Image models and the definition of image entropy applied to the problem of unsupervised segmentation

Anton David Brink

A thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, March 1994.

Abstract

Region segmentation of digital images by unsupervised thresholding is a common, conceptually simple and important branch of image processing and analysis. Its applications range from that of simple binarization to serving as a useful pre–processing stage for operations such as pattern recognition and image restoration. While many different algorithms have been proposed for the automatic selection of the "correct" threshold the results vary widely in their general usefulness. A class of selection schemes is based on the principle of maximum entropy. This formalism, while effective, is usually invoked without reference to its origins or its relationship to images.

This thesis attempts to clarify the definition of what is meant by the entropy of an image, to which end various image and image segmentation models are discussed and proposed. Some apparent shortcomings related to the use of the Shannon entropy formula are addressed and the outcome of the research is applied to the problem of threshold selection. The results indicate a marked improvement in performance of methods using some form(s) of context–related information over those which simply apply the entropy formula without regard to its spatially insensitive nature.

Evaluation of results and processes is usually based on aesthetic preference. Some quantitative evaluation techniques are investigated and implemented to evaluate the above threshold selection processes. While some agreement is found between these methods and qualitative evaluations, it is clear that much work still needs to be done in this particular field.

In conclusion, it is found that the results obtainable by invoking the maximum entropy formalism are strongly dependent on how effectively the particular strengths of this formalism are exploited. The inclusion of even very little additional pixel–related information, such as the mean value or variance within a small neighbourhood, can drastically improve performance. In order to get the best out of such a process it is clearly useful to explicitly consider how we are viewing or modelling the image. Instead of disregarding a fundamentally correct formalism given less than satisfactory results, the reasons for the poor results should be determined and, by taking into consideration such additional information as may be available, the implementation should be optimized.

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

Anton David Brink

23rd day of March, 1994

Image models and the definition of image entropy applied to the problem of unsupervised segmentation

Anton David Brink (90-05635/V

ERRATA

The following corrections, as required by the examiners of the thesis, have been made. Page numbers quoted refer to the page numbers of the original thesis as handed in on 23 March 1994. All errors, including typographical errors, have been addressed.

Preface.

p(v): At the end of the third line "insights" should read "insight".

List of Figures.

p(xii): The second line of the caption for Figure 6.10, "...(a) method G and (b) method H." should read "...(a) method H and (b) method I.".

p(xii): The second line of the caption for Figure 6.11, "...(a) method I and (b) method J." should read "...(a) method L and (b) method M.".

p(xii): The second line of the caption for Figure 6.12, "...(a) method K and (b) method L." should read "...(a) method N and (b) method O.".

1. Introduction.

p4: The ninth line of the second paragraph, "...so that entire process of segmentation from..." should read "...so that the entire process of segmentation, from...".

p5: At the top of the page, "...we use the theorem to determine the most likely "posterior distribution" (solution)." now reads "...we use the theorem to determine the "posterior distribution" of probabilities over the solution space. The aim is then to select the most likely solution: the posterior distribution and hence the selection of a solution is directly dependent on the data (evidence) and information (prior distribution) that we have.".

2. The maximum entropy method.

p18: In the second line of the second-last paragraph, "...Field [Pratt, 1991; Geman and Geman, 1984; Mardia, 1989; Marqués and Gasull, 1993]." now reads "...Field [Pratt,

1991; Geman and Geman, 1984; Mardia, 1989; Marqués and Gasull, 1993; Besag, 1986].".

p18: An additional paragraph has been inserted between the second–last and last paragraphs on this page. It reads "While the methods considered for the purposes of this thesis are not themselves Bayesian, the potential power of methods using the Bayesian approach, specifically those which take spatial relationships into account by means of random field models [Dubes and Jain, 1989; Johnson, 1994] cannot be ignored. Such an approach requires prior and posterior distributions defined over image space, and can serve as the next, more sophisticated step after selection of the prior distribution by the Maximum Entropy principle. The methods described are not limited to global threshold selection, but seek rather to accurately binarize the grey–scale image according to local information content. Another advantage of a Bayesian approach is that one is not left with a single "take it or leave it" solution, but a family of possible solutions which enable one either to view them all as a kind of "probabilistic movie" or, having selected one "best" solution, to determine an error bar on the selection indicating its quality. This goes much further than mere threshold evaluation, which tends itself to be rather *ad hoc*. The main drawback of such techniques is that they can be computationally extremely expensive and time–consuming: while this may be acceptable for critical applications requiring absolute accuracy, this thesis is primarily concerned with the fast and simple selection of thresholds by means of entropy maximization."

The additional references cited here are included in the errata for the References section.

3. Digital image models and entropy.

p34: The equation below the second paragraph,

$$p_{gg'} = \Pr\{x(\mathbf{k}) \in g, x(\mathbf{k}+\lambda) \in g'\},\ g,g' = 0,\cdots,n-1$$

should read

$$p_{gg'} = \Pr\{x(\mathbf{k}) = g, x(\mathbf{k}+\lambda) = g'\},\ g,g' \in \{0,\cdots,n-1\}$$

p36: The second line of the second–last paragraph, "...[Journel and Deitsch, 1993]..." should read "...[Journel and Deutsch, 1993]...".

4. Maximum entropy threshold selection.

p53: The equation at the bottom of the page,

$$\sum_{i=1}^{N} g_i = \left[\sum_{i=1}^{N} \mu_0(T)\right]_{g_i \le T} + \left[\sum_{i=1}^{N} \mu_1(T)\right]_{g_i > T}$$

is less ambiguously expressed as

$$\sum_{i=1}^{N} g_i = \sum_{i=1_{g_i \le T}}^{N} \mu_0(T) + \sum_{i=1_{g_i > T}}^{N} \mu_1(T)$$

p54: Equation (4.36),

$$\begin{aligned} H_\times(T) &= \sum_{i=1}^{N} q_i(T) \log \frac{q_i(T)}{p_i} \\ &= \left[\sum_{i=1}^{N} \mu_0(T) \log \frac{\mu_0(T)}{g_i} \right]_{g_i \le T} + \left[\sum_{i=1}^{N} \mu_1(T) \log \frac{\mu_1(T)}{g_i} \right]_{g_i > T} \\ &= \sum_{g=0}^{T} f_g \mu_0(T) \log \frac{\mu_0(T)}{g} + \sum_{g=T+1}^{n-1} f_g \mu_1(T) \log \frac{\mu_1(T)}{g} \end{aligned} \tag{4.36}$$

is less ambiguously expressed as

$$\begin{aligned} H_\times(T) &= \sum_{i=1}^{N} q_i(T) \log \frac{q_i(T)}{p_i} \\ &= \sum_{i=1_{g_i \le T}}^{N} \mu_0(T) \log \frac{\mu_0(T)}{g_i} + \sum_{i=1_{g_i > T}}^{N} \mu_1(T) \log \frac{\mu_1(T)}{g_i} \\ &= \sum_{g=0}^{T} f_g \mu_0(T) \log \frac{\mu_0(T)}{g} + \sum_{g=T+1}^{n-1} f_g \mu_1(T) \log \frac{\mu_1(T)}{g} \end{aligned} \tag{4.36}$$

p54: The line below equation (4.36), "...corresponding do the..." should read "...corresponding to the...".

p57: The third line below equation (4.41), "...follows similar resoning to..." should read "..follows similar reasoning to...".

p60: Equation (4.48),

$$
\begin{aligned}
H_0(T) &= -\sum_{j=1}^{G} i_j^0(T) \log i_j^0(T) \\
&= -\sum_{i=1}^{N} \sum_{i \in 0} g_i \, i \log i \\
H_1(T) &= -\sum_{j=1}^{G} i_j^1(T) \log i_j^1(T) \\
&= -\sum_{i=1}^{N} \sum_{i \in 1} g_i \, i \log i
\end{aligned}
\tag{4.48}
$$

is more clearly expressed as

$$
\begin{aligned}
H_0(T) &= -\sum_{j=1}^{G} i_j^0(T) \log i_j^0(T) \\
&= -\sum_{i=1}^{N} \sum_{i \in C_0} g_i \, i \log i \\
H_1(T) &= -\sum_{j=1}^{G} i_j^1(T) \log i_j^1(T) \\
&= -\sum_{i=1}^{N} \sum_{i \in C_1} g_i \, i \log i
\end{aligned}
\tag{4.48}
$$

and the following explanatory sentence has been added below the equation: "The below- and above-threshold classes are denoted C_0 and C_1, so that $C_0 = \{i \| \; g_i \leq T\}$ and $C_1 = \{i \| \; g_i > T\}$."

pp60,61: Equation (4.49),

$$
\begin{aligned}
H_0(T) &= -\sum_{j=1}^{G} p_j^0(T) \log p_j^0(T) \; , \quad p_j^0(T) = \frac{i_j^0(T)}{I_0} \\
&= -\sum_{i=1}^{N} \sum_{i \in 0} g_i \, p_0 \log p_0 \; , \quad p_0 = \frac{i}{I_0} \; , \; i \in 0 \\
H_1(T) &= -\sum_{j=1}^{G} p_j^1(T) \log p_j^1(T) \; , \quad p_j^1(T) = \frac{i_j^1(T)}{I_1} \\
&= -\sum_{i=1}^{N} \sum_{i \in 1} g_i \, p_1 \log p_1 \; , \quad p_1 = \frac{i}{I_1} \; , \; i \in 1
\end{aligned}
\tag{4.49}
$$

where

$$I_0 = \sum_{i=1}^{N}{}_{i \in 0} \; i\, g_i$$

$$I_1 = \sum_{i=1}^{N}{}_{i \in 1} \; i\, g_i$$

is more clearly written as

$$\begin{aligned} H_0(T) &= -\sum_{j=1}^{G} p_j^0(T) \log p_j^0(T) \;, \quad p_j^0(T) = \frac{i_j^0(T)}{I_0} \\ &= -\sum_{i=1}^{N}{}_{i \in C_0} \; g_i\, p_0 \log p_0 \;, \quad p_0 = \frac{i}{I_0} \;, \; i \in C_0 \\ H_1(T) &= -\sum_{j=1}^{G} p_j^1(T) \log p_j^1(T) \;, \quad p_j^1(T) = \frac{i_j^1(T)}{I_1} \\ &= -\sum_{i=1}^{N}{}_{i \in C_1} \; g_i\, p_1 \log p_1 \;, \quad p_1 = \frac{i}{I_1} \;, \; i \in C_1 \end{aligned} \tag{4.49}$$

where

$$I_0 = \sum_{i=1}^{N}{}_{i \in C_0} \; i\, g_i$$

$$I_1 = \sum_{i=1}^{N}{}_{i \in C_1} \; i\, g_i$$

5. Results.

p81: In the third line from the end of the first paragraph, "...the result is appears..." should read "...the result appears...".

6. Evaluation methods.

p98: In the caption for Figure 6.10, "...(a) method G and (b) method H." should read "...(a) method H and (b) method I.".

p99: In the caption for Figure 6.11, "...(a) method I and (b) method J." should read "...(a) method L and (b) method M.".

p99: In the caption for Figure 6.12, "...(a) method K and (b) method L." should read "...(a) method N and (b) method O.".

References.

p104: An additional reference is inserted between those for Albregtsen (1993) and Bretthorst (1990), which reads

Besag, J (1986). "On the statistical analysis of dirty pictures", *J. Royal Statistical Society B* **48 (3)**, 259–312.

p105: An additional reference is inserted between those for Donoho *et al* (1992) and Edwards (1979), which reads

Dubes, R C and A K Jain (1989). "Random field models in image analysis", *J. Applied Statistics* **16 (2)**, 131–164.

p107: An additional reference is inserted between those for Jaynes (1988) and Jones (1989), which reads

Johnson, E J (1994). "A model for segmentation and analysis of noisy images", *J. American Statistical Association* **89 (425)**, 230–241.

In Memory of my Grandfather

Daniël Brink

06 July 1905 – 25 October 1993

Preface

The research undertaken here is to some extent an extension and a consequence of the research I performed for my Master of Science degree at Rhodes University. I have long been interested in the remarkably subjective problems posed by the segmentation of an image and the insights it gives to the way we ourselves think. The maximum entropy formula was no stranger to me, having come across it during the aforementioned MSc research. However, I remained blissfully unaware of its theoretical and philosophical implications, and its meaning and derivation, for many years. Having finally realized what it is about, I was naturally interested in determining just how powerful and useful this concept may be by applying it to my old hobby–horse of threshold selection. While the most obvious aim of this research is, of course, to determine a reliable method of automatic threshold selection, the additional aim of optimizing the implementation of the maximum entropy formalism for this purpose is by no means secondary and has led to some understanding of the importance and relevance of *all* the information we have available.

I have attempted to stick quite closely to these aims without getting too sidetracked by philosophical issues, but also without ignoring them as I believe that they are of fundamental importance in their own right. The purpose of this research is mainly the determination of the role of entropy in image segmentation: how (and why) it should be implemented, its strengths and weaknesses, its theoretical and philosophical implications and how they affect what we understand by "a digital image". I try to determine what is really meant by the term "image entropy", both in the context of the thresholding operation and that of more general image–related processes.

I would like to thank the following for their support and encouragement throughout my research, particularly during the last few frantic weeks of completion.

- my supervisor, Dr Neil Pendock, for his advice, assistance and, not least, for waking me up to the potential of the principle of maximum entropy in the first place.
- Prof Hank Miller for his interest and for proof–reading the previous draft of this thesis.
- Mr André Bester, Mr Neil Davidson, Ms Chantal Friedman and Mr Peter Hawkes for their suggestions and for providing walls to bounce ideas off.

Finally, I wish to thank the FRD for financial assistance during the first two years of this research.

Contents

Illustrations

Tables

Chapter 1

Introduction

This research is concerned with the process of splitting an image into its constituent parts, such as "background" and "object", based on their relative component "brightness", or *grey-level*, distributions. Such processes belong to a branch of image processing referred to, for obvious reasons, as *segmentation*. The images in question are digital, both spatially and in terms of their grey-level arrangement. A standard black-and-white digital image consists of a two-dimensional array of discrete integer values representing shades of grey from black (0) to white (maximum possible grey-level). A common convention uses 1 byte to represent the brightness of each element (*pixel*) of this array, giving a range of 256 discrete grey-levels from 0 to 255. Colour images are often formed by superimposing 3 such grey-scale images, one each for the video colour display bands red, green and blue. Alternatively, the superimposed images can represent the characteristics "hue", "saturation" and "intensity". Other types of multiband image formats are also used: an example is LANDSAT's 7-band images, where each band (grey-scale image) represents a specific subrange of the electromagnetic spectrum. For the purposes of this thesis, by "image" I will refer specifically to single-band grey-scale images. The techniques used in this case can be relatively easily extended to multiband images.

A statistical measure commonly associated with the image is its *grey-level histogram*. This is simply a discrete *a posteriori* distribution indicating the frequency of occurrence of each grey-level in the image, i.e. the number of pixels having each grey-level. It is often converted to a grey-level probability histogram by normalizing the grey-level frequencies:

$$p_g = \frac{f_g}{N} , \quad \sum_{g=0}^{n-1} p_g = 1$$

where p_g is the *a posteriori* probability of grey-level g, f_g is its frequency of occurrence, N

is the total number of pixels in the image and n is the number of grey–levels.

A shortcoming of the histogram is that spatial information present in the image is not taken into account. An alternative measure which takes limited spatial information into account is the *grey–level co–occurrence matrix*. In general, it is a two–dimensional matrix whose entries indicate the number of times that pixels with grey–level g' occur in the immediate neighbourhood of pixels with grey–level g, say: the number of co–occurrences of g and g'. Specific definitions of what constitutes a neighbourhood of a pixel vary according to application. The matrix entries can be normalized in a similar fashion to those of the histogram to produce a discrete co–occurrence probability distribution.

1.1 Image segmentation

Image segmentation is a broad field of research which can be roughly divided into two main streams [Pratt, 1991; Gonzalez and Woods, 1992; Haralick and Shapiro, 1985]. *Feature detection* is concerned primarily with the detection of primitive features such as spots, lines and edges. *Region segmentation* is concerned with splitting up the image into relatively homogeneous regions, each having some property that distinguishes it from other regions in the image: a simple example is that of a bright object on a dark background. These two branches can be used in combination to achieve more sophisticated effects, e.g. simple region segmentation methods are prone to errors caused by non–uniform illumination in the image. An edge–detection method (less prone to this problem) can be used to guide a local region classifier. Following this in turn with a region–growing scheme results in illumination–independent grey–level segmentation of the image, usually referred to as *variable* or *dynamic thresholding* [Brink, 1991a].

For the purposes of this thesis, by "segmentation" I will refer only to *region segmentation*. Clearly segmentation can to some extent be considered a primitive form of pattern recognition. In terms of human visual perception the two seem to correspond roughly to right and left brain hemisphere processes, respectively [Edwards, 1979; Hofstadter, 1979]. In the discussion that follows, it should be pointed out that the breaking up of tasks into "right" and "left" hemisphere operations constitutes a gross simplification of the true operation of the human brain.

Consider the painting *The use of words I* by the Belgian surrealist René Magritte [Gablik, 1970] in Figure 1.1. One's initial reaction to this picture is "But it *is* a pipe!". This is the result of initially segmenting the image (right hemisphere), followed by recognition, labelling and reading the caption (mainly left hemisphere). The right brain simply divides up the image into "background", "object 1" (the region corresponding to the pipe) and "object 2" (the region corresponding to the caption). The left (literal) hemisphere then identifies the

regions in terms of shapes and patterns, ignores the background and reads the caption, which says "This is not a pipe", hence the reaction. Incidentally, the statement is of course correct: it is only a picture of a pipe.

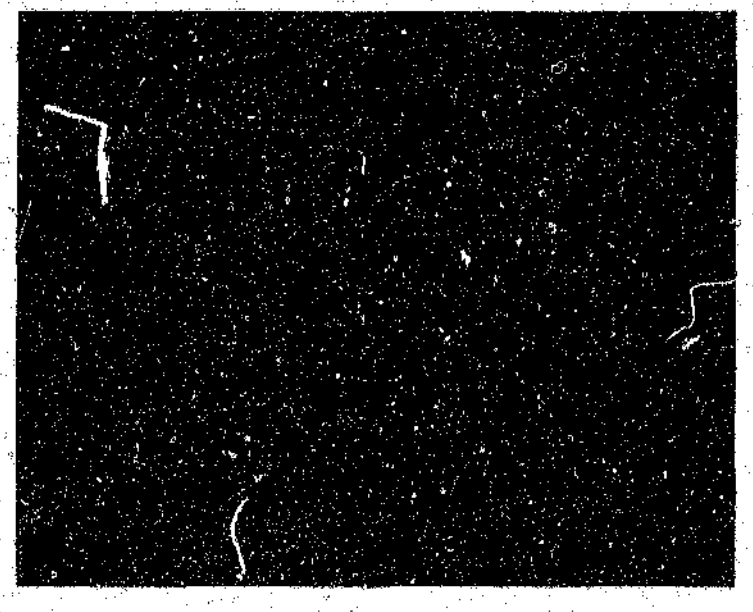


Figure 1.1. The use of words I, *René Magritte, 1928–29.*

Philosophy aside, this example serves to illustrate the difference between pattern recognition and segmentation and also demonstrates how segmentation can serve as a pre-process for more sophisticated operations.

1.2 Automatic threshold selection

Thresholding is intuitively the simplest form of segmentation. This involves the selection of a grey–scale value or set of values to split the range of grey–levels into subranges corresponding to the constituent regions of the image, subject to these regions being distinguishable on the basis of brightness. This approach can of course be equally well applied to other pixel–related features of the image, such as local gradient, colour (in the case of multiband images), variance, etc., depending on individual requirements, but it is most commonly a grey–level operation: most of the features mentioned (essentially texture measures) can in any case be easily re–expressed as grey–level values.

The selection of optimal thresholds has long been a problem of interest to researchers in image processing. Naturally one would wish to automate a process as conceptually primitive as this, necessitating the definition of some criterion of "goodness" of threshold that can be optimized. The proposal of and motivation for such *criterion functions* have resulted in an enormous amount of literature [Pratt, 1991; Gonzalez and Woods, 1992; Weszka, 1978;

Haralick and Shapiro, 1985; Sahoo *et al*, 1988].

The simplest form of thresholding is binary thresholding, whereby the grey–scale image is simply dichotomized into two distinct classes by means of a single grey–level threshold. These classes would usually correspond to "background" and "object/s", indicated by black (0) and white (maximum grey–level) in the output segmented image. It is usually a straightforward matter to extend this process to multi–threshold selection, so for the purposes of this thesis I confine my comments to binary thresholding. It must be noted, however, that binary thresholding is far from being generally applicable and is obviously limited in its applications. In general it would be most desirable to develop a method whereby the number of classes present in the image were also automatically determined, so that the entire process of segmentation from the decision of how many thresholds are required to the selection of the actual threshold values could be accomplished without human intervention. However, such a decision is often subjective and related to the specific application that the user has in mind.

1.3 Criterion functions and entropy

Some techniques, particularly older ones, are *ad hoc* and based largely on heuristics. Most current methods involve the optimization of a criterion function, as mentioned above, based on some property of the image or its grey–level histogram. These functions usually take the form of some measure of the error incurred by a particular partitioning or, alternatively, the improvement in some desirable quality of the regions (such as uniformity or entropy) resulting from the segmentation.

The criterion function on which this research is focused is the entropy of the image. Entropy in this context is usually the information–theoretic entropy measure of Shannon [Shannon and Weaver, 1949]. A simple interpretation of entropy is that it represents "information": by maximizing the entropy, then, the information content or gain is optimized. However, this relatively unsophisticated interpretation can lead to misunderstandings both in the application of the method and in its more subtle implications. The principle of maximum entropy is closely related to Bayesian analysis [Jaynes, 1988] and can be derived in a number of ways: one common approach is to derive it from the more general principle of maximum degeneracy [Jaynes, 1968; Frieden, 1980], leading to yet another interpretation.

In image processing this principle is most commonly applied to the problems of image restoration and reconstruction. For example, in the case of image restoration many possible "clear" images (solutions) exist which could have given rise to the degraded image we wish to restore. We wish to select the most probable solution. To this end, we invoke Bayes' Theorem which necessitates selection of a suitable "prior distribution" [Mood *et al*, 1974], quantifying what we know or expect about the solution prior to collecting any data. After gathering

data we use the theorem to determine the most likely "posterior distribution" (solution). The assignment of a prior distribution often involves making assumptions based on what we may "feel" is reasonable. The maximum entropy principle indicates selection of that solution which makes no unwarranted assumptions, such as smoothness or flatness, about the form of the restored image. Based on the axiomatic derivation of Shore and Johnson [Shore and Johnson, 1980], as well as the vast body of work produced by Skilling and others [Gull and Skilling, 1985; Skilling, 1988, 1989; Jones, 1989], it has become accepted by many that the maximum entropy solution (and the related minimum cross–entropy solution) to such problems of inductive inference is the only correct one.

If we view segmentation as a process similar to restoration (i.e. a binary image has been "degraded" to produce the grey–scale image), similar reasoning can be applied to motivate selection of a maximum entropy threshold. Based on arguments such as those outlined above, such a choice of threshold should be optimal.

Unfortunately, the task is not that simple. The definition of image entropy is a controversial issue. Techniques based on the image histogram, co–occurrence matrices and the image itself have been proposed. The choice of probability distribution results from the way the digital image is modelled. While the image is not itself a distribution of probabilities, it is clearly positive and additive: Skilling [Skilling, 1988] has shown that the entropy holds in general for such distributions and is not therefore limited to dealing only with probabilities.

To add to these problems, the form of the entropy expression itself is in dispute, with different forms being appropriate to different applications. These forms are also to some extent model dependent. What is therefore required is a suitable model of the image and/or the segmentation process to determine unambiguously a measure of entropy unique to a given image, which can then be used to optimize the segmentation or choice of threshold.

1.4 Thesis outline

The different forms of entropy, its relationship to Bayesian analysis and arguments for and against the principle are presented in Chapter 2.

Specifically in connection with image processing, different image models have been developed in an attempt to define what is meant by "image entropy". These are discussed in Chapter 3. Models which are hoped to be more effective for segmentation are also developed and improvements to existing models are proposed.

Many entropy–based threshold selection methods using various models have been proposed in the past [Pal and Pal, 1989a; Abutaleb, 1989; Kapur *et al*, 1985]. In Chapter 4 entropic threshold selection methods are investigated and new methods based on the models developed

in Chapter 3 are described. The results obtained using the various methods to binarize sample grey–scale images are shown in Chapter 5.

The problem of quantitative evaluation of thresholding methods is dealt with in Chapter 6. In general, evaluation to date has been largely subjective, based on criteria such as usefulness of results for a specific purpose or even aesthetic appeal. The results of Chapter 5 and the outcomes of quantitative evaluation techniques are discussed in Chapter 7.

Chapter 2

The maximum entropy method

The maximum entropy method (MaxEnt or MEM) is a controversial subject in many fields of research today. Common questions raised are "What do we mean by entropy?", "What form should this entropy take?", "Why maximize entropy and not some other function?" and "Is there more tea?". The discussion is often heated and has given rise to a great deal of confusing and, at times, contradictory literature. I by no means pretend to give a comprehensive overview here, nor to be an expert on the subject: the aim is simply to try to present a clear and simple description of the method and the arguments surrounding it, and to justify its selection as a criterion for image segmentation and related processes.

There are two main forms of the entropy expression, due to Shannon [Shannon and Weaver, 1949] and Burg [Burg, 1967; Frieden, 1980], respectively. Some controversy exists about which form is more "correct". However, the choice of entropy expression is usually dependent on the application and is often quite apparent. Kikuchi and Soffer have in fact proposed a model reconciling the two forms as extremes of a more general degeneracy principle [Kikuchi and Soffer, 1977]. This model is one of those described in Chapter 3. Other expressions for entropy have recently been proposed, but these seem to either add nothing new to the theory or lack the support of a convincing argument. The various forms of entropy are briefly discussed in Section 2.1 below.

The principle of entropy maximization is often mentioned in the same breath as Bayesian statistics [Jaynes, 1988; Press *et al*, 1992]. MaxEnt is in fact usually touted as the *best* choice for selection of a prior probability distribution in order to solve the problem of Bayesian estimation of the "correct" result (posterior probability distribution), the Bayesian solution being regarded as the most desirable [Skilling, 1990; Wernecke and D'Addario, 1977; Zellner, 1988]. By "prior probability distribution" we essentially mean a model of what we expect

our solution to comply with (i.e. constraints) based on such information as we may have available regarding the system generating the data. The link between the principle of maximum entropy and Bayesian statistics is discussed in Section 2.2.

It has been indicated above that the maximum entropy prior is the "best" one (it leads to the optimal Bayesian solution). This claim has been supported by both qualitative and quantitative arguments [Jaynes, 1968; Skilling, 1988, 1989, 1990; Shore and Johnson, 1980; Gull and Skilling, 1984, 1985]. Arguments against this claim have generally conceded that MaxEnt yields good results, but refute that it is the "best" (or the *only* correct) result, and that other measures might do as well [Press *et al*, 1992; Titterington, 1984]. Some of the arguments for and against entropy are presented in Section 2.3.

2.1 The entropy expression

Shannon entropy

The most generally accepted form of entropy was derived by Shannon [Shannon and Weaver, 1949] in connection with information theory. Given a discrete probability distribution p_i, $i = 1, \cdots, n$, the entropy is given by

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{2.1}$$

The i refers to a set of possible outcomes or states of a discrete information source modelled as a Markov process [Grimmett and Stirzaker, 1982]. Shannon's measure is intended as a measure of information gain, choice and uncertainty. Incidentally, the form of (2.1) is also that of the Gibbs entropy as defined and used in the field of statistical mechanics: theoretical analysis of this concept and the (to some extent) parallel of the Boltzmann entropy are presented in the excellent papers by Jaynes and by Gull [Jaynes, 1965; Gull, 1991].

Shannon points out a number of interesting properties of (2.1), including:

1. $H = 0$ if and only if $p_i = 0 \quad \forall\, i \neq j$, $p_j = 1$, where j can indicate any position in the distribution (note that $0 \log 0$ is defined to be zero). Otherwise H is positive. This makes intuitive sense, since $p_j = 1$ indicates certainty of the outcome, and the information gained by the occurrence of event j is thus zero.

2. For a given number of discrete states n, H is a maximum when all the p_i are equal, i.e. $p_i = 1/n \quad \forall\, i = 1, \cdots, n$. Intuitively this is the most uncertain situation: all outcomes are equally likely, making accurate prediction impossible.

