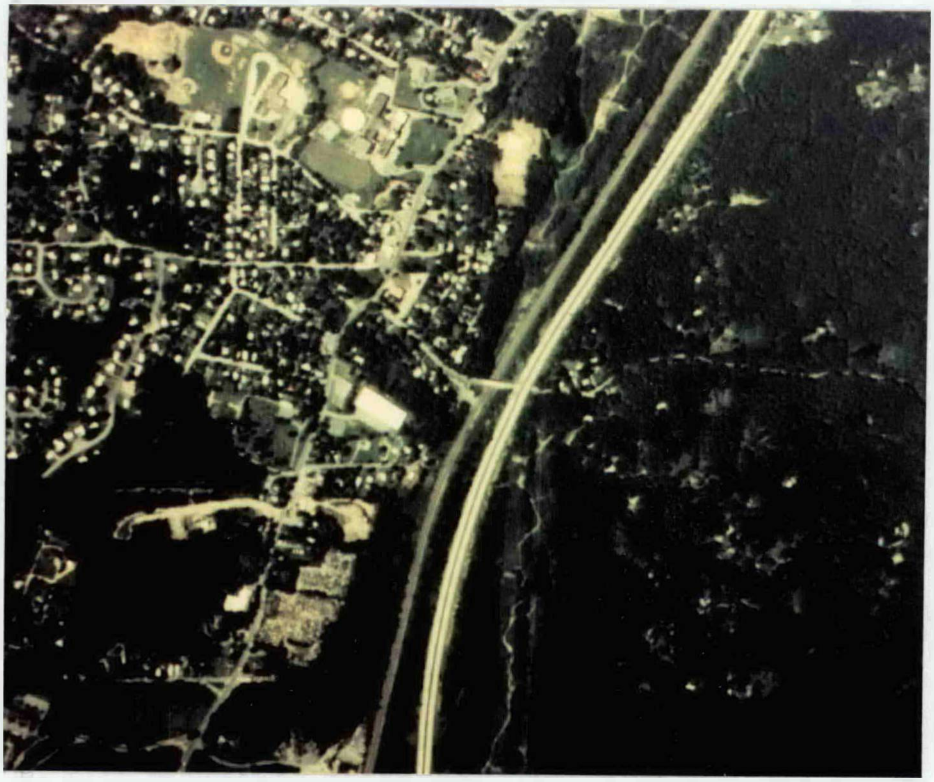


Figure 4.1 Digitized air photo containing railroad system (RR1)

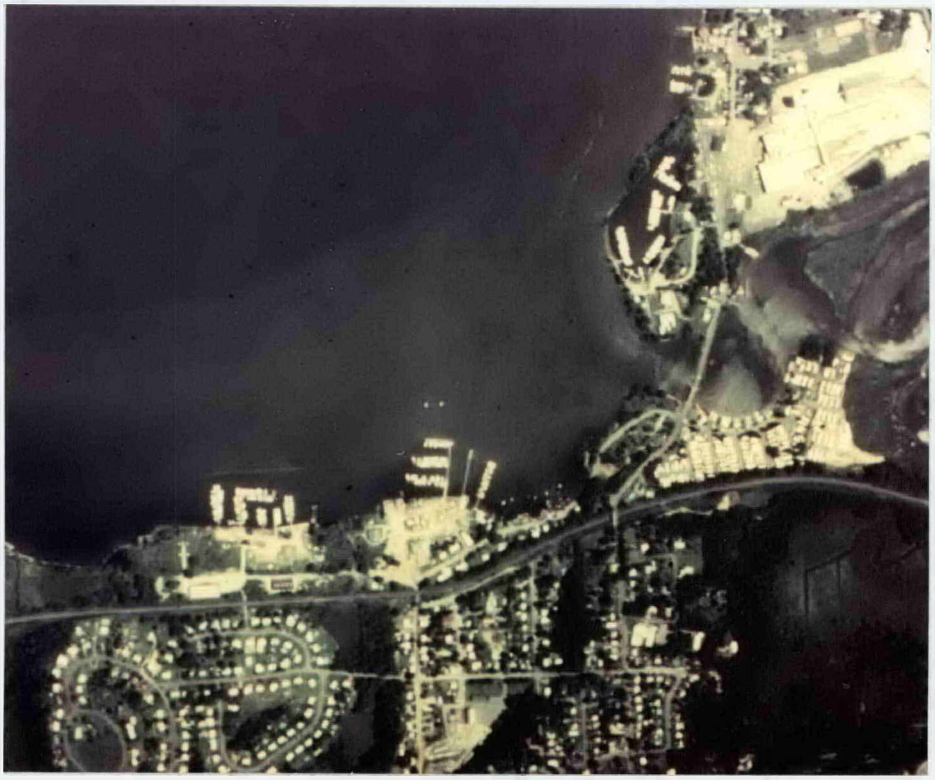


Figure 4.2 Digitized air photo containing railroad system (RR2)

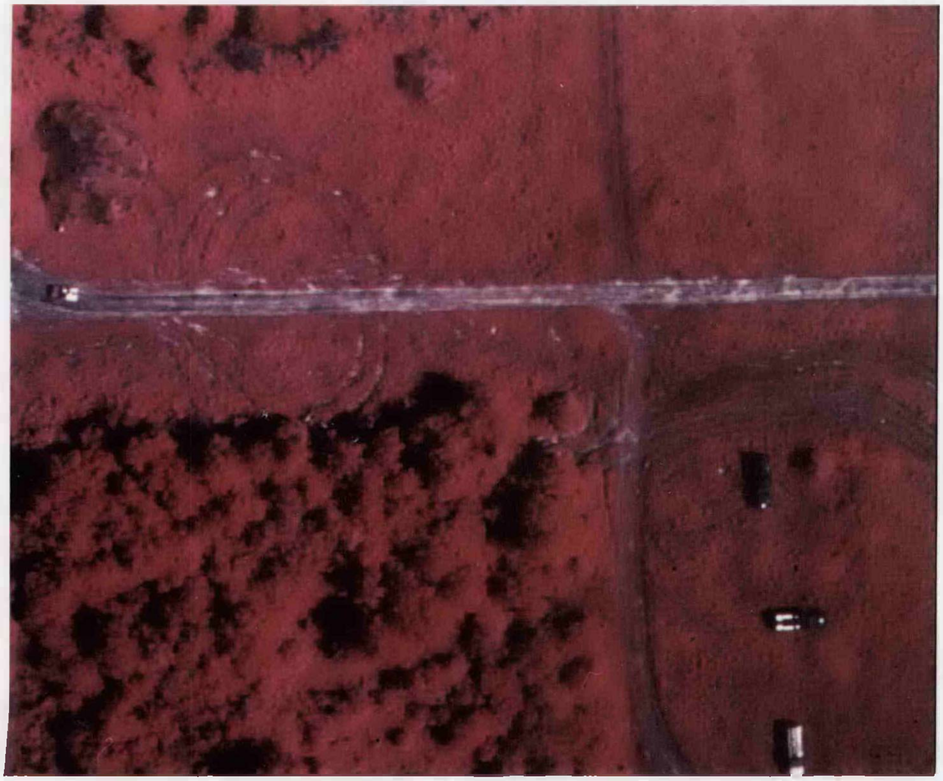


Figure 4.3 Digitized air photo containing camouflaged targets (ROME1)

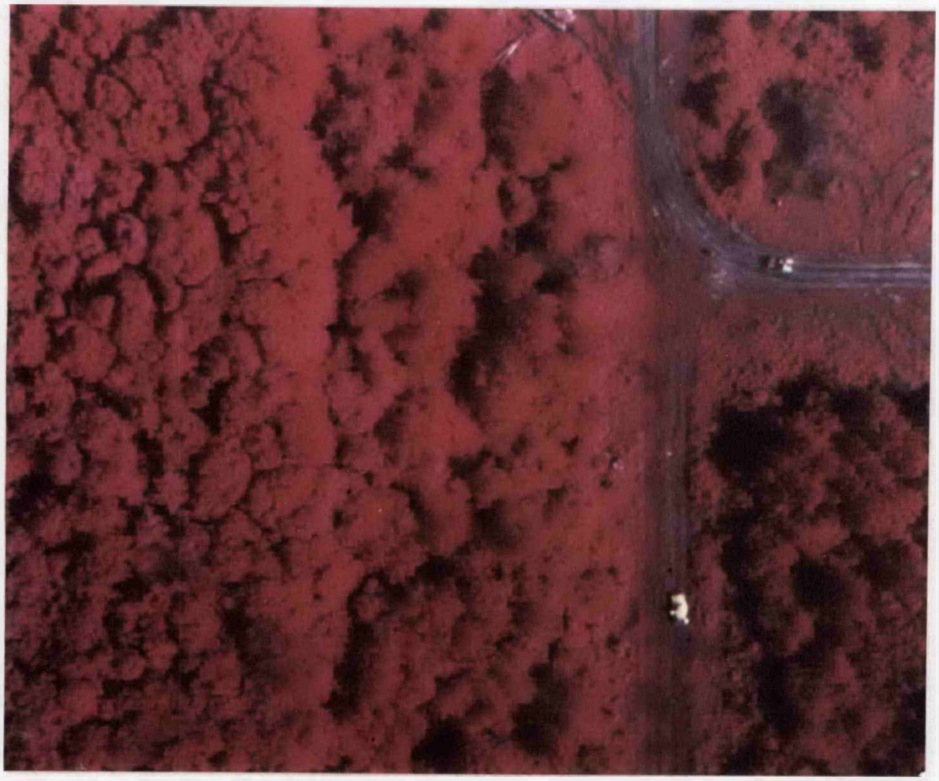


Figure 4.4 Digitized air photo containing camouflaged targets (ROME2)

4.1 Optimal Features Chosen as a Function of Scene Type

As a result of the analysis to determine the number of features to use for classification described in Section 3.4, the six optimal features were chosen for each of the four images. Table 4.1 lists the names of the features chosen for each of these four images and Table 4.2 list the relative occurrences of like features across image type. Figure 4.5 illustrates the feature set chosen for image RR1. Appendix A contains a mathematical description of all the features listed in Table 2.1 and Appendix B contains images of these features derived from image RR1 as a reference.

