Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Image Processing for Remote Sensing

J. Kittler

Phil. Trans. R. Soc. Lond. A 1983 **309**, 323-335
doi: 10.1098/rsta.1983.0045

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. Lond. A* go to: **http://rsta.royalsocietypublishing.org/subscriptions**

This journal is © 1983 The Royal Society

Downloaded from rsta.royalsocietypublishing.org

Phil. Trans. R. Soc. Lond. A **309**, 323–335 (1983)
Printed in Great Britain

Image processing for remote sensing

By J. Kittler
S.E.R.C. Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, U.K.

In this paper a number of approaches to multispectral image segmentation and classification are considered. The methods range from the simple Bayesian decision rule for classification of image data on pixel-by-pixel basis, to sophisticated algorithms using contextual information. Both the spatial pixel category dependencies and the two-dimensional correlation-type contextual information have been incorporated in decision-making schemes. The aim of these algorithms is to achieve a greater reliability in the process of interpretation of remote-sensing data.

1. Introduction

Remote sensing is concerned with the problem of detecting the nature of an object and of its monitoring by means of external non-contact observations. In the study of Earth resources the observations are made by multispectral cameras and other passive and active sensors (infrared scanner, scanning radiometer, gamma-ray spectrometer, radar scatterometer, radar imager). The acquired multispectral–multisensor image data are then analysed by electronic means.

Image processing of remotely sensed data involves image correction, image preprocessing, image segmentation and, finally, image interpretation. These processing stages are depicted in figure 1.

The role of the first stage is to correct the recorded data for geometric distortions, the effects of the propagation medium, atmospheric variations, etc. In general, image data to be processed will be multichannel. If the data in individual channels are acquired by using different sensors, or at different instants, it is essential to massage it so that all picture elements (pixels) representing a particular point on the ground are brought into exact correspondence. This process of image alignment involves congruencing, rectification or simple image registration depending on the configuration of the sensing system. For detailed discussion of these topics the reader is referred to Bernstein (1978) and Haralick (1976).

Image preprocessing techniques have been developed to minimize variability in recorded data due to noise and various artefacts such as viewing-angle variations. A detailed treatment of digital filtering of two-dimensional data can be found in Huang (1979). Grey tone normalization for atmospheric, viewing-angle and intensity variations is reviewed in Haralick (1976). Image compression and coding for storage and transmission purposes also constitute an integral part of the preprocessing stage. Pratt (1978) provides a broad coverage of various approaches to this problem.

The ultimate goal of remote-sensing data processing is the interpretation of image segments that exhibit similar statistical properties. The lowest level of interpretation is the segment classification obtained at the output of the third stage of the model. However, this form of output rarely suffices because we are usually interested in extracting additional information from the data, such as the shape and size of objects that these segments represent and their spatial relation with other objects. Methodology for these higher levels of interpretation can be found in Fu (1977, 1982*a*, *b*), Pavlidis (1977), Devijver & Kittler (1982) and Kittler (1983). Any ambiguity

The Royal Society is collaborating with JSTOR to digitize, preserve, and extend access to
Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences.
www.jstor.org

Downloaded from rsta.royalsocietypublishing.org

in interpretation usually has to be resolved by using ancillary data such as topographic, soil-type and geopolitical maps, general knowledge about cultivation practices, rainfall conditions, seasonal characteristics – briefly, world models.

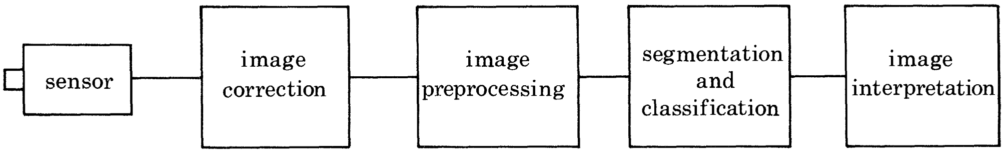


FIGURE 1. Image-processing system.

This paper concentrates on the problem of image segmentation and classification. In § 2 the problem of classification of multispectral pixel data will be formulated and a simple solution based on the Bayes classification rule introduced. Section 3 deals with object classification by using spectral and textural features. The effect of one-dimensional spatial correlation between neighbouring pixels on the classification performance is studied in § 4. Classification algorithms exploiting two-dimensional spatial dependences of pixel data are the subject of § 5. Finally, § 6 discusses probabilistic relaxation labelling algorithms that incorporate contextual information.

2. Pixel-by-pixel classification

The first step in the analysis of remotely sensed data is to identify homogeneous segments in the image. Each homogeneous segment is then associated with one of the possible classes $\omega_i (i = 1, 2, \ldots, m)$ of land cover. This image segmentation and segment association process can be carried out in one of the following ways. Either we can first classify all the data on a pixel-by-pixel basis and then link identically labelled pixels to form connected segments. These segments are then associated with the class corresponding to the pixel label. Alternatively, we can first detect segments of pixels exhibiting similar properties. These homogeneous segments are then classified on a segment basis to the appropriate categories. Because of this duality, I shall consider the image segmentation and classification problems together.

I shall now describe a model for the multispectral remotely sensed data for analysis. For each point (or site) in the image we have a multivariate observation

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^{\mathrm{T}}, \tag{1}$$

where x_i is the measurement obtained by the ith spectral channel of the sensor. We shall assume that the data generation process can be modelled statistically. In other words $\boldsymbol{x}$ is considered to be a random variable having a unique conditional probability distribution for each class ω_i. It will be convenient to characterize the probability model for $\boldsymbol{x}$ in terms of the set of class conditional probability density functions $p(\boldsymbol{x}|\omega_i)$ and the *a priori* class probabilities $P(\omega_i), \forall i$.