In the case of continuous distributions, Jaynes [Jaynes, 1968] has pointed out that the con-

tinuous form of (2.1),

$$H = -\int p(x) \log p(x)\, dx$$

does not hold, as it is not the result of any derivation and furthermore lacks invariance under a change of variables $x \to y(x)$. The corresponding expression in the continuous case was shown to be

$$H = -\int p(x) \log \frac{p(x)}{m(x)}\, dx$$

where $m(x)$ is in general an "invariant measure" function proportional to the limiting density of discrete points, i.e. related to our sampling of x. The integrals in both expressions are assumed to cover the full range over which the distributions are defined. Skilling [Skilling, 1986] has since discussed a "discretized" form of this expression

$$H = -\sum_{i=1}^{n} p_i \log \frac{p_i}{m_i} \tag{2.2}$$

Entropy is a relative measure. In our case, m can be interpreted as a model based on known constraints (prior information), relative to which the entropy, given "data" (posterior information) yielding a new distribution p, is measured. This is a more general expression, (2.1) representing the special case where m is a uniform prior distribution.

The maximum entropy formalism attempts to maximize (2.1) or (2.2), subject to known constraints about the parameters we are trying to estimate ($p_i,\ i = 1, \cdots, n$). Effectively, by maximizing the entropy, we are attempting to maximize our information gain or, equivalently, arrive at a solution which fits our prior knowledge but makes no assumptions beyond what is known. In the absence of any prior information, the maximum entropy distribution is simply the uniform distribution, as indicated by Shannon's point (2) above. It can thus be aptly viewed as an application of Laplace's principle of insufficient reason, which states that the uniform distribution is the most unbiased when one has no prior knowledge regarding a probabilistic event.

Kullback cross-entropy

Closely related to the relative entropy (2.2) is the cross-entropy [Kullback, 1959; Shore and Johnson, 1980], also sometimes referred to as the Kullback–Leibler number, discrimination information, directed divergence or even relative entropy. This is used as a non-metric form of distance measure, measuring the difference between two probability distributions:

$$H_\times = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i} \tag{2.3}$$

where, for our purposes, q and p denote the posterior and prior distributions, respectively. In a sense, it measures the amount of information required to change our prior model of the image, p, into a new model, q, given additional information (data) [Shore and Johnson, 1980]. The similarity to (2.2) is clear. Since this takes the form of a distance measure, the guiding principle underlying the use of (2.3) is that of minimum cross-entropy. This is clearly equivalent to maximization of relative entropy, albeit with slightly different philosophical implications.

Burg entropy

Another common expression for entropy is the so-called Burg entropy [Burg, 1967; Frieden, 1980],

$$H = \sum_{i=1}^{n} \log p_i \tag{2.4}$$

This form is related directly to problems in radio astronomy and is simply the physical entropy of an electromagnetic field in the limit of many photons per mode. In radio astronomy, the signal obtained is the Fourier transform of the actual data consisting of two-dimensional autocorrelation values spaced over the antenna's bandpass region, i.e. the power spectrum of the data. Burg did not set out to seek a maximum entropy solution, but rather assumed one at the outset. This was justified by the reasoning that the autocorrelation data should have occurred from a maximally smooth electromagnetic field (see Shannon's point (2) above). In addition, maximum entropy implies maximum admitted ignorance about the process and thus a maximally unbiased or conservative representation of the electromagnetic field.

As indicated before, the forms (2.1) and (2.4) can be reconciled as extremes of a more general principle, based on an image model discussed in Chapter 3 [Kikuchi and Soffer, 1977].

Entropy information

Pan and Harris [Pan and Harris, 1990] have defined a measure of "entropy information" based on the Shannon equation (2.1), in connection with the specific application of discretizing (or segmenting) a measurement. The essential details of the theory are presented here, the specific application being addressed in Chapter 4.

Suppose we have obtained a measurement $\mathbf{x} \in [a, b]$, $a < b$, on a sample of size N (e.g. an image consisting of N pixels). A qualitative variable $\mathbf{y} \in \{y_1, \cdots, y_s\}$ is selected as an "objective feature" (e.g. correlations between samples taking on any of s discrete values) observed on the same sample as our measurement. The entropy, $H(\mathbf{y})$, of $\mathbf{y}$ and the conditional

entropy, $H_x(\mathbf{y})$, of $\mathbf{y}$ on $\mathbf{x}$ are defined as

$$H(\mathbf{y}) = -\sum_{j=1}^{s} p(y_j) \log p(y_j)$$

$$H_x(\mathbf{y}) = -\sum_{j=1}^{s} p(y_j|\mathbf{x}) \log p(y_j|\mathbf{x}) \tag{2.5}$$

where $p(a|b)$ should be read as "the probability of a given b". The relative entropy information on $\mathbf{y}$ contained in $\mathbf{x}$ is defined as

$$\rho(\mathbf{x} \to \mathbf{y}) = \frac{\Delta I(\mathbf{x} \to \mathbf{y})}{H(\mathbf{y})} \tag{2.6}$$

where $\Delta I(\mathbf{x} \to \mathbf{y}) = H(\mathbf{y}) - H_x(\mathbf{y})$ is referred to as the absolute entropy information. It measures how much of the uncertainty about $\mathbf{y}$ is reduced when $\mathbf{x}$ is realized. The mean the relative entropy information can be calculated as

$$\overline{\rho}(\mathbf{x} \to \mathbf{y}) = \int \rho(\mathbf{x} \to \mathbf{y}) f(x)\, dx \tag{2.7}$$

where $f(x)$ is the probability density function of $\mathbf{x}$. If $\mathbf{x}$ is a digital image, a discrete approximation to $f(x)$ could be obtained from the grey-level histogram. If we discretize the (usually continuous) measurement into n discrete subranges (e.g. we digitize the continuous image, resulting in a set of discrete grey-levels), we can write the discrete form

$$\overline{\rho}(\mathbf{x} \to \mathbf{y}) = 1 + \frac{\sum_{i=1}^{n} \sum_{j=1}^{s} p(y_j|x_i) p(x_i) \log p(y_j|x_i)}{H(\mathbf{y})}$$

where $p(x_i)$ is a discrete approximation to $f(x)$. Note that the notation can be a little confusing: x_i is the discrete value representing subrange i of the continuous measurement. It is **not** the ith spatial sample of $\mathbf{x}$. Using Bayes' Theorem (see Section 2.2),

$$p(y_j|x_i) = \frac{p(x_i|y_j) p(y_j)}{p(x_i)}$$

we can write

$$\overline{\rho}(\mathbf{x} \to \mathbf{y}) = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{s} p(x_i|y_j) p(y_j) \log \frac{p(x_i|y_j) p(y_j)}{p(x_i)}}{\sum_{j=1}^{s} p(y_j) \log p(y_j)} \tag{2.8}$$

This function has the following properties [Pan and Harris, 1990]:

(1) $0 \le \overline{\rho}(\mathbf{x} \to \mathbf{y}) \le 1$.

(2) $\overline{\rho}(\mathbf{x} \to \mathbf{y}) = 0$ if and only if $\mathbf{y}$ is statistically independent of $\mathbf{x}$.

(3) $\overline{\rho}(\mathbf{x} \to \mathbf{y}) = 1$ if and only if $s \leq n$ and $\mathbf{y}$ is a deterministic function of $\mathbf{x}$, $\mathbf{y} = \phi(\mathbf{x})$

(4) $\overline{\rho}(\mathbf{y} \to \mathbf{x}) = 1$ if and only if $n \leq s$ and $\mathbf{x}$ is a deterministic function of $\mathbf{y}$, $\mathbf{x} = \psi(\mathbf{y})$

(5) $\overline{\rho}(\mathbf{x} \to \mathbf{y}) = \overline{\rho}(\mathbf{y} \to \mathbf{x}) = 1$ if and only if $n = s$ and there exists a deterministic function γ such that $\mathbf{y} = \gamma(\mathbf{x})$ and $\mathbf{x} = \gamma^{-1}(\mathbf{y})$

Generalized entropy

Kesavan and Kapur [Kesavan and Kapur, 1989] introduce a "generalized maximum entropy principle" which does not limit itself necessarily to the use of the logarithmic forms above. Instead, a general convex function $\phi(\cdot)$ is defined, so that

$$H_G = -\sum_{i=1}^{n} \phi(p_i) \tag{2.9}$$

becomes the measure of "entropy" to be maximized. Interestingly, on deriving the precise form of $\phi(\cdot)$ under different statistical models, the forms (2.1), (2.4) or the general principle of Kikuchi and Soffer (Chapter 3) result nonetheless.

Exponential entropy

Pal and Pal [Pal and Pal, 1989b] have proposed an exponential form for the entropy expression,

$$H_{PP} = \frac{1}{n} \sum_{i=1}^{n} p_i e^{1-p_i} \tag{2.10}$$

specifically for application to images. However, their choice of an exponential form seems to be based more on distaste for the usual logarithmic forms (2.1) and (2.4) than a sound theoretical basis. Their arguments against the logarithmic form, briefly, are that: (1) as $p_i \to 0$, $\log p_i \to \infty$ and (2) if $p_i = 1$, then $\log p_i = 0$, where $\log p_i$ is viewed as a measure of information gain. On a purely intuitive basis it should be fairly acceptable that there is no information to be gained in the occurrence of an event whose probability is unity and, conversely, the occurrence of an event whose probability is assumed to be zero should at least raise an eyebrow.

2.2 Bayes' Theorem and its relationship to the principle of maximum entropy

In Bayesian analysis probability is given a broader interpretation than that of being regarded merely as a kind of normalized frequency of occurrence [Press *et al*, 1992; Mood *et al*, 1974;

Loredo, 1989]. Instead, it is often regarded as a measure of the "likelihood", on a scale from 0 to 1, of the truth of a proposition or hypothesis rather than a repeatable *event*. This likelihood can then be modified when given new data pertaining to the process about which the hypothesis is made. Thus the problem is that of making inferences about parameters describing a model (hypothesis), not directly observable, given data D, say, which are indirectly related to the parameters we wish to determine. Let us denote the possible sets of parameters $A, B, C, \cdots$. The only way to express such inferences is in terms of the conditional probabilities $P(A|D), P(B|D), \cdots$ [Skilling, 1990], where we read $P(X|Y)$ as "the probability of X given data or information Y".

Let M represent any particular hypothesis $A, B, C, \cdots$. We need to determine the probability distribution $P(M|D)$ as a function of M. Unfortunately, all the data will give us is the "likelihood" $P(D|M)$, i.e. the likelihood that the data could be produced by M. If we were "frequentist" statisticians we could embed the data D in a sample space of other (non–existent) data sets $\{D_1, \cdots, D_N\}$ which could have been (but weren't) observed. We would then try to estimate the set of probabilities $P(D_i|M)$ that the data sets D_i would be observed if the hypothesis M were true. Loredo [Loredo, 1989; Pendock, 1993b] presents a critique of frequentist methods, one of the many criticisms being that the random variable in this case is the hypothesis, not the data. The Bayesian approach is almost exactly the opposite: we consider the class of all hypothetical distributions $\{M_1, \cdots, M_K\}$ consistent with the received data D and attempt to find the most probable solution.

We represent the probability that "X *and* Y are true" by the notation $P(X \cap Y)$, given by

$$P(X \cap Y) = P(X)P(Y|X) \tag{2.11}$$

By Aristotelian logic the statements "X and Y are true" and "Y and X are true" are identical, so

$$P(X \cap Y) = P(X)P(Y|X) = P(Y)P(X|Y) = P(Y \cap X)$$

which leads us directly to Bayes' Theorem

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)} \tag{2.12}$$

Bretthorst [Bretthorst, 1990] points out that in Bayesian probability theory *all* probabilities are conditional, implying that a probability such as $P(X)$ standing for the "likelihood" of X does not make sense unless the evidence (the prior information, I) on which it is based is given. The correct form, then, is $P(X|I)$ and we can write Bayes' Theorem [Loredo, 1989] as

$$P(X|Y \cap I) = \frac{P(X|I)P(Y|X \cap I)}{P(Y|I)} \tag{2.13}$$

The meaning of (2.13) is no different to that of (2.12): we have simply explicitly included our prior information.

By invoking this theorem (and including our prior information explicitly), we can solve the inference problem outlined above as

$$P(M|D \cap I) = \frac{P(M|I)P(D|M \cap I)}{P(D|I)} \tag{2.14}$$

This tells us how to manipulate probabilities once they have been assigned. However, there is nothing within the theory to tell us how to determine or assign our probabilities.

The individual interpretation of each of the probabilities in (2.14) sheds some light on this. Pendock [Pendock, 1993a] expresses the theorem in words as

$$Posterior = \frac{Prior \times Likelihood}{Evidence}$$

For any single data set D, the total "evidence" (data and information) $P(D|I)$ is a constant and may be ignored [Pendock, 1993b; Skilling, 1990; Loredo, 1989]. We are given the likelihood $P(D|M \cap I)$ [Skilling, 1990]: in problems such as image restoration, it is related to the error incurred by assuming M to be the "correct" result. This often takes the form of a discrepancy measure, such as the Pearson χ^2, between the received data D and a version of M modified by what we know of the system, I [Jaynes, 1986; Gull and Skilling, 1984]. A similar interpretation can be applied to segmentation. The real problem now is to determine the prior distribution (the likelihood of our hypothesis before considering any data) $P(M|I)$, which has been left unspecified: Jaynes [Jaynes, 1986] has pointed out that this is due to our still being caught up in the "bad habits of orthodox statistics" whereby we effectively claim more knowledge than we really have for the likelihood above (e.g. knowledge of noise statistics) and less than we really have for the prior. This would make it easy to dismiss $P(M|I)$ as being uninformative. However, the strength of the Bayesian approach lies precisely in its ability to combine the data with the prior information.

In many problems the prior distribution takes the form of constraints, e.g. we may desire the smoothest solution for some reason. When we assume an *entropic prior*, we are assuming maximum ignorance about the system. Thus any concrete knowledge is built in and no assumptions are made. The form of the entropic prior is related to the Shannon entropy (2.1). In fact, Skilling [Skilling, 1989; Pendock, 1993a] has presented a convincing argument

for a *general* entropic prior of the form

$$\begin{aligned} P(M|I) &= P(M|m,\alpha,I) \\ &= P(p_1,\cdots,p_n|m,\alpha,I) \\ &\propto \exp\left(\alpha\sum_{i=1}^{n} p_i - m_i - p_i\log\frac{p_i}{m_i}\right) \\ &= \exp\big(\alpha H(M,m)\big) \end{aligned} \tag{2.15}$$

where $H(M,m)$ denotes the relative entropy (compare this with (2.2)), m is a model defining a component–wise magnitude for M and α is the inverse of the expected spread of M about m [Pendock, 1993a]. The p_i are the discrete components of the distribution M. Exponentiating the entropy effectively gives us the degeneracy or multiplicity of the prior distribution (the number of ways in which it can be realized), subject to our known constraints. Thus by maximizing the entropy we reach the prior that is "most likely" in the sense that we are choosing the most degenerate solution [Jaynes, 1968, 1986]. This is one of the most common qualitative arguments in favour of using maximum entropy.

The order in which we apply the entropy principle and Bayes' Theorem is sometimes reversed, depending on application and the kind of information we have available. Pan and Harris [Pan and Harris, 1990] do this in their threshold selection technique for the optimal discretization (segmentation) of geological data, where a measure of relative entropy information or correlation (2.8) is maximized. The measure has been described in Section 2.1 and its application to thresholding appears in Chapter 4.

Bayes' Theorem allows one to *determine* an unknown probability distribution by re–expressing it in terms of other more easily obtainable distributions. The maximum entropy principle *assigns* a probability distribution to a variable, based on selection of the most likely distribution that fits our prior knowledge (not belief). Conceptually these are clearly very different. However, mathematically they have much in common [Jaynes, 1988] and should be seen as being complementary: the entropy enables one to select a "sensible" prior in order that Bayes' Theorem may determine the best possible posterior distribution to fit both the data and the prior information. Alternatively, Bayes' Theorem allows one to express an unknown probability distribution in terms of others that *are* known, which can then be substituted into the entropy expression.

2.3 Why maximum entropy?

Maximum entropy, when understood intuitively as a maximization of degeneracy or information or as a means to avoid making unfounded assumptions, seems quite a compelling choice.

Many such qualitative arguments for it (and against it) have been put forward [Narayan and Nityananda, 1986]. There is little doubt that in using this principle we get good results for philosophically correct reasons. It is often just the philosophy that swings one in favour of maximum entropy: e.g. the "kangaroo problem" as discussed extensively by Gull, Skilling and others [Jaynes, 1986; Gull and Skilling, 1984; Skilling, 1986] can be solved from a standard statistical viewpoint: however, it involves making unwarranted assumptions about the data (we assume that the data are uncorrelated) whereas the maximum entropy/Bayesian approach reaches the same result *without* any such assumptions (the data cannot be assumed to be correlated). The difference here is subtle but significant.

The arguments against entropy and Bayesian methods as a whole are often presented in a similar hand–waving fashion. Press *et al*, for example, suggest that the solution of inverse problems has gained a Bayesian flavour largely as an "accident of history" [Press *et al*, 1992]. They suggest that "an outsider looking only at the equations that are actually solved, and not at the accompanying philosophical justifications, would have a difficult time separating the so–called Bayesian methods from the so–called empirical ones". This is true, of course. The outsider will also gain no insight into or understanding of the problem and will have every right to remain unconvinced about the correctness of the results: the point is, again, that the correct results are obtained *for the right reasons.*

Uniqueness of entropy as a prior distribution selector

The most controversy rages about the statement that the general form of the relative entropy (the exponent in (2.15))

$$H(M,m) = \sum_{i=1}^{n} p_i - m_i - p_i \log \frac{p_i}{m_i} \tag{2.16}$$

gives us the only correct choice of prior distribution for problems of inference [Skilling, 1989, 1990, 1901; Shore and Johnson, 1980]. Press *et al* [Press *et al*, 1992] argue that there is nothing universal about the form (2.1) (a special case of (2.16)), holding up the example of the Burg entropy (2.4) as a working alternative. The origins of this form and the way it is used are neglected. It is suggested that the usefulness of the entropy measure stems primarily from its nonlinearity. This could equally well be achieved by a function like $f(p) = -\sqrt{p}$ which has no information–theoretic justification. This is precisely the point: entropy *is* justifiable.

Shore and Johnson showed that both the principle of maximum entropy and Kullback's related principle of minimum cross–entropy are uniquely correct methods for inductive inference by means of an axiomatic derivation of these principles [Shore and Johnson, 1980]. Based on this work, Skilling has since determined the more general form (2.16) as the true (and only correct) guiding principle for such methods of inference [Skilling, 1989, 1990, 1991, 1988].

Donoho *et al* [Donoho *et al*, 1992] have also taken issue with the maximum entropy formalism, claiming to show that it is *only* useful in the case of near–blackness (e.g. astronomical images) for both signal–to–noise enhancements and superresolution (capability to resolve to better than the Rayleigh limit for linear methods). They also point out that other methods may exploit near–blackness to a greater extent, implying therefore that entropy may be good in certain situations but there is no real reason to prefer it.

In his reply in the same paper, Gull has indicated that some misunderstandings have crept in through translating the maximum entropy principle into "the language used by statisticians". In translating the findings back to a Bayesian framework, the following come to light: the authors treat maximum entropy as a regularization method for inversion, a superresolution method and a noise reduction (smoothing) process. As Gull points out, it is none of the above, although it can have these properties. The analysis of its performance in these roles is therefore valuable but says nothing about its role as a selector of prior distributions. The misunderstanding leads to inappropriate generalizations such as the attempt to generalize the form of the entropy expression to allow use of, for example, an l_1 norm, as well as *ad hoc* choices for the regularization constant.

The assumption of data independence

Another argument against entropy is based on more practical grounds: the standard form of the entropy expression (2.1) is based on the assumption that samples from the prior distribution (pixels, grey–level probabilities, etc.) are independent of each other [Titterington, 1984; Press *et al*, 1992]. Considering our particular application, this means that the image pixels can be randomly rearranged without affecting the entropy value. Clearly one would not usually expect adjacent pixels to be spatially independent of each other: this is counter–intuitive simply by virtue of the fact that *we* see and make sense of the image as a whole. Unfortunately, the precise nature of this dependence or correlation varies from image to image. If we cannot quantify this inter–dependence, the maximum entropy formalism comes to our rescue by suggesting that we cannot assume correlation if we don't know its precise form: therefore we do *not* assume correlation. Of course, in such a case it is numerically (but *not* philosophically) equivalent to assuming the data to be uncorrelated.

While this argument may make us feel a little better about our independent pixels, it would nevertheless be useful to attempt to build such spatial dependence as we can discern into our formula. This could be done in a variety of ways. Skilling, in his reply to Titterington [Titterington, 1984], suggests directly incorporating spatial information in the measure m_i of (2.2) or (2.16) in the form of inter–pixel correlations or other context–dependent information. This is essentially the method used with some success by Liang [Liang, 1988], although it is

not explicitly formulated as a maximum entropy method. Similarly, in the method proposed by Pan and Harris [Pan and Harris, 1990] the "objective feature" chosen (see Section 2.1) may be a context-related feature such as local grey-level variance or mean in an image.

Journel and Deutsch [Journel and Deutsch, 1993] address the issue by developing an expression for spatial entropy as a measure of spatial disorder similar to spatial variance. The resulting formula is reminiscent of the form of an autocorrelation function: a bivariate entropy measure is proposed, with the entropy value dependent on the spatial distance or *lag*, λ, between two random variables (images) $x(k)$ and $x(k+\lambda)$. The measure is normalized to give a relative measure lying in the range 0 to 1. The image entropy is defined as the average value of this relative measure over all possible lags. A more detailed description of the technique is given in Chapter 3.

Another approach currently gaining favour is to characterize the image as a Markov Random Field [Pratt, 1991; Geman and Geman, 1984; Mardia, 1989; Marqués and Gasull, 1993]. Geman and Geman have presented a comprehensive introduction to the theory and underlying concepts of this approach with a view to Bayesian image restoration [Geman and Geman, 1984]. Stated simply, such an image model allows us to take into account pixel inter-relationships within a pre-defined neighbourhood: thus we are neither assuming spatial independence, nor unnecessarily taking account of weak correspondences between pixels that are far apart. Thus the advantage of using a Markov model is that we can determine prior distributions which include spatial information based on neighbourhoods small enough to retain computational feasibility, but still rich enough to model and process interesting classes of images. Mardia [Mardia, 1989] presents a brief discussion of these methods applied to image analysis and, in particular, a form of segmentation.

The general similarity of the spatial methods outlined above [Journel and Deutsch, 1993; Geman and Geman, 1984] to methods using, for example, the entropy of the image autocorrelation function, co-occurrence matrices or a distributions quantifying the relationship between pixel grey-level and some local property defined over a small neighbourhood is apparent. Models and threshold selection techniques using various context-sensitive methods such as these are introduced in the next two chapters.

Chapter 3

Digital image models and entropy

Models of digital images have been widely used as the basis for image processing techniques such as restoration, enhancement and segmentation [Frieden, 1972, 1973, 1980; Jain and Angel, 1974; Marqués and Gasull, 1993]. In particular, statistical models have played an important role in providing a justifiable basis for the use of the principles of maximum entropy (MaxEnt) [Skilling, 1986, 1989, 1991; Jones, 1989; Pal and Pal, 1991a; Seth and Kapur, 1990] and minimum cross–entropy (MinxEnt) [Kullback, 1959; Shore and Johnson, 1980; Kapur and Kesavan, 1990] in these techniques.

Various models of images are used in image processing and analysis, with the assumption of one model or another often being implicit. In investigating and applying the principle of maximum entropy, it is useful to consider the model and the assumptions one is making about the image explicitly. It should be stressed that a digital image is only a rough approximation to the ideal continuous image, particularly with regard to its grey–scale. In Section 2.1 a model based on the physical process of image capture is described and the effects of digitization are discussed.

The maximum entropy principle has been most commonly applied to problems such as the restoration and reconstruction of images. The models used, rather than being founded on the imaging process, are based directly on the nature of the received digital image, i.e. a two–dimensional discrete distribution of positive integer grey–levels. These fall mostly into one of two categories: (1) the image itself is regarded as a probability distribution, where each pixel's grey–level is a measure of the likelihood of illumination reaching that point in the image, and (2) the image grey–level histogram is used, with each bin simply representing the probability of occurrence of a particular grey–level. The former appears more desirable in the sense that it seems to take information about the spatial arrangement of pixels into

account: clearly, many different images could have identical histograms, although it could be argued that such images then correctly have the same entropy. However, examination of the entropy expression (see Chapter 2) shows that the entropy is independent of the order in which the individual samples of a discrete signal or distribution are taken, i.e. the pixels or histogram bins are not assumed to be spatially dependent. This apparent shortcoming was addressed at some length in Chapter 2 and will henceforth only be considered as a motivation for attempts to incorporate contextual information.

Statistical models of digital images can usually be based on parallel models in the fields of statistical thermodynamics and mechanics. The classical methods were largely developed independently by Boltzmann in Germany and Gibbs in the USA. Modifications were introduced with the advent of the quantum theory by Bose, Einstein, Fermi and Dirac. Details of these methods can be found in most standard references on statistical physics [Sears and Salinger, 1975; Tolman, 1979; Toda *et al*, 1983]. Some models based on these have been described by Frieden [Frieden, 1980]. An overview of these models and brief outlines of their physical and/or statistical derivations are presented and discussed in Section 3.2. Models based on other distributions or giving a different perspective on the models of Section 3.2 are described in Section 3.3. The issue of the incorporation of spatial information is also addressed here.

3.1 The Poisson model

Many current image processing techniques implicitly assume a Gaussian model for the illumination of receptors in an imaging device. Based on this assumption, under uniform illumination conditions one would expect the distribution of light quanta (photons) over the receptor array to be Gaussian. Simply equating quanta with grey–levels in a digital image, one would expect a Gaussian grey–level histogram centred at the mean image grey–level. Such a model has been previously used to address the image thresholding problem [Kittler and Illingworth, 1986; Albregtsen, 1993]. However, the assumption that grey–levels can be directly equated with photon–counts is not strictly correct. This is dealt with in more detail below.

An older model which is currently regaining favour [Pal and Pal, 1991a; Skilling, 1990] is based on the Poisson distribution. Assuming that we are dealing in this case with an optical image, it has been found that from a quantum viewpoint the number of incident light quanta (photons) on an imaging array is governed by Poisson statistics with a sufficient degree of validity for our purposes [Dainty and Shaw, 1974]. If the average number of quanta per receptor is q, say, then the number of receptors with r quanta is proportional to

$$p_r = \frac{q^r e^{-q}}{r!} \tag{3.1}$$

and

$$\sum_{r=0}^{\infty} p_r = 1$$

where p_r is the probability of r quanta reaching a receptor.

In practice the number of quanta registered by the receptor is subject to its sensitivity and its saturation level. There exists a threshold number of quanta T below which the receptor is unable to detect any incident light. If the image field is too bright the receptor saturates, corresponding to an upper threshold S above which no additional quanta can be sensed. The mean recorded number of quanta is thus

$$q' = \sum_{r=0}^{\infty} r' \frac{q^r e^{-q}}{r!} \tag{3.2}$$

where

$$r' = \begin{cases} 0 & r < T \\ r - T & T \le r < S \\ S - T & r \ge S \end{cases}$$

as opposed to the true mean

$$q = \sum_{r=0}^{\infty} r \frac{q^r e^{-q}}{r!} \tag{3.3}$$

In the case of a digital image each pixel can be assumed to correspond to a receptor. The observed grey-level is a measure of the incoming light intensity (number of quanta) at each pixel position. Since there are only a relatively small number of distinct grey-levels (usually 256 for a 1 byte-per-pixel digitization) compared with the number of quanta making up the illumination, it is reasonable to assume that many quanta are required to form a single "unit" of grey-level. The grey-level is thus a measure of the number of "packets" of quanta reaching each pixel: a low-resolution approximation to the actual quantum count. If we denote the number of quanta per grey-level by Δr then the mean recorded grey-level is given by

$$\mu = \sum_{r=0}^{\infty} g \frac{q^r e^{-q}}{r!} \tag{3.4}$$

where the grey-level g (number of quantum packets counted) is

$$g = \begin{cases} 0 & r < T \\ 0 & T \le r < T + \Delta r \\ 1 & T + \Delta r \le r < T + 2\Delta r \\ 2 & T + 2\Delta r \le r < T + 3\Delta r \\ \vdots & \vdots \\ n-1 & T + (n-1)\Delta r \le r < T + n\Delta r \\ n-1 & r \ge T + n\Delta r \end{cases}$$

The upper saturation level is given by $S = T + n\Delta r$, n the number of discrete grey–levels.