4.2 Dependent and Independent Classification Accuracies

Using the features shown in Table 4.1 dependent and independent classification accuracies were determined for each of the images. These accuracies represent the ability with which the classifier can place the training data (that data used to develop the classifier) into the proper coverage categories and the ability with which the classifier properly categorizes a selected subset of known pixels, not used in training. The method used to report these data is in the form of a confusion matrix. This matrix represents the known categories as rows and the classifier assigned categories as columns. In an ideal classification this matrix would have values only along the major diagonal with zeros elsewhere. Any deviation from perfect classification will place values in these off-diagonal terms. Tables 4.3 through 4.6 contain the confusion matrices developed for the dependent data of the four images and Tables 4.7 through 4.10 contain those matrices developed for the independent data sets.

Table 4.1
List of the features chosen for each scene used in this study

RR1	RR2	ROME1	ROME2
Sum Variance Range	Contrast Average	Sum Variance Range	Contrast Range
Mean Brightness	Contrast Range	Mean Brightness	Sum Variance Range
Variance	Sum Variance Range	Variance	Mean Brightness
Red Spectral Band	Mean Brightness	Contrast	Brightness
Green Spectral Band	Green Spectral Band	Infrared Spectral Band	Red Spectral Band
Blue Spectral Band	Blue Spectral Band	Green Spectral Band	Green Spectral Band

Table 4.2
Compilation of the occurrence rate of selected features within the optimal set for the four images used in this effort - A first-order measure of robustness

Sum Variance Range	4 of 4	Contrast Range	2 of 4
Mean Brightness	4 of 4	Contrast Average	1 of 4
Green Spectral Band	4 of 4	Infrared Spectral Band	1 of 4
Variance	2 of 4	Brightness	1 of 4
Red Spectral Band	2 of 4	Contrast	1 of 4
Blue Spectral Band	2 of 4		

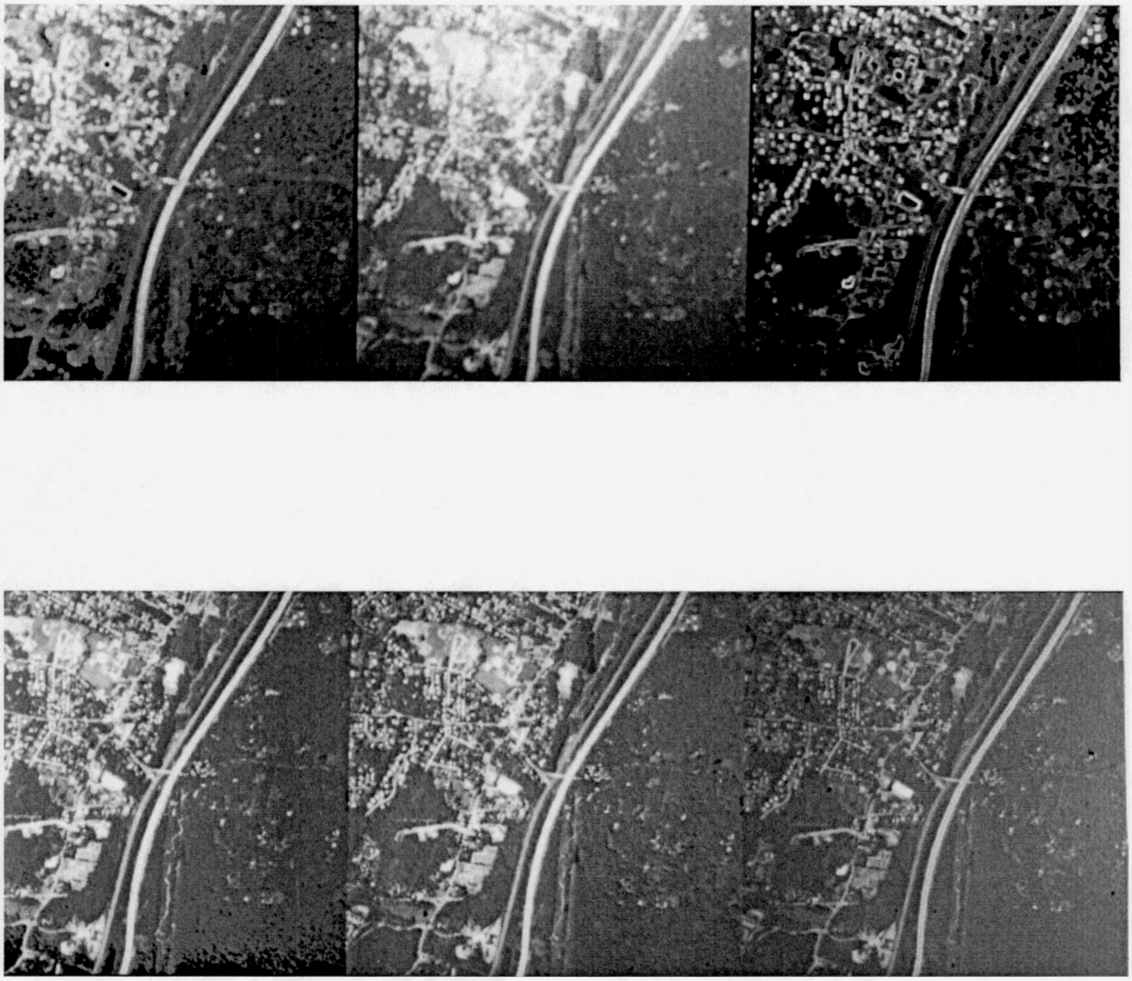


Figure 4.5 Chosen optimal monochrome-derived image features for air photo containing railroad system (RR1)

Table 4.3
Confusion matrix developed for the dependent data analyzed in image RR1

	Grass	Highway	Houses	Roads	Railways	Trees
Grass	72	0	0	0	2	0
Highway	0	213	0	0	0	0
Houses	0	0	151	10	0	0
Roads	0	0	4	162	0	0
Railways	0	0	0	2	98	0
Trees	0	0	0	0	0	155

Table 4.4
Confusion matrix developed for the dependent data analyzed in image RR2

	Grass	Houses	Roads	Railways	Trees
Grass	67	0	0	0	0
Houses	0	135	8	0	0
Roads	0	1	150	9	0
Railways	0	0	3	175	0
Trees	0	0	0	0	168

Table 4.5
Confusion matrix developed for the dependent data analyzed in image ROME1

	Grass	Trees	Road(1)	Road(2)	Camouflage
Grass	265	0	0	0	3
Trees	0	256	0	0	22
Road(1)	0	0	163	0	0
Road(2)	0	0	0	369	4
Camouflage	0	109	0	2	281

Table 4.6
Confusion matrix developed for the dependent data analyzed in image ROME2

	Grass	Trees	Road(1)	Road(2)	Camouflage
Grass	186	0	0	6	0
Trees	0	451	0	0	0
Road(1)	0	0	107	0	0
Road(2)	3	0	0	169	0
Camouflage	0	0	0	0	53

Table 4.7
Confusion matrix developed for the independent data analyzed in image RR1

	Grass	Highway	Houses	Roads	Railways	Trees
Grass	36	0	0	0	1	0
Highway	0	106	0	0	0	0
Houses	0	0	75	5	0	0
Roads	0	0	3	80	0	0
Railways	0	0	0	1	50	0
Trees	0	0	0	0	0	77

Table 4.8
Confusion matrix developed for the independent data analyzed in image RR2

	Grass	Houses	Roads	Railways	Trees
Grass	33	0	0	0	0
Houses	0	71	0	0	0
Roads	0	0	75	6	0
Railways	0	0	2	86	0
Trees	0	0	0	0	84

Table 4.9
Confusion matrix developed for the independent data analyzed in image ROME1

	Grass	Trees	Road(1)	Road(2)	Camouflage
Grass	133	0	0	0	0
Trees	0	127	0	0	12
Road(1)	0	0	82	0	0
Road(2)	0	0	0	184	2
Camouflage	0	52	0	2	144

Table 4.10
Confusion matrix developed for the independent data analyzed in image ROME2

	Grass	Trees	Road(1)	Road(2)	Camouflage
Grass	94	0	0	2	0
Trees	0	225	0	0	0
Road(1)	0	0	53	0	0
Road(2)	1	0	0	86	0
Camouflage	0	0	0	0	26

As can be seen from these matrices, both the dependent and independent classification accuracies are very high for most classes. Table 4.11 represents overall classification accuracies across all classes for each of the images. These values are straight averages of the individual class accuracies with no attention given to relative number of pixels trained for each class.

Table 4.11
Overall classification accuracy across class for images analyzed

Image	Dependent	Independent
RR1	97.9 %	97.7 %
RR2	97.1 %	97.8 %
ROME1	90.5 %	91.0 %
ROME2	99.1 %	99.4 %

4.3 Image Classification

The railroad system image, RR1, and it's accompanying optimal feature set was used as an input to a maximum likelihood classifier to produce a land cover map. *A priori* probabilities for this classification were set equal since no information of this sort was known. Therefore classification was carried out according to Equation 10. Figure 4.6 shows the color-coded land cover map produced.