The above probabilities constitute all the essential ingredients for making decisions about class membership of patterns $\boldsymbol{x}$, provided that it is appropriate to weigh all classification errors equally. The corresponding optimal decision rule, known as the Bayes minimum error rule, states (Devijver & Kittler 1982)

$$\text{assign} \quad \boldsymbol{x} \quad \text{to} \quad \omega_i \quad \text{if} \quad P(\omega_i|\boldsymbol{x}) = \max_j P(\omega_j|\boldsymbol{x}). \tag{2}$$

Thus pattern $\boldsymbol{x}$ is assigned to class ω_i if the *a posteriori* probability $P(\omega_i|\boldsymbol{x})$ of class ω_i given $\boldsymbol{x}$ is

Downloaded from rsta.royalsocietypublishing.org

greater than the *a posteriori* probability of any other class. The functions $P(\omega_j|\boldsymbol{x})$ can be computed by using the Bayes formula relating the conditional probabilities as

$$P(\omega_j|\boldsymbol{x}) = \{P(\omega_j)\,p(\boldsymbol{x}|\omega_j)\}/p(\boldsymbol{x}), \tag{3}$$

where $p(\boldsymbol{x})$ is the mixture density given as

$$p(\boldsymbol{x}) = \sum_{j=1}^{m} P(\omega_j)\,p(\boldsymbol{x}|\omega_j). \tag{4}$$

The Bayes decision rule (2) has been in existence in one form or another for more than two centuries. The recent and current research in statistical pattern recognition relating to this topic is concerned with the problems of its implementation. The problems lie first of all in the fact that we do not know the probability functions $p(\boldsymbol{x}|\omega_j)$. The only information assumed to be available for the recognition system design is a set of training patterns of which the class membership is known, that is ground truth data.

Second, the enormous amount of data that must be processed in one image alone has debarred the use of computationally involved classification schemes. Very simple decision-making algorithms are frequently found in commercially available systems. A typical example is the slicer classifier where a pixel is assigned to class ω_i if each component of vector $\boldsymbol{x}$ falls within a specific interval corresponding to this class.

Alternatively we can assume that the probability distribution of vector $\boldsymbol{x}$ has a parametric form. Under such an assumption the decision rule in (2) also becomes parametric. Apart from the resulting computational simplicity of the decision rule, this has the additional benefit that the probability distribution functions can be estimated more accurately because we need to infer only the parameters of these distributions.

The popular assumption of normality (Devijver & Kittler 1982), i.e.

$$p(\boldsymbol{x}|\omega_i) = [(2\pi)^d|\Sigma_i|]^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\,\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\}, \tag{5}$$

where Σ_i is the covariance matrix of the ith class and $\boldsymbol{\mu}_i$ is the mean vector, leads to the following parametric rule:

$$\text{assign} \quad \boldsymbol{x} \quad \text{to} \quad \omega_i \quad \text{if} \quad (\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\,\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+K_i = \min_j\{(\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\,\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)+K_j\}. \tag{6}$$

In (6) K_i denotes

$$K_i = \log[(2\pi)^d\,|\Sigma_i|]-\log P^2(\omega_i). \tag{7}$$

Further, if the covariance matrices Σ_i and class *a priori* probabilities $P(\omega_i)$ are identical, i.e. $\Sigma_i = \Sigma, P(\omega_i) = 1/m, \forall i$, then we can achieve even greater simplification, for now $K_i = K_j$, $\forall_j$ and (6) can be rewritten as

$$\text{assign} \quad \boldsymbol{x} \quad \text{to} \quad \omega_i \quad \text{if} \quad (2\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\,\Sigma^{-1}\boldsymbol{\mu}_i = \max_j\,(2\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\,\Sigma^{-1}\boldsymbol{\mu}_j. \tag{8}$$

It follows that the decision rule in (8) is linear in $\boldsymbol{x}$.

Finally when $\Sigma_i = I$ for all classes, the resulting decision rule is the well known nearest mean (minimum distance) classifier, i.e.

$$\text{assign} \quad \boldsymbol{x} \quad \text{to} \quad \omega_i \quad \text{if} \quad \delta(\boldsymbol{x},\boldsymbol{\mu}_i) = \min_j \delta(\boldsymbol{x},\boldsymbol{\mu}_j), \tag{9}$$

where $\delta(\boldsymbol{x},\boldsymbol{\mu}_j)$ is the Euclidean distance between vectors $\boldsymbol{x}$ and $\boldsymbol{\mu}_j$, defined as

$$\delta(\boldsymbol{x},\boldsymbol{\mu}_j) = \{[\boldsymbol{x}-\boldsymbol{\mu}_j]^{\mathrm{T}}\,[\boldsymbol{x}-\boldsymbol{\mu}_j]\}^{\frac{1}{2}}. \tag{10}$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

The various simplifying assumptions can often be justified on the basis of computational involvement only. Although this will inevitably introduce approximation errors, some consolation can be drawn from the fact that with a given data base, simpler models can be estimated more accurately than more complicated ones.

In this paper I shall not impose any engineering constraints that would preclude the discussion of more sophisticated classification schemes. I shall adopt the view that the rapidly advancing technology will make it possible to implement such schemes in the near future. The recent commercial availability of parallel processors such as CLIP4 provides supporting evidence for this view. Moreover, the greatly improved data quality achieved with the launch of Landsat D/D makes it more important than ever before to use algorithms that can fully exploit the information content of the remotely sensed imagery.