If we assume the number of quanta per grey–level Δr to be very much larger than the sensitivity threshold of the receiving device, $\Delta r \gg T$, then this lower threshold can be ignored. The problem of saturation should be addressed at the image capture stage (adjustment of aperture or shutter speed to avoid overexposure) and cannot usually be addressed after digitization. The grey–level is then given by

$$g = \begin{cases} 0 & r < \Delta r \\ 1 & \Delta r \leq r < 2\Delta r \\ 2 & 2\Delta r \leq r < 3\Delta r \\ \vdots & \vdots \\ n-1 & r \geq (n-1)\Delta r \end{cases}$$

with *a priori* grey–level probabilities

$$p_g = \sum_{r=g\Delta r}^{(g+1)\Delta r-1} \frac{q^r e^{-q}}{r!} \tag{3.5}$$

It has been incorrectly assumed in some earlier work [Pal and Pal, 1991a] that the Poisson model can be extended to the grey–levels themselves. The digital image corresponding to a uniformly illuminated receptor array would then have grey–level probabilities

$$p_g = \frac{\mu^g e^{-\mu}}{g!}$$

where p_g is the probability of occurrence of grey–level g and μ is the mean grey–level. Clearly, under uniform illumination we would not expect the variance of the quanta per receptor to be high enough to cause a significant variation in grey–level over the image. The grey–level histogram should be regarded as a low–resolution approximation to the quantum histogram and it is thus reasonable to expect a single grey–level probability spike at $g = \mu$ approximating the corresponding Poisson probability distribution of quanta centred on $r = q$.

It should be emphasized that this applies to a uniformly illuminated image only. In a general image, each pixel has an associated mean illumination q and a Poisson distribution which becomes apparent when multiple images of the same scene are captured over time.

3.2 The image as a distribution

Unlike the Poisson image model, based on the image capture process, digital image models are generally based directly on the statistics of the data representing the image after digitization. Some of these models are described below. Note that in this discussion the term "photons" is

often used to indicate units of grey–level rather than the actual light quanta of the physical process. It will be explicitly indicated when this is not the case.

The monkey model

A digital image can be regarded as a probability distribution with each pixel's grey–level representing an estimate of the number (or probability) of photons reaching that point [Frieden, 1980]. We imagine the scene to be imaged as being made up of a fixed number of photons (unit grey–levels) G and our initial image as an empty grid or raster consisting of N cells (pixels). The G photons are allocated one at a time among the N cells with uniform spatial probability [Frieden, 1972]. This model has been dealt with in some depth and is often referred to as the "monkey model" due to the analogy of a group of monkeys randomly throwing G balls at a 2–dimensional array of N boxes to form the image [Gull and Daniell, 1978; Skilling, 1986; Jaynes, 1986]. Note that for the purposes of this particular model equating grey–levels and photon–counts will not affect its performance. A diagram of the model for $G = 20$ photons is shown in Figure 3.1: for clarity it is represented in one dimension.

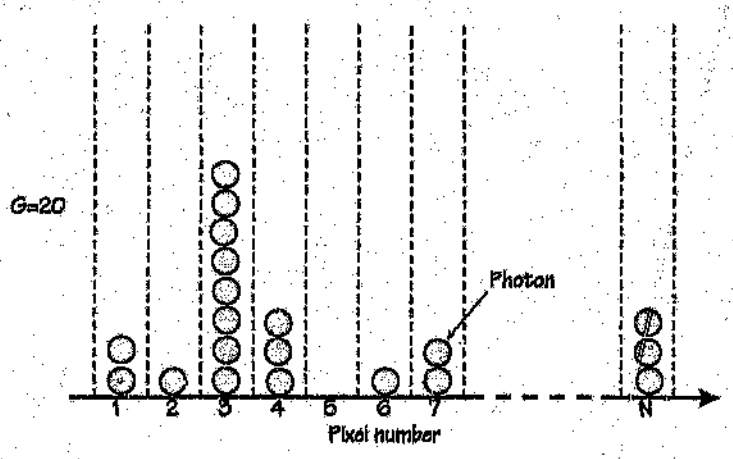


Figure 3.1 *Photon/unit grey–level allocation model ("monkey model") in 1 dimension. G photons are dealt out among N cells (pixels) with uniform spatial probability. The model obeys a simple positivity constraint.*

Jaynes' *principle of maximum degeneracy* [Jaynes, 1968] states that, subject to *a priori* constraints, the most degenerate image to be formed in this fashion is the most likely to occur. In other words, subject to any prior knowledge we may possess about the system, we wish to maximize the degeneracy of our estimate of the image. Our only *a priori* constraint

for this model is that of non–negativity, i.e. $g_i \geq 0, \quad 1 \leq i \leq N$, where g_i is the grey–level of pixel i. The photons can be considered to be indistinguishable and any number can occupy a given cell. Any given image has a corresponding non–unique grey–level histogram. The number of ways W that a general image $(g_1, g_2, \cdots, g_N)$ can be formed is given by the Boltzmann law

$$W(g_1, \cdots, g_N) = \frac{G!}{g_1! \cdots g_N!} \tag{3.6}$$

Following Jaynes' rationale, we set $W = \max$ or, equivalently, $\log W = \max$. Using Stirling's approximation $\log m! \simeq (0.5 + m)\log m - m + 0.5 \log 2\pi$ we find that the quantity to be maximized has the same form as the Shannon entropy. We maximize the expression

$$H(g_1, \cdots, g_N) = -\sum_{i=1}^{N} g_i \log g_i \tag{3.7}$$

Although the g_i are not probability values as such, the entropy expression has been shown to hold for positive, additive distributions in general [Skilling, 1988]. Note that in order to compare the entropies of different images the grey–levels should be normalized to "probability" values by dividing each pixel value by $G = \sum_i g_i$ or, equivalently, imposing the constraint that $\sum_i g_i = G = \text{constant } \forall$ images. In terms of probabilities, the expression for entropy [Shannon and Weaver, 1949] is given by (2.1), or

$$H = -\sum_{i=1}^{N} p_i \log p_i \quad , \; p_i = \frac{g_i}{G}$$

Jaynes [Jaynes, 1968; Skilling, 1986] has pointed out that entropy is a *relative* measure given by (2.2), i.e.

$$H = -\sum_{i=1}^{N} p_i \log \frac{p_i}{m_i}$$

where m_i is a model describing the prior information available. (2.1) implicitly assumes a uniform prior whereby all events are equally likely. This assumption is valid when we have no position–specific prior knowledge.

Skilling [Skilling, 1986] presents the following argument against this model: we can represent the selection of the maximum entropy image pictorially (Figure 3.2 shows this selection for linear constraints on a 3–pixel image).

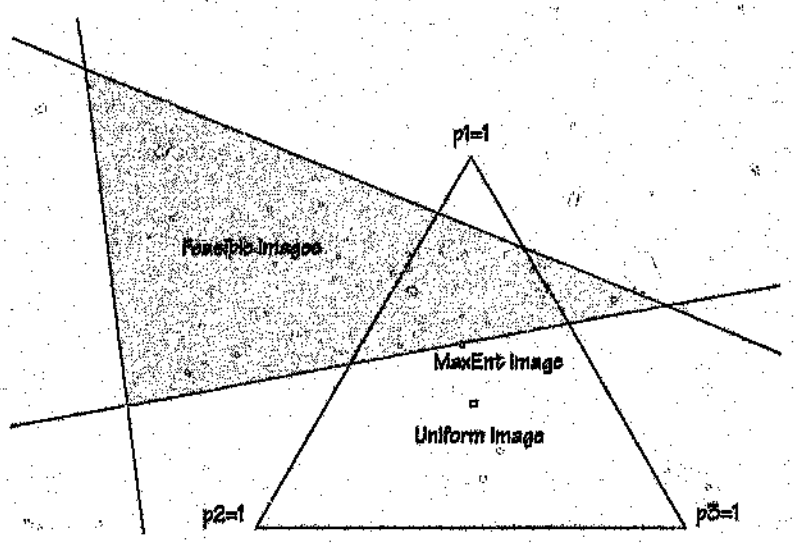


Figure 3.2. *Maximum entropy with hard linear constraints.*

When the constraints are non–linear, there may be more than one local maximum of entropy if the set of feasible images is not convex (Figure 3.3).

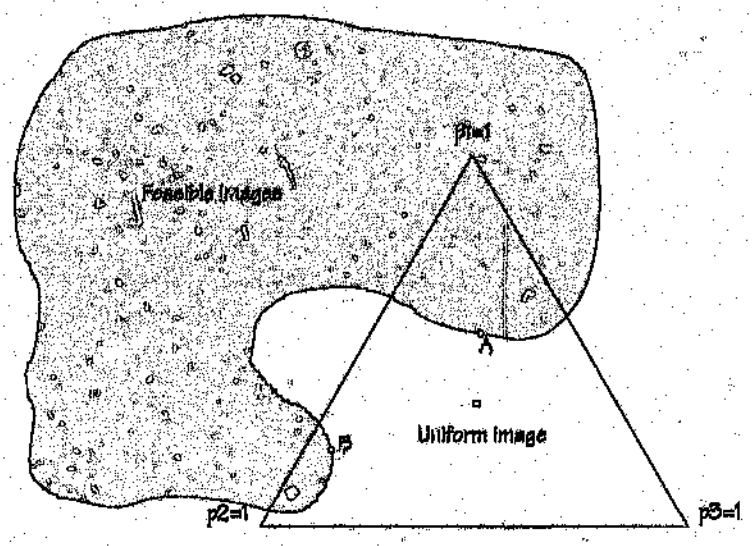


Figure 3.3. *Maximum entropy with hard non–linear constraints.*

In this case images A and B could both be used as alternative estimates, with preference shown for A, the image with larger entropy. However, this preference is hard to quantify. Skilling shows that if we let the number of photons G become large, image B is contradicted. If G is finite, the degeneracy is no longer an absolute selector of the image: it is merely a probability modulator and we still need to choose an image where selection of the overall

maximum entropy image is merely *ad hoc*.

Digital image bounded above and below

Frieden [Frieden, 1973, 1980] introduced a model for digital images based on Fermi–Dirac statistics. These apply to particles (photons) which are indistinguishable and obey the Pauli exclusion principle, meaning that there can be no more than one particle in each permitted energy state [Sears and Salinger, 1975]. A digital image is effectively bounded above and below (there is an upper limit on the grey–level of each pixel, in addition to the lower limit described above). Hence $a \leq g_i \leq b$, where a and b are the *a priori* limits, e.g. $a = 0$ and $b = 255$ for an image with one byte per pixel. To satisfy these constraints, we assume each cell (pixel) to be limited to b possible sites for photons, a of which are already filled. Drawing a parallel with the mechanical system described above, we have $b - a$ possible "energy states" per "energy level" (pixel) i, among which are distributed $g_i - a$ "particles" (unit grey–levels). Figure 3.4 depicts this model for a 1–dimensional image.

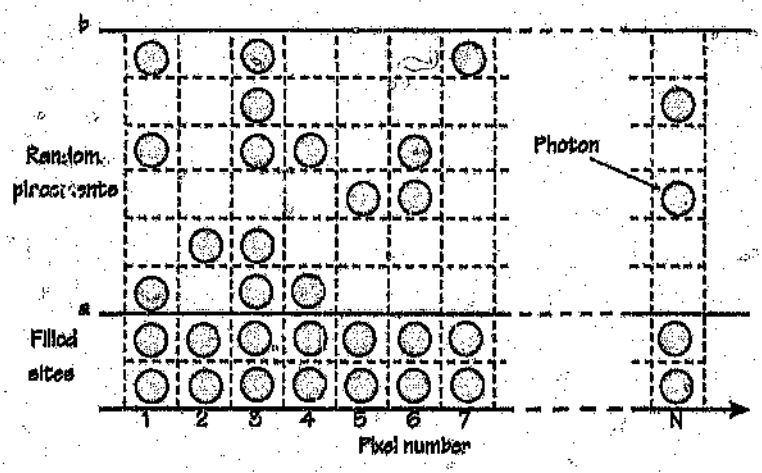


Figure 3.4. *1–dimensional representation of Frieden's model obeying upper and lower bound constraints. The G photons or unit grey–levels are dealt out randomly among the sites above level a. All sites below level a are pre–filled.*

If cell i contains g_i photons, the number of ways w_i in which it can have been formed is given simply by the number of ways to distribute $(g_i - a)$ indistinguishable particles among $(b - a)$ distinguishable locations, i.e.

$$w_i = {}_{(b-a)}C_{(g_i-a)} = \frac{(b-a)!}{(g_i-a)!(b-g_i)!} \tag{3.8}$$

and, assuming independence (or lack of dependence) between cells, the degeneracy of the image is given by

$$W(g_1, \cdots, g_N) = \prod_{i=1}^{N} w_i \tag{3.9}$$

Maximizing this expression is again equivalent to (using Stirling's approximation) maximizing a form of image entropy

$$H(g_1, \cdots, g_N) = -\sum_{i=1}^{N} (g_i - a)\log(g_i - a) - \sum_{i=1}^{N} (b - g_i)\log(b - g_i) \tag{3.10}$$

or, equivalently

$$H(g_1, \cdots, g_N) = -\sum_{g=a}^{b} f_g(g - a)\log(g - a) - \sum_{g=a}^{b} f_g(b - g)\log(b - g) \tag{3.11}$$

where f_g is the frequency of occurrence of grey-level g obtained from the image histogram.

Frieden's use of Fermi–Dirac statistics implies that the degeneracy of each pixel is important in a digital image. A number $(b-a)$ of photon sites is assumed to exist at each pixel location: this is not true, since for any real digital image such "sites" are not distinguishable and hence the use of Fermi–Dirac statistics is not indicated. The only information we have is the number of filled sites (photons) at a given location, not which sites are in fact filled: we do not care how many ways the pixel's grey-level can be achieved, only that we know what that level is. This would imply that the pixel degeneracy is unity, in which case the model becomes trivial. While the model is not "wrong" per sé, it contains redundant elements.

Kikuchi–Soffer "reconciliation" model

There are two main forms of the entropy equation: the Shannon entropy [Shannon and Weaver, 1949] as given by equation (2.1) and the Burg entropy [Burg, 1967] given by (2.4) or, equating grey-levels with probabilities,

$$H(g_1, \cdots, g_N) = \sum_{i=1}^{N} \log g_i \tag{3.12}$$

which relates primarily to problems of spectral analysis. As indicated in the previous chapter, there has been some controversy surrounding the question of which form is correct. Kikuchi and Soffer [Kikuchi and Soffer, 1977; Frieden, 1980] have proposed a model reconciling these two forms as special cases of a more general form by taking a quantum physics approach.

The image is modelled as a random assortment of photons and the degeneracy is maximized on the assumption that they are Bose particles.

Each cell is assumed to have a finite number z of subsites for photon allocation. These subsites correspond to the finite number of degrees of freedom present within an image pixel due to physical reasons such as cell area, bandwidth, aperture size and detection time interval (shutter speed). A 1–dimensional representation of this model is shown in Figure 3.5. Bose particles (photons) obey Bose–Einstein statistics, meaning that the particles are considered indistinguishable and there is no restriction on the number that may occupy a single energy state. The energy states (subsites or degrees of freedom), however, are distinguishable [Sears and Salinger, 1975].

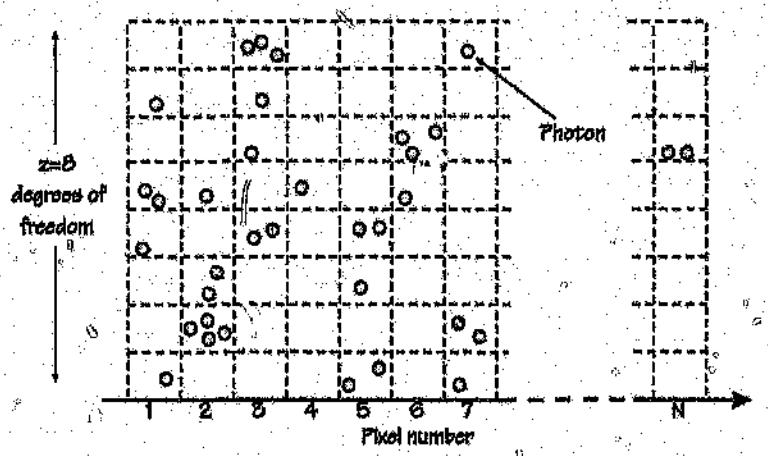


Figure 3.5. *1–dimensional representation of the model due to Kikuchi and Soffer. G photons are dealt out among N pixels. Each pixel has z subsites, into each of which any number of photons may jam.*

Similar reasoning to that used in the two previous models is followed, solving for the image which maximizes the degeneracy W. Following the Bose–Einstein statistics, the number of ways w_i in which a pixel grey–level g_i can be formed, assuming the photons to be indistinguishable and distributed over z distinguishable subsites, is given by [Sears and Salinger, 1975]

$$w_i = \frac{(g_i + z - 1)!}{g_i!(z-1)!} \tag{3.13}$$

Statistical independence between cells is again assumed and hence total image degeneracy is

again given by (3.9). A solution $\log W = \max$ is again sought:

$$\log W(g_1, \cdots, g_N) = \sum_{i=1}^{N} \log \frac{(g_i + z - 1)!}{g_i!(z-1)!} = \max \tag{3.14}$$

The reason for the term "reconciliation model" now becomes apparent. We examine (3.14) in two different limits of particle number relative to number of degrees of freedom, $g_i \ll z$ and $g_i \gg z$. In the first limit the numerator of (3.14) approximates $(z-1)!$ and we get

$$\log W(g_1, \cdots, g_N) \simeq -\sum_{i=1}^{N} \log g_i! = \max$$

and using Stirling's approximation this yields the Shannon entropy form (3.7).

Evaluating (3.14) in the opposite limit [Frieden, 1980] yields the principle

$$\log W(g_1, \cdots, g_N) \simeq \sum_{i=1}^{N} \log g_i = \max$$

which is recognized as the Burg entropy form (3.13) above. It is apparent then that Kikuchi and Soffer have reconciled these two disparate forms of maximum entropy principle as limiting forms of a more general maximum degeneracy principle.

Using Stirling's approximation and (3.14) we obtain an expression

$$\begin{aligned} \log W = &\sum_{i=1}^{N} (g_i + z - 1) \log(g_i + z - 1) \\ &- \sum_{i=1}^{N} g_i \log g_i \\ &- \sum_{i=1}^{N} (z-1) \log(z-1) \end{aligned} \tag{3.15}$$

which can be viewed as the sum of three "entropies".

As a physical model of image formation the definition of pixel degeneracy based on cell-size, shutter speed and other considerations is attractive, but such *a priori* information is rarely available and again our only real knowledge about an individual pixel is its grey-level, which for our purposes is simply reached by letting "enough" light fall on the cell in question. It is interesting to note here that in any practical application, due to physical reasons such as

photon dimensions relative to cell size, we can reasonably assume the case $g_i \ll z$. Hence the Shannon entropy (2.1) again appears to be the correct form for image processing purposes.

3.3 Other distributions and the problem of independence

The models described in Section 3.2 use the image itself as a probability distribution. Another commonly used distribution is the image grey–level histogram [Kapur *et al*, 1985; Brink, 1991b], without explicit assumption of a model justifying its use. The co–occurrence matrix is also used on occasion [Pal and Pal, 1989a]: this can be seen as being related to the use of the histogram, but with a context–dependent measure added for improved performance. In this section a model justifying use of the histogram is developed. Another interpretation of the monkey model, using similar reasoning to that justifying the use of the histogram, is also described.

These models, like those of Section 3.2, still assume independence (or non–dependence) of adjacent "events" in the probability distribution. Context–related information can be included in a number of ways. Some suggestions were outlined in the previous chapter. Clearly, a measure such as the cross–entropy (2.3) (or relative entropy (2.2)) would include a contextual component by virtue of the fact that corresponding samples of the prior and posterior distributions are directly compared. This correspondence can be exploited for threshold selection, as shown in Chapter 4 where a thresholding scheme using the cross–entropy is described. Methods such as this are related to the processing or analysis of the image rather than determination of the image entropy. Other measures, such as the spatial entropy described by Journel and Deutsch [Journel and Deutsch, 1993], are aimed more directly at a description of the image itself. This model is described below. Two new models are presented. The first is based on the autocorrelation function of the histogram or image. This measure should incorporate spatial information by virtue of the fact that correlation values quantify inter–sample relationships. The second is conceptually more difficult to grasp: pixel positions or coordinates relative to the image origin are themselves viewed as "probabilities" in the sense that they are used in the entropy expression. In being explicitly included in the calculation, a unique entropy should be found to correspond to a given image.

Entropy of the image histogram

Grey–level histograms are often used in image processing techniques such as segmentation (thresholding) [Otsu, 1979; Brink, 1989, 1990] and image enhancement [Gonzalez and Woods, 1992]. In particular, methods based on the entropy of the histogram have been proposed in the past [Kapur *et al*, 1985; Brink, 1991b] without any real justification, other than the fact that the histogram is a form of *a posteriori* grey–level probability distribution and can be used as an *a priori* estimate for future images of the same scene.

The grey–level histogram of an image can be obtained simply by determining the frequency (number of pixels), f_g, of each grey–level g and dividing this by the total number of pixels N in the image to give a measure of "probability" of occurrence, $p_g = f_g/N$, of that grey–level. The entropy of the histogram is defined as

$$H = -\sum_{g=a}^{b} p_g \log p_g \tag{3.16}$$

where a and b are the minimum and maximum grey–levels present in the image. The non–normalized frequency histogram can also be used, subject to the condition that the number of pixels $N = \sum_g f_g = \text{constant}\ \forall$ images involved in a computation.

The justification for maximizing the histogram entropy parallels that for the image entropy using the monkey model. The histogram degeneracy (the number of distinct images which give rise to the same histogram) is, according to the Boltzmann law,

$$W(f_a, \cdots, f_b) = \frac{N!}{f_a! \cdots f_b!} \tag{3.17}$$

Maximizing $\log W$ as before reduces to maximization of the entropy (3.16). Another approach can be used whereby histogram entropy is derived from a model of the image itself, as follows.

We assume the image to be analogous to a thermodynamical system: let the grey–levels g correspond to energy levels and their locations in the image (i.e. pixels) correspond to distinguishable energy states. The number of particles at level g (the number of occupied states) is given by the frequency f_g. Here the analogy becomes tenuous, since if a pixel (state) is occupied at one grey–level (energy level), it cannot be occupied at any other. There is thus some dependence between histogram bins built into this, in that we have to take account of the fact that if, say, there are N pixels in total and f_a pixels have level $g = a$ (the lowest grey–level), then the number of free pixels (states) available at $g = a+1$ is $N - f_a$. In general we find that the number of available pixels N_{k+1} at level $g = k + 1$ is given by

$$N_{k+1} = N - \sum_{g=a}^{k} f_g \tag{3.18}$$

with $N_a = N$. Using Fermi–Dirac statistics we find the degeneracy w_g of a grey–level g analogous to that of an energy level [Sears and Salinger, 1975], given by

$$w_g = {}_{N_g}C_{f_g} = \frac{N_g!}{(N_g - f_g)! f_g!} \tag{3.19}$$

and the total degeneracy of the image (assuming statistical independence between grey–levels) is

$$W = \prod_{g=a}^{b} w_g \tag{3.20}$$

We again want to maximize W or, equivalently, $\log W$. Again using Stirling's approximation

$$\begin{aligned}
\log W &= \sum_{g=a}^{b} N_g \log N_g - (N_g - f_g)\log(N_g - f_g) - f_g \log f_g \\
&= N \log N - (N_b - f_b)\log(N_b - f_b) \\
&+ \sum_{g=a}^{b-1} \left[N_{g+1} \log N_{g+1} - (N_g - f_g)\log(N_g - f_g) \right] - \sum_{g=a}^{b} f_g \log f_g \\
&= N \log N - \sum_{g=a}^{b} f_g \log f_g
\end{aligned}$$

since $N_{g+1} = N_g - f_g$ and $N_b - f_b = 0$. Since the first term above is constant, the quantity to be maximized is

$$H = -\sum_{g=a}^{b} f_g \log f_g$$

or

$$H = -\sum_{g=a}^{b} p_g \log p_g$$

in the case of the normalized histogram.

We have thus derived an image model which in a sense justifies the use of histogram entropy as a measure of the entropy of the image itself. Clearly, a drawback of the histogram is that all "image" (i.e. context dependent) information is completely ignored without recourse to the philosophical panacea that entropy gives us by "not assuming dependence". As a criterion function for threshold selection, where segmentation usually takes place without regard to such information anyway, it has its merits and should not be disregarded [Kapur *et al*, 1985].

Entropy of the image

If we adopt a similar approach using the image, we obtain the following model. Given total image "brightness" $G = \sum_i g_i = \sum_g g f_g$. If pixel $i = 1$ has grey–level g_1 then the maximum number of levels available for pixel $i = 2$ would be $G - g_1$, and in general the number of

available grey–levels G_{k+1} for pixel $i = k+1$ is, as in (3.13),

$$G_{k+1} = G - \sum_{i=1}^{k} g_i \tag{3.21}$$

with $G_1 = G$. The degeneracy of pixel i is then given by

$$w_i = {}_{G_i}C_{g_i} = \frac{G_i!}{(G_i - g_i)!g_i!} \tag{3.22}$$

If there is a non–zero lower grey–level limit a for the image, corresponding to pre–fogging of a photographic plate, we can substitute the values $G_i - a$ and $g_i - a$ for G_i and g_i, similar to Frieden's model (Figure 3.4). The assumption of independence gives the overall degeneracy measure

$$W = \prod_{i=1}^{N} w_i$$

again and we maximize $\log W$ given by

$$\begin{aligned}
\log W &= \sum_{i=1}^{N} G_i \log G_i - (G_i - g_i)\log(G_i - g_i) - g_i \log g_i \\
&= G\log G - (G_N - g_N)\log(G_N - g_N) \\
&+ \sum_{i=1}^{N-1} \left[G_{i+1} \log G_{i+1} - (G_i - g_i)\log(G_i - g_i) \right] - \sum_{i=1}^{N} g_i \log g_i \\
&= G\log G - \sum_{i=1}^{N} g_i \log g_i
\end{aligned}$$

since $G_{i+1} = G_i - g_i$ and $G_N - g_N = 0$. Since the first term above is constant, the quantity to be maximized is

$$H = -\sum_{i=1}^{N} g_i \log g_i$$

i.e. the Shannon entropy resulting from assumption of the monkey model (3.7). Note that here I have assumed a similar situation to Frieden's, i.e. provision for lower and upper bounds, with the modification that the upper bound for each pixel is dependent on the grey–levels of previous pixels instead of being fixed. In effect this is simply a restatement of the monkey model making explicit a constraint which was previously only implied.

Spatial entropy

Journel and Deutsch address the issue of sample inter–dependence in a distribution by developing an expression for spatial entropy, viewed as a measure of spatial disorder [Journel and Deutsch, 1993]. Consider a random variable $x(\mathbf{k})$ which can take on n outcome values with probabilities p_g, $g = 0, \cdots, n-1$ such that $\sum_{g=0}^{n-1} p_g = 1$. We may view the random variable $x(\mathbf{k})$ as the grey–level of each pixel $\mathbf{k} = (i, j)$ in an image, for example, with the p_g indicating grey–level probabilities. The entropy associated with this probability set (the histogram entropy) is clearly

$$H = -\sum_{g=0}^{n-1} p_g \log p_g \geq 0$$

This measure, as indicated before, takes no account of inter–relationships between samples of $x(\mathbf{k})$. Journel and Deutsch propose a bivariate entropy measure, with the entropy value dependent on the spatial distance or *lag*, λ, between samples $x(\mathbf{k})$ and $x(\mathbf{k}+\lambda)$. We can now define bivariate probabilities

$$p_{gg'} = \Pr\{x(\mathbf{k}) \in g, x(\mathbf{k}+\lambda) \in g'\},\ g, g' = 0, \cdots, n-1$$

which, regarding the random variable as an image, would correspond to the probability that grey–level g occurs a distance λ from grey–level g', say. The bivariate entropy is given by

$$H(\lambda) = -\sum_{g=0}^{n-1} \sum_{g'=0}^{n-1} p_{gg'}(\lambda) \log p_{gg'}(\lambda) \geq 0 \tag{3.23}$$

For lag $\lambda = 0$, we have $p_{gg'}(0) = 0$, $\forall g \neq g'$ and $p_{gg}(0) = p_g$, giving us simply the univariate (Shannon) entropy (2.1) of the histogram, i.e.

$$H(0) = -\sum_{g=0}^{n-1} p_g \log p_g = H$$

For lag $\lambda = \infty$ we can regard $x(\mathbf{k})$ and $x(\mathbf{k}+\infty)$ as independent, giving $p_{gg'}(\infty) = p_g p_{g'}$, $\forall g, g'$. If the probability distribution p_g is assumed to be stationary (i.e. it is independent of position, $\mathbf{k}$, so that for any given lag, λ, $p_g = p_{g'}$ if $g = g'$) the bivariate entropy has an upper bound given by

$$H(\infty) = -\sum_{g=0}^{n-1} \sum_{g'=0}^{n-1} p_g p_{g'} (\log p_g + \log p_{g'}) = 2H(0)$$

Assuming that we have a well-behaved non-periodic probability density function, $p_{gg'}(\lambda)$ decreases continuously as $|\lambda|$ increases and consequently $H(\lambda)$ increases with $|\lambda|$ over the interval $[H(0), 2H(0)]$. A standardized relative measure of bivariate entropy can be written

$$H_R(\lambda) = \frac{H(\lambda) - H(0)}{H(0)} \in [0, 1] \tag{3.24}$$

The image entropy could then be defined as the average value of this relative measure over all possible lags. In general the coordinate k and the lag λ are vector quantities defining multi-dimensional spatial coordinates.