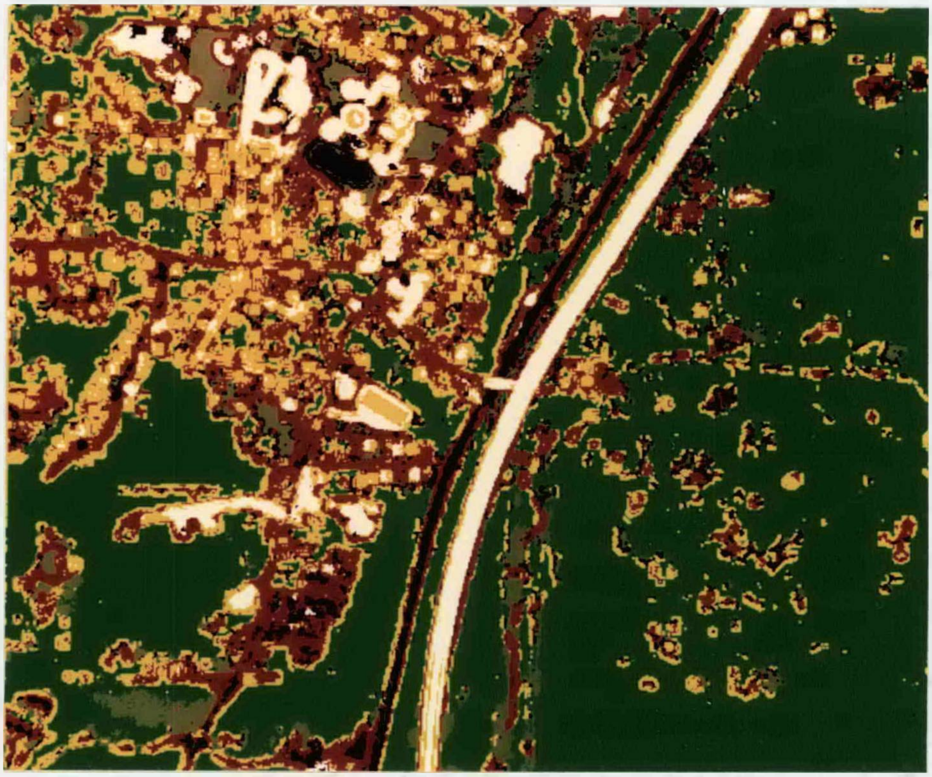


Figure 4.6 Land cover classification map produced from maximum likelihood classification (red-minor roads, green-trees, olive-grass, white-highway, black-railways, yellow-houses)

Visually, this map appears to categorize the original scene content. A more rigorous evaluation of the effectiveness of the classification was carried out according to the random pixel selection method described in Section 3.5.

4.4 Classification Accuracies from Random Point Analysis

Using the method of random point analysis the land cover map shown above was compared to "ground truth" obtained from air photo data. Fifty randomly selected points were chosen from each class shown in Figure 4.6 and presented to the analyst. The confusion matrix shown below (Table 4.12) illustrates the accuracy with which the 50 randomly selected scene elements in each class were categorized. The overall accuracy obtained from this analysis is significantly lower than that obtained using the independent classification accuracies of Table 4.11. This is expected since the "independent" data used to arrive at the value in Table 4.11 was closely associated in proximity with the training data. The data in Table 4.12 better describes the actual classification accuracy of the scene since no proximity ties with the training data are implied. The second overall accuracy figure shown is a weighted value incorporating percentages of the entire scene classified into each class and will provide a better feel for the overall image classification accuracy.

Table 4.12
Confusion matrix developed for the random point analysis of image RR1

	Grass	Highway	Houses	Roads	Railways	Trees	# of pixels in final map
Grass	29	0	1	2	3	15	17230
Highway	5	35	5	4	0	1	12148
Houses	2	3	21	6	0	18	37706
Roads	10	5	6	15	0	14	41065
Railways	16	5	6	9	6	8	22396
Trees	1	0	2	0	0	47	131599

Overall Accuracy = 51%
Overall Accuracy Weighted by #'s in Final Map = 66%

5.0 CONCLUSIONS AND RECOMMENDATIONS

This effort has demonstrated how large families of image-derived textural features can be reduced to a small number of useful images needed to perform scene segmentation. The emphasis on this effort was not on overall classification accuracy but rather on the development and testing of tools for the selection of the most appropriate features for scene segmentation. This effort demonstrated that a small set of image derived features could be selected from a candidate set of nearly 50 and used to achieve high classification accuracy on independent data sets (approximately 96%). This reduction in the number of features required is a significant since feature generation can be very compute intensive therefore shortening run times.

The approach pursued here required user-assisted training procedures to facilitate the selection of classes and the isolation of appropriate features for performing scene segmentation. However, one of the objectives of this effort was to determine whether any of the image-derived features were robust enough to be pre-identified as useful in scene segmentation. If this were the case then these features might be useful in the development of unsupervised or automated scene segmentation algorithms. While a rigorous treatment was beyond the scope of this study, the results for the four scenes studied are very encouraging. It was observed that three features proved optimal on all four images used. These results must be interpreted as preliminary due to the small data set , however, they suggest that at least within some confines of content and scale a robust family of image-derived textural features may be identifiable for use in scene segmentation.

Future efforts should consider an expanded set of image-derived features including the effects of varying radiometric resolution (*e.g.* number of grey levels in cooccurrence feature calculations) and kernel size. Also the effects of image type (scale, orientation and content) should be more rigorously evaluated. A particularly promising use of the tools developed here would be in evaluation of the trade-offs between spectral and spatial resolution. The need for multiple spectral channels when high resolution texture data are available from a monochrome images could be evaluated.

In summary, an image processing tool has been developed to facilitate the selection of an optimized set of image-derived features for scene segmentation. Efforts should be made to attempt to identify improved features so that better classifiers can be built. The tool can be effectively used to determine whether a new feature or family of features is of value for the image types of interest. We believe that this approach can be a powerful tool in developing improved image classification procedures for both supervised and unsupervised classification.

6.0 REFERENCES

Conners, R., Towards a set of statistical features which measure visually perceivable qualities of texture, *CH1428-2/79/0000-0382, IEEE*, 1979, pp. 382-389.

Duda, R.O. and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

Fu, K.S., Pattern recognition in remote sensing of the Earth's resources, *IEEE Transactions on Geoscience Electronics*, GE-14, 1, January 1976, pp. 10-18.

Galloway, M.M., Texture classification using grey-level run lengths, *Computer Graphics and Image Processing*, 4, June 1975, pp. 172-179.

Haralick, R.M., A texture-context feature extraction algorithm for remotely sensed imagery, *Proceedings of the IEEE Decision and Control Conference (1971)*, Gainesville, Florida, December 15-17, 1971, pp. 650-657.

Haralick, R.M., K. Shanmugam and I. Dinstein, Textural features for image classification, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3, 6, November 1973, pp. 610-621.

Haralick, R.M., Statistical and structural approaches to texture, *Proceedings of the IEEE*, 67, 5, May 1979, pp. 786-804.

Hsu, S., Texture-tone analysis for automated land use mapping, *Photogrammetric Engineering and Remote Sensing*, 44, 1978, pp. 1393-1404.

Keltig, R.L. and D.A. Landgrebe, Classification of multispectral image data by extraction and classification of homogeneous objects (ECHO), *IEEE Transactions on Geoscience Electronics*, GE-4, 1, January 1976, pp. 19-26.

Queriros, C.E. and E.S. Gelsema, On feature selection, IEEE International Conference on Pattern Recognition, 1, 1984, pp. 128-130.

Richards, J.A., *Remote Sensing Digital Image Analysis, An Introduction*, Springer-Verlag, New York, 1986.

Robert, D.J., Selection and analysis of optimal textural features for accurate classification of monochrome digitized image data, MS Thesis, Rochester Institute of Technology, Rochester, New York, 1989.

Rosenfeld, A. and M. Thurston, Edge and curve detection for visual scene analysis, *IEEE Transactions on Computers*, C-20, 1971, pp. 562-569.

Rosenfeld, A., A note on automatic detection of texture gradients, *IEEE Transactions on Computers*, C-24, 11, October 1975, pp. 988-991.

Salvaggio, C., D.J. Robert and J.R. Schott, Generation of textural features from monochromatic imagery for land cover classification, Rochester Institute of Technology, RIT/DIRS Report #89/90-63-130, January 1990.

Schott, J.R., E. Kraus and C. Salvaggio, Optimum spectral band selection, Rochester Institute of Technology, RIT/DIRS Report #88/89-54-117, July 1988.

Sutton, R. and E. Hall, Texture measures for Automatic classification of pulmonary disease, *IEEE Transactions on Computers*, C-21, 7, July 1972, pp. 667-676.

Weszka, J.S., C.R. Dyer and A. Rosenfeld, A comparative study of texture measures for terrain classification, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-6, 4, April 1976, pp. 269-285.

APPENDIX A Mathematical Description of Textural Features Used in this Study

SPECTRAL FEATURES

These features consist of the grey levels of individual pixels from the different bands of the digital image.

TEXTURAL FEATURES

These features are measures of the interaction between neighboring pixels in a single band. They can be calculated according to several different methods as explained below.

COOCCURRENCE MATRIX FEATURES

Fourteen textural features are defined below as calculated from gray-level cooccurrence matrices. Specifications that go along with this feature are the distance between the two pixels compared, the orientation between the comparison and the size of the window (which decides how many pixels will be compared). Because these features are dependent on the angle over which they are calculated, the actual features values calculated will be the average over all four angles (0, 45, 90, 135 degrees) and the range over all four angles. Therefore, 28 out of the final 46 textural features are calculated from gray-level cooccurrence matrices.

The notation used to describe the calculation of these features is as follows.

Ng is the number of gray levels in the quantized image.

R is the number of gray levels after quantization (also the dimension of the cooccurrence matrix)

p(i,j) is the (i,j) the entry in a quantized gray-tone spatial dependence, matrix, it is equal to P(i,j)/R.