The additional information that can be used is conveyed by spatial characteristics of the data. In particular, in classification of groups of pixels rather than one pixel at a time, we can take advantage of class homogeneity of land surface covers. Other useful information is contained in texture, shape, structural relations between land cover objects, context, and general dependences between pixels due to instrumental scanning errors, atmospheric turbulence, noise, weather conditions, etc. Algorithms that take some of these factors into consideration will be the subject of discussion in the following sections.

3. Object classification

(*a*) *Spectral features*

The classification of data on a pixel-by-pixel basis discussed in the previous section can be affected by noise. If we consider that the size of land cover objects is usually considerably larger than the area corresponding to a single pixel, then the neighbouring pixels are likely to have similar properties. We can take advantage of this observation and identify homogeneous segments in the image that can be subsequently classified on a pixel-group (object) basis. In the context of the statistical estimation theory the expected improvement in performance is based on the premise that the larger the sample from which a statistic is inferred, the better its estimate.

Before we can proceed with object classification we need a procedure for determining homogeneous segments in the image. Kettig & Landgrebe (1976) and Landgrebe (1980) suggest partitioning the image into small regions. Those that have similar properties are then merged to create homogeneous segments.

Segmentation procedure

1. Group pixels into cells of 2 × 2 pixels (see figure 2).
2. Test statistical homogeneity. If the test fails then classify each singular pixel individually.
3. Check adjacent cells of non-singular pixels for statistical similarity. If similar, then merge the cells into one segment.

By using this procedure homogeneous segments will grow to their natural boundaries.

I now denote the set of pixels belonging to one of the segments by X, i.e.

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}. \tag{11}$$

By analogy with (2), I shall assign all the pixels in the set to class ω_i satisfying

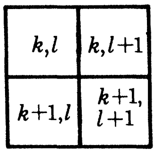
$$P(\omega_i|X) = \max_j P(\omega_j|X). \tag{12}$$

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The *a posteriori* probability $P(\omega_j|X)$ can be expressed in terms of the joint conditional and mixture probability density functions, i.e.

$$P(\omega_j|X) = \frac{p(x_1, \ldots, x_n|\omega_j)\, P(\omega_j)}{p(x_1, \ldots, x_n)}. \tag{13}$$



k, l	$k, l+1$
$k+1, l$	$k+1,$ $l+1$

FIGURE 2. A pixel cell.

If the variables x_k $(k = 1, 2, \ldots, n)$ are conditionally statistically independent then the numerator in (13) can be evaluated as a product of pixel densities, i.e.

$$p(x_1, \ldots, x_n|\omega_j) = \prod_{k=1}^{n} p(x_k|\omega_j). \tag{14}$$

In general, conditionally independent variables will be unconditionally dependent and the denominator cannot therefore be simplified in a similar manner. However, as the denominator is identical for all classes and only finding the most probable class is of interest, the following rule can be used:

$$\text{assign} \quad X \quad \text{to} \quad \omega_i \quad \text{if} \quad P(\omega_i) \prod_{k=1}^{n} p(x_k|\omega_i) = \max_{j} P(\omega_j) \prod_{k=1}^{n} p(x_k|\omega_j). \tag{15}$$

(*b*) *Texture measures*

In many cases the spectral characteristics of data varies from pixel to pixel. If these spatial variations are (locally) stationary, they will give rise to an apparent regular spatial pattern, which is referred to as texture. Texture can be characterized by quantitative textural measures. Many such measures have been suggested in the literature but it is beyond the scope of this paper to discuss them in detail. Broadly speaking, the following categories of texture measures have been proposed (Haralick 1979; Pratt 1978; Niemann 1981; Weszka *et al.* 1976): first-order statistics of grey level differences; second-order statistics of the grey tone spatial dependence matrix; run length statistics; two-dimensional power spectrum measurements; edgeness per unit area. However, this list is by no means exhaustive.

In general there are no guidelines as to what texture measures are best suited in given circumstances. For given data, answers to these questions can be found by using feature-selection methods. More specifically, each texture measure is considered as a candidate feature; suppose we extract N such measures. Then feature selection is concerned with the problem of selecting a subset, $d < N$, of these candidate measures that allow for best discrimination between various classes of textures. The selected subset of features forms a feature vector x, which can be classified by using the decision rule in (2). A recent discussion of this topic can be found in Devijver & Kittler (1982).

It is apparent that texture measures can be computed only for groups of neighbouring pixels (not for a single pixel). The most appropriate procedure is to specify a window of $k_x \times k_y$ pixels and base the texture measure computation on the resulting group of $n = k_x k_y$ pixels. Again there are two options. Either this window can be centred at successive pixels and for each a vector of texture measurement can be obtained. This vector can then be assigned to an appropriate land cover category.

Downloaded from rsta.royalsocietypublishing.org

Alternatively the image can be segmented first by using the procedure described in the previous subsection. Note that here the initial cells are $k_x \times k_y$ instead of 2×2 pixels. Once homogeneous segments are detected, the texture statistics can be recomputed for the complete segments and the segments then classified.

4. One-dimensional spatial dependences

With the exception of texture classification discussed in the previous subsection, so far I have assumed that pixel data are conditionally independent. In other words, observation $\boldsymbol{x}_{kl}$ corresponding to the (k, l)th pixel is independent of the measurements on pixels in its neighbourhood. In practice the value of $\boldsymbol{x}_{kl}$ is likely to be correlated with those of the surrounding pixels, $\boldsymbol{x}_{k+r,\,l+s}$, where r and s are small positive and negative integers. In general, these correlations should be modelled by two-dimensional spatial random processes and I shall consider such models in the next sections. Here I shall assume that there are correlations only between successive variables in each scan line. This assumption has been shown to be realistic in a number of experimental studies. In particular, low-order autoregressive and autoregressive moving-average models appear to fit the Landsat data reasonably well (Craig 1979; Tubbs & Coberly 1978; Tubbs 1980). It allows us to model the data by using simple one-dimensional spatial models, which have an analogy in the time-series analysis.