In practice, the $p_{gg'}$ are effectively grey-level co-occurrence probabilities over single-neighbour pixel neighbourhoods defined by the lag λ. The calculation of a single spatial entropy thus involves the calculation of an $n \times n$ co-occurrence matrix for each possible lag value: clearly, unless we severely restrict the range of lag distances over which we assume dependence (i.e. assume a kind of Markov property), the computational task is enormous.

The autocorrelation model

Given a discrete signal or distribution, the inter-dependence of different points in the distribution is described by its autocorrelation function. For simplicity I will confine myself for the present to the image histogram. Clearly the grey-level frequencies f_g of the histogram can usually be regarded as being inter-related. We can quantify the grey-level inter-dependence by determining the autocorrelation function of the histogram

$$\rho_{gg} = (r_{1-n}, \cdots, r_{-2}, r_{-1}, r_0, r_1, r_2, \cdots, r_{n-1}) \tag{3.25}$$

where the coefficients r_k are given by

$$r_k = \sum_{g=0}^{n-1} p_{k+g} p_g \qquad -(n-1) \leq k \leq n-1 \tag{3.26}$$

where n is the number of grey-levels and

$$p_g = \frac{f_g}{N}$$

is the probability of occurrence of grey-level g in the N-pixel image. Note that $p_g = 0$ outside the range $0 \leq g \leq n-1$. If we normalize the autocorrelation coefficients we can interpret the resulting distribution as the probability of general grey-level inter-dependence:

$$p_k^* = \frac{r_k}{\sum_{k=1-n}^{n-1} r_k} \tag{3.27}$$

In other words, p_k^* quantifies the likelihood that pixels differing by k grey-levels are related. The (Shannon) entropy of this distribution, based on similar arguments as those justifying the monkey model [Brink, 1994], is

$$H_r = -\sum_{k=1-n}^{n-1} p_k^* \log p_k^* \tag{3.28}$$

In order to take spatial relationships into account, we should ideally work with the two-dimensional autocorrelation function of the image itself. In this case, we quantify the inter-dependence of the actual image pixels, or the likelihood that pixels a given distance apart are related.

In determining the entropy of an autocorrelation function, we are tacitly assuming independence of the discrete autocorrelation values: e.g. the relationship between pixels k units apart is unrelated to the relationship between pixels l units apart, say. While this assumption is convenient, as before it is not founded on rigid theory. On the other hand, the independence question can be taken to ridiculous extremes. What is required is a distribution based on the image data which takes into account the spatial organization of the image, while remaining unaffected by the assumption of independence.

Using the autocorrelation function of a distribution as "input" to the entropy expression to some extent parallels the approach of Journel and Deutsch [Journel and Deitsch, 1993] using measures of bivariate entropy, described above. Such a measure of "spatial entropy" has the added advantage that we are not worried about inter-dependences of inter-dependences. However, the applicability of such an approach to the specific problem of threshold selection or general segmentation is not clear as, by definition, the image is split up into regions: do we find the spatial entropy of each region in isolation of groups of regions having the same classification or of some measure on the segmented image as a whole?

The Markov modelling of images mentioned in Chapter 2 could be used to determine context-dependent probabilities for the image pixels. These could then be plugged into the entropy expression (2.1) or (2.2): we do not then worry about the entropic non-assumption of spatial relationships as these would have effectively been built in prior to the calculation. This is the motivation behind the use of image-related distributions such as the grey-level co-occurrence matrix [Pal and Pal, 1989a] and (grey-level, local mean grey-level) scatterplot [Abutaleb, 1989; Brink, 1992a]. As the methods and "models" described here do not involve physical models of the digital image, they have been included as required with the descriptions of the thresholding or segmentation techniques to which they apply in the next chapter.

The "labelled photon" model

The following model is based on a consideration of the monkey model as applied to either the image itself or its histogram. Effectively, when assuming the monkey model for the image, the entropy value obtained is not unique to the image, for the reasons outlined earlier in this chapter. It is instead a measure of the degeneracy of the image, i.e. the number of photon distributions which could have given rise to the image. Consider the image entropy

$$H = -\sum_{i=1}^{N} g_i \log g_i$$

This uses information available from the histogram alone, as can be seen when we rewrite the formula as

$$H = -\sum_{g=0}^{n-1} f_g \, g \log g \tag{3.29}$$

Similarly the histogram entropy, related to the number of distinct images which share the same grey-level distribution, which is given by

$$H = -\sum_{g=0}^{n-1} f_g \log f_g$$

is based on information related only to grey-level frequencies: if a "histogram of frequencies" were drawn up, we could again rewrite the expression as

$$H = -\sum_{f=0}^{F} f_f \, f \log f \tag{3.30}$$

where F is the highest grey-level frequency of the histogram and f_f represents the number of occurrences of grey-level frequency f.

We are interested in an entropy measure related to the degeneracy of the pixel positions themselves, i.e. the number of ways in which a photon (or grey-level unit) destined for a given coordinate in the image can be "chosen". If the image is viewed as a kind of histogram, then this "super-image" would consist only of the pixel positions viewed as *probabilities* for the purpose of using them in the Shannon entropy expression (2.1). We could regard it as being made up of "labelled" photons (unit grey-levels), the received image being a plot of the frequencies of occurrence of these labels. Then the image entropy is given by

$$H = -\sum_{j=1}^{G} i_j \log i_j \tag{3.31}$$

where i_j is the *label* (relative to the image origin) associated with the j–th photon received. G is the total illumination of the image (number of photons) as before. Expressed in terms of "probabilities" (3.31) becomes

$$H = -\sum_{j=1}^{G} p_j \log p_j \quad , p_j = \frac{i_j}{I} \tag{3.32}$$

where

$$I = \sum_{j=1}^{G} i_j$$

$$= \sum_{i=1}^{N} i\, g_i$$

As in (3.29) and (3.30), this can be expressed as

$$H = -\sum_{i=1}^{N} g_i\, i \log i \tag{3.33}$$

or

$$H = -\sum_{i=1}^{N} g_i\, p' \log p' \quad , p' = \frac{i}{I} \tag{3.34}$$

This measure has the advantage that position information is included directly in its expression and that a unique entropy value can be determined for a given image. It has the disadvantages of being intuitively difficult to grasp and mathematically *ad hoc*. It will, however, be interesting to see how well it performs in an application such as segmentation.

Chapter 4

Maximum entropy threshold selection

Grey-level thresholding is conceptually the simplest form of image segmentation. However, the automatic selection of a particular grey-level as a threshold requires the definition and use of some suitable criterion function which is optimized in some sense by the choice of threshold. The choice of criterion function is based on an understanding of the requirements of the problem, i.e. the model of the process and/or image justifying why optimization of a given function may be expected to yield "correct" results. An additional requirement is obviously that the results obtained are suited to a particular application or at least make some sort of aesthetic sense. The latter requirement is often the basis for heuristic techniques. It is also clearly important in the evaluation of techniques for a given application.

Many forms of criterion function have been proposed in the past [Pratt, 1991; Gonzalez and Woods, 1992; Weszka, 1978; Haralick and Shapiro, 1985; Sahoo et al, 1988]. These functions usually take the form of an error measure [Kittler and Illingworth, 1986; Brink and Pendock, 1993], a "goodness-of-fit" measure [Brink, 1989, 1990] or some measure quantifying a desirable property of the result, such as image entropy. There are, of course, some methods such as the preservation of moments of the image [Tsai, 1985] which do not fit into these general categories, but they are rare.

As indicated, entropic methods belong mainly to the third category mentioned above. Why maximization of image entropy is desirable should be apparent from the previous chapters. How we define the image entropy is, however, a matter of contention. In this chapter some earlier methods applying the maximum entropy principle to threshold selection are briefly described and new methods based on some of the models and ideas described in Chapters 2 and 3 are introduced. Unless otherwise indicated, the term "thresholding" should be taken to mean specifically *global, binary* thresholding as distinct from multi-level or dynamic

thresholding processes.

4.1 Background: an entropic criterion function

The methods described in this section all use the Shannon entropy formula (2.1) as a criterion function for automatic threshold selection. Various image–related distributions are used to represent the image in these methods, with varying degrees of success. In most cases the maximum entropy formalism has been applied without reference to its origins or derivation in terms of an explicit image model. Possible models to fit these requirements have been described in Section 3.3 of the previous chapter.

Entropy of the histogram

Kapur *et al* [Kapur *et al*, 1985] have proposed a threshold selection technique based on maximization of the histogram entropy (3.16), inspired partly by a method previously implemented by Pun [Pun, 1980]. When an image is thresholded, the histogram is essentially divided into two distributions representing *class 0* and *class 1*. These classes can be interpreted as "background" and "foreground". Class 0 consists of the grey–levels $0 \leq g \leq T$ and class 1 of levels $(T+1) \leq g \leq (n-1)$, where T is the threshold value. The new binary image resulting from this thresholding operation consists of two grey–levels with (histogram) probabilities $P_0(T)$ and $P_1(T)$ given by

$$P_0(T) = \sum_{g=0}^{T} p_g \qquad (4.1)$$

$$P_1(T) = 1 - P_0(T)$$

These are usually referred to as the *a posteriori* class probabilities. The entropy of the binary image histogram, from (3.16), is then given by

$$H_b(T) = -P_0(T)\log P_0(T) - P_1(T)\log P_1(T) \qquad (4.2)$$

To find the optimum threshold τ, we could maximize $H_b(T)$ with respect to T, i.e.

$$\tau = \arg\Big\{\max_{0 \leq T < n-1} \{H_b(T)\}\Big\} \qquad (4.3)$$

This approach results in the trivial selection of a threshold that yields $P_0(\tau) = P_1(\tau) = 0.5$ as closely as possible given the discrete nature of the histogram and image. Pun's approach is based on this method, but defines and uses upper bounds of the components constituting (4.2) [Pun, 1980]. This is done as follows:

Clearly, we can write

$$\sum_{g=0}^{T} p_g \log p_g \leq \sum_{g=0}^{T} p_g \log\{\max\{p_0, p_1, \cdots, p_T\}\}$$
$$= P_0(T) \log\{\max\{p_0, p_1, \cdots, p_T\}\}$$

Hence

$$-\sum_{g=0}^{T} p_g \log p_g \geq -P_0(T) \log\{\max\{p_0, p_1, \cdots, p_T\}\} \tag{4.4}$$

We can define sub–entropies corresponding to the two classes, using (3.16), as

$$H_0'(T) = -\sum_{g=0}^{T} p_g \log p_g$$
$$H_1'(T) = -\sum_{g=T+1}^{n-1} p_g \log p_g \tag{4.5}$$

and so, from (4.4) and (4.5), it can be shown that

$$-P_0(T) \log P_0(T) \leq \frac{H_0'(T) P_0(T)}{\log\{\max\{p_0, p_1, \cdots, p_T\}\}}$$
$$= F_0(T)$$

Similarly

$$-P_1(T) \log P_1(T) \leq \frac{H_1'(T) P_1(T)}{\log\{\max\{p_{T+1}, p_{T+2}, \cdots, p_{n-1}\}\}}$$
$$= F_1(T)$$

Pun then defines the quantity

$$F(T) = F_0(T) + F_1(T) \tag{4.6}$$

as an upper bound on the binary image's histogram entropy (4.2), i.e. $H_b(T) \leq F(T)$. $F(T)$ is maximized with respect to T. Kapur *et al* present several arguments against this measure [Kapur *et al*, 1985]. The prime argument is that the value of T where the maximum of $F(T)$ occurs is arbitrary and tends in any case to lie close to the point where $P_0(T) = P_1(T) = 0.5$.

The approach adopted by Kapur *et al* is relatively simple. When the histogram is divided into two classes by a threshold T, each class distribution is viewed as an *independent* distribution.

To this end, the grey–level probabilities p_g corresponding to each class must be normalized to define the respective class entropies

$$H_0(T) = -\sum_{g=0}^{T} \frac{p_g}{P_0(T)} \log \frac{p_g}{P_0(T)}$$
$$H_1(T) = -\sum_{g=T+1}^{n-1} \frac{p_g}{P_1(T)} \log \frac{p_g}{P_1(T)} \tag{4.7}$$

The aim is to select that threshold which maximizes the entropies of both distributions. Kapur *et al* define a measure

$$\Psi(T) = H_0(T) + H_1(T) \tag{4.8}$$

and find the optimum threshold, τ, as

$$\tau = \arg\{\max_{0 \le T < n-1} \{\Psi(T)\}\} \tag{4.9}$$

An alternative approach to finding a suitable threshold using the class entropies (4.7) was proposed by Brink [Brink, 1991b]. The process of (4.9) maximizes the sum of the class entropies. It is conceivable that this maximum could occur at a point where the entropy of one class is much larger than that of the other. This would mean that the threshold would be biased towards one distribution at the expense of the other. We would ideally like to maximize the entropies of both distributions. This involves a trade–off to find a threshold value which will *most nearly* maximize each entropy without undue discrimination against the other. To this end, a *maximin* approach can be adopted [Brink, 1991b] whereby the threshold τ maximizes the smaller of $H_0(T)$ and $H_1(T)$ over all values T, i.e.

$$\tau = \arg\{\max_{0 \le T < n-1} \{\min(H_0(T), H_1(T))\}\} \tag{4.10}$$

Entropy of the (grey–level, local mean grey–level) scatterplot

In an attempt to include context–dependent information about the image in the entropic threshold selection process, Abutaleb [Abutaleb, 1989] has proposed the use of the (grey–level, local mean grey–level) scatterplot instead of the grey–level histogram. This is also sometimes referred to as a two–dimensional histogram. The distribution is essentially a co–occurrence matrix displaying the frequencies of co–occurrence of pixel grey–levels k with local pixel neighbourhood mean grey–levels l. Abutaleb has not specified how the neighbourhood of a pixel is defined: a common approach is to consider this as being made up of a 3×3 block of pixels with its origin at the centre. The neighbourhood of the pixel with image coordinates (i, j) is shown in Figure 4.1.

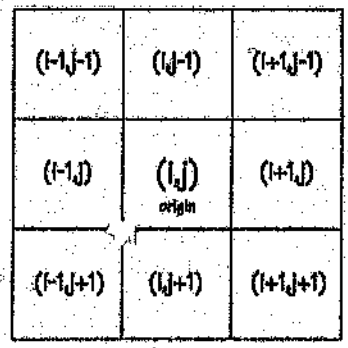


Figure 4.1 3×3 *neighbourhood of a pixel centred at image coordinates* (i, j).

The mean grey–level of such a neighbourhood is simply given by

$$\bar{g} = \frac{1}{9} \sum_{y=j-1}^{j+1} \sum_{x=i-1}^{i+1} g_{xy}$$

where g_{xy} is the grey–level g of pixel (x, y) in the neighbourhood. This was the method adopted here to determine the local neighbourhood means l. The frequencies f_{kl} of the scatterplot are normalized by dividing by the total number of pixels in the image, N, to give a plot of (grey–level, local mean grey–level) probabilities

$$p_{kl} = \frac{f_{kl}}{N}$$

Clearly, regions of relatively uniform grey–level will contribute mainly to scatterplot entries on and near the leading diagonal of the matrix, i.e. the diagonal passing through quadrants 0 and 1 of Figure 4.2. The entries far off the diagonal correspond to more rapid changes in grey–level in the image, such as edges and noise. Abutaleb explicitly assumes that the off–diagonal quadrants (shaded regions in Figure 4.2) make little or no contribution to the image or its information content and can therefore be ignored in the entropy calculations to follow.

The scatterplot is split into four quadrants by a threshold pair (T, S), where T is the pixel grey–level threshold and S is the neighbourhood mean threshold. The quadrants 0 and 1 of Figure 4.2 correspond to the two classes ("background" and "foreground") into which the image is to be divided. The shaded regions are ignored, as explained above.

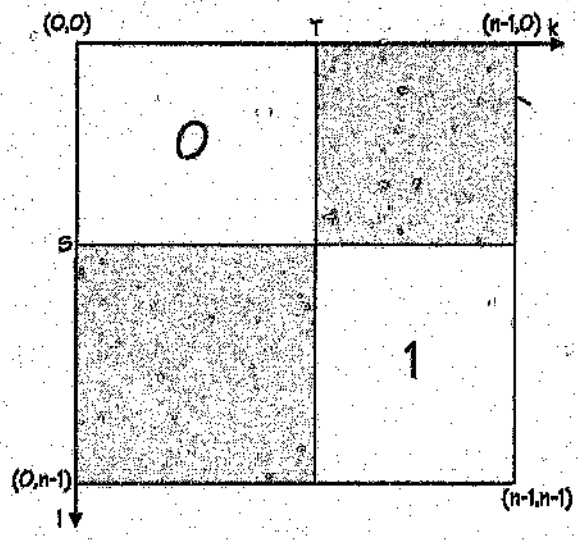


Figure 4.2. *Quadrants of the (grey-level, local mean grey-level) scatterplot.*

Following the approach of Kapur *et al* above, Abutaleb defines two class entropies based on the class-normalized co-occurrence probabilities:

$$\begin{aligned} H_0(T,S) &= -\sum_{k=0}^{T}\sum_{l=0}^{S}\frac{p_{kl}}{P_0(T,S)}\log\frac{p_{kl}}{P_0(T,S)} \\ H_1(T,S) &= -\sum_{k=T+1}^{n-1}\sum_{l=S+1}^{n-1}\frac{p_{kl}}{1-P_0(T,S)}\log\frac{p_{kl}}{1-P_0(T,S)} \end{aligned} \tag{4.11}$$

where

$$P_0(T,S) = \sum_{k=0}^{T}\sum_{l=0}^{S} p_{kl} \tag{4.12}$$

is the class probability of quadrant 0. The formula for $H_1(T,S)$ in (4.11) is valid only if the assumption that the probability contributions from the shaded quadrants are negligible is correct. The strict correct form should of course be

$$H_1(T,S) = -\sum_{k=T+1}^{n-1}\sum_{l=S+1}^{n-1}\frac{p_{kl}}{P_1(T,S)}\log\frac{p_{kl}}{P_1(T,S)} \tag{4.13}$$

where

$$P_1(T,S) = \sum_{k=T+1}^{n-1}\sum_{l=S+1}^{n-1} p_{kl} \tag{4.14}$$

is the class probability of quadrant 1. Still following Kapur *et al*, a measure of binarization entropy Ψ is defined as

$$\Psi(T,S) = H_0(T,S) + H_1(T,S) \tag{4.15}$$

which is maximized to determine the optimum threshold pair (τ, σ) in a similar fashion to (4.9).

The main motivation behind Abutaleb's assumption regarding the off–diagonal entries was to reduce computational time. It is found in practice that this assumption is not valid, and that use of the correct forms (4.13) and (4.14) results in threshold selections which often differ substantially from those using the uncorrected formulae (4.11).

Following Brink's [Brink, 1991b, 1992a] reasoning, an alternative threshold selection scheme would be to trade off the maxima of the individual class entropies (4.11) and (4.13) in order to find a suitable compromise threshold pair (τ, σ) corresponding to the *maximin* of the class entropies, i.e.

$$(\tau, \sigma) = \arg\{\max_{0 \le T < n-1,\ 0 \le S < n-1} \{\min(H_0(T, S), H_1(T, S))\}\} \tag{4.16}$$

where the corrected form of $H_1(T, S)$ (4.13) is used [Brink, 1992a].

Neither of these methods actually take into account the information present in the shaded quadrants of Figure (4.2), nor is any suggestion put forward regarding which classes pixels corresponding to these should be assigned to. This is usually done arbitrarily or by creating a third class corresponding to "busy regions". Ideally, a threshold contour through the scatterplot should be defined rather than the single point as defined above. This should result in a segmentation making full use of the contextual information present in the plot, corresponding to a dynamic thresholding scheme which to some extent avoids the problems associated with global methods. Unfortunately such a process would be very difficult to implement and would also be computationally very expensive.

Entropy of the grey–level co–occurrence matrix

The co–occurrence matrix differs from the scatterplot described above mainly in that each entry f_{kl} represents the number of times (frequency) that a pixel of grey–level l occurs in the immediate neighbourhood of a pixel with grey–level k. The neighbourhood of a given pixel at a position (i, j) is often defined as the 8–neighbourhood of Figure 4.1.

Again with the aim of including contextual information in the threshold selection process, Pal and Pal [Pal and Pal, 1989a] based an entropic criterion function on a 2–neighbour co–occurrence matrix. This is formed by considering at each position (i, j) only the immediate neighbours to the right, $(i + 1, j)$, and below, $(i, j + 1)$. A single threshold τ is found, as opposed to the threshold pair (τ, σ) determined using the scatterplot. This applies to both "axes" of the matrix, dividing it into four quadrants A, B, C and D as shown in Figure 4.3. The assymetry of the neighbourhood defined here is important. If a symmetrical

neighbourhood such as that of Figure 4.1 is used, the resulting matrix is symmetric about the leading diagonal and the information contained in quadrants B and D is identical. The 2–neighbourhood defined above lends a useful interpretation to these quadrants: B corresponds to "foreground" pixels having "background" pixels as neighbours and D to "background" pixels with "foreground" neighbours.

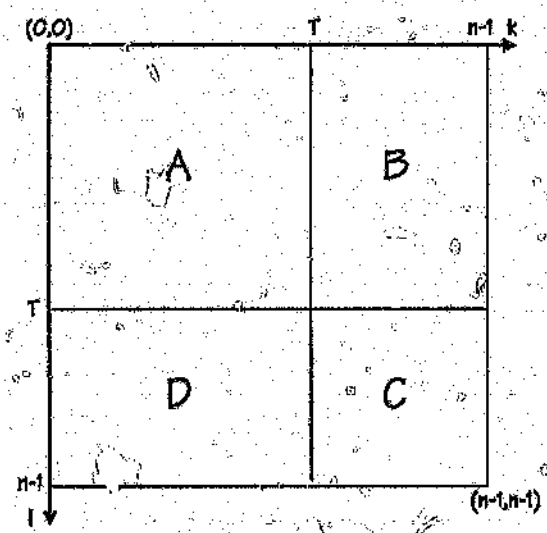


Figure 4.3. *Quadrants of the co–occurrence matrix resulting from thresholding at T.*

The frequencies of co–occurrence f_{kl} are normalized by division by the overall sum of the frequencies to give co–occurrence probabilities

$$p_{kl} = \frac{f_{kl}}{\sum_{k=0}^{n-1} \sum_{l=0}^{n-1} f_{kl}}$$

The class probabilities of the four quadrants are given respectively by

$$\begin{aligned} P_A(T) &= \sum_{k=0}^{T} \sum_{l=0}^{T} p_{kl} \\ P_B(T) &= \sum_{k=T+1}^{n-1} \sum_{l=0}^{T} p_{kl} \\ P_C(T) &= \sum_{k=T+1}^{n-1} \sum_{l=T+1}^{n-1} p_{kl} \\ P_D(T) &= \sum_{k=0}^{T} \sum_{l=T+1}^{n-1} p_{kl} \end{aligned} \tag{4.17}$$

These class probabilities are used in the same way as before [Kapur *et al*, 1985; Abutaleb, 1989; Brink, 1991b, 1992a] to normalize the overall co-occurrence probabilities in order to determine the quadrant entropies

$$\begin{aligned} H_A(T) &= -\sum_{k=0}^{T}\sum_{l=0}^{T}\frac{p_{kl}}{P_A(T)}\log\frac{p_{kl}}{P_A(T)} \\ H_B(T) &= -\sum_{k=T+1}^{n-1}\sum_{l=0}^{T}\frac{p_{kl}}{P_B(T)}\log\frac{p_{kl}}{P_B(T)} \\ H_C(T) &= -\sum_{k=T+1}^{n-1}\sum_{l=T+1}^{n-1}\frac{p_{kl}}{P_C(T)}\log\frac{p_{kl}}{P_C(T)} \\ H_D(T) &= -\sum_{k=0}^{T}\sum_{l=T+1}^{n-1}\frac{p_{kl}}{P_D(T)}\log\frac{p_{kl}}{P_D(T)} \end{aligned} \tag{4.18}$$

Pal and Pal propose two algorithms for threshold determination based on these quadrant entropies. The first is similar to Abutaleb's approach, whereby a total entropy of the regions corresponding to A and C is maximized, i.e.

$$H(T) = H_A(T) + H_C(T) \tag{4.19}$$

The threshold τ is thus

$$\tau = \arg\{\max_{0\le T<n-1}\{H(T)\}\} \tag{4.20}$$

The second algorithm is based on what Pal and Pal refer to as *conditional* entropies. Consider an image composed of two regions 0 and 1 ("background" and "foreground", say), made up of the respective sets of grey-levels $\{g_0\}$ and $\{g_1\}$. The conditional entropy of the foreground (1), given the background (0), can be defined as

$$H(1|0) = -\sum_{g_1}\sum_{g_0} p(g_1|g_0)\log p(g_1|g_0) \tag{4.21}$$

and, similarly,

$$H(0|1) = -\sum_{g_0}\sum_{g_1} p(g_0|g_1)\log p(g_0|g_1) \tag{4.22}$$

is the conditional entropy of the background given the foreground. These are equivalent to the entropies of the off-diagonal quadrants $H_B(T)$ and $H_D(T)$. Pal and Pal claim, therefore, that in order to obtain the optimum threshold, these conditional entropies should be maximized. This is done by again defining their sum as a measure of "overall" conditional entropy

$$H'(T) = H_B(T) + H_D(T)$$

so that

$$\tau = \arg\{\max_{0 \le T < n-1}\{H'(T)\}\} \tag{4.23}$$

The results presented by Pal and Pal [Pal and Pal, 1989a] indicate average performance for the first algorithm and occasionally remarkable performance on the part of the conditional entropy approach. Unfortunately, the performance of the latter appears also to be quite erratic, resulting in either very good or very bad thresholds, based on a purely aesthetic evaluation. Pal and Pal have retained both of these methods in later publications [Pal and Pal, 1989b, 1991b], simply introducing a different definition of entropy (2.10), briefly described in Chapter 2.

The fact that these algorithms select a single threshold means that all pixels can be directly classified, unlike Abutaleb's method which leaves two quadrants undecided.

Threshold selection using the relative entropy information

The measure of "entropy information" (2.8) was developed by Pan and Harris [Pan and Harris, 1990] specifically in connection with the discretization of a measurement of a physical (in their case, geological) signal or variable into a finite set of subranges. For our purposes, this exactly corresponds to the problem of thresholding a grey-scale image (partitioning the data into two finite subranges of grey-level). The derivation of the relative entropy information has been given in Chapter 2. The formula is reproduced here for convenience:

$$\overline{p}(\mathbf{x} \to \mathbf{y}) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{s} p(x_i|y_j)p(y_j)\log\frac{p(x_i|y_j)p(y_j)}{p(x_i)}}{\sum_{j=1}^{s} p(y_j)\log p(y_j)}$$

where, for our purposes, x_i represents the ith grey-level (the ith discrete subrange of the continuous brightness function $\mathbf{x}$ defined over the image as a function of position) and the y_j represent related, possibly context-dependent, qualitative information also obtained from the data or the measurement process. Usually this process would apply in discretizing continuous data: the measure above is in fact the relative entropy information resulting from the discretization of the continuous illumination distribution to produce the digital grey-scale image which is our starting point. We want to discretize this image further to m, $m < n$, levels. We may proceed as follows.