$p_x(i)$ is the ith entry in the marginal-probability matrix which is obtained by summing the rows of p(i,j) where

$$p(i,j) = \sum_{j=1}^{Ng} P(i,j)$$

$$p_y(j) = \sum_{i=1}^{Ng} p(i,j)$$

$$p_{x+y}(k) = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) \qquad \text{for } k = 2,3,\ldots,2Ng \quad \text{and } i+j = k$$

$$p_{x-y}(k) = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) \quad \text{for } k = 0,1,\ldots,Ng-1 \quad \text{and } |i-j| = k$$

1) Angular Second Moment

$$\mathbf{f1} = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} [p(i,j)]^2$$

2) Contrast

$$\mathbf{f2} = \sum_{n=0}^{Ng-1} n^2 \left[\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) \right] \quad \text{for } |i-j|=n$$

3) Correlation

$$\mathbf{f3} = \frac{\left[\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (i,j)\ p(i,j) - \mu_x \mu_y \right]}{\sigma_x \sigma_y}$$

4) Variance

$$\mathbf{f4} = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (i-\mu)^2\ p(i,j)$$

5) Inverse Difference Moment

$$\mathbf{f5} = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} \left[\frac{p(i,j)}{1+(i-j)^2} \right]$$

6) Sum Average

$$\mathbf{f6} = \sum_{i=2}^{2Ng} i\ p_{x+y}(i)$$

7) Sum Variance

$$\mathbf{f7} = \sum_{i=2}^{2Ng} (i-\mathbf{f8})^2\ p_{x+y}(i)$$

8) Sum Entropy

$$\mathbf{f8} = -\sum_{i=2}^{2Ng} p_{x+y}(i)\ \log[p_{x+y}(i)]$$

9) Entropy

$$\mathbf{f9} = -\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j)\ \log[p(i,j)]$$

10) Difference Entropy

$$\mathbf{f10} = -\sum_{i=0}^{Ng-1} p_{x-y}(i)\ \log[p_{x-y}(i)]$$

11), 12) Information Measures Of Correlation

$$\mathbf{f11} = \frac{HXY - HXY1}{\max(HX, HY)}$$

$$\mathbf{f12} = \left[1 - e^{-2(HXY2 - HXY)}\right]^{1/2}$$

$$\text{where:}\quad HXY = -\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j)\ \log[p(i,j)]$$

$$HXY1 = -\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j)\ \log[px(i)\ py(j)]$$

$$HXY2 = -\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} px(i)\ py(j)\ \log[px(i)\ py(j)]$$

13) Difference Variance

f13 = variance of px-y.

14) Maximal Correlation Coefficient

$$\mathbf{f14} = [\text{Second largest eigenvalue of } Q]^{1/2}$$

$$Q(i,j) = \sum_{k} \frac{p(i,k)\ p(j,k)}{p_x(i)\ p_y(k)}$$

FIRST-ORDER STATISTICS FEATURES

These are basic features which are simple statistical measures on groups of pixels.

Gradient is a measure of the edgeness in a window defined as

$$G(d) = \sum_{i,j=N} \Big[\left|I(i,j)-I(i+d,j)\right| + \left|I(i,j)-I(i-d,j)\right| + \left|I(i,j)-I(i,j+d)\right| + \left|I(i,j)-I(i,j-d)\right| \Big]$$

d = the distance between pixels for the sample
I(i,j) = point i,j in the image window I
N = dimension of the window

Mean Brightness is the mean gray value over a window of pixels the same size as was used for the cooccurrence calculations.

Variance is the variance of the gray values within the window.

Brightness is simply the gray value of each pixel, or the original monochrome image

RUN LENGTH STATISTICS FEATURES

Given a block of pixels (the same size as the windows over which the cooccurrence features were calculated), run length features are based on the lengths and orientations of groups of linearly connected pixels of identical gray level. Let p(i,j) be be the number of runs of length j and gray level i. A matrix can then be made with i rows and j columns, with its entries being the value of p(i,j) for orientations of 0°, 45°, 90°and 135°.

N_r = the number of runs
N_g = the number of gray levels
P = the number of points in the window

The features are as follows

$$\text{Short Runs Emphasis} = \frac{\displaystyle\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} \frac{p(i,j)}{j^2}}{\displaystyle\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)}$$

This divides the number of runs by the length of the run squared and tends to emphasize short runs. The denominator is the total number of runs and acts as a normalizing factor.

$$\text{Long Runs Emphasis} = \frac{\displaystyle\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} j^2\, p(i,j)}{\displaystyle\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)}$$

This multiplies the number of runs by the length of the run squared, emphasizing long runs.

$$\text{Gray Level Nonuniformity} = \frac{\displaystyle\sum_{i=1}^{Ng}\left[\sum_{j=1}^{Nr} p(i,j)\right]^2}{\displaystyle\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)}$$

This squares the number of run lengths for each gray level. When runs are equally distributed through gray levels, the function has a low value.

$$\text{Run Length Nonuniformity} = \frac{\displaystyle\sum_{j=1}^{Ng}\left[\sum_{i=1}^{Nr} p(i,j)\right]^2}{\displaystyle\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)}$$

This squares the number of runs for each length. If runs are equally distributed in length, the function takes on its lowest value.

$$\text{Run Percentage} = \frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)}{P}$$

This ratios the total number of runs to the total number pixels in the window. The function has its lowest value for a window with highly linear structure.

GRAY-LEVEL DIFFERENCE STATISTICS FEATURES

Another approach to defining features is to use matrices with entries based on pairs of gray levels taken d distance apart. The absolute value of the difference between any of these two pixels a distance d apart is computed as

$$f_\delta(x,y) = \left| f(x,y) - f(x+\Delta x,\ y+\Delta y) \right|$$

The probability $p_d(i)$ is the probability density of $f_d(x,y)$ where i is the range of values possible for $f_d(x,y)$the number of gray-levels - 1. Based upon these calculations, the four following features are defined.

$$\text{Contrast} = \sum_{i}^{Ng-1} i^2\, p_\delta(i)$$

$$\text{Angular Second Moment} = \sum_{i}^{Ng-1} p_\delta(i)^2$$

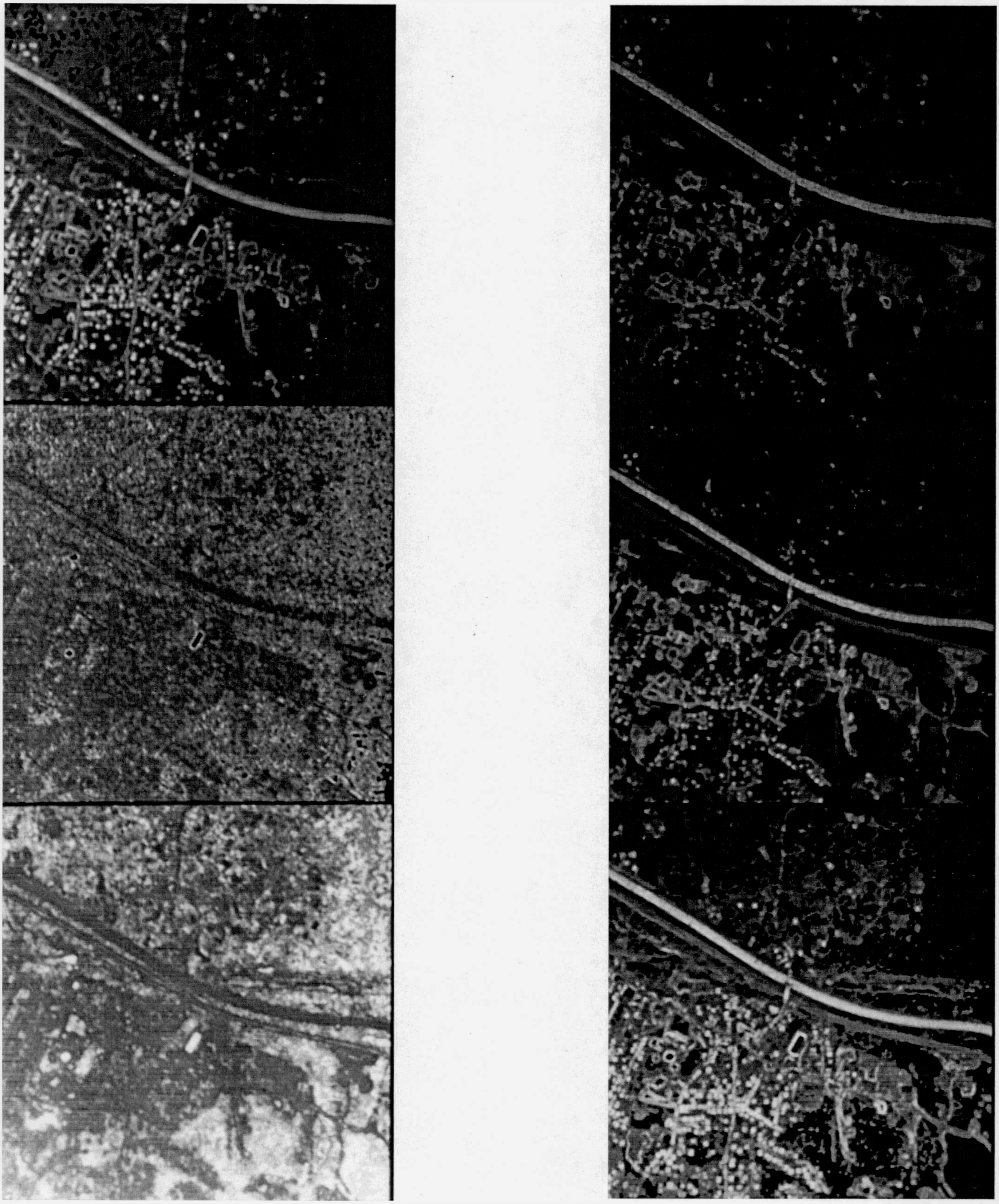
$$\text{Entropy} = -\sum_{i}^{Ng-1} p_\delta(i)\, \log[p_\delta(i)]$$