For simplicity I shall further assume that observations in one spectral channel are not spatially dependent on observations in other channels. Thus for modelling purposes I consider only one spectral channel at a time. Now suppose the sensor is scanning the kth line of the current image frame in some spectral channel. According to the ARMA model the lth observation x_l is related to the previous observations as

$$\sum_{t=0}^{p} \psi_t x_{l-t} = \sum_{t=0}^{q} \varphi_t(\epsilon_{l-t} + \nu_{l-t}), \tag{16}$$

where p and q denote the autoregressive and moving-average processes respectively, ψ_t and φ_t are the parameters of the process and ϵ_l is a normally distributed independent random variable with zero mean and variance σ^2; ν_l is the true (uncorrupted) intensity at the lth pixel, characteristic of a particular class of land cover. Note that $\psi_0 = \varphi_0 = 1$.

The model in (16) depicts the situation where due to crosstalk or a cell–spot overlap, the measured output x_l at the lth pixel in each scan line depends on the outputs at pixels x_{l-t} ($t = 1, 2, \ldots, p$) and the history of the noise.

This general type of model may also be found useful in describing correlation between the noise variables. For instance, atmospheric turbulence will give rise to spatial dependence of the noise, which strictly speaking is two-dimensional. However, because the data are acquired in a scanning fashion and the atmospheric conditions change as a function of time, the noise components in one scan line can be considered independent of those in the previous line.

I shall now analyse the effect on pixel classification of a particular type of model with $p = 1$ and $q = 0$. A model of this order has been suggested for remotely sensed data by Tubbs & Coberly (1978). Ignoring the transient effect in the border regions between homogeneous segments, from (16)

$$y_l + \psi_1 y_{l-1} = \epsilon_l, \tag{17}$$

where y_l,

$$y_l = x_l - \mu_l, \tag{18}$$

Downloaded from rsta.royalsocietypublishing.org

is the centralized observation on the lth pixel and μ_l is the corresponding mean. From (17) the variance, τ^2, of y_l satisfies

$$\tau^2 = E\{y_l^2\} = \sigma^2 + \psi_1^2 \tau^2, \tag{19}$$

which can be rearranged as

$$\tau^2 = \sigma^2/(1-\psi_1^2). \tag{20}$$

From the stability point of view parameter ψ_1 must satisfy $|\psi_1| < 1$. Thus the variance of x_l is larger than that of ϵ_l. It follows that the AR(1) autoregressive process introduces between-variable dependences that give rise to observations with a considerably larger variance than the noise process ϵ_l.

Usually the larger the variance, the more difficult it is to achieve a low-error pixel classification performance. It is therefore desirable to base the classification process on a filtered variable that exhibits a lower variance than τ^2.

In the context of model (17) this filtered observation z_l can be computed as

$$z_l = x_l + \psi_1 x_{l-1}. \tag{21}$$

Variable z_l in (20) has class-conditional mean ν_{li} ($i = 1, 2, \ldots, m$) and variance σ^2. It is independent of previous outcomes z_{l-t}, $t = 1, 2, \ldots$, and, therefore, it should afford a more reliable classification of pixel data. However, this conjecture will now be scrutinized to verify its validity.

Assume that the pixel data is normally distributed with the class mean μ_{li} and equal variance τ^2. Further, suppose that *a priori* probabilities of classes ω_i are equal. Then our ability to discriminate between classes ω_i and ω_j on the basis of observation x_l can be measured by using the Mahalanobis distance (Devijver & Kittler 1982), defined as

$$J(\omega_i, \omega_j) = (\mu_{li} - \mu_{lj})^2/\tau^2. \tag{22}$$

Suppose that pixel x_l is classified on the basis of the filtered measurement z_l. Then

$$J'(\omega_i, \omega_j) = (\nu_{li} - \nu_{lj})^2/\sigma^2. \tag{23}$$

I shall now compare the criteria J and J' for model (17). From (16), (17) and (18),

$$\mu_l = -\psi_1 \mu_{l-1} + \nu_l, \tag{24}$$

which for a homogeneous region where $\mu_l = \mu_{l-1}$ leads to

$$\mu_l = \nu_l/(1+\psi_1). \tag{25}$$

Substituting from (20) and (25) into (23),

$$J'(\omega_i, \omega_j) = \{(1+\psi_1)/(1-\psi_1)\}\, J(\omega_i, \omega_j). \tag{26}$$

Thus in this case the classification performance for the filtered image will improve or deteriorate depending on the sign of parameter ψ_1, i.e. whether the model is basically a low-pass or high-pass filter.

In view of these results it is important to study the model effects very carefully before a commitment to a specific classification strategy is made. It is possible that the combined data generation and acquisition process is such that better results can be achieved in the original observation space. On the other hand, if the model of the process indicates that it is advantageous to classify pixel data by using filtered observations, an appropriate filtering scheme must be derived from the model. The extension of these results to multivariate observations is straightforward.