The relative entropy information for discretizing our data into m classes is given by

$$\overline{p}(\mathbf{x} \to \mathbf{y}) = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{s} p(x_i|y_j)p(y_j)\log\frac{p(x_i|y_j)p(y_j)}{p(x_i)}}{\sum_{j=1}^{s} p(y_j)\log p(y_j)} \tag{4.24}$$

In order to incorporate "qualitative" context-related information, let y_j represent the discretized local grey-level means within a 3×3 neighbourhood (Figure 4.1) of each of the N

pixels (samples). The local mean values are discretized to take on any of the s discrete values y_j, $j = 1, \cdots, s$.

The probabilities in (4.24) can be estimated in two ways: the maximum likelihood method or the Bayesian robust method [Pan and Harris, 1990]. Let us denote f_{ij} as the number of pixels for which $\mathbf{x}$ takes values within subrange i *and* $\mathbf{y}$ has the value y_j: in other words, a frequency of co-occurrence. The maximum likelihood estimates of the relevant probabilities are then

$$\begin{aligned} \hat{p}(x_i|y_j) &= \frac{f_{ij}}{N_j} \\ \hat{p}(y_j) &= \frac{N_j}{N} \\ \hat{p}(x_i) &= \frac{N_i}{N} \end{aligned} \tag{4.25}$$

where $N_j = \sum_{i=1}^{m} f_{ij}$ is the number of pixels in neighbourhoods having local mean y_j, $N_i = \sum_{j=1}^{s} f_{ij}$ is the number of pixels falling within each class and $N = \sum_{i=1}^{m} \sum_{j=1}^{s} f_{ij}$ is the total number of pixels. In this case, these quantities can be obtained directly from the (grey-level, local mean grey-level) scatterplot described above [Abutaleb, 1989].Substituting in (4.24) gives the maximum likelihood estimate

$$\hat{\bar{p}}(\mathbf{x} \rightarrow \mathbf{y}) = 1 - \frac{\sum_{i=1}^{m} \sum_{j=1}^{s} f_{ij} \log(f_{ij}/N_i)}{\sum_{j=1}^{s} N_j \log(N_j/N)} \tag{4.26}$$

For robustness, the Bayesian estimators for these probabilities should be used. These are [Pan and Harris, 1990]

$$\begin{aligned} \hat{p}(x_i|y_j) &= \frac{f_{ij}+1}{N_j+m} \\ \hat{p}(y_j) &= \frac{N_j+1}{N+s} \\ \hat{p}(x_i) &= \frac{N_i+1}{N+m} \end{aligned} \tag{4.27}$$

yielding the Bayesian estimate of the mean of the relative entropy information as

$$\hat{\bar{p}}(\mathbf{x} \rightarrow \mathbf{y}) = 1 - \frac{\sum_{i=1}^{m} \sum_{j=1}^{s} (N_j+1) \frac{f_{ij}+1}{N_j+m} \log\left[\left(\frac{f_{ij}+1}{N_j+m}\right)\left(\frac{N_j+1}{N+s}\right)\left(\frac{N+m}{N_i+1}\right)\right]}{\sum_{j=1}^{s} (N_j+1) \log\left(\frac{N_j+1}{N+s}\right)} \tag{4.28}$$

We set $m = 2$ since we are only concerned with segmenting the image into two grey-level subranges. A simple iterative process is used to select the optimum threshold: the optimum threshold r is selected as the value which maximizes the entropy information (4.26) or (4.28).

4.2 Entropic threshold selection based on other models and distributions

Most of the methods described in the previous section have not motivated the choice of the Shannon entropy as a criterion function, nor has any explicit mention of a model of the image or the thresholding process been made, except in the case of Pan and Harris [Pan and Harris, 1990]. This is not to say that the methods lack validity: they simply lack theoretical or philosophical motivation. The methods introduced below are based more explicitly on image models. Some attempts are also made to include measures of spatial inter-dependence to improve performance.

The monkey model

Given the number of times this model has come up in this thesis, it is surprising that it has not yet been used in the context of threshold selection. This oversight is remedied here.

Given our aim in thresholding an image, i.e. we wish to split it into two relatively flat, homogeneous regions as far as possible, the application of entropy based on a monkey model of the image classes viewed independently should yield good results. In the absence of prior information, entropy will give us the flattest solution without contradicting the data (the input image). The development follows similar lines to those expressed in the previous section. The entropy of the image itself, viewing pixel grey-levels as probabilities of illumination, is given by (3.7) or, equivalently,

$$\begin{aligned} H &= -\sum_{i=1}^{N} p_i \log p_i \quad , \; p_i = \frac{g_i}{G} \\ &= -\sum_{g=0}^{n-1} f_g \frac{g}{G} \log \frac{g}{G} \end{aligned} \tag{4.29}$$

where $G = \sum_{i=1}^{N} g_i = \sum_{g=0}^{n-1} g\, f_g$ is the total illumination of the image. Thresholding the image results in two classes. If viewed as independent components of the image, we can define

corresponding class entropies

$$\begin{aligned} H_0(T) &= -\sum_{i=1(i\in 0)}^{N} \frac{g_i}{G_0(T)} \log \frac{g_i}{G_0(T)} \\ &= -\sum_{g=0}^{T} f_g \frac{g}{G_0(T)} \log \frac{g}{G_0(T)} \\ H_1(T) &= -\sum_{i=1(i\in 1)}^{N} \frac{g_i}{G_1(T)} \log \frac{g_i}{G_1(T)} \\ &= -\sum_{g=T+1}^{n-1} f_g \frac{g}{G_1(T)} \log \frac{g}{G_1(T)} \end{aligned} \tag{4.30}$$

where

$$G_0(T) = \sum_{g=0}^{T} g\, f_g$$

and

$$G_1(T) = \sum_{g=T+1}^{n-1} g\, f_g$$

As before, both the threshold selection schemes of Kapur *et al* [Kapur *et al*, 1985] ((4.8) and (4.9)) and Brink [Brink, 1991b] (4.10) can be used to determine the optimum threshold τ.

It was indicated above that the uninformative prior assuming a flat distribution could be expected to yield good results. However, segmentation is equally dependent on clear, sharp discontinuities and similar information: if such information is more important than our tacit assumption of uniformity, we can expect an improvement in performance on taking such information into account. To this end a modified process using the relative entropy (2.2) may be used.

A measure of local grey-level variance defined over a 3×3 neighbourhood (Figure 4.1), N_3, can be defined as

$$\sigma_i^2 = \sum_{i\in N_3} \frac{(g_i - \mu_{N_3})^2}{9}$$

where μ_{N_3} is the mean grey-level in the pixel neighbourhood. The relative entropy (2.2) of the image is then given by

$$H = -\sum_{i=1}^{N} p_i \log \frac{p_i}{m_i}\ ,\ p_i = \frac{g_i}{G}\ ,\ m_i = 1 + \sigma_i^2 \tag{4.31}$$

The measure m_i is set to $1 + \sigma_i^2$ as there is no guarantee that the variance will be non-zero in all neighbourhoods. The class entropies become

$$H_0(T) = -\sum_{i=1(i\in 0)}^{N} g_i/G_0(T) \log \frac{g_i/G_0(T)}{m_i}$$
$$H_1(T) = -\sum_{i=1(i\in 1)}^{N} g_i/G_1(T) \log \frac{g_i/G_1(T)}{m_i} \tag{4.32}$$

Note that these can no longer be expressed purely in terms of the histogram, as in (4.30), due to the context-related nature of the m_i. The process of threshold selection remains unchanged.

Frieden model of a digital image bounded above and below

This model was described in Section 3.2 of the previous chapter. According to this model, the entropy of a digital image with lower grey-level bound a and upper bound b is given by (3.10) as

$$H = -\sum_{i=1}^{N} (g_i - a) \log(g_i - a) - \sum_{i=1}^{N} (b - g_i) \log(b - g_i)$$

or, equivalently, by (3.11) as

$$H = -\sum_{g=a}^{b} f_g(g - a) \log(g - a) - \sum_{g=a}^{b} f_g(b - g) \log(b - g)$$

Brink [Brink, 1992b] has proposed an implementation of this model for threshold selection. Briefly, the image is again split into two subimages corresponding to background and foreground at each threshold iteration. The class entropies, from (3.11), are then given by

$$H_0(T) = -\sum_{g=a}^{T} f_g(g - a) \log(g - a) - \sum_{g=a}^{T} f_g(T - g) \log(T - g)$$
$$H_1(T) = -\sum_{g=T+1}^{b} f_g(g - (T+1)) \log(g - (T+1)) - \sum_{g=T+1}^{b} f_g(b - g) \log(b - g) \tag{4.33}$$

The optimum threshold τ can again be found either by maximizing the sum of $H_0(T)$ and $H_1(T)$ or by determining the maximin threshold (4.10). The interpretation of a and b in this case [Brink, 1992b] is that they are the respective *a posteriori* limits of the original image, i.e. the minimum and maximum grey-levels present in the image. The prior limits $g_{\min} = 0$

and $g_{\max} = n - 1$ could also be used. However, the arguments against this model, presented in Chapter 3, should be borne in mind.

Minimum cross-entropy or maximum relative entropy threshold selection

The cross-entropy [Kullback, 1959] (2.3) can be used as a non-metric distance measure between two probability distributions p_i and q_i. For our purposes, p_i can be interpreted as a prior distribution and q_i as a posterior distribution. If we regard the original image as the prior distribution (or, in the case of relative entropy, a spatial measure or guide for the thresholded image) and the thresholded result as the posterior distribution [Brink and Pendock, 1993], we can minimize this "distance" to determine the optimum threshold.

Ideally, p_i and q_i should behave as probability distributions, each summing to unity. In practice, for the application of the principles of maximum entropy or minimum cross-entropy, it is sufficient that the distributions are positive, additive and, for comparison of entropies or for determination of the correct relative or cross-entropy values, that they have the same cumulative sums, i.e.

$$\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} q_i \tag{4.34}$$

In order to apply (2.3) to the threshold selection problem, the resulting binary image should have levels given by the below- and above-threshold means, $\mu_0(T)$ and $\mu_1(T)$, of the original image, obtained from its histogram as follows:

$$\begin{aligned} \mu_0(T) &= \sum_{g=0}^{T} \frac{g\, p_g}{P_0(T)} \\ \mu_1(T) &= \sum_{g=T+1}^{n-1} \frac{g\, p_g}{P_1(T)} \end{aligned} \tag{4.35}$$

Using these as the binary image levels ensures that condition (4.34) is satisfied, i.e.

$$\sum_{i=1}^{N} g_i = \left[\sum_{i=1}^{N} \mu_0(T)\right]_{g_i \le T} + \left[\sum_{i=1}^{N} \mu_1(T)\right]_{g_i > T}$$

where g_i represents the grey-level of pixel i of the original image.

At any threshold T the cross–entropy, using (2.3), can be written

$$\begin{aligned} H_\times(T) &= \sum_{i=1}^{N} q_i(T)\log\frac{q_i(T)}{p_i} \\ &= \left[\sum_{i=1}^{N}\mu_0(T)\log\frac{\mu_0(T)}{g_i}\right]_{g_i\le T} + \left[\sum_{i=1}^{N}\mu_1(T)\log\frac{\mu_1(T)}{g_i}\right]_{g_i>T} \\ &= \sum_{g=0}^{T} f_g\mu_0(T)\log\frac{\mu_0(T)}{g} + \sum_{g=T+1}^{n-1} f_g\mu_1(T)\log\frac{\mu_1(T)}{g} \end{aligned} \tag{4.36}$$

The value of T corresponding do the minimum of this function is selected as the optimum threshold τ, i.e.

$$\tau = \arg\left\{\min_{0\le T<n-1}\{H_\times\}\right\} \tag{4.37}$$

The cross–entropy is not a metric distance measure as it is not symmetric with respect to the two distributions p_i and q_i, i.e.

$$\sum_{i=1}^{N} q_i\log\frac{q_i}{p_i} \neq \sum_{i=1}^{N} p_i\log\frac{p_i}{q_i}$$

If desired, symmetry could be imposed by simply adding the two distances to give

$$H'_\times = \sum_{i=1}^{N} q_i\log\frac{q_i}{p_i} + \sum_{i=1}^{N} p_i\log\frac{p_i}{q_i} \tag{4.38}$$

It is interesting to note that there is some similarity of form between (4.38) and that of a correlation coefficient. Consider

$$\begin{aligned} q\log(q/p) + p\log(p/q) &= \log(q/p)^p + \log(p/q)^q \\ &= \log\frac{q^q p^p}{p^q q^p} \end{aligned}$$

The expression inside the logarithm resembles *(variance/covariance)*, an inverted measure of correlation, where here "variance" is given by $\sigma^x \propto \sum x^x$ instead of the usual $\sigma^2 \propto \sum x^2$.

The same approach can be used as above to implement this symmetric measure of cross–entropy (4.38) for threshold selection. Following (4.36), (4.38) becomes

$$\begin{aligned} H'_\times(T) = &\sum_{g=0}^{T} f_g\left[\mu_0(T)\log\frac{\mu_0(T)}{g} + g\log\frac{g}{\mu_0(T)}\right] \\ &+ \sum_{g=T+1}^{n-1} f_g\left[\mu_1(T)\log\frac{\mu_1(T)}{g} + g\log\frac{g}{\mu_1(T)}\right] \end{aligned} \tag{4.39}$$

and the optimum threshold is determined as before using (4.37).

In Section 2.1 the similarity between the relative entropy measure (2.2) and the cross–entropy (2.3) was noted. Effectively, minimization of the cross–entropy is identical to maximization of the relative entropy, the only difference between (2.2) and (2.3) being the sign in front of the summation. The discussion that follows therefore has equal bearing on both expressions.

If the continuous distributions $p(x)$ and $q(x)$ obey Gaussian statistics, maximization of the entropy of the posterior distribution, $q(x)$, relative to the prior distribution, $p(x)$, or, equivalently, minimization of the cross–entropy, can be shown to be related to maximization of the squared correlation between the distributions as defined by Otsu [Otsu, 1979]. Consider the Gaussian distributions

$$q(x) = \frac{1}{\sqrt{2\pi}\sigma'} \exp \frac{-(x-\mu')^2}{2\sigma'^2}$$

and

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

The cross–entropy of $q(x)$ and $p(x)$ is given by

$$\begin{aligned} H_\times(x) &= \int q(x) \log \frac{q(x)}{p(x)}\, dx \\ &= \log\left(\frac{\sigma}{\sigma'}\right) \int q(x)\, dx - \int q(x) \frac{(x-\mu')^2\sigma^2 - (x-\mu)^2\sigma'^2}{2\sigma^2\sigma'^2}\, dx \\ &= \log\frac{\sigma}{\sigma'} - \frac{1}{2} + \frac{1}{2}\frac{\int (x-\mu)^2 q(x)\, dx}{\int (x-\mu)^2 p(x)\, dx} \end{aligned}$$

If the first moment is preserved ($\mu = \mu'$) then this becomes

$$H_\times(x) = \log\frac{\sigma}{\sigma'} - 0.5 + 0.5\frac{\sigma'^2}{\sigma^2} \tag{4.40}$$

If, as is the case for the problem of threshold selection above, the prior distribution corresponds to the original image and the posterior distribution to the binary image with the first moment of the original image preserved, then the third term in (4.40) is simply an inverted form of Otsu's measure [Otsu, 1979]

$$\begin{aligned} \kappa &= \frac{\sigma_T^2}{\sigma_W^2} \\ &= \frac{\sigma^2}{\sigma'^2} \end{aligned}$$

Maximization of this measure was shown by Otsu to be equivalent to maximization of his so–called "measure of class separability", ν, which in turn has been shown to be the square of

the correlation between the original and thresholded images [Cseke and Fazekas, 1990; Brink, 1990]. The result (4.40) therefore implies therefore that the use of cross–correlation is a special case of the more general cross–entropy or relative entropy principle and is dependent on the distributions being Gaussian in nature. Seth and Kapur [Seth and Kapur, 1990] have also linked this principle to that of the minimum Pearson's chi–squared (χ^2) method of estimation.

Segmentation based on spatial entropy minimization

The concept of spatial entropy [Journel and Deutsch, 1993] was introduced in Chapter 3. This was proposed as a measure of image entropy, not specifically as a criterion function for threshold selection. As suggested in Chapter 3, the calculation of a single value of spatial entropy would involve the determination of an $n \times n$ co–occurrence matrix for each possible lag: for an image consisting of N pixels, taking negative lags into account, this would mean that up to $2N$ such matrices could be required. For even a relatively small image (100×100 pixels, say) this is an enormous task. Applying this scheme to determine the maximum entropy thresholded image could require us to perform this task up to $n - 1$ times, once for each threshold value. The first simplification, therefore, is to treat the image as a type of Markov Random Field, with each pixel dependent only on pixels in its immediate neighbourhood, e.g. 3×3 or 5×5. We then consider only values of λ lying within this neighbourhood. Note that $\lambda = (\lambda_i, \lambda_j)$ for our purposes. In a 3×3 neighbourhood, $-1 \leq \lambda_i \leq +1$ and $-1 \leq \lambda_j \leq +1$. For simplicity we will consider only lags λ defined over the 3×3 neighbourhood (Figure 4.1), N_3, here. A feasible threshold selection scheme is suggested below.

The requirement of stationarity of the probability distribution function p_g (i.e. invariance of the grey–level histogram with position) is observed. For our purposes this implies "wrap–around" of the image: the image is assumed to be periodic, repeating infinitely in each dimension. Without this assumption, the relative spatial entropy would not necessarily be limited to the range $[0, 1]$ as determined in Chapter 3 (3.24).

Stepping iteratively through all possible threshold values, the image is binarized at each threshold, T. The binary (2×2) co–occurrence matrices are calculated for each (bivariate) lag, λ, giving

$$p_{gg'}(\lambda) = p_{00}(\lambda),\ p_{01}(\lambda),\ p_{10}(\lambda),\ p_{11}(\lambda),\ \forall \lambda = (\lambda_i, \lambda_j),\ \lambda_i, \lambda_j = -1, 0, +1$$

Using (3.23) and (3.24) we determine

$$H(T, \lambda) = -\sum_{g=0}^{1} \sum_{g'=0}^{1} p_{gg'}(\lambda) \log p_{gg'}(\lambda)$$

and

$$H_R(T,\lambda) = \frac{H(T,\lambda) - H(T,0)}{H(T,0)}$$

The mean over all lags in the 3×3 neighbourhood, N_3, is

$$\overline{H}_R(T) = \frac{1}{9} \sum_{\lambda \in N_3} H_R(T,\lambda) \tag{4.41}$$

It should be noted that the use of negative as well as positive lags resulting from use of the symmetrical neighbourhood N_3 effectively weights the measure $\overline{H}_R(T)$ in favour of the $H_R(\lambda)\ \forall \lambda \neq 0$. This follows similar resoning to that justifying the use of an asymmetrical neighbourhood by Pal and Pal [Pal and Pal, 1989a] in defining the co–occurrence matrix. In the present approach it should be clear from the symmetry of N_3 that the co–occurrence matrix corresponding to a negative lag is simply a "flipped" version (reversed axes) of that corresponding to the symmetrically opposed positive lag of the same magnitude, i.e. $p_{gg'}(+\lambda) = p_{g'g}(-\lambda)$. This implies $H(+\lambda) = H(-\lambda)$. Besides the biased weighting that this causes, many of the calculations are redundant. We therefore consider only positive (or only negative) lags, reducing the neighbourhood further to include only the origin, its immediate neighbours to the right and below and its diagonal neighbours below which fall within the original 3×3 neighbourhood. The modified neighbourhood, N_3', is shown in Figure 4.4.

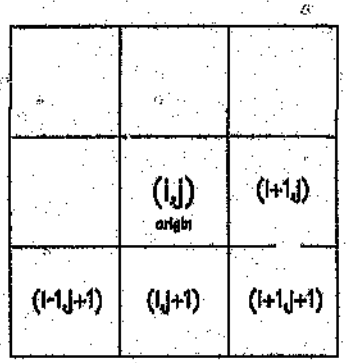


Figure 4.4 *The asymmetrical 4–neighbourhood ("future" samples or positive lags) of a pixel centred at image coordinates* (i,j).

This unbiased version yields

$$\overline{H}_R(T) = \frac{1}{5} \sum_{\lambda \in N_3'} H_R(T,\lambda) \tag{4.42}$$

Since it is a measure of spatial disorder, maximizing the relative spatial entropy with respect to the threshold would yield the most disordered binary result: this is not usually desirable.

The guiding principle here is therefore that of spatial entropy *minimization*. Since the trivial single-level result of thresholding in such a way that all pixels belong to the same class could be seen as displaying a minimum of spatial disorder, it is necessary (and not unreasonable) to limit the range of allowable threshold values to lie within the grey-scale range of the input image, i.e. $a < T < b$ where $a \geq 0$ and $b \leq n-1$ are the *a posteriori* limits of the original image.

Note that this method deals directly with the binary results of thresholding the image, not the original grey-level arrangement of each class as other methods have done. Determination of separate relative spatial class entropies would require stationarity of the grey-level distribution of each class, i.e. wrap-around within each class as opposed to the overall image. Since the regions corresponding to each class are usually to some extent fragmented there is no obvious computationally efficient way to achieve this. An absolute measure of spatial entropy, not dependent on an assumption of a stationary probability density function, could be used in such a case: this is effectively what has been done by Pal and Pal and others [Pal and Pal, 1989a; Abutaleb, 1989; Brink, 1992a]. Another approach is to determine the relative spatial entropies of the two subranges of the histogram resulting from its partitioning by a threshold: a histogram wrap-around could be used to ensure stationarity and the resulting relative spatial entropies of the histogram may be used to select the optimum threshold in terms of histogram partitioning. The following method based on the histogram autocorrelation function is based on such an argument.

Segmentation using the autocorrelation function of the histogram

The method described here uses the autocorrelation function of the histogram [Brink, 1994]. The entropy of this distribution is given by (3.28) as

$$H_r = -\sum_{k=1-n}^{n-1} p_k^* \log p_k^*$$

where

$$p_k^* = \frac{r_k}{\sum_{k=1-n}^{n-1} r_k}$$

is the probability that pixels differing by k grey-levels are related, r_k being the correlation coefficient for a delay of k signal (histogram) samples.

When the image is thresholded, the histogram is split into (in the case of binary thresholding) two subranges which are ideally distinct and homogeneous, allowing them to be regarded as separate distributions. At the optimum threshold it can be expected that these distributions would be at their most uniform, resulting in a high degree of "inter-relatedness".

Given an image with a grey-level range $a \leq g \leq b$, the threshold T splits this into two ranges $a \leq g \leq T$ and $T+1 \leq g \leq b$. Following (3.26), the coefficients of the autocorrelation functions for these subranges can be determined as

$$r_k^0(T) = \sum_{g=a}^{T} p_{k+g}^0(T)\, p_g^0(T) \qquad -(T-a) \leq k \leq T-a \tag{4.45}$$

where

$$p_g^0(T) = \begin{cases} p_g / P_0(T) & a \leq g \leq T \\ 0 & g < a,\ g > T \end{cases}$$

and

$$r_k^1(T) = \sum_{g=T+1}^{b} p_{k+g}^1(T)\, p_g^1(T) \qquad -(b-T-1) \leq k \leq b-T-1 \tag{4.46}$$

where

$$p_g^1 = \begin{cases} p_g / P_1(T) & T+1 \leq g \leq b \\ 0 & g < T+1,\ g > b \end{cases}$$

$P_0(T)$ and $P_1(T)$ are the class probabilities (4.1) as before. The autocorrelation values are normalized to get "probabilities" of inter-relatedness

$$p_k^{*0}(T) = \frac{r_k^0(T)}{\sum_{k=a-T}^{T-a} r_k^0(T)}$$

$$p_k^{*1}(T) = \frac{r_k^1(T)}{\sum_{k=T+1-b}^{b-T-1} r_k^1(T)}$$

The corresponding class entropies can be defined as

$$\begin{aligned} H_0^r(T) &= -\sum_{k=a-T}^{T-a} p_k^{*0}(T) \log p_k^{*0}(T) \\ H_1^r(T) &= -\sum_{k=T+1-b}^{b-T-1} p_k^{*1}(T) \log p_k^{*1}(T) \end{aligned} \tag{4.47}$$

As before [Kapur *et al*, 1985], the sum of these entropies can be maximized to find the optimum threshold (4.9) or, alternatively, the threshold corresponding to the maximin [Brink, 1991b] of $H_0^r(T)$ and $H_1^r(T)$ (4.10) can be implemented.

Ideally we would want to apply such a scheme to the image itself, taking into account the relationships between actual pixels. However, in thresholding the image the resulting regions

(classes) are not necessarily connected or continuous: determining the sample positions (coordinates) of pixels relative to their region origins is therefore not possible. In other words, in image space the regions cannot be taken as two separate, independent distributions.

Threshold selection based on the "labelled photon" model

This model is described in Section 3.3. Effectively the sample coordinate of the digital signal (image) is itself used as a probability in the entropy expression (3.31) or (3.32). In this approach, the disjointedness of the sub-images resulting from thresholding is ignored. The two class entropies, from (3.31) and (3.32), are thus given by

$$\begin{aligned} H_0(T) &= -\sum_{j=1}^{G} i_j^0(T) \log i_j^0(T) \\ &= -\sum_{i=1}^{N}{\sum_{i\in 0}} g_i\, i \log i \\ H_1(T) &= -\sum_{j=1}^{G} i_j^1(T) \log i_j^1(T) \\ &= -\sum_{i=1}^{N}{\sum_{i\in 1}} g_i\, i \log i \end{aligned} \tag{4.48}$$

where $i_j^0(T)$ denotes the image coordinate i of all photons j such that the grey-level $g_i \leq T$, and $i_j^1(T)$ denotes those of photons contributing to areas where $g > T$. Again a precondition for comparison of these distributions is that we require

$$\sum_{j=1}^{G} i_j^0(T) = \sum_{j=1}^{G} i_j^1(T)$$

The simplest solution is to normalize both distributions, as in (3.32) and (3.34), to get

$$\begin{aligned} H_0(T) &= -\sum_{j=1}^{G} p_j^0(T) \log p_j^0(T)\,, \quad p_j^0(T) = \frac{i_j^0(T)}{I_0} \\ &= -\sum_{i=1}^{N}{\sum_{i\in 0}} g_i\, p_0 \log p_0\,, \quad p_0 = \frac{i}{I_0}\,, \; i \in 0 \\ H_1(T) &= -\sum_{j=1}^{G} p_j^1(T) \log p_j^1(T)\,, \quad p_j^1(T) = \frac{i_j^1(T)}{I_1} \\ &= -\sum_{i=1}^{N}{\sum_{i\in 1}} g_i\, p_1 \log p_1\,, \quad p_1 = \frac{i}{I_1}\,, \; i \in 1 \end{aligned} \tag{4.49}$$

where

$$I_0 = \sum_{i=1}^{N} \sum_{i \in 0} i g_i$$

$$I_1 = \sum_{i=1}^{N} \sum_{i \in 1} i g_i$$

are the class coordinate sums. As is with most methods dealt with, either maximization of the sum of $H_0(T)$ and $H_1(T)$ or maximization of the smaller of these can be used to determine optimum thresholds.

In the next chapter the results of applying some of the methods of Sections 4.1 and 4.2 are shown. While one can often evaluate results of this nature on the basis of their appearance, it is desirable to quantify this in some way in order to evaluate techniques in an unbiased manner. This subject is broached in Chapter 6.

Chapter 5

Results

The results of applying many of the thresholding processes described in the previous chapter are presented here. Three sample images have been used to demonstrate their performance, all of them reasonably suited to bi-level thresholding. All the images have one byte (8 bits) per pixel, allowing a grey-scale range of 256 discrete levels ($g = 0, \cdots, 255$). The image dimensions vary. The first is a scanned image of René Magritte's painting *The use of words I* (Figures 1.1 and 5.1), consisting of 300×236 pixels ($N = 70800$) The next is a low contrast video image of part of a printed page (Figure 5.2) of 126×126 pixels ($N = 15876$) and the final image (Figure 5.3) is an infrared image of a warm tarred road in the early evening, made up of 256×256 pixels, i.e. $N = 65536$. Other images have been used in this research, but are in general not well suited to binary thresholding, yielding results which, while sometimes aesthetically pleasing, have little application.

In Section 5.1 the results of using the techniques described in Section 4.1 are presented. The methods used are those based on the histogram [Kapur *et al*, 1985; Brink, 1991b], the (grey-level, local mean grey-level) scatterplot [Abutaleb, 1989; Brink, 1992a], the 2-neighbour grey-level co-occurrence matrix [Pal and Pal, 1989a] and the relative entropy information [Pan and Harris, 1990].