$$\text{Mean} = \frac{1}{N_g}\left[\sum_{i}^{Ng-1} i\, p_\delta(i)\right]$$

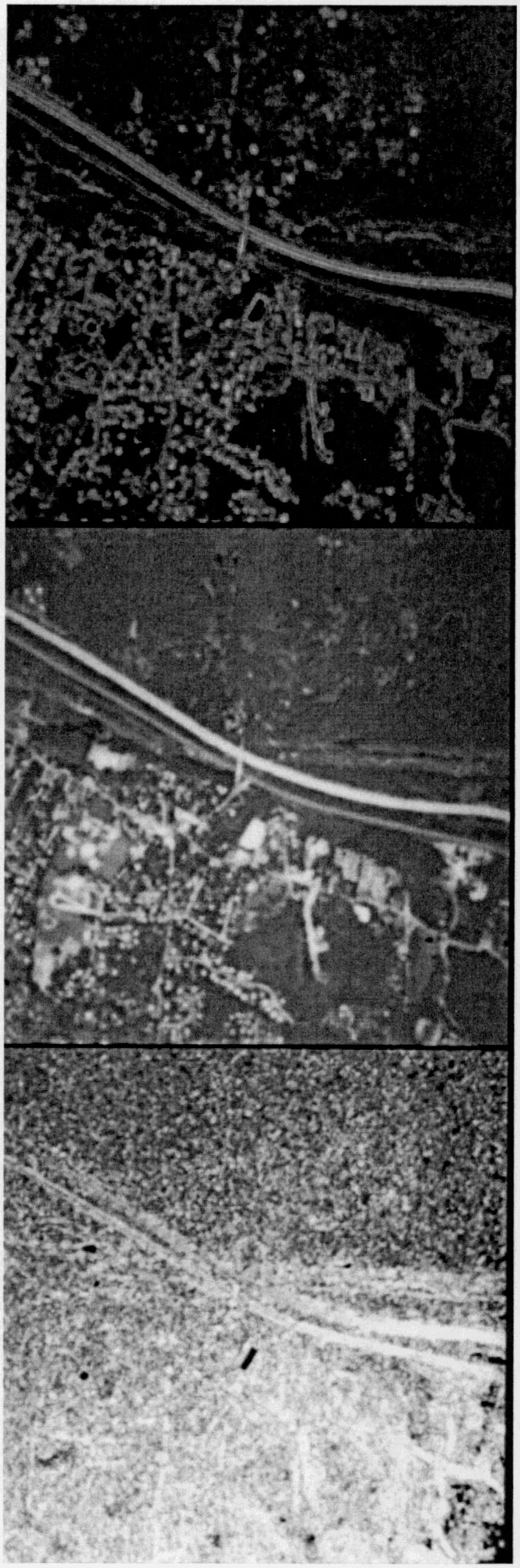
APPENDIX B Pictorial Description of Textural Features Used in this Study (RR1)

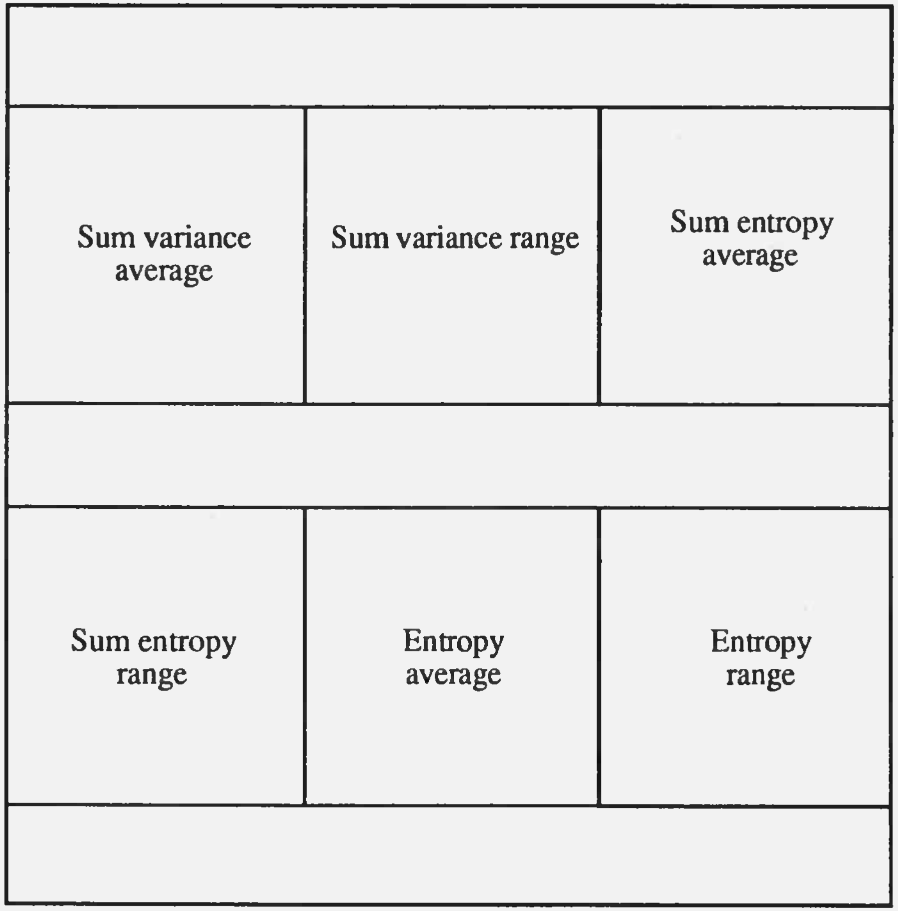
The following pages contain full resolution (512x512 pixels) images of the textural features derived on image RR1 shown in Figure 4.1. The cover sheets contain a layout sketch indicating the position of the particular features on the following page. Consult Appendix A for a mathematical description of these features.

Angular second moment average	Angular second moment range	Contrast average
Contrast range	Correlation average	Correlation range

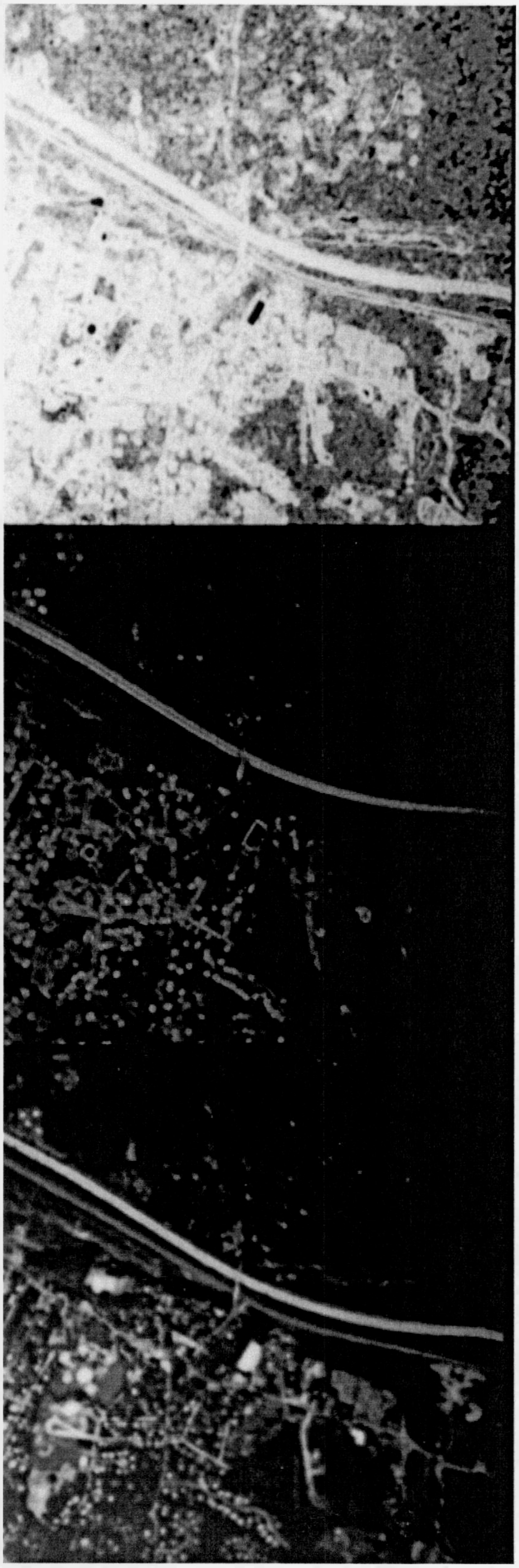


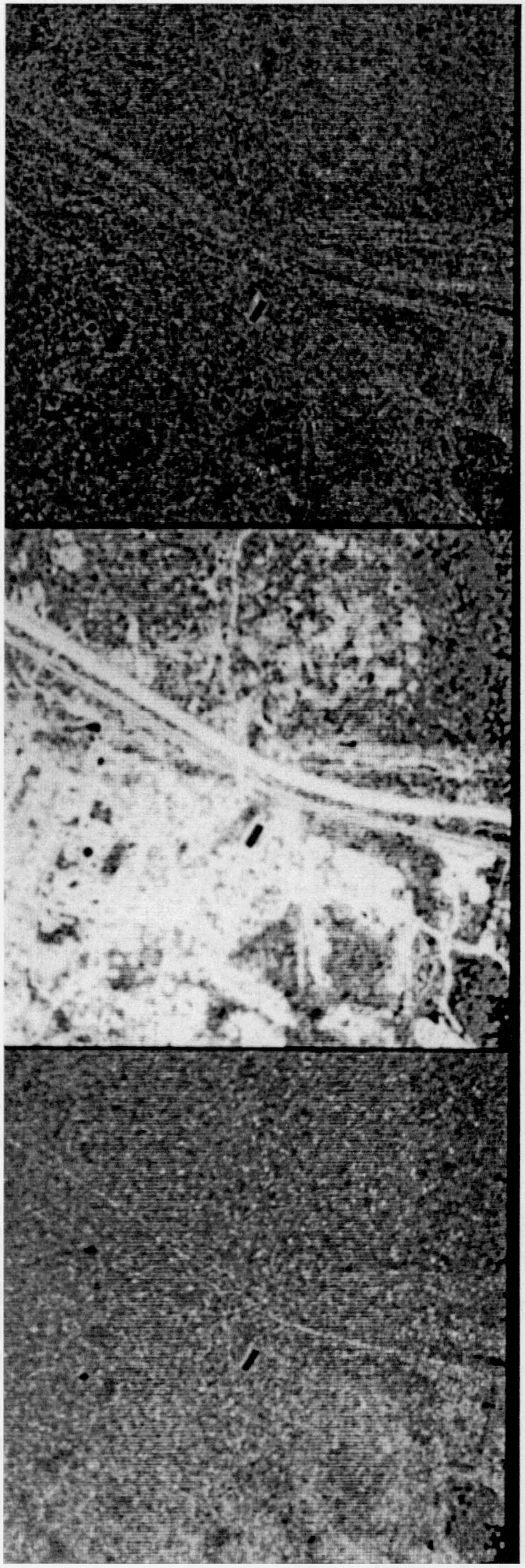
Variance average	Variance range	Inverse difference moments average
Inverse difference moments range	Sum average average	Sum average range