Downloaded from rsta.royalsocietypublishing.org

5. Two-dimensional spatial dependences

In this section I shall consider how the contextual information imbedded both in pixel category relations and in two-dimensional correlations can be exploited to improve classification performance. In the preceding sections I made use of both types of information either explicitly or implicitly. The particular contextual information exploited is that neighbouring pixels are likely to belong to the same category. This relation, which I have termed homogeneity, led to the object classification rule (15) which has been shown to yield a lower classification error than the basic rule in (2) (Landgrebe 1980).

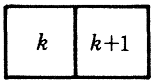


Figure 3. Neighbouring pixels.

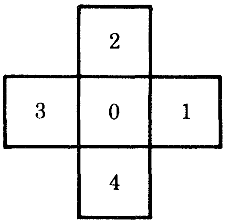


Figure 4. The four-neighbourhood.

Before any observations are made, the pixel category relation that gives rise to homogeneity can be characterized in terms of joint probabilities. Adopting the notation in figure 3, the dependence of the class membership θ_k of the kth pixel on the class membership θ_{k+1} of its neighbour $k+1$ can be quantified by the probability of joint occurrence of θ_k and θ_{k+1}, $P(\theta_k, \theta_{k+1})$. For homogeneity this is $P(\theta_k = \omega_i, \theta_{k+1} = \omega_i)$. (Here I am using the same notation for the random variable θ_k and its realization.)

In principle, the probability that label θ_k will take a particular value could depend on all the pixels in the image, and certainly on more than one of the neighbours of pixel k. For simplicity and not without justification I shall, however, assume that the region of mutual influence extends only to the four pixels 1, 2, 3 and 4 as shown in figure 4. (Index k has been dropped for convenience.) In other words the effect of more distant pixels and of the pixels in the diagonal directions will be ignored. Then formally homogeneity can be characterized by the joint probability $P(\theta_0 = \omega_i, \theta_1 = \omega_i, \theta_2 = \omega_i, \theta_3 = \omega_i, \theta_4 = \omega_i)$.

Homogeneity, of course, is not the only pixel category relation that can occur and is of interest in remotely sensed data. A particular land cover at one pixel may provide strong contextual evidence for another class at a neighbouring pixel; for instance, the land cover on each side of a road is likely to be similar and in any case its special signature will be different from that of the road. Objects on a river are more likely to be boats rather than cars. In general, therefore, what is of interest is the probability of particular spatial relations between pixel categories. The contextual information provided by these relations is embodied in the joint *a priori* class probabilities $P(\theta_0, \theta_1, \ldots, \theta_4)$. Introducing a vector notation $\boldsymbol{\theta}$ for the class labels, i.e.

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \ldots, \theta_4)^{\mathrm{T}}, \tag{27}$$

this joint probability can be written as $P(\boldsymbol{\theta})$.

In the following subsections I shall consider how these joint probabilities can be incorporated in the decision-making process.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

(a) *Decision rules based on contextual information*

Similarly to object classification discussed earlier, decisions are to be based on a set of observations rather than on one pixel at a time. As distant pixels are assumed to have no bearing on the classification of pixel $\boldsymbol{x}_0$, the set of interest, X, is defined by

$$X = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_4\}. \tag{28}$$

Note that here we use X to classify only $\boldsymbol{x}_0$, not all the elements of the set as in § 3*a*.

By analogy with (12) $\boldsymbol{x}_0$ should be assigned to the most likely class, i.e.

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad P(\omega_l|X) = \max_i P(\omega_i|X). \tag{29}$$

The application of the Bayes formula for calculating conditional probabilities would transform the decision rule (29) into a scheme defined in terms of probability density functions $p(X|\omega_i)$. However, because of the complex dependences assumed between the variables of the four-neighbourhood, this probability density function cannot in general be simplified. Inference of the joint density function $p(X|\omega_i)$ would lead to estimation and computational problems, because working with X amounts to increasing the dimensionality of the random variable by a factor of five. I shall therefore attempt to simplify the decision rule by considering the original *a posteriori* functions $P(\omega_i|X)$ direct.

Expand $P(\omega_i|X)$ as

$$P(\omega_i|X) = \sum_{S_i} P(\boldsymbol{\theta}|X) = \{1/p(X)\} \sum_{S_i} p(X|\boldsymbol{\theta})\, P(\theta), \tag{30}$$

where S_i is the set of $\boldsymbol{\theta}$ such that $\theta_0 = \omega_i$. Because the denominator term $p(X)$ is identical for all the classes it can be ignored, thus leading to the decision rule

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad \sum_{S_l} p(X|\boldsymbol{\theta})\, P(\boldsymbol{\theta}) = \max_i \sum_{S_i} p(X|\boldsymbol{\theta})\, P(\boldsymbol{\theta}). \tag{31}$$

Moreover, if the class labels are independent, i.e. the spatial configuration of pixel classes conveys negligible contextual information, the rule becomes

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad \sum_{S_l} p(X|\boldsymbol{\theta}) \left\{\prod_{j=1}^{4} P(\theta_j)\right\} P(\omega_l) = \max_i P(\omega_i) \sum_{S_i} p(X|\boldsymbol{\theta}) \prod_{j=1}^{4} P(\theta_j). \tag{32}$$

These classification schemes involve estimation of the joint probability density functions $p(X|\boldsymbol{\theta})$. In general these functions will be very difficult to obtain. A notable exception is when the data can be adequately modelled as an autonormal scheme discussed by Besag (1974), Bartlett (1975) and Fu & Yu (1980). The joint density function of such a scheme for univariate observations is defined as

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = (2\pi\sigma^2)^{-\frac{5}{2}}\, |B|^{\frac{1}{2}} \exp\{-\tfrac{1}{2}\sigma^{-2}(\boldsymbol{X}-\boldsymbol{u})^{\mathrm{T}}\, B(\boldsymbol{X}-\boldsymbol{u})\}, \tag{33}$$

where $\sigma^2 B^{-1}$ is the variance–covariance matrix, $\boldsymbol{u}$ is the vector of means of classes $\theta_0, \theta_1, \ldots, \theta_4$ and $\boldsymbol{X}$ is a vector of univariate observations corresponding to the four-neighbourhood in figure 4. Methods for estimating σ^2 and B are documented in Fu & Yu (1980).