The results due to the new methods described in Section 4.2 are shown in Section 5.2. These include the standard and the context-dependent (modified) methods based on the monkey model, each using both threshold determination schemes (4.9) and (4.10), the model with both upper and lower *a posteriori* grey-level bounds [Brink, 1992b], minimum cross-entropy threshold selection [Brink and Pendock, 1993], threshold selection by relative spatial entropy minimization, thresholding based on the autocorrelation function of the histogram [Brink, 1994] and, finally, the results of implementing the "labelled photon" model.

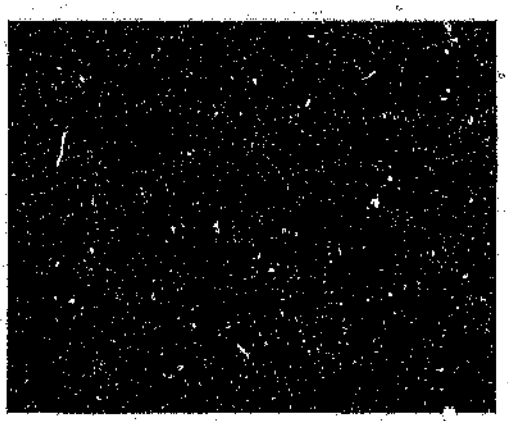


Figure 5.1. The use of words I, *René Magritte, 1928-29. 1 byte per pixel,* 300 × 236 *pixels, grey-level range* $0 \leq g \leq 200$.

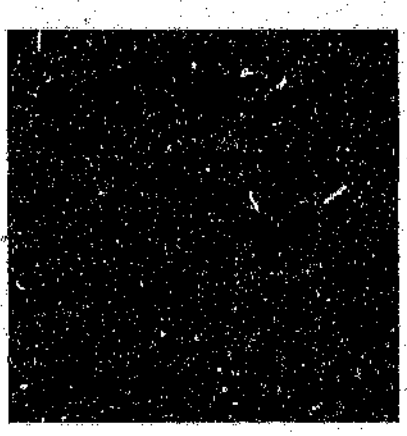


Figure 5.2. *Video image of a portion of a printed page. 1 byte per pixel,* 126 × 126 *pixels, grey-level range* $33 \leq g \leq 63$.

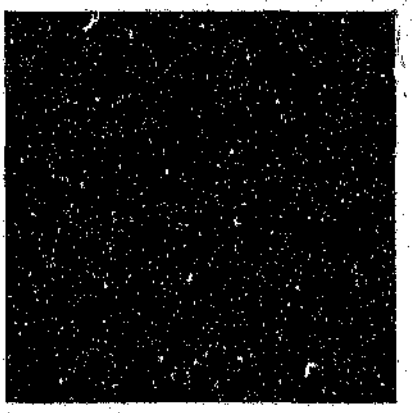


Figure 5.3 *Infrared image of a warm tarred road in the early evening. 1 byte per pixel,* 256 × 256 *pixels, grey-level range* $18 \leq g \leq 178$.

The results are presented with minimal comment. Some remarks regarding implementation, relative computation times and limited subjective evaluations appear in Section 5.3.

5.1 Thresholded images resulting from the methods of Chapter 4.1

In the figures that follow results of the implementation of the threshold selection methods described in Chapter 4.1 are presented for each of the sample images (Figures 5.1–5.3). To facilitate comparison the results for each image are presented together and numbered as Figure 5.x (a) to (g). Each figure has been broken across two pages due to size constraints. The letters refer to the following methods:

(a) Results of applying the method based on the histogram entropy proposed by Kapur *et al* [Kapur *et al*, 1985].

(b) Results of using the modified method based on the histogram entropy, due to Brink [Brink, 1991b], where the *maximin* of the class entropies is used to select the threshold.

(c) The uncorrected method based on the (grey–level, local mean grey–level) scatterplot, due to Abutaleb [Abutaleb, 1989]. Note that a 2–dimensional threshold vector is selected here.

(d) Corrected version of the above method due to Brink [Brink, 1992a], again using a *maximin* entropy threshold selection criterion.

(e) Results of using the first algorithm proposed by Pal and Pal [Pal and Pal, 1989a], based on the class (quadrant) entropies of quadrants A and C of the 2–neighbour co–occurrence matrix (Figure 4.3).

(f) The second algorithm due to Pal and Pal [Pal and Pal, 1989a], using the entropies of quadrants B and D of the co–occurrence matrix.

(g) The method of Pan and Harris [Pan and Harris, 1990] maximizing the relative entropy information measure: the results are identical for both the maximum likelihood (4.26) and the Bayesian (4.28) estimates.

The actual threshold value (or vector) selected in each case is given in the relevant caption.

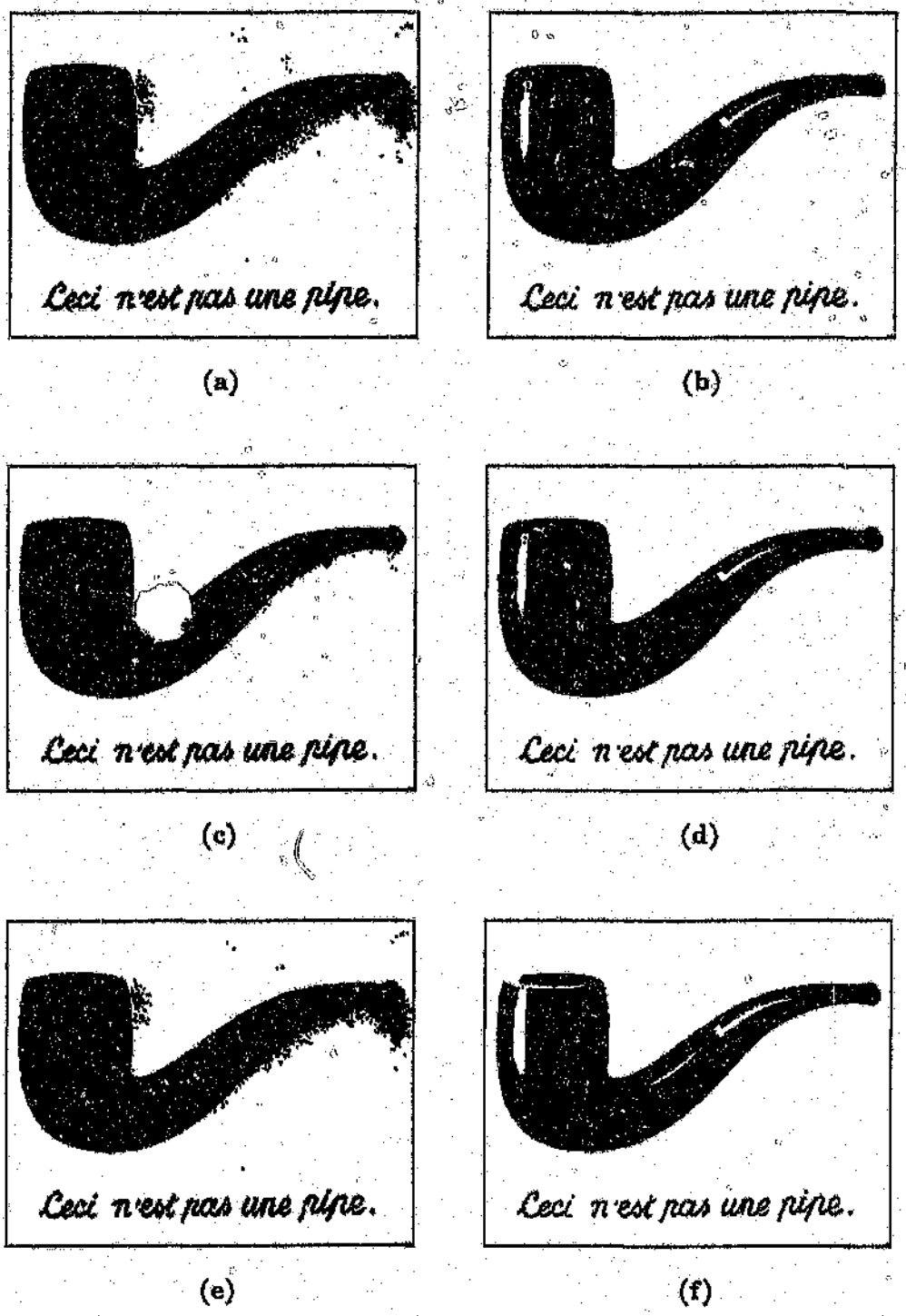


Figure 5.4 *Results of thresholding the image of Figure 5.1:* (a) $\tau = 133$, (b) $\tau = 95$, (c) $(\tau, \sigma) = (120, 111)$, (d) $(\tau, \sigma) = (94, 91)$, (e) $\tau = 132$ *and* (f) $\tau = 81$.

Continued overleaf...

(g)

Figure 5.4 (continued) *Results of thresholding the image of Figure 5.1:* (g) $\tau = 137$.

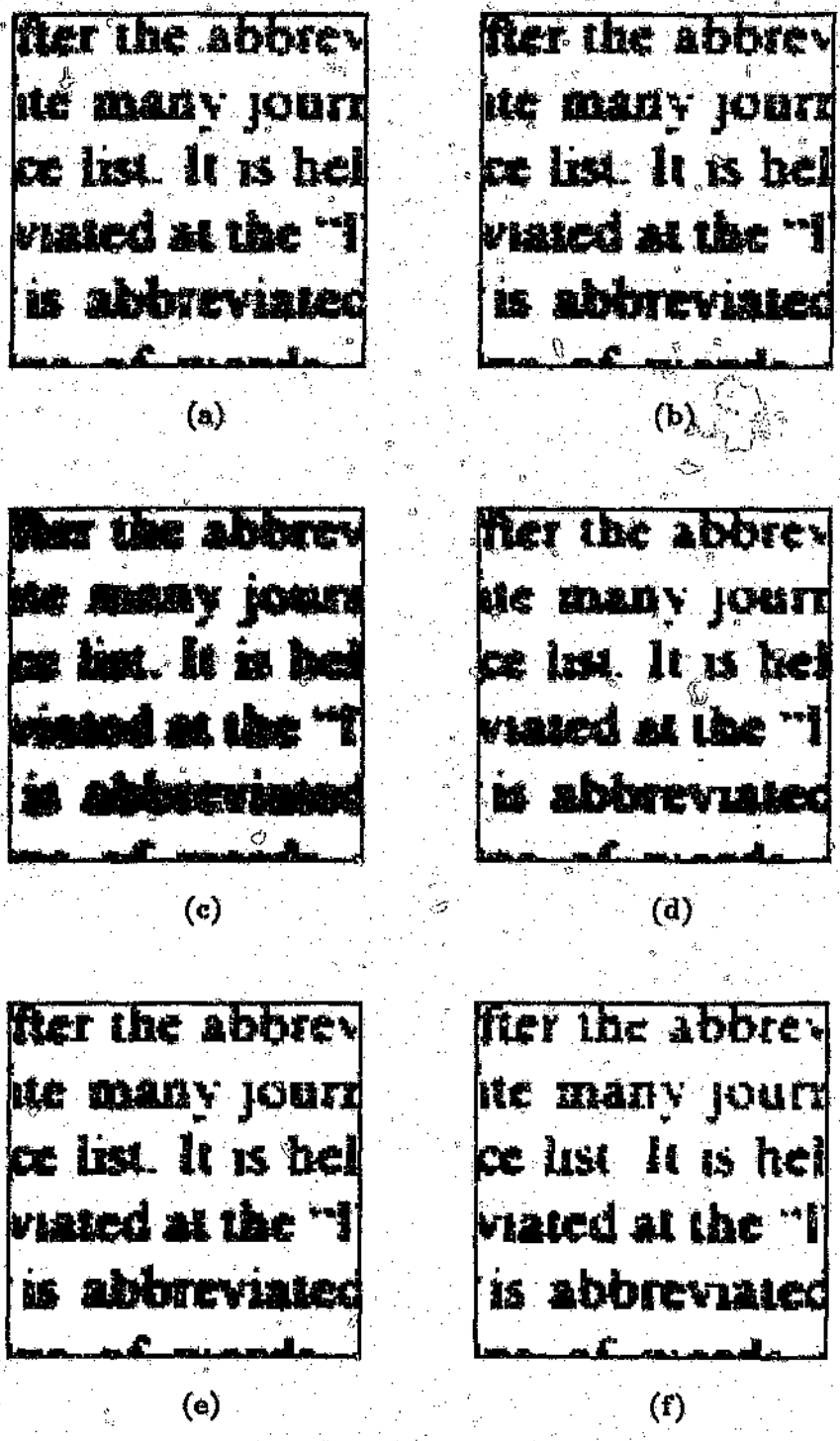
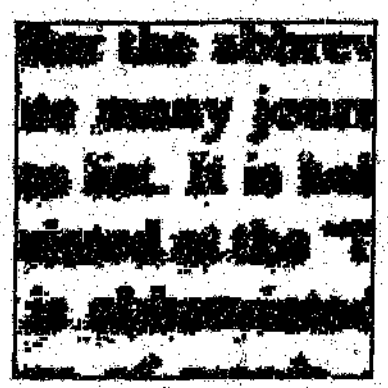


Figure 5.5 *Results of thresholding the image of Figure 5.2:* (a) $\tau = 48$, (b) $\tau = 48$, (c) $(\tau, \sigma) = (51, 50)$, (d) $(\tau, \sigma) = (46, 47)$, (e) $\tau = 48$ *and* (f) $\tau = 45$.

Continued overleaf...



(g)

Figure 5.5 (continued) *Results of thresholding the image of Figure 5.2:* (g) $\tau = 53$.

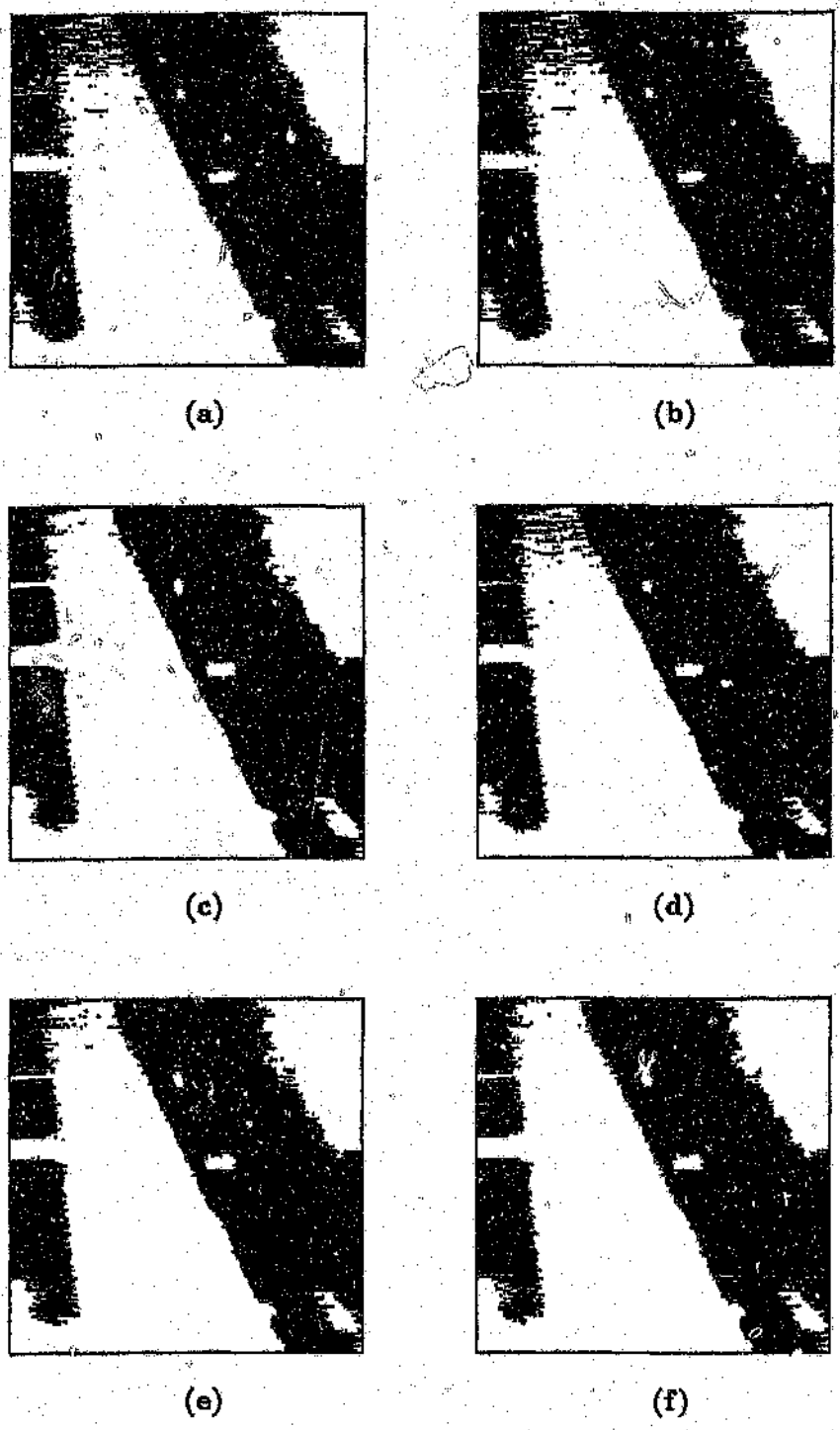


Figure 5.6 *Results of thresholding the image of Figure 5.3:* (a) $\tau = 94$, (b) $\tau = 94$, (c) $(\tau, \sigma) = (86, 87)$, (d) $(\tau, \sigma) = (92, 91)$, (e) $\tau = 87$ *and* (f) $\tau = 85$.

Continued overleaf...

(g)

Figure 5.6 (continued) *Results of thresholding the image of Figure 5.3:* (g) $r = 86$.

5.2 Thresholded images resulting from the methods of Chapter 4.2

In the figures that follow results of the implementation of the new thresholding algorithms introduced in Chapter 4.2 are presented. To facilitate comparison the results for each of the images of Figures 5.1–5.3 are presented together as far as possible. These have been numbered as Figure 5.x (a) to (m). Due to page size constraints, each figure is broken across three pages. The letters (a) to (m) refer to the following methods:

(a) Results of applying the first method based on the entropy derived from the monkey model of the image, determining the optimum threshold as that value which maximizes the sum of the class entropies.

(b) Results of applying the second method based on the entropy derived from the monkey model of the image, where the optimum threshold is the value which maximi the smaller of the class entropies (*maximin* approach).

(c) The modified method based on the monkey model, incorporating local variance data as a measure in the relative entropy expression (4.31) and (4.32). The sum of the class entropies (4.32) is maximized to determine the threshold.

(d) The modified method based on the monkey model, maximizing the smaller of the class entropies (4.32) to determine the threshold.

(e) The method based on Frieden's [Frieden, 1980; Brink, 1992b] model of an image with known upper as well as lower bounds. The sum of the class entropies is maximized to determine the threshold.

(f) As above, but using the *maximin* method to determine the threshold.

(g) Results of using the principle of minimum cross-entropy [Brink and Pendock, 1993].

(h) Results of implementing the symmetric version of the cross-entropy formula (4.31) to determine the optimum threshold.

(i) Minimization of the relative spatial entropy (4.42) [Journel and Deutsch, 1993] of the binary result in order to select a suitable threshold.

(j) Maximization of the sum of the class entropies as determined from the autocorrelation functions of the class histograms [Brink, 1994].

(k) *Maximin* of the class entropies as determined from the autocorrelation functions of the class histograms [Brink, 1994].

(l) Results arising from implementation of the method based on the "labelled photon" model of the image. The sum of the class entropies is maximized to determine the threshold.

(m) As above, but using the *maximin* of the class entropies as a threshold selector.

The actual threshold value (or vector) selected in each case is given in the relevant caption.

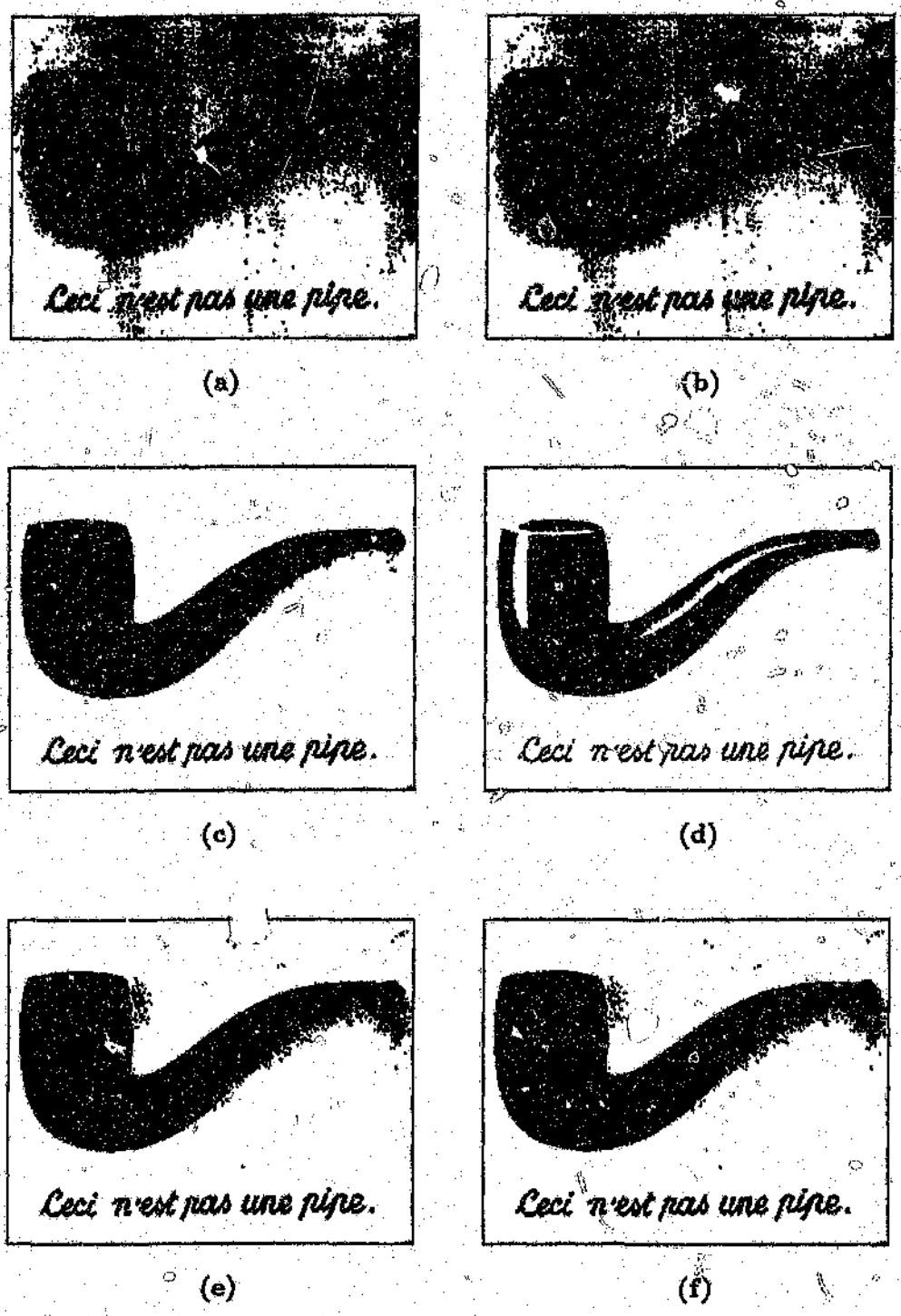


Figure 5.7 *Results of thresholding the image of Figure 5.1:* (a) $\tau = 153$, (b) $\tau = 153$, (c) $\tau = 120$, (d) $\tau = 65$, (e) $\tau = 132$, (f) $\tau = 132$.

Continued overleaf...

(g)

(h)

(i)

(j)

(k)

Figure 5.7 (continued) *Results of thresholding the image of Figure 5.1:* (g) $\tau = 67$, (h) $\tau = 64$, (i) $\tau = 119$, (j) $\tau = 139$, (k) $\tau = 70$.

Continued overleaf...

(l) (m)

Figure 5.7 (continued) *Results of thresholding the image of Figure 5.1:* (l) $\tau = 156$, (m) $\tau = 157$.

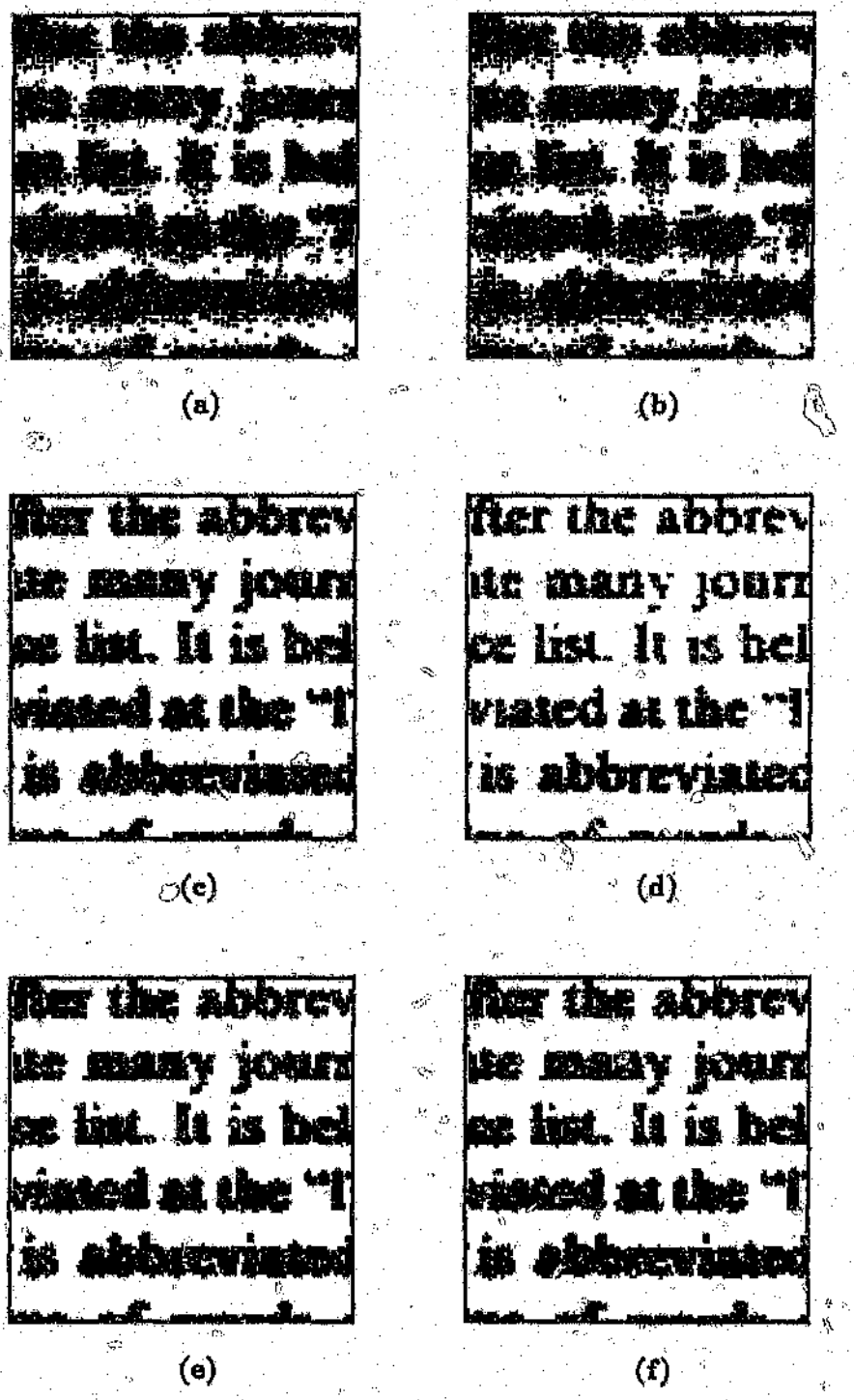
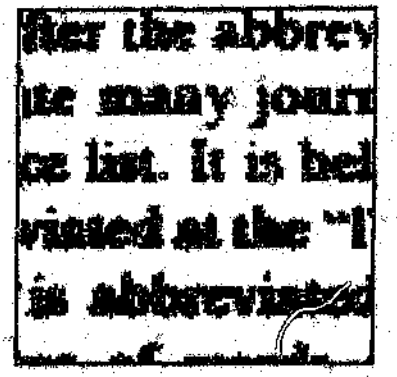
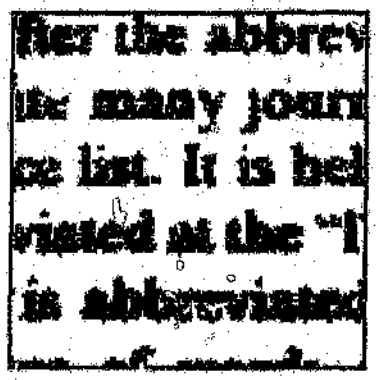


Figure 5.8 *Results of thresholding the image of Figure 5.2:* (a) $\tau = 56$, (b) $\tau = 56$, (c) $\tau = 51$, (d) $\tau = 48$, (e) $\tau = 51$, (f) $\tau = 51$.