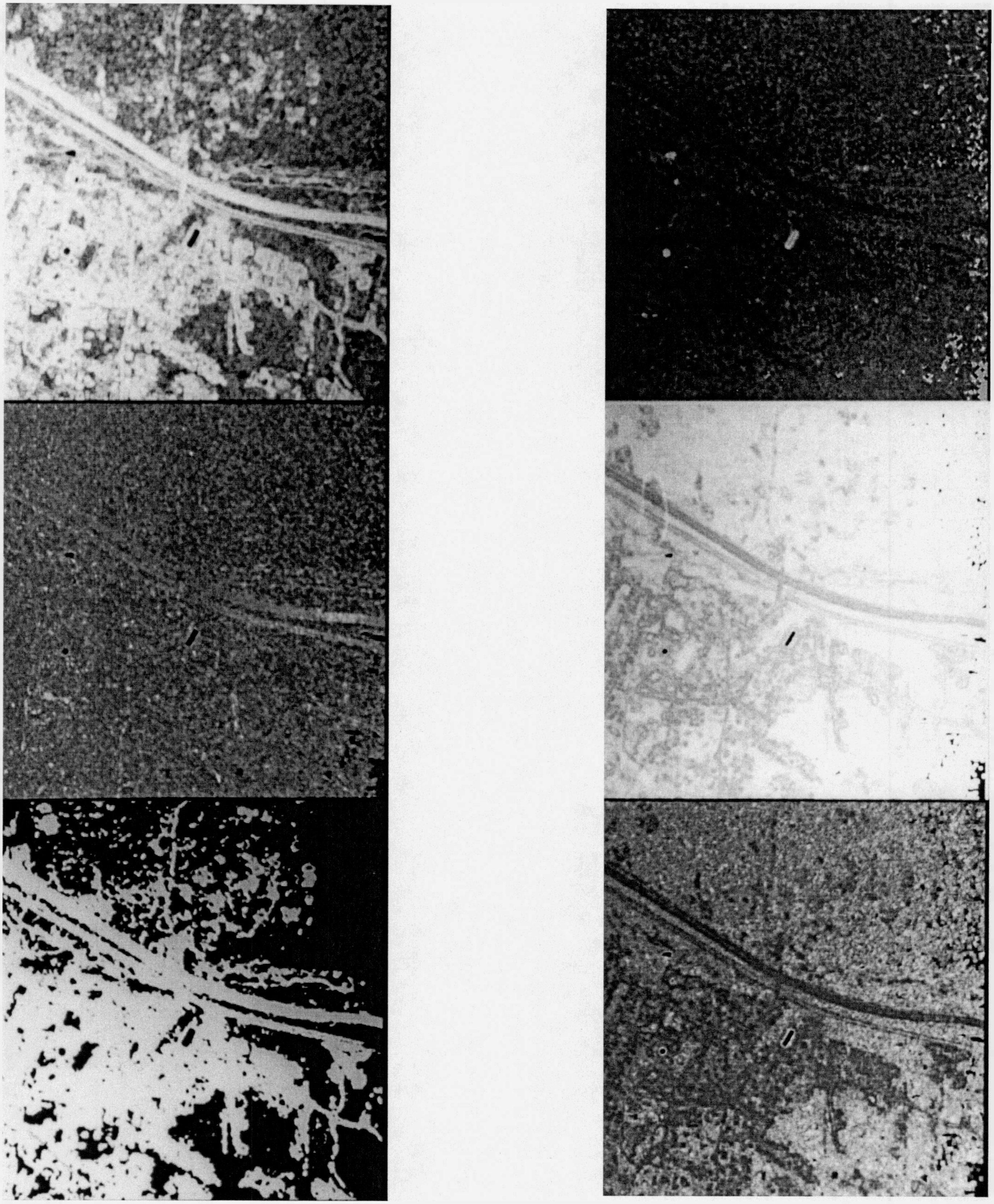


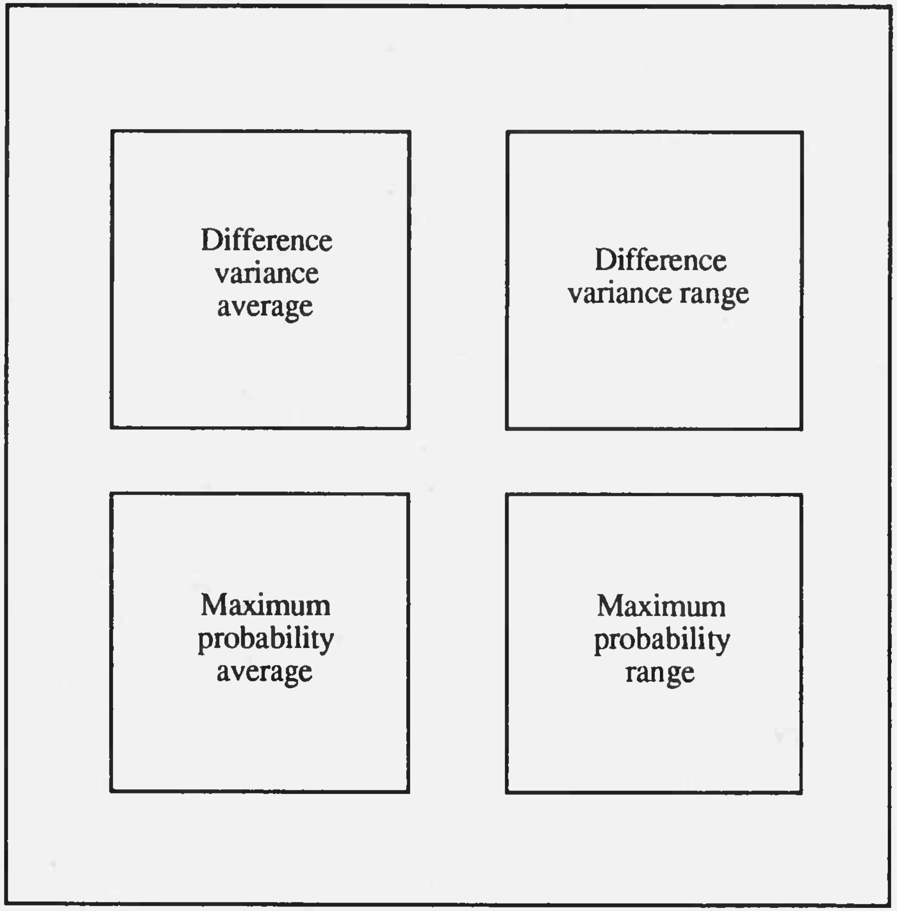
Sum variance average	Sum variance range	Sum entropy average
Sum entropy range	Entropy average	Entropy range





Difference entropy average	Difference entropy range	Information measure of correlation A average
Information measure of correlation A range	Information measure of correlation B average	Information measure of correlation B range