Under the assumption that the appearance of $\boldsymbol{x}_j$ is a function of θ_j only, which implies that

$$p(\boldsymbol{x}_j|\theta_j, \boldsymbol{x}_l, \theta_l, l \neq j) = p(\boldsymbol{x}_j|\theta_j), \tag{34}$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

$p(X|\boldsymbol{\theta})$ in (31) can be simplified as

$$p(X|\boldsymbol{\theta}) = p(\boldsymbol{x}_0|\omega_i) \prod_{j=1}^{4} p(\boldsymbol{x}_j|\theta_j). \tag{35}$$

From (31) and (35) the joint class probability decision rule becomes

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad p(\boldsymbol{x}_0|\omega_l) \sum_{S_l} P(\boldsymbol{\theta}) \prod_{j=1}^{4} p(\boldsymbol{x}_j|\theta_j) = \max_i p(\boldsymbol{x}_0|\omega_i) \sum_{S_i} P(\boldsymbol{\theta}) \prod_{j=1}^{4} p(\boldsymbol{x}_j|\theta_j). \tag{36}$$

In addition to the density function $p(\boldsymbol{x}|\omega_i)$, the decision rule (36) requires the estimation of *a priori* joint probability function $P(\boldsymbol{\theta})$. This will require the availability of a large quantity of ground truth data.

Finally, it should be noted that a pixel can be classified by determining the most probable combination of class labels $\boldsymbol{\theta}$ for the appropriate neighbourhood. Accordingly,

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad p(X|\boldsymbol{\theta}^*)\, P(\boldsymbol{\theta}^*) = \max_{S_i} p(X|\boldsymbol{\theta})\, P(\boldsymbol{\theta}) \quad \text{and} \quad \theta_0^* = \omega_l. \tag{37}$$

This scheme has been proposed in Fu & Yu (1980) but in my view it is associated with some conceptual problems. In particular each pixel ends up with five labels that may differ. Even if all but one are ignored, potential inconsistencies in the resulting labelling cannot be eliminated.

(*b*) *The compound decision rule*

In addition to assumption (34) I shall further assume that the contextual classification between any non-adjacent cells is negligible. For $l > 0$ this implies that

$$p(\boldsymbol{x}_j|\boldsymbol{x}_l, \theta_l, l \neq j) = p(\boldsymbol{x}_j|\theta_0), \tag{38}$$

because pixel 0 is the only adjacent cell to pixels $\boldsymbol{x}_j (j = 1, 2, \ldots, 4)$. Under these two assumptions $P(\omega_i|X)$ in (29) can be expressed as

$$P(\omega_i|X) = p(X|\omega_i)\frac{P(\omega_i)}{p(X)} = \frac{P(\omega_i)}{p(X)} \prod_{j=0}^{4} p(\boldsymbol{x}_j|\theta_0 = \omega_i). \tag{39}$$

To avoid the need for estimating conditional probability functions $p(\boldsymbol{x}_j|\theta_0)$ $(j \geqslant 1)$, I shall expand $p(\boldsymbol{x}_j|\theta_0)$ as

$$p(\boldsymbol{x}_j|\theta_0) = \sum_{r=1}^{m} p(\boldsymbol{x}_j, \theta_j = \omega_r|\theta_0) = \sum_{r=1}^{m} p(\boldsymbol{x}_j|\omega_r)\, P(\omega_r|\theta_0), \tag{40}$$

provided that the probabilities $P(\omega_r|\theta_0)$ are isotropic. Substituting (40) into (39) the compound decision rule is obtained (Welch & Salter 1971):

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad P(\omega_l)\, p(\boldsymbol{x}_0|\omega_l) \prod_{j=1}^{4} \sum_{r=1}^{m} p(\boldsymbol{x}_j|\omega_r)\, P(\omega_r|\omega_l) = \max_i P(\omega_i)\, p(\boldsymbol{x}_0|\omega_i) \prod_{j=1}^{4} \sum_{r=1}^{m} p(\boldsymbol{x}_j|\omega_r)\, P(\omega_r|\omega_i). \tag{41}$$

$P(\omega_r|\omega_i)$ are called transition probabilities. If these probabilities exhibit directional dependence, $P(\omega_r|\omega_i)$ must be replaced by $P(\theta_j = \omega_r|\omega_i)$. For brevity I shall denote these directional probabilities by $P_j(\omega_r|\omega_i)$.

Thus in comparison with (2) the only additional information needed to implement decision rule (41) are the class transition probabilities $P(\omega_r|\omega_i), \forall i, r$, or directional transition probabilities $P_j(\omega_r|\omega_i)$ $(j = 1, 2, \ldots, 4)$. This is computationally less involving than the estimation of the joint probability function $P(\boldsymbol{\theta})$ of the method in (36).