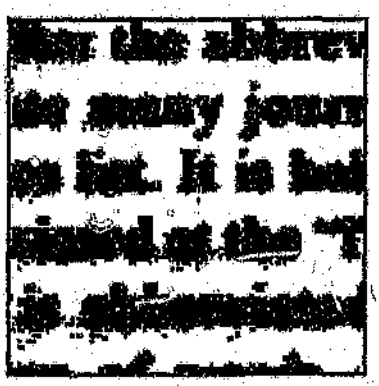
Continued overleaf...



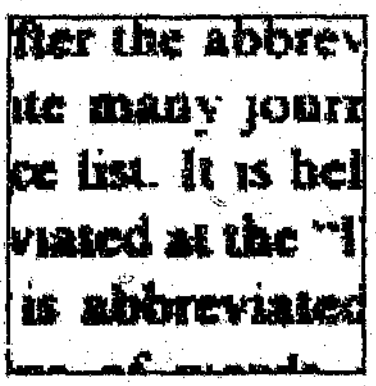
(g)



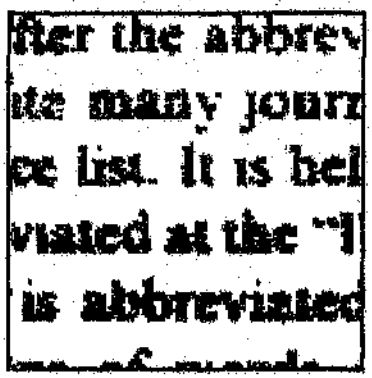
(h)



(i)



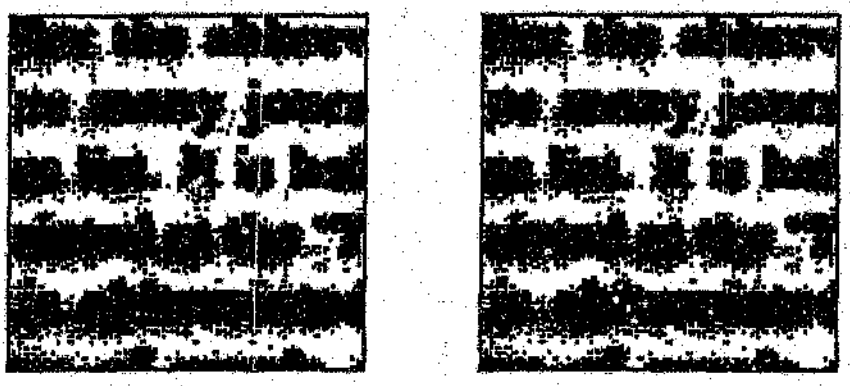
(j)



(k)

Figure 5.8 (continued) *Results of thresholding the image of Figure 5.2:* (g) $\tau = 49$, (h) $\tau = 49$, (i) $\tau = 53$, (j) $\tau = 48$, (k) $\tau = 48$.

Continued overleaf...



(l) (m)

Figure 5.8 (continued) *Results of thresholding the image of Figure 5.2:* (l) $\tau = 56$, (m) $\tau = 56$.

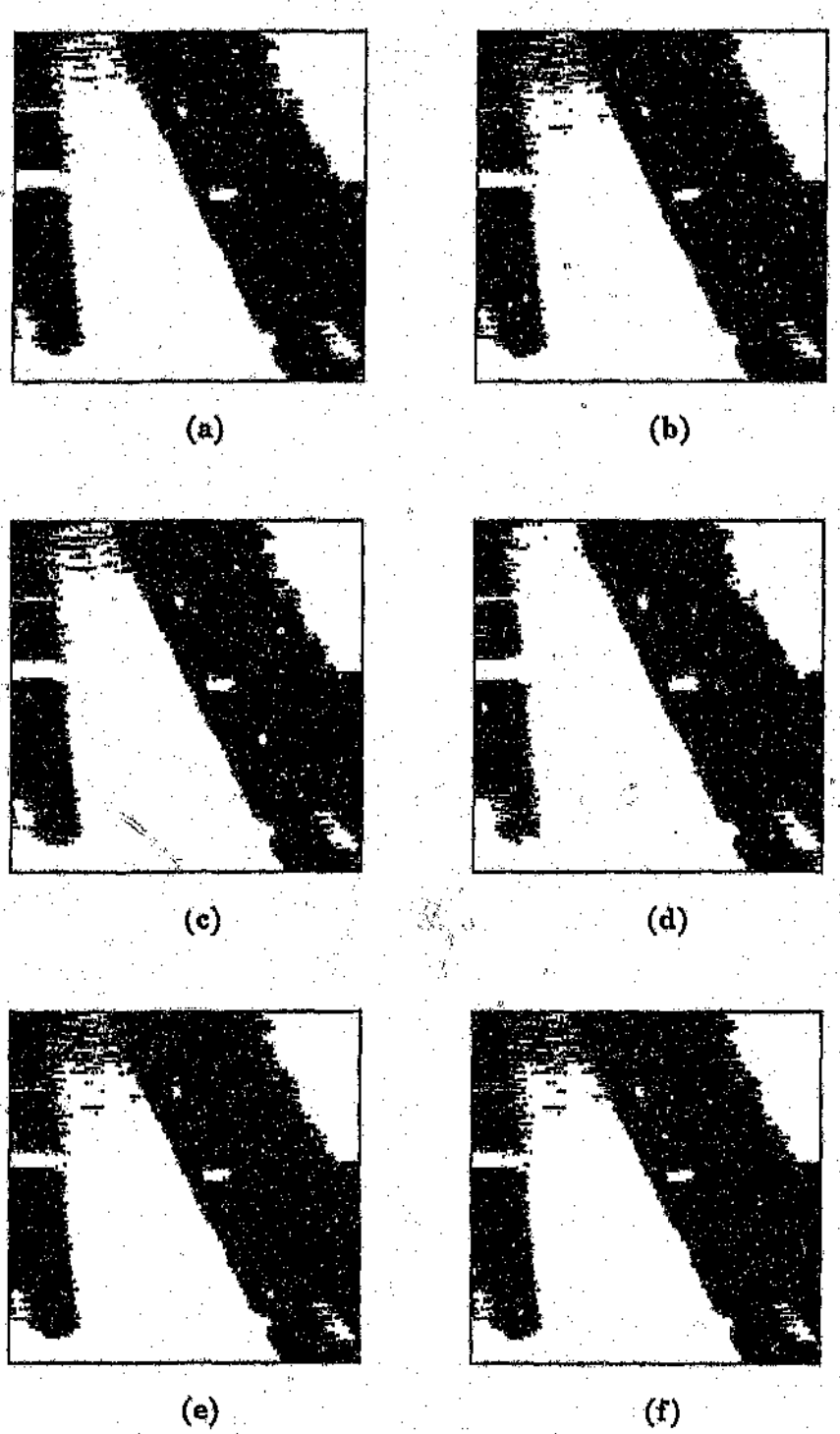
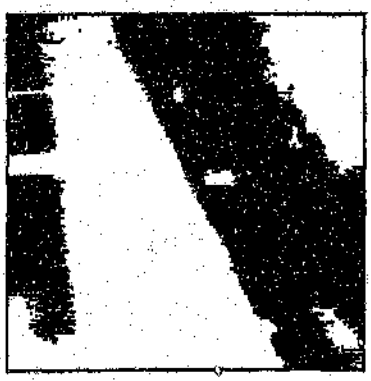


Figure 5.9 *Results of thresholding the image of Figure 5.3:* (a) $\tau = 90$, (b) $\tau = 94$, (c) $\tau = 91$, (d) $\tau = 85$, (e) $\tau = 94$, (f) $\tau = 95$.

Continued overleaf...

(g)

(h)



(i)

(j)

(k)

Figure 5.9 (continued) *Results of thresholding the image of Figure 5.3:* (g) $\tau = 84$, (h) $\tau = 83$, (i) $\tau = 85$, (j) $\tau = 100$, (k) $\tau = 95$.

Continued overleaf...

(l) (m)

Figure 5.9 (continued) *Results of thresholding the image of Figure 5.3:* (l) $\tau = 117$, (m) $\tau = 116$.

5.3 Remarks

The results of Section 5.1 indicate an aesthetic improvement when the maximin approach (Figures 5.4–6 (b) and (d)) is used. In particular, the results of Figures 5.4 (a) and (c) (based on summation of class entropies) appear to be more successful in actually segmenting the image into strict "background" and "foreground" regions without regard for illumination effects such as reflections, while the maximin approach (Figures 5.4 (b) and (d)) appears rather to "binarize" the image in an attempt to approximate the grey–scale image with only two ouput levels. However, from this set of examples it is clear that the grey–levels corresponding to the "background" class at the mouthpiece end of the pipe in Figure 5.1 are similar to those constituting the "foreground" class in the reflection on the pipe bowl and stem. Global threshold selection cannot avoid such errors. The method of Abutaleb [Abutaleb, 1989] (Figure 5.4(c)) comes closest to correctly distinguishing between these classes, due to the use of limited contextual information to select two thresholds, one for grey–level and one for local mean grey–level. Unfortunately, one is left in the dark as to how to automatically classify those pixels which remain unclassified, shown as an intermediate grey–level in Figures 5.4–6 (c) and (d). This could possibly be accomplished by a region growing process, but such methods can be time consuming. The methods using the grey–level co–occurrence matrix [Pal and Pal, 1989a] (Figures 5.4–6 (e) and (f)) appear to yield useful results from both perspectives. Use of the actual class quadrants A and C (Figure 4.3) gives a relatively good separation of "objects" from "background" and the conditional entropies from quadrants B and D yield results that are aesthetically pleasing. Finally, as one would expect from the measure's context–sensitive nature, the results of using the relative entropy information [Pan and Harris, 1990] (Figures 5.4–6 (g)) also display a tendency towards actual segmentation rather than mere binarization.

The results of applying the new methods introduced in Chapter 4.2 (Section 5.2) are as varied as those above. Among these, those obtained by implementation of both the standard (non-metric) cross-entropy or relative entropy and the modified metric version stand out as being among the most aesthetically pleasing (Figures 5.7-9 (g) and (h)). This measure, while not taking specific account of pixel positions in the image, does use context-related information by assuming the original grey-scale image as a prior model of the thresholded result. It can also be regarded as a relation of the monkey model, which simply assumes a uniform prior distribution. The direct implementation of a monkey model for threshold selection (Figures 5.7-9 (a) and (b)) gives quite disappointing results. These are drastically improved by incorporation of the local variance as a measure on the image (Figures 5.7-9 (c) and (d)). As could be expected by its *ad hoc* nature, the results of using the "labelled photon" model (Figures 5.7-9 (l) and (m)) appear to be unusable for any normal application. The methods based on the Frieden model of an image bounded above and below (Figures 5.7-9 (e) and (f)), as well as the approach based on minimization of the relative spatial entropy (4.42) (Figures 5.7-9 (i)), have yielded results comparable to those of the methods of Kapur *et al* [Kapur *et al*, 1985] and Abutaleb [Abutaleb, 1989] (Figures 5.4-6 (a) and (c)) and appear to have separated the image classes as well as possible given the global nature of the thresholding process. The entropy of the autocorrelation function of the histogram [Brink, 1994] has yielded similar results to the methods based directly on the histogram [Kapur *et al*, 1985; Brink, 1991b]: when the sum of the class entropies is maximized, the result is appears to try to separate the classes (Figure 5.7-9 (j)), while the maximin process results in a more aesthetically pleasing output (Figure 5.7-9(k)).

It will have been noticed that many of the techniques used, while purporting to be based on the image itself as a probability distribution, can easily be (and in many cases have been) expressed in terms of the image histogram alone. While this obviously means that only information from this rather limited distribution is used, it has also enabled the processes concerned to be speeded up substantially in terms of computation time. Often more than one threshold may be required, as many images consist of more than just two classes. Most of the methods described can relatively easily be extended to allow selection of such multiple thresholds, at the cost of increased computation time. The processing speed is directly dependent on the number of thresholds desired, the number of required calculations being proportional to

$$ {}_{(b-a)}C_r = \frac{(b-a)!}{r!(b-a-r)!} \tag{5.1} $$

where $(b-a)+1$ is the number of grey-levels in the original image ($(b-a)$ is thus also the number of possible threshold candidates) and r is the number of thresholds desired.

The evaluation of thresholding methods and results is often a subjective process. In Chapter 6 m[illegible]ods of quantitative threshold evaluation are presented and discussed.

Chapter 6

Evaluation methods

The evaluation of thresholding methods appears to have received little attention in the past. While subjective evaluation of results for a specific application is in general quite effective, it would be far more satisfactory to be able to quantify one's preference. Most methods suggested to date have dealt more specifically with the evaluation of the *results* of using various threshold selection methods [Weszka and Rosenfeld, 1978; Sieracki *et al*, 1989], as opposed to the techniques themselves. This is hardly more satisfactory than subjective evaluation. In any case, as is borne out by the work of Weszka and Rosenfeld [Weszka and Rosenfeld, 1978], a method that can be used to evaluate thresholded results can itself also be used as a criterion function for the selection of a supposedly optimum threshold: how do we evaluate the evaluation techniques? Other methods have been proposed to evaluate one aspect or another of the relative performance of different selection techniques [Brink and De Jager, 1987; Albregtsen, 1993]. A possible solution may be to evaluate any given technique on a combination of various criteria since good performance for one application may be poor for another, as was indicated in Section 5.3 of the previous chapter.

6.1 Heuristic methods

Weszka and Rosenfeld [Weszka and Rosenfeld, 1978] addressed the problem of threshold evaluation using two criteria to measure the "cost" of a given threshold applied to an image. These criteria are based on discrepancy (error) and busyness, respectively and were developed from earlier measures for the evaluation of image *smoothing* methods.

Discrepancy measurement

A simple, direct measure of image discrepancy can be obtained if the thresholded image levels are set to the *a priori* below– and above–threshold means of the original grey–scale image,

$\mu_0(T)$ and $\mu_1(T)$, given by (4.27) as

$$\mu_0(T) = \sum_{g=0}^{T} \frac{g\,p_g}{P_0(T)}$$

and

$$\mu_1(T) = \sum_{g=T+1}^{n-1} \frac{g\,p_g}{P_1(T)}$$

The discrepancy $D(T)$ between the images is then simply their absolute difference [Brink, 1987]

$$D(T) = \sum_{g=0}^{T} |g - \mu_0(T)|\,p_g + \sum_{g=T+1}^{n-1} |g - \mu_1(T)|\,p_g \tag{6.1}$$

The levels of the binary image can alternatively be given by $\mu \pm \sigma$ [Weszka and Rosenfeld, 1978] where μ is the original image's mean grey-level

$$\mu = \sum_{g=0}^{n-1} g\,p_g$$

and σ is its grey-level standard deviation

$$\sigma = \sqrt{\sum_{g=0}^{n-1} (g-\mu)^2\,p_g}$$

An alternative approach, adopted by Weszka and Rosenfeld, is to measure the *classification error* incurred by using a given threshold. This is based on the assumption that the image is made up of two distinct components ("background" and "object" again), each having a specified range of grey-levels, and then computing the probability of misclassifying object pixels as background and vice versa for any given threshold. This is then a measure of the discrepancy between the thresholded image and some "ideal" result. Weszka and Rosenfeld do not explicitly state the form of this measure, which requires exact prior knowledge of the background and object distributions.

Sieracki *et al* evaluate various threshold selection schemes, most of them *ad hoc*, for the specific application of segmenting microscopic fluorescent cells from their background [Sieracki *et al*, 1989]. The requirement is that the cell sizes obtained from the segmented image should be accurate. Based on test images where cell sizes were known *a priori*, the errors incurred by each method were evaluated by simply subtracting the known area (in pixels) from the

area obtained by segmentation. This was then expressed as a percentage of the known area, referred to as "normalized error". The error was measured for each cell and the means and standard deviations of these measurements were used to evaluate each technique.

Clearly, discrepancy measures such as those described above can themselves be used as criterion functions for the selection of "optimum" thresholds. This makes their use as evaluation techniques somewhat superfluous, since one could effectively get to the heart of the matter immediately by simply implementing them directly. It also implies that the perfect threshold selection technique is already known. The method described above [Sieracki *et al*, 1989] is clearly limited to a particular type of application and furthermore does not appear to take into account errors incurred by misclassifying background pixels as object. This could conceivably reduce the apparent error in some cases while clearly doing the opposite in terms of actual error.

Measurement of "busyness"

Weszka and Rosenfeld also make use of a measure of busyness (noise or roughness) in the thresholded image as a measure of cost. This is based on the assumption that the ideal objects and background are not strongly textured and have simple, compact shapes: i.e. we expect the thresholded image to look smooth rather than busy. The measure that is used [Weszka and Rosenfeld, 1978] is based on the 4–neighbour co–occurrence matrix. The neighbourhood so defined consists of the pixels $(i, j-1)$, $(i-1, j)$, (i, j), $(i+1, j)$ and $(i, j+1)$, with the neighbourhood origin defined at (i, j), as shown in Figure 6.1.

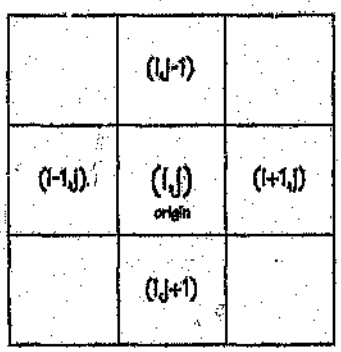


Figure 6.1 *The 4–neighbourhood of a pixel centred at image coordinates* (i, j).

A threshold T partitions the co–occurrence matrix into three non–overlapping regions defined by the quadrants A to D (see Figure 6.2). Assuming a bright object on a dark background, these correspond to (1) co–occurrences of background pixels (quadrant A), (2) co–occurrences

of object pixels (quadrant C) and (3) co-occurrences of object pixels with background pixels (quadrants B and D).

A measure of busyness can then be defined by simply summing those matrix entries representing the proportion of object–background adjacencies in the image. To avoid the trivial result where T is such that all input grey-levels are mapped to a single output level (thus giving zero busyness), the threshold value is constrained to lie between the object and background grey-level means. This again implies that we have *a priori* knowledge about the classes which is not usually available. As before in the case of the discrepancy measures above, such a technique can itself be used to determine an "optimum" threshold by finding T such that busyness is minimized.

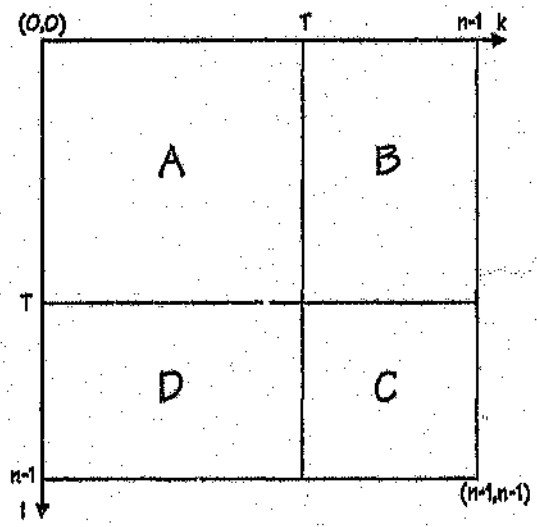


Figure 6.2. *Quadrants of the co-occurrence matrix resulting from thresholding at T. The quadrants A and C on the leading diagonal correspond to "background" and "objects", respectively, while B and D are related to busyness as evidenced by object–background adjacencies.*

In general, any single method that can be used to evaluate the *results* of a general input grey-scale image can be used to select thresholds itself. However, a combination of such methods can lead to a meaningful evaluation. The combined evaluations based on various pre-defined desirable features of the result can give one a semi-quantitative indication of the suitability of a given technique for a particular application: one may require, for example, maximum smoothness coupled with minimum discrepancy, indicating a choice of threshold which will yield the best possible trade-off between these requirements. Of course, such combinations can again themselves be used to determine the best trade-off threshold, but such an operation may well be more time consuming than a method based on a single criterion function which nevertheless comes close in terms of performance.

6.2 Evaluation using synthetic images

Brink and De Jager [Brink and De Jager, 1987] describe an evaluation technique which effectively requires thresholding methods to *restore* a synthetic binary image which has been degraded by contrast reduction, aperture blur and Gaussian noise to produce a "grey-scale" image. Four different images were used with differing levels of blur and noise. The correlation between the binary thresholded result and the original "clean" image was used to evaluate the performance of each threshold selection method for differing signal-to-noise ratios (SNR). The correlation is given by

$$\rho(X,Y) = \frac{E\{X \cdot Y\} - E\{X\}E\{Y\}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \tag{6.2}$$

where X denotes the original (undegraded) image and Y the thresholded (restored) result. $E\{\cdot\}$ denotes expected values and σ_X^2 and σ_Y^2 are the respective image variances.

This approach works relatively well, but poses some problems. The correlation coefficient is not very sensitive to small discrepancies. As a result, threshold selection methods which manage to restore small but important features of the original image at the expense of some excess "noise" on larger features are often rated as being less reliable than methods which correctly restore the large features but fail completely to extract smaller details. Such details should perhaps therefore be suitably weighted. Design of the synthetic image is also problematic, since we ideally need to test threshold selection methods on various types of feature, including

1. small dark features on a light background, and vice versa,

2. vertical, horizontal and diagonal edges, and

3. rounded contours and edges. A test image which should satisfy these requirements is shown in Figure 6.3.

The blurring imposed [Brink and De Jager, 1987] was fairly severe and hence quite unrealistic for this application as features such as lines and edges all but disappeared. One would normally attempt to restore such an image before segmentation. It is proposed therefore to degrade the image of Figure 6.3 by contrast reduction and the addition of varying amounts of Gaussian noise only. For testing dynamic methods a synthetic illumination gradient can be multiplied with the image as well.

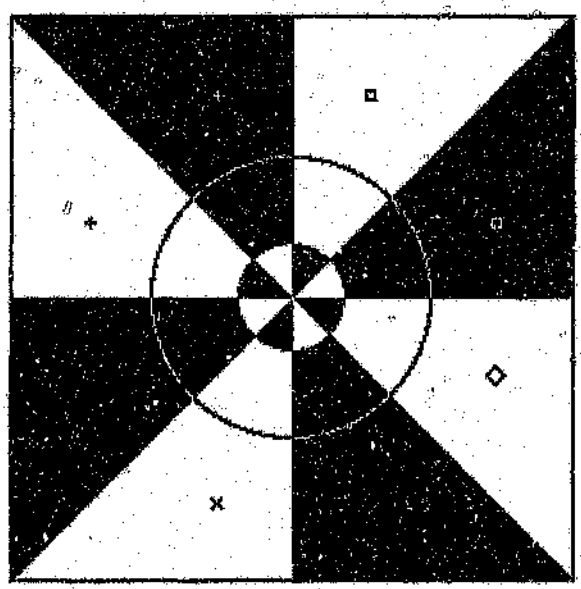


Figure 6.3. *Synthetic binary test image for evaluation of thresholding methods.*

Alternatives to the correlation coefficient should also be considered. Two methods that could be used are the cross-entropy (or relative entropy), as used for threshold selection in Chapter 4, and a simple χ^2 (chi-squared) test. The latter can in fact be derived from the cross-entropy principle [Seth and Kapur, 1990]. The cross-entropy between the original and the thresholded images is given by

$$H_X(X,Y) = \sum_{i=1}^{N} q_i \log \frac{q_i}{p_i} \tag{6.3}$$

where

$$q_i = \frac{g_i}{\sum_{i=1}^{N} g_i}, \quad i = 1, \cdots, N \in Y$$

and

$$p_i = \frac{g_i}{\sum_{i=1}^{N} g_i}, \quad i = 1, \cdots, N \in X$$

The Pearson's χ^2 distance between the two images is given by

$$\chi^2(X,Y) = \sum_{i=1}^{N} \frac{(g_i^{(Y)} - g_i^{(X)})^2}{g_i^{(X)}} \tag{6.4}$$

where the grey-level superscripts refer to the images to which they respectively belong. The grey-levels of the undegraded image in this case are interpreted as the "expected frequencies of occurrence" and those of the thresholded image as the "observed frequencies".

An *ad hoc* scheme for evaluating the weighted classification error associated with methods applied to this image would be to simply count the number of misclassified pixels, each pixel being assigned a weight inversely proportional to the total number of pixels constituting the particular image region to which it belongs. The error associated with the misclassification of a pixel from a small feature is then proportionally greater than that associated with the misclassification of a pixel from a large region.

6.3 Evaluation using synthetic histograms

The method described by Albregtsen [Albregtsen, 1993] involves assessing the accuracy with which techniques using the histogram for threshold selection manage to divide the histogram into background and foreground distributions. To this end, a set of histograms each consisting of the superposition of two Gaussian distributions centred at different means is used. The relative sizes of the distributions and their distance apart are varied while the standard deviations are fixed. These synthetic probability histograms are given by

$$p_g = \frac{P_0}{\sigma_0\sqrt{2\pi}} e^{-\frac{(g-\mu_0)^2}{2\sigma_0^2}} + \frac{P_1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(g-\mu_1)^2}{2\sigma_1^2}}, \quad 0 \le g \le n-1 \tag{6.5}$$

where the "background" distribution is given by

$$p_g^{(0)} = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(g-\mu_0)^2}{2\sigma_0^2}}$$

and the "foreground" is

$$p_g^{(1)} = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(g-\mu_1)^2}{2\sigma_1^2}}$$

P_0 and P_1 are the *a priori* class proportions of the image (the relative sizes of the two Gaussian distributions), not to be confused with the *a posteriori* class probabilities $P_0(T)$ and $P_1(T)$ resulting from a threshold T, given by equation (4.1) in Chapter 4. μ_0 and μ_1 are the respective class means and σ_0 and σ_1 the corresponding standard deviations. For simplicity we set $\sigma_0 = \sigma_1 = \sigma$ [Albregtsen, 1993].

When a histogram-based threshold selection scheme operates on such a synthetic histogram, the selected threshold T attempts to partition it into its constituent background and foreground distributions. Since there is usually a certain amount of overlap, such a partitioning will inevitably result in the misclassification of some background pixels as foreground and vice versa. The total classification error is thus given by

$$E(T) = P_0 \sum_{g=T+1}^{n-1} p_g^{(0)} + P_1 \sum_{g=0}^{T} p_g^{(1)} \tag{6.6}$$

The best possible threshold in terms of this measure (minimizing $E(T)$ with respect to T) is, assuming equal standard deviations,

$$\tau = \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2}{\mu_0 - \mu_1} \log \frac{P_1}{P_0} \tag{6.7}$$

A "foreground error" is also defined,

$$E_1(T) = \sum_{g=0}^{T} p_g^{(1)} \tag{6.8}$$

this being effectively the probability that a foreground pixel is misclassified as background.

Methods using the histogram for threshold selection are tested, calculating the overall error (6.6) and the foreground error (6.8) for background-to-foreground ratios P_0/P_1 ranging from 1 (both distributions the same size) to 100 (foreground small relative to background). For each method four separate graphs are obtained for class means differing by 1, 2, 3 and 4 standard deviations, i.e. $\mu_1 - \mu_0 = d \cdot \sigma$, $d = 1, 2, 3, 4$. These are displayed as error versus $\log_{10}(P_0/P_1)$.