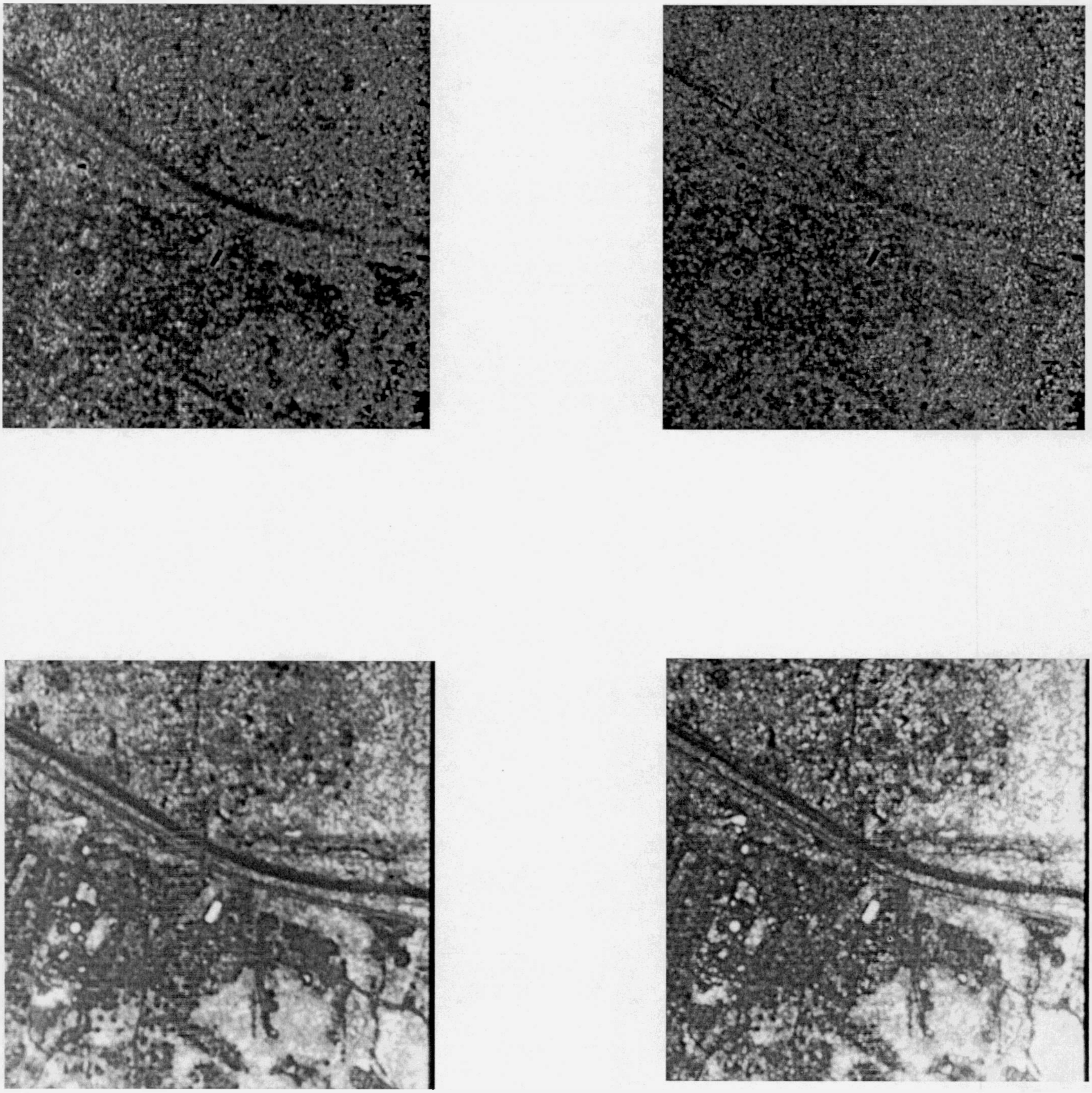


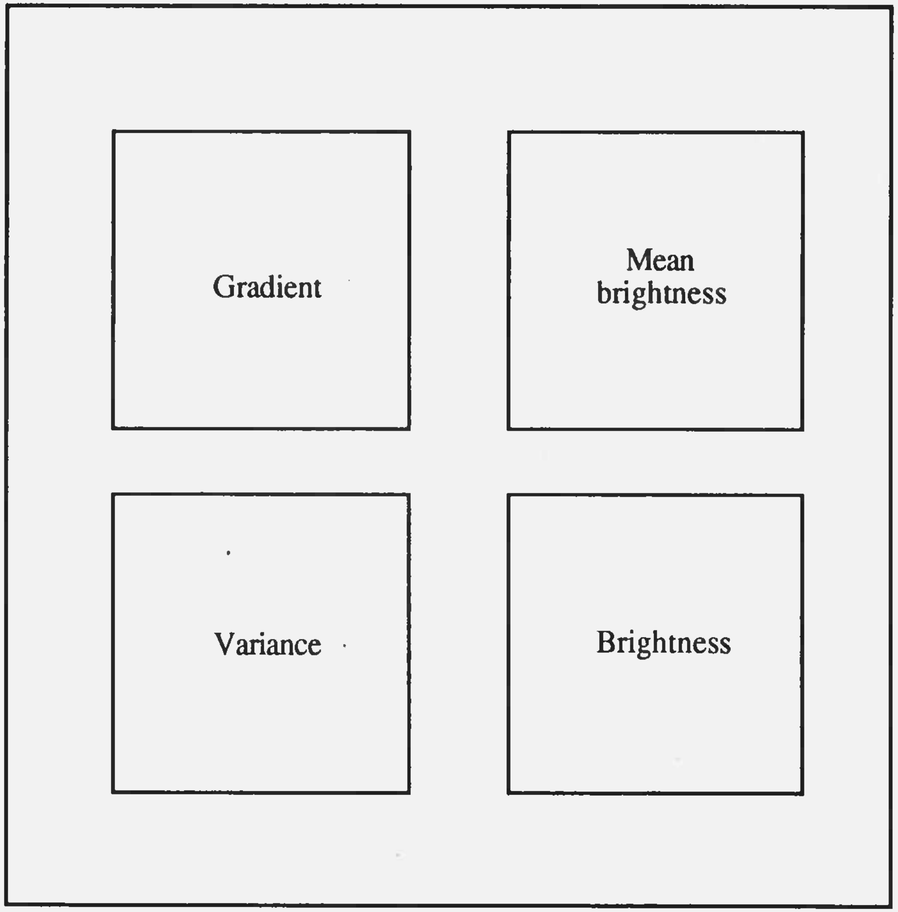
Difference variance average

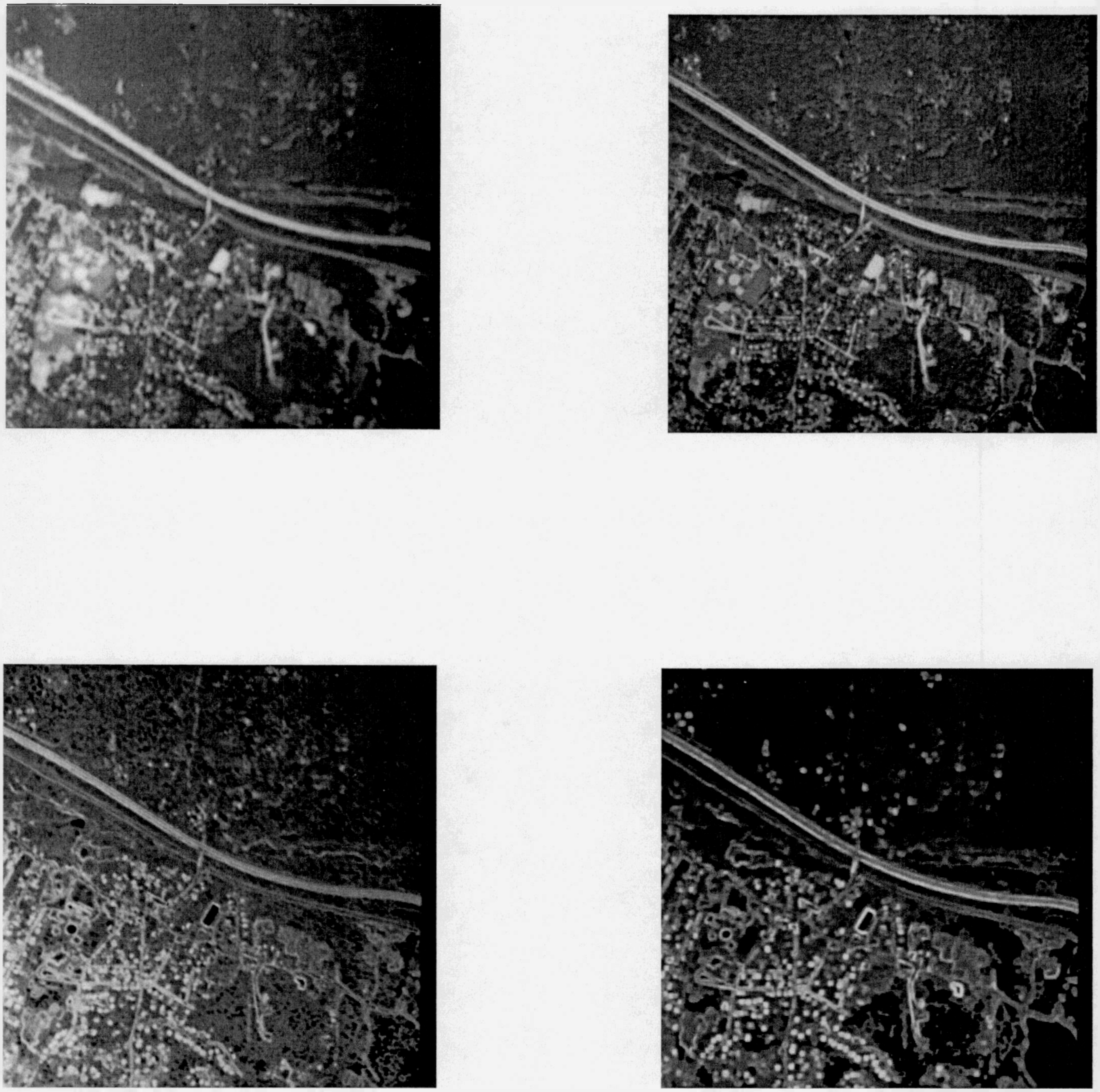
Difference variance range

Maximum probability average

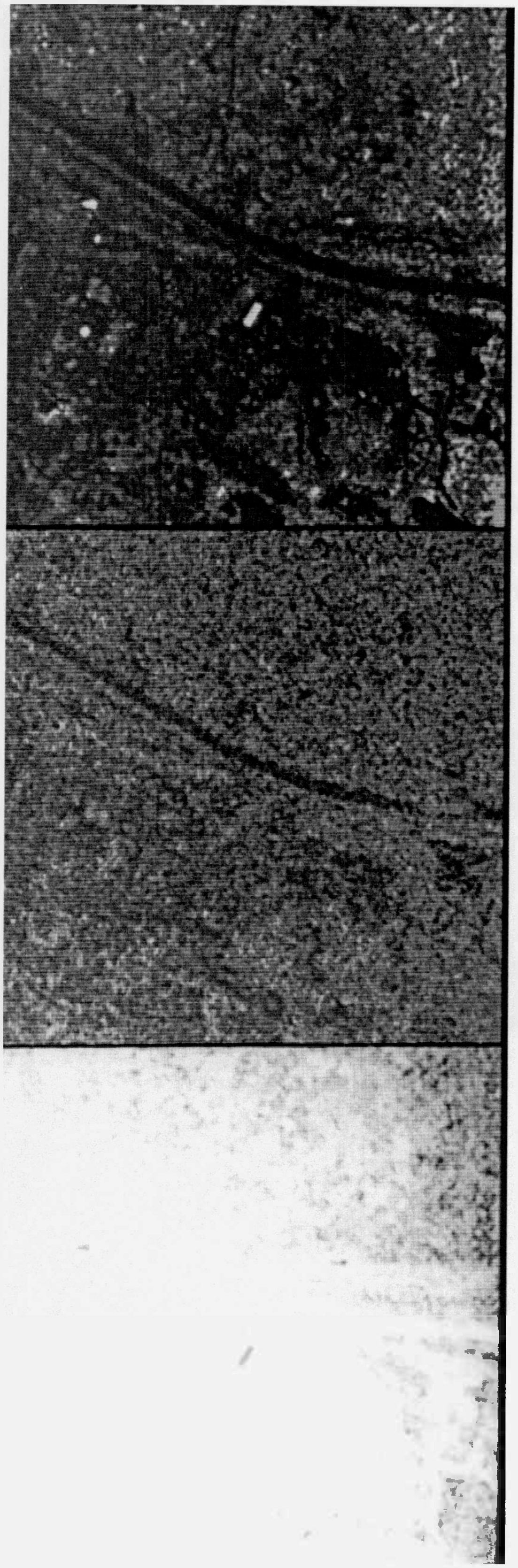
Maximum probability range

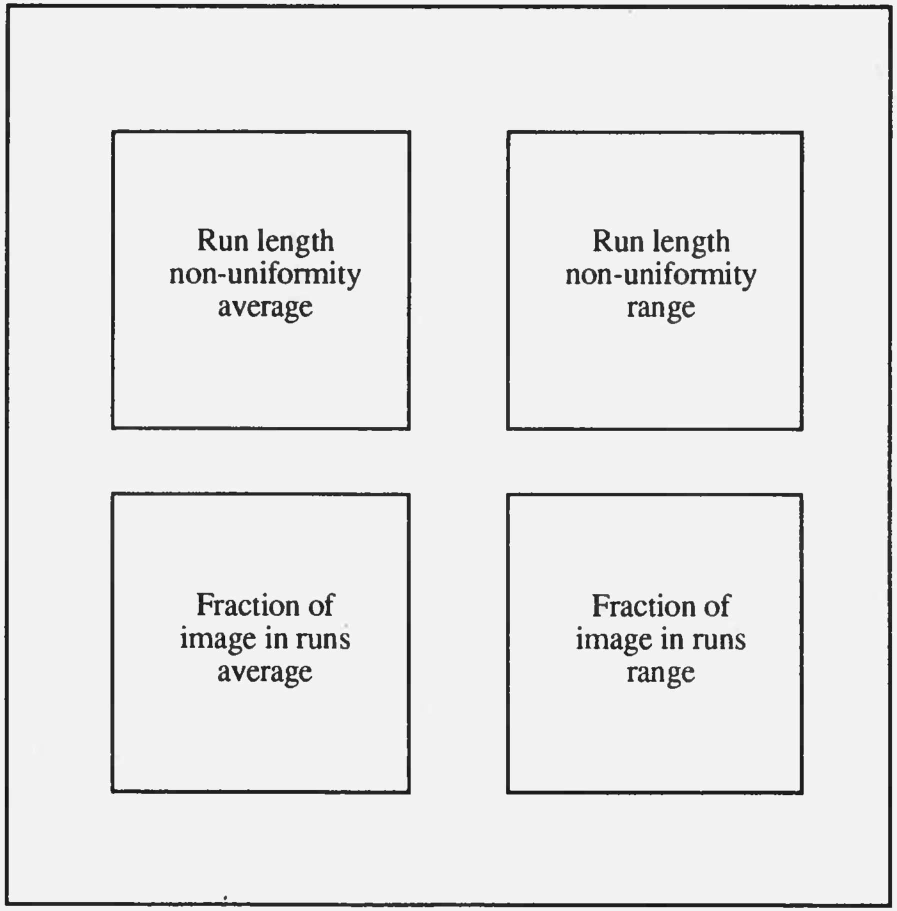



Gradient
Mean brightness
Variance
Brightness



Short run emphasis inverse moment average	Short run emphasis inverse moment range	Long run emphasis inverse moment average