Downloaded from rsta.royalsocietypublishing.org

6. Relaxation labelling

Up to now I have discussed non-iterative decision-making schemes incorporating contextual information. In all cases the aim has been to determine the class membership of a pixel by computing the *a posteriori* probabilities $P(\omega_i|X)$ of class ω_i, given observations on the pixel and its neighbours. An alternative approach is to compute these probabilities recursively. The basic updating formula for the probability $P(\omega_i, \boldsymbol{x}_0)$ of joint occurrence of ω_i and $\boldsymbol{x}_0$ is

$$P^n(\omega_i, \boldsymbol{x}_0) \Rightarrow P^{n-1}(\omega_i, X). \tag{42}$$

The functions $P^n(\omega_i, \boldsymbol{x}_0)$ appear as arguments in $P^n(\omega_i, X)$ and so on.

Note that strictly speaking $P^n(\omega_i, \boldsymbol{x}_0)$ is no longer the joint probability of class ω_i and pixel $\boldsymbol{x}_0$. It is a variable that should eventually be equivalent to $P(\omega_i, X)$. Initially $P^0(\omega_i, \boldsymbol{x}_0)$ is set equal to

$$P^0(\omega_i, \boldsymbol{x}_0) = P(\omega_i|\, \boldsymbol{x}_0)\, p(\boldsymbol{x}_0). \tag{43}$$

As before, the actual form of the function $P(\omega_i, X)$ will depend on the assumptions made for the dependence between variables in X and their classes. Suppose that the spatial dependence of pixels can be characterized by the joint probability $P(\omega_i, X)$ in (30) with the *a priori* class probabilities being independent. Then, from (32),

$$P(\omega_i, X) = \sum_{S_i} p(X|\boldsymbol{\theta})\, P(\omega_i) \prod_{j=1}^{4} P(\theta_j). \tag{44}$$

Under the assumption of known $p(X|\boldsymbol{\theta})$, the expression (44) can be used to drive the relaxation process (42). For univariate observations the density function in (33) is of practical importance in many applications. After the initialization stage the *a priori* probabilities in (44) lose their identities and are replaced by the current estimates of the conditional probabilities $P^{n-1}(\omega_j|\, \boldsymbol{x}_0)$, i.e. (Kittler & Föglein 1983*a*)

$$P^n(\omega_i, \boldsymbol{x}_0) \Rightarrow P^{n-1}(\omega_i|\, \boldsymbol{x}_0) \sum_{S_i} p(X|\boldsymbol{\theta}) \prod_{j=1}^{4} P^{n-1}(\theta_j|\, \boldsymbol{x}_j). \tag{45}$$

The probabilities $P^n(\omega_i, \boldsymbol{x}_0)$ may have to be normalized to ensure that

$$\sum_{i=1}^{m} P^n(\omega_i|\, \boldsymbol{x}_0) = 1. \tag{46}$$

Yu & Fu (1983) proposed an algorithm that can be interpreted as a simplification of the relaxation process in (45). I shall approximate the probabilities $P^n(\theta_j|\, \boldsymbol{x}_j)$ $(j = 1, 2, \ldots, 4)$ as follows:

$$P^n(\theta_j = \omega_r|\, \boldsymbol{x}_j) = 1 \quad \text{if} \quad P^n(\theta_j = \omega_r|\, \boldsymbol{x}_j) = \max_l P^n(\theta_j = \omega_l|\, \boldsymbol{x}_j); \tag{47}$$

$$P^n(\theta_j = \omega_r|\, \boldsymbol{x}_j) = 0 \quad \text{otherwise}. \tag{48}$$

I shall denote the vector $\boldsymbol{\theta}$ at the nth iteration with $\theta_0 = \omega_i$ and θ_j such that (47) holds, by $\tilde{\boldsymbol{\theta}}_n$. Then the summation in (45) becomes

$$\sum_{S_i} p(X|\boldsymbol{\theta}) \prod_{j=1}^{4} P^{n-1}(\theta_j|\, \boldsymbol{x}_j) = p(X|\tilde{\boldsymbol{\theta}}_n) \tag{49}$$

and the recursive formula simplifies to

$$P^n(\omega_i, \boldsymbol{x}_0) = P^{n-1}(\omega_i|\, \boldsymbol{x}_0)\, p(X|\tilde{\boldsymbol{\theta}}_{n-1}). \tag{50}$$

The corresponding algorithm can now be stated as follows.

Downloaded from rsta.royalsocietypublishing.org

(i) Classify pixels by using the minimum error Bayes rule and set

$$P^0(\omega_i|\boldsymbol{x}_0) = P(\omega_i|\boldsymbol{x}_0).$$

(ii) For every pixel $\boldsymbol{x}_0$ at the stage n form vector $\tilde{\boldsymbol{\theta}}_{n-1}$ and compute $p(X|\tilde{\boldsymbol{\theta}}_{n-1})$.

(iii) Use (50) to update class probabilities, i.e.

$$P^n(\omega_i|\boldsymbol{x}_0) \Rightarrow P^{n-1}(\omega_i|\boldsymbol{x}_0)\, p(X|\tilde{\boldsymbol{\theta}}_{n-1})/p(X),$$

and normalize them to satisfy condition (46).

(iv) Reclassify pixels according to the minimum error Bayes rule, i.e.

$$\text{assign} \quad \boldsymbol{x}_0 \quad \text{to} \quad \omega_l \quad \text{if} \quad P^n(\omega_l|\boldsymbol{x}_0) = \max_i P^n(\omega_i|\boldsymbol{x}_0).$$

(v) If no pixel is reassigned then terminate the algorithm, else return to step 2.

If the spatial dependence of pixels satisfies the assumptions made in §5*b*, it can be shown (Kittler & Föglein 1983*b*) that the corresponding algorithm for updating $P^n(\omega_i|\boldsymbol{x}_0)$ is the conventional relaxation algorithm (Zucker *et al.* 1978; Rosenfeld *et al.* 1976). It permits the determination of pixel labels that are compatible in the sense of the transition probabilities.

I shall denote the modified *a posteriori* probability of class ω_i given an arbitrary pixel $\boldsymbol{x}_0$ after n iterations by $P^n(\omega_i|\boldsymbol{x}_0)$. Then at the next stage this probability is updated according to

$$P^{n+1}(\omega_i|\boldsymbol{x}_0) = \frac{P^n(\omega_i|\boldsymbol{x}_0)\, Q^n(\omega_i)}{\sum_{r=1}^{m} P^n(\omega_r|\boldsymbol{x}_0)\, Q^n(\omega_r)}, \tag{51}$$

where $Q^n(\omega_r)$ is the so-called neighbourhood function. It is defined as

$$Q^n(\omega_r) = \sum_{j=1}^{4} d_j \sum_{l=1}^{m} P_j(\omega_r|\omega_l)\, P^n(\omega_l|\boldsymbol{x}_j), \tag{52}$$

where d_j is a neighbourhood weight.

The stability and convergence of this relaxation algorithm has been studied by Zucker *et al.* (1978). The effect of weights d_j and of the transition probabilities on the algorithm performance has been investigated by Richards *et al.* (1980). It has been shown that a suitable choice of d_j can prevent the loss of corners, lines, endpoints and other degradations in pixel labelling.

7. Conclusions

In the paper a number of approaches to multispectral image segmentation and classification have been considered. The methods range from the simple Bayesian decision rule for classification of image data on pixel-by-pixel basis, to sophisticated algorithms using contextual information. Both the spatial pixel category dependences and the two-dimensional correlation-type contextual information have been incorporated in decision-making schemes. The algorithms have been developed to achieve a greater reliability in the process of the interpretation of remote-sensing data.

References

Bartlett, M. S. 1975 *The statistical analysis of spatial pattern*. London: Chapman & Hall.

Bernstein, R. (ed.) 1978 *Digital image processing for remote sensing*. New York: IEEE Press.

Besag, J. E. 1974 *Jl R. stat. Soc.* B **36**, 192–236.

Craig, R. G. 1979 In *Proc. 13th Int. Symp. Remote Sensing of Environment, Ann Arbor, Michigan*, pp. 1517–1524.

Devijver, P. & Kittler, J. 1982 *Pattern recognition: a statistical approach*. Englewood Cliffs: Prentice-Hall.

Fu, K. S. (ed.) 1977 *Syntactic pattern recognition, applications*. Berlin: Springer-Verlag.

Fu, K. S. 1982*a* *Syntactic pattern recognition and applications*. Englewood Cliffs: Prentice-Hall International.

Downloaded from rsta.royalsocietypublishing.org

Fu, K. S. 1982*b* In *Pattern recognition theory and applications* (ed. J. Kittler, K. S. Fu & L. F. Pau), pp. 139–155. Dordrecht: D. Reidel.

Fu, K. S. & Yu, T. S. 1980 *Statistical pattern classification using contextual information*. Chichester: John Wiley & Sons.

Haralick, R. M. 1976 In *Digital picture analysis* (ed. A. Rosenfeld), pp. 5–63. Berlin: Springer-Verlag.

Haralick, R. M. 1979 *Proc. IEEE* **67**, 786–804.

Huang, T. S. (ed.) 1979 *Picture processing and digital filtering*. Berlin: Springer-Verlag.

Kettig, R. L. & Landgrebe, D. A. 1976 *IEEE Trans. Geosci. Electron.* **GE-4**, 19–26.

Kittler, J. 1983 In *Physical and biological processing of images* (ed. O. J. Braddick and A. C. Sleigh), pp. 232–243. Berlin: Springer-Verlag.

Kittler, J. & Föglein, J. 1983*a* In *Proc. 3rd Scandinavian Conf. Image Analysis, Copenhagen.*

Kittler, J. & Föglein, J. 1983*b* (Submitted.)

Landgrebe, D. A. 1980 *Pattern Recogn.* **12**, 165–175.

Niemann, H. 1981 *Pattern analysis*. Berlin: Springer-Verlag.

Pavlidis, T. 1977 *Structural pattern recognition*. Berlin: Springer-Verlag.

Pratt, W. K. 1978 *Digital image processing*. New York: John Wiley and Sons.

Richards, J. A., Landgrebe, D. A. & Swain, P. H. 1980 In *Proc. 5th Int. Conf. Pattern Recognition, Miami*, pp. 61–65.

Rosenfeld, A., Hummel, R. & Zucker, S. 1976 *IEEE Trans. Syst. Mgmt Cybern.* **SMC-6**, 420–433.

Tubbs, J. D. 1980 *IEEE Trans. Syst. Mgmt Cybern.* **SMC-10**, 177–180.

Tubbs, J. D. & Coberly, W. A. 1978 In *Proc. 12th Int. Symp. Remote Sensing of Environment, Manila, Philippines.*

Welch, J. R. & Salter, K. G. 1971 *IEEE Trans. Syst. Mgmt Cybern.* **SMC-1**, 24–30.

Weszka, J., Dyer, C. & Rosenfeld, A. 1976 *IEEE Trans. Syst. Mgmt Cybern.* **SMC-6**, 269–285.

Yu, T. S. & Fu, K. S. 1983 *Pattern Recogn.* **16**, 89–108.

Zucker, S., Krishnamurthy, E. & Haar, R. 1978 *IEEE Trans. Syst. Mgmt Cybern.* **SMC-8**, 41–48.