Long run emphasis inverse moment range	Gray level non-uniformity average	Gray level non-uniformity range



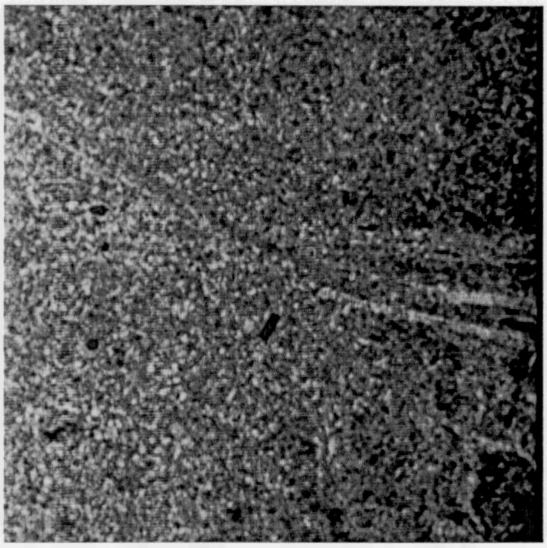


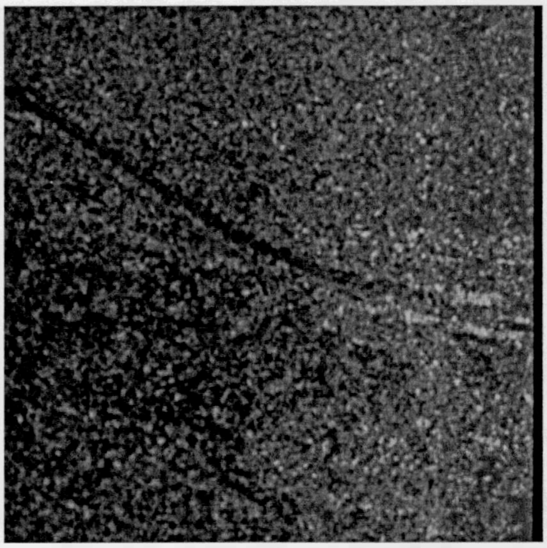
Run length non-uniformity average

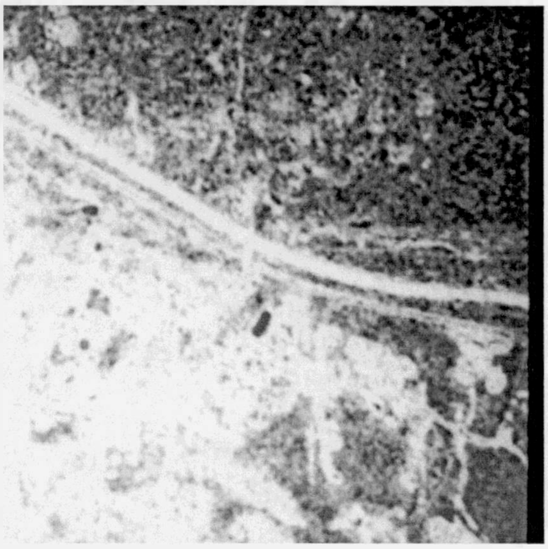
Run length non-uniformity range

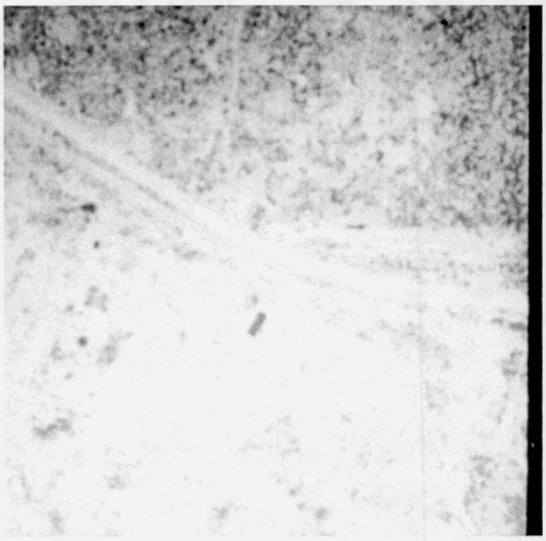
Fraction of image in runs average

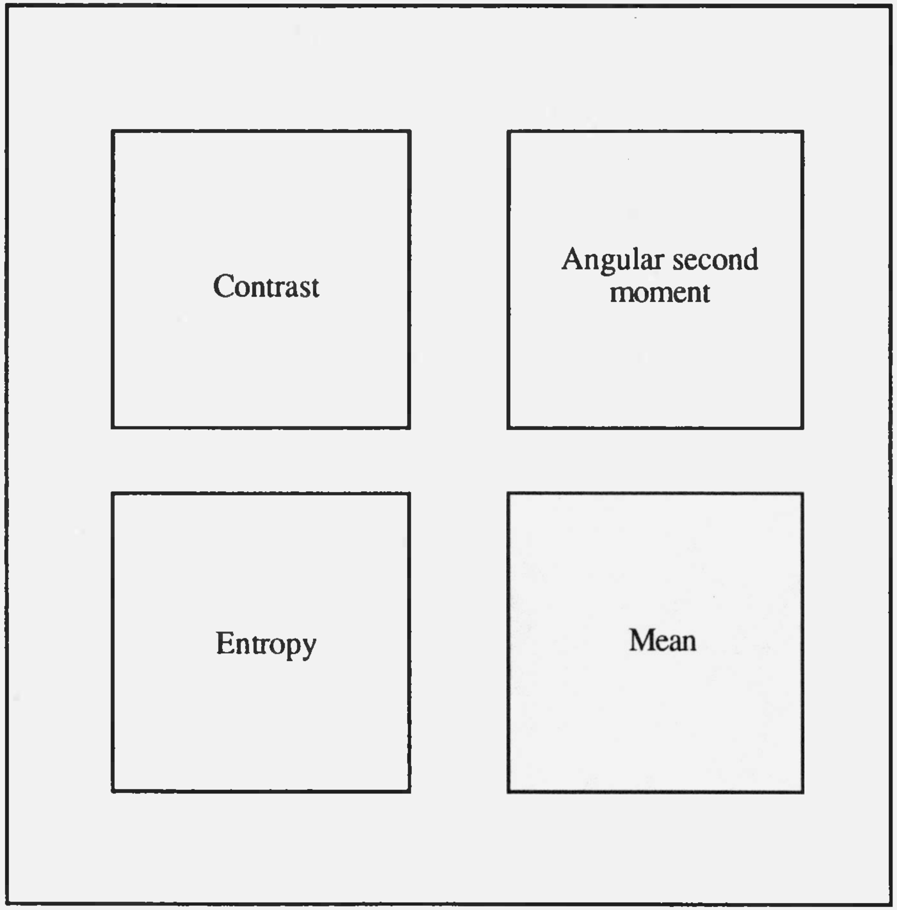
Fraction of image in runs range










Contrast
Angular second moment
Entropy
Mean