The reason given [Albregtsen, 1993] for including foreground error (6.8) is that in the case of low object-to-background ratios some methods may obtain a low overall error simply by placing the threshold too high, thus losing foreground information. However, only methods based directly on minimization of the total error (6.6) (or a similar criterion) are likely to produce this effect. In practice, with increasing ratio P_0/P_1 the threshold often tends to move towards the mean value of the larger histogram mode, μ_0. This means that the so-called foreground error *decreases* at the cost of an enormous increase in background misclassification. (6.8) is in any case an incomplete measure, as logically the error in the foreground should include background pixels incorrectly identified as also belonging to the foreground. Thus for our purposes only (6.6) is of particular interest. An alternative approach is to define the misclassification error as the larger of (6.8) and its equivalent "background error",

$$E_0(T) = \sum_{g=T+1}^{n-1} p_g^{(0)}$$

i.e.

$$E'(T) = \max_{0 \le T < n} \{E_0(T), E_1(T)\} \tag{6.9}$$

A related measure unbiased by relative mode sizes is given simply by the sum of these measures

$$E''(T) = E_0(T) + E_1(T) \tag{6.10}$$

which unfortunately evaluates thresholds relative to a "best" choice equal to the trivial solution

$$T'' = \frac{\mu_0 + \mu_1}{2}$$

The plots on which Albregtsen's evaluations are based show absolute error versus $\log_{10}(P_0/P_1)$. The evaluation could be made clearer if the error relative to the minimum possible error (resulting from the threshold given by (6.7)) were plotted.

Methods using the (grey-level, local mean grey-level) scatterplot [Abutaleb, 1989; Brink, 1992a] and the grey-level co-occurrence matrix [Pal and Pal, 1989a] could be evaluated along similar lines using a two-dimensional distribution (matrix) consisting of the superposition of two bivariate Gaussian distributions with means centred along the leading diagonal. The error evaluation for the methods of Pal and Pal [Pal and Pal, 1989a] would follow directly from the histogram after selection of the threshold, while the methods due to Abutaleb and Brink [Abutaleb, 1989; Brink, 1992a] should determine the error not only in terms of misclassified pixels, but also those pixels remaining unclassified.

6.4 Evaluation results using synthetic images

The methods described in Section 6.2, using the synthetic binary image of Figure 6.3, were implemented to obtain relative performance figures for the threshold selection methods described in Chapter 4. These techniques were applied to two degraded versions of the test image. In each of these, the original image contrast was substantially reduced from having grey-levels 0 and 255 (contrast range of 256 grey-levels) to having levels 96 and 160 (contrast of 64 grey-levels). Zero-mean Gaussian noise was then added to create the two different images, one with a standard deviation of 8 grey-levels (Figure 6.4) and the other with a 16 grey-level standard deviation (Figure 6.6). The histograms of these images appear in Figures 6.5 and 6.7, respectively.

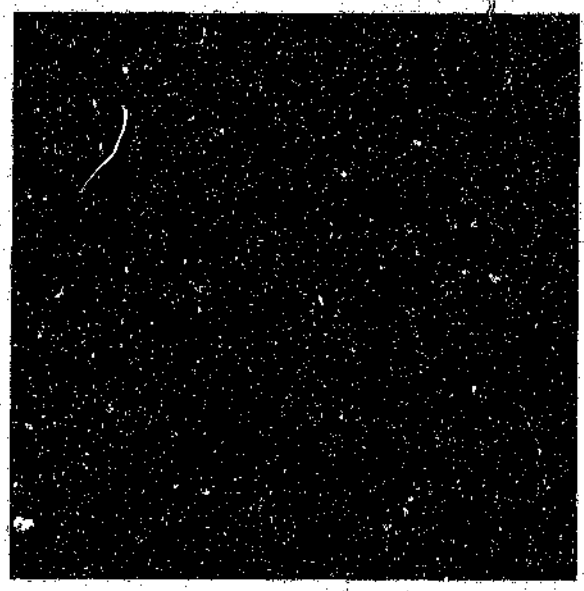


Figure 6.4. *The binary test image of Figure 6.3 with reduced contrast and added Gaussian noise with a standard deviation of 8 grey-levels.*

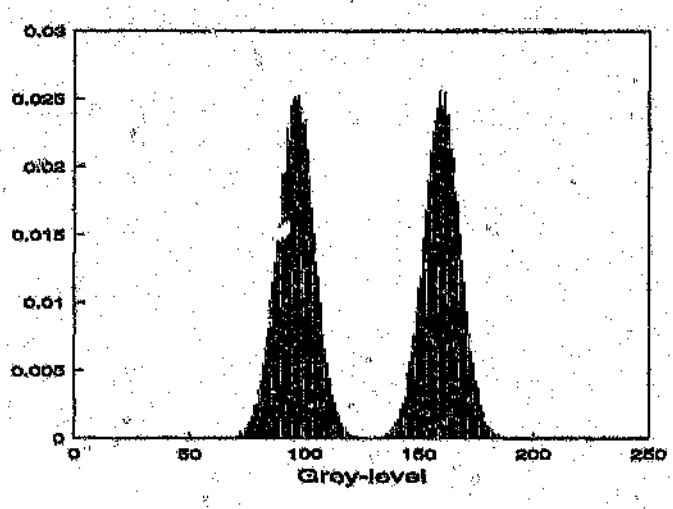


Figure 6.5. *Grey-level histogram of the image of Figure 6.4.*

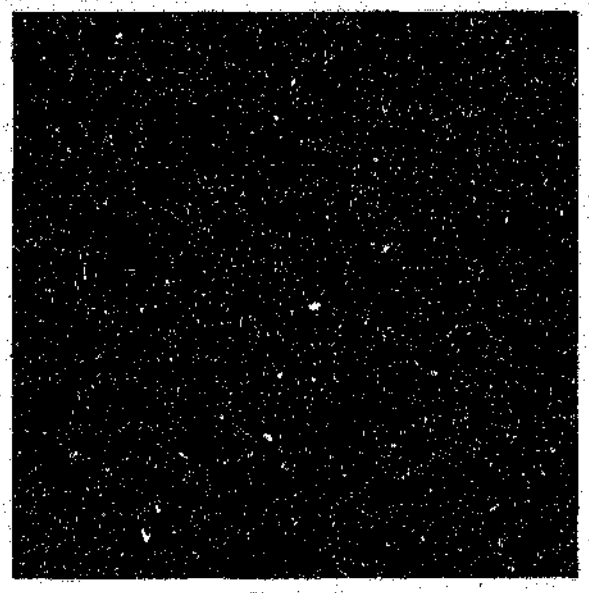


Figure 6.6. *The binary test image of Figure 6.3 with reduced contrast and added Gaussian noise with a standard deviation of 16 grey-levels.*

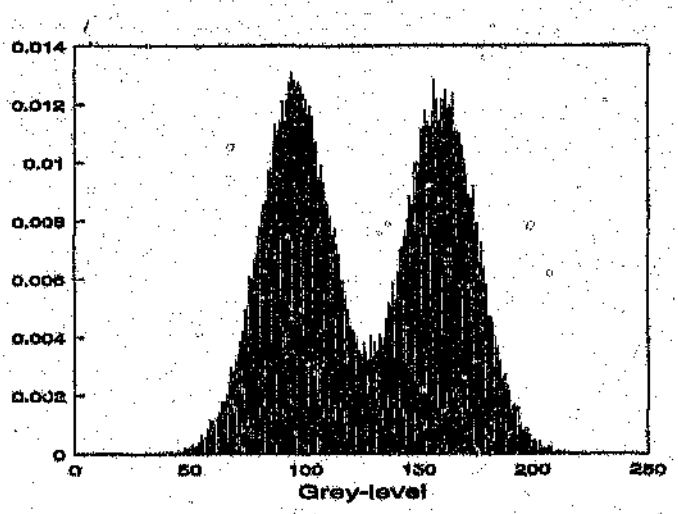


Figure 6.7. *Grey-level histogram of the image of Figure 6.6.*

The discrepancy measures used, described in Section 6.2, are the cross–entropy (6.3), Pearson's chi–squared test (6.4) and a weighted classification error as briefly described at the end of Section 6.2. The results are tabulated separately for each degraded image in Tables 6.1 and 6.2. The threshold selection methods have been referred to by letters, as follows:

A. The histogram–based maximum entropy measure of Kapur *et al* [Kapur *et al*, 1985], using equations (4.7), (4.8) and (4.9).

B. The maximin class entropy of Brink [Brink, 1991b] using (4.8) and (4.10).

C. Abutaleb's [Abutaleb, 1989] method using the (grey–level, local mean grey–level) scatterplot, (4.11) and (4.15).

D. Brink's [Brink, 1992a] modification of the above method using a maximin approach (4.16) and the corrected "foreground" class entropy (4.13), (4.14).

E. Algorithm 1 proposed by Pal and Pal [Pal and Pal, 1989a] based on the 2–neighbour grey–level co–occurrence matrix, using (4.18), (4.19) and (4.20).

F. The second method due to Pal and Pal [Pal and Pal, 1989a], using conditional entropies (4.18), (4.23).

G. The method using the relative entropy information (4.26) or (4.28) [Pan and Harris (1990)].

H. Threshold selection based on the monkey model, as described in Section 4.2, using (4.25) and the maximization of the sum of the class entropies (4.9).

I. Using the maximin of the class entropies resulting from the monkey model.

J. The modified version of **H** above incorporating contextual information in the form of local variances as a measure on the image (4.32), based on maximization of the sum of the class entropies.

K. As above, but using the maximin of the class entropies to select the threshold.

L. Result of implementing the method based on Frieden's [Frieden, 1973, 1980] model of an image bounded above and below, maximizing the sum of the class entropies (4.26).

M. As above, except that the maximin of the class entropies is used [Brink, 1992b].

N. Minimum cross–entropy (or maximum relative entropy) threshold selection using (4.29) [Brink and Pendock, 1993].

O. Implementation of the symmetricised cross–entropy method (4.32) [Brink and Pendock, 1993].

P. Threshold selection based on minimization of the relative spatial entropy (4.42).

Q. Thresholding using the autocorrelation function of the histogram [Brink, 1993]. The sum of the class entropies (4.36) is used here.

R. As above, but using the maximin of the class entropies.

S. The "labelled photon" approach described in Section 4.2. The sum of the class entropies (4.38) is used here.

T. As above using the maximin of the class entropies.

Method	$H_{\times}(X,Y)$	$\chi^2(X,Y)$	$E_w(X,Y)$
A	0.00032	82.0	0.02476
B	0.00015	37.0	0.00910
C	0.00107	171.5	14.71412
D	0.00081	130.5	15.11958
E	0.00000	0.0	0.00000
F	0.00099	158.0	0.25112
G	0.00000	0.0	0.00000
H	0.00015	37.0	0.00910
I	0.00052	133.0	0.10814
J	.00143	228.5	0.32882
K	0.00103	264.0	0.16164
L	0.00000	0.0	0.00000
M	0.00001	2.0	0.00145
N	0.00001	2.0	0.00145
O	0.00001	2.0	0.00145
P	0.00000	0.0	0.00000
Q	0.02993	9488.0	5.56093
R	0.00010	26.0	0.00743
S	0.01968	3305.5	4.49133
T	0.01968	3305.5	4.49133

Table 6.1. *Evaluation results of threshold selection methods applied to the degraded image of Figure 6.4.*

Method	$H_\times(X,Y)$	$\chi^2(X,Y)$	$E_w(X,Y)$
A	0.00537	1153.0	0.76796
B	0.00536	1115.0	0.80433
C	0.00309	501.0	15.67320
D	0.00393	636.0	15.74693
E	0.00545	1213.0	0.78897
F	0.00537	1153.0	0.76796
G	0.00545	1213.0	0.78897
H	0.00545	1213.0	0.78897
I	0.00537	1153.0	0.76796
J	0.00596	1083.0	0.90443
K	0.00536	1115.0	0.80433
L	0.00563	1296.5	0.79801
M	0.00594	1405.0	0.95435
N	0.00594	1405.0	0.95435
O	0.00594	1405.0	0.95435
P	0.00545	1213.0	0.78897
Q	0.03021	5267.0	6.75700
R	0.00536	1115.0	0.80433
S	0.02293	3894.0	4.73658
T	0.02293	3894.0	4.73658

Table 6.2. *Evaluation results of threshold selection methods applied to the degraded image of Figure 6.6.*

It is interesting to compare the evaluations in Tables 6.1 and 6.2, particularly where some disparity exists between evaluation methods. For example, while all the methods indicate that the results obtained using the "labelled photon" model are not particularly good (rows S and T of the tables), the weighted error method rates it as much better than any methods based on the (grey-level, local mean grey-level) scatterplot (C and D). This contradicts the relative evaluations of both the χ^2 and $H_\times$ methods. The reason for this lies in the fact that the latter two methods evaluate overall error while the weighted error measure attaches greater weight to small "objects" within the image. It should also be noted in the case of the scatterplot-based methods that pixels that remain unclassified by virtue of being in the wrong quadrants are regarded as "errors" by the weighted error method. There appears to be a consensus that the most generally useful techniques, certainly as far as thresholding of images with relatively low noise levels is concerned, are those based on Frieden's [Frieden, 1973, 1980] bounded image model (rows L and M) and the methods based on cross-entropy minimization [Brink and Pendock, 1993] in rows N and O. Methods that work reasonably well at higher noise levels are those based on both forms of the monkey model (rows H, I, J and K) and the co-occurrence methods (rows E and F) of Pal and Pal [Pal and Pal, 1989a]. Interestingly, the context-dependent method based on the monkey model (rows J and K)

is rated very similarly to the standard form (rows H and I), even displaying an apparently worse performance at times. The relative entropy information [Pan and Harris, 1990] (row G) and the minimum relative spatial entropy [Journel and Deutsch, 1993] (row P) methods both perform consistently well, regardless of noise.

6.5 Evaluation results using synthetic histograms

Albregtsen's [Albregtsen, 1993] technique for evaluating threshold selection methods on the basis of how well they partition a synthetic histogram has been implemented here. Only those methods whose decision process could be expressed directly in terms of the histogram have been tested. As indicated in Section 6.3, a two-dimensional version could be developed as a simple extension for the evaluation of methods using scatterplots and co-occurrence matrices. However, such a process would be extremely slow as a vast number of slightly different 2-D distributions would need to be generated. In the graphs that follow (Figures 6.8 to 6.12), only total classification error (6.6) is shown. The "best" threshold error (based directly on minimization of the classification error) is depicted in Figure 6.8. For brevity, the techniques are again referred to by the letters used in Section 6.4. The graphs depict error versus $\log(P_0/P_1)$, where the mode size ratio (P_0/P_1) is varied over the range 1 to 100. Four graphs appear on each plot, corresponding to $(\mu_1 - \mu_0) = D \cdot \sigma$, $D = 1, 2, 3, 4$. The standard deviations of both modes are fixed at σ =32 grey-levels.

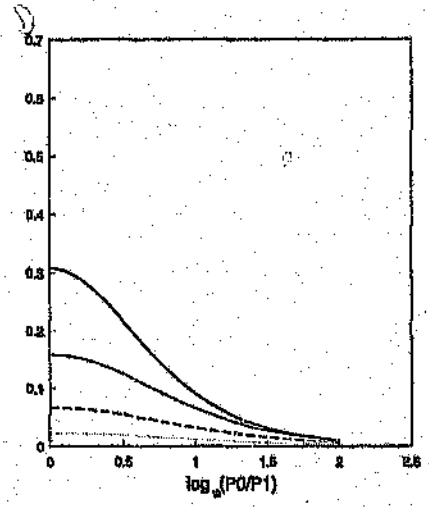
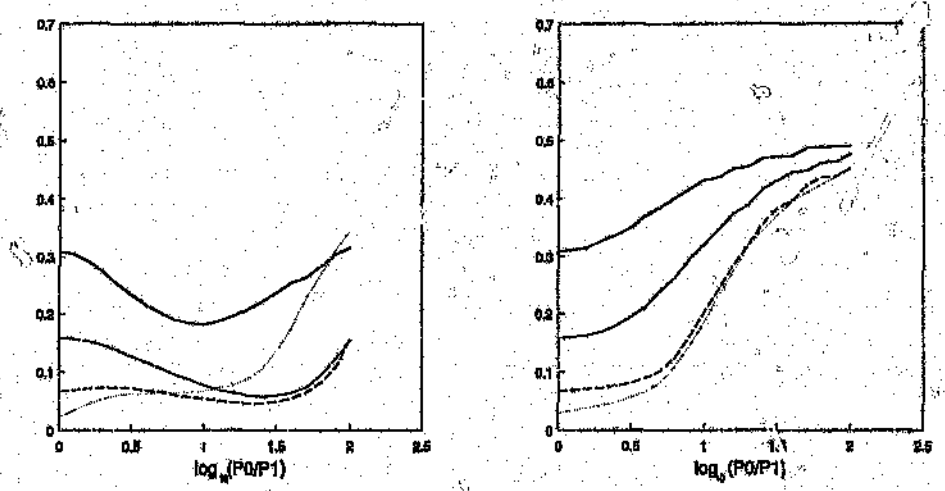
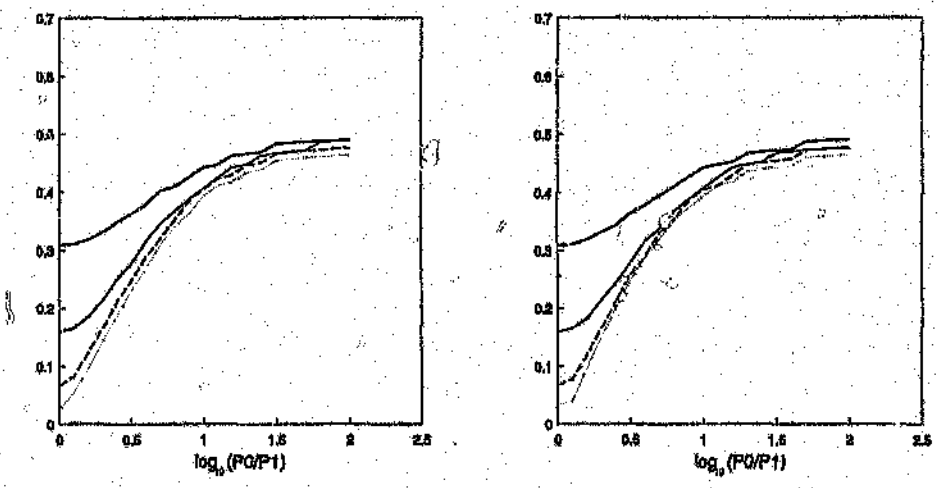


Figure 6.8. *Minimum possible classification error. Histogram mode size ratio* (P_0/P_1) *is varied from* 1 *to* 100. *The four curves shown are for the 4 different values of* $\mu_1 - \mu_0 = D \cdot \sigma$, $D = 1$ *(thick solid line),* $D = 2$ *(thin solid line),* $D = 3$ *(dashed line) and* $D = 4$ *(dotted line).*



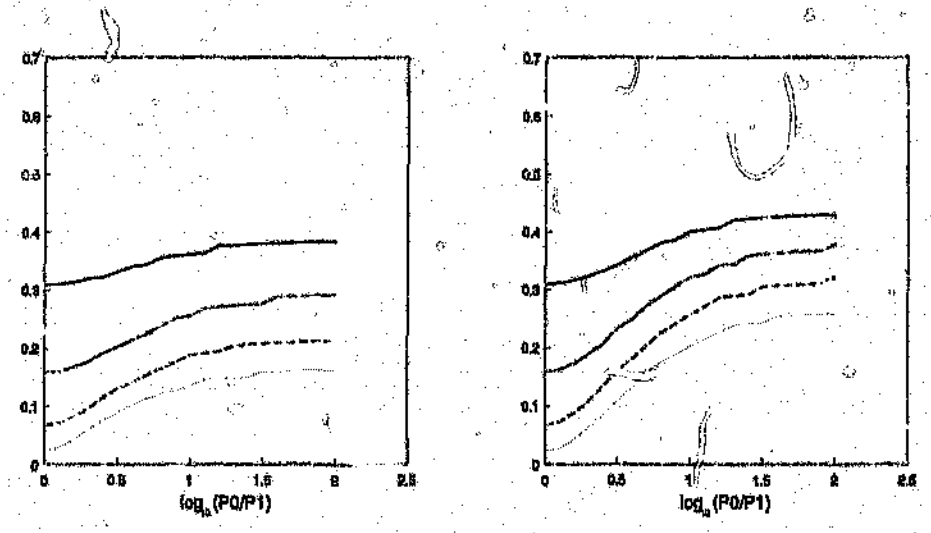
(a) (b)

Figure 6.9. *Classification errors resulting from threshold selection using* (a) *method A and* (b) *method B. Histogram mode size ratio* (P_0/P_1) *is varied from* 1 *to* 100. *The four curves shown are for the 4 different values of* $\mu_1 - \mu_0 = D \cdot \sigma$, $D = 1$ *(thick solid line),* $D = 2$ *(thin solid line),* $D = 3$ *(dashed line) and* $D = 4$ *(dotted line).*



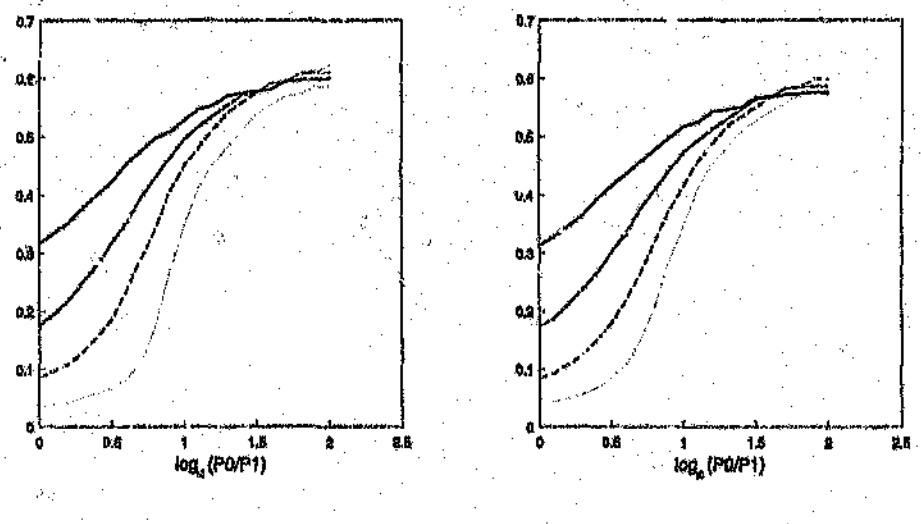
(a) (b)

Figure 6.10. *Classification errors resulting from threshold selection using* (a) *method G and* (b) *method H. Histogram mode size ratio* (P_0/P_1) *is varied from* 1 *to* 100. *The four curves shown are for the 4 different values of* $\mu_1 - \mu_0 = D \cdot \sigma$, $D = 1$ *(thick solid line),* $D = 2$ *(thin solid line),* $D = 3$ *(dashed line) and* $D = 4$ *(dotted line).*



(a) (b)

Figure 6.11. *Classification errors resulting from threshold selection using* (a) *method I and* (b) *method J. Histogram mode size ratio* (P_0/P_1) *is varied from* 1 *to* 100. *The four curves shown are for the 4 different values of* $\mu_1 - \mu_0 = D \cdot \sigma$, $D = 1$ *(thick solid line),* $D = 2$ *(thin solid line),* $D = 3$ *(dashed line) and* $D = 4$ *(dotted line).*



(a) (b)

Figure 6.12. *Classification errors resulting from threshold selection using* (a) *method K and* (b) *method L. Histogram mode size ratio* (P_0/P_1) *is varied from* 1 *to* 100. *The four curves shown are for the 4 different values of* $\mu_1 - \mu_0 = D \cdot \sigma$, $D = 1$ *(thick solid line),* $D = 2$ *(thin solid line),* $D = 3$ *(dashed line) and* $D = 4$ *(dotted line).*

Most methods appear to perform progressively worse as the "foreground" mode shrinks in size, as would be expected. The method based on Frieden's bounded image model [Brink, 1992b] is a notable exception (Figure 6.11 (a) and (b)). The cross-entropy method, which was evaluated as quite reliable using synthetic images (Section 6.4), performs extremely badly here (Figures 6.12 (a) and (b)).

It should be noted that the result obtained using the histogram entropy method of Kapur *et al* (Figure 6.9(a)) does not match that obtained by Albregtsen for this same method. The other methods evaluated by Albregtsen [Albregtsen, 1993] were therefore also implemented and the results obtained appear to match. My own implementation of this method has been thoroughly checked and rechecked, and it must therefore be assumed that Albregtsen's implementation is at fault.

In the next chapter the results of Chapter 5 and the quantitative evaluation methods introduced in this chapter are discussed.

Chapter 7

Discussion and conclusions

The aim of this work has been to develop an effective model for the segmentation of images using threshold selection techniques based on image entropy. Some of the motivations for using the principle of maximum entropy and the arguments surrounding it have been presented and ways of addressing its apparent shortcomings have been investigated. As Skilling [Titterington, 1984] puts it, "I hold that these more complicated situations should ideally be handled by extending a technique which is basically correct rather than by invoking *ad hoc* alternatives.".

The models serving as a basis for the definition of what is meant by "image entropy" are widely different. Basing entropy directly on the data (image) alone leads us to the problem that inter-relationships are ignored (e.g. the monkey model), often to the detriment of both our results and our credibility in lending this method what sometimes appears to be dogmatic support. The use of co-occurrence and autocorrelation [Journel and Deutsch, 1993; Pal and Pal, 1989a; Brink, 1994] and the simpler expedients of using cross-entropy [Brink and Pendock, 1993] or building variances into the monkey model (Chapter 4) are heading in the right direction, as the results clearly show.

Some limited discussion of both the results (Chapter 5) and the quantitative evaluations of thresholding methods (Chapter 6) has already taken place in the relevant chapters. However, it will be informative to balance the subjective evaluations of Chapter 5 with the quantitative evaluations of Chapter 6.

In the evaluation of the various methods, one should also take into account computational complexity and time. Obviously methods requiring the use of scatterplots [Abutaleb, 1989; Brink, 1992a] and co-occurrence matrices [Pal and Pal, 1989a] require a lot of extra computation simply to generate the distributions in question. Hence, while the methods of

Pal and Pal [Pal and Pal, 1989a] yield excellent results (Figures 5.4-6 (e) and (f)) both aesthetically and quantitatively (Tables 6.1 and 6.2, rows E and F), their computational cost is high. This would have to be considered if real-time applications were required. In terms of sheer quality of results, however, measures of image entropy based on co-occurrence matrices, autocorrelation and local variance information must obviously be taken seriously.

The results arising from implementation of the image model bounded above and below, due to Frieden [Frieden, 1973, 1980; Brink, 1992b] (Figures 5.7-9 (e) and (f)) also give both subjectively and quantitatively good results (Tables 6.1 and 6.2, rows L and M). However, while this model treats the image itself as a probability distribution, the pixels' relative position information is lost due to the assumption of pixel independence inherent in the entropy measure. For this reason the formulae can be expressed purely in terms of the image histogram and evaluated as histogram partitioning methods, using Albregtsen's [Albregtsen, 1993] evaluation technique. Interestingly, based on this evaluation, these methods appear to be the most stable with respect to relative class mode sizes of the histogram (Figures 6.11 (a) and (b)), with performance being very little affected by an enormous reduction of relative object size!

Visually the results of applying the principle of minimum cross-entropy are the most pleasing (Figures 5.7-9 (g) and (h)). However, the quantitative evaluations using synthetic images indicate only average performance when noise levels are high (Table 6.2, rows N and O) and exceptionally poor performance in terms of overall classification error with increasing mode size ratio (Figures 6.12 (a) and (b)).

The methods based directly on the histogram [Kapur *et al*, 1985; Brink, 1991b] appear to achieve average to good performance based on subjective and quantitative evaluations (Figures 5.4-6 (a) and (b), Tables 6.1 and 6.2: rows A and B, Figures 6.9 (a) and (b)). The maximin method of threshold selection [Brink, 1991b] (Figures 5.4-6 (b)) yields better results on subjective and synthetic image evaluations, while the sum of the class entropies [Kapur it et al, 1985] gives better histogram partitioning performance (Figure 6.9 (a)). This latter result is to some extent expected: if the results of Figures 5.4-6 (a) are compared with those of Figure 5.4-6 (b), it appears that the extraction of actual image classes is more "correct" in the former, while the results of Figure 5.4-6 (b) *look* better in terms of their resemblance to the original images of Figures 5.1 to 5.3.

The worst results throughout are those of the application of the "labelled photon" model. This model poses a problem in terms of its interpretation: viewing a coordinate or position in space as a probability is difficult. The model is best understood from its formulation outlined in Chapter 3, as viewing the image as a "histogram" arising from some kind of "meta-image".

However, its validity is questionable since neither its performance nor its philosophical and theoretical bases are particularly sound. It has therefore been included here mainly as an example of why care needs to be taken to avoid *ad hoc* solutions.

The quantitative evaluation techniques of the previous chapter show some interesting biases, which lead to their relatively different evaluations of the threshold selection methods. The method based on synthetic images is to some extent the more pleasing, in the sense that we are after all attempting to binarize an image. Albregtsen's [Albregtsen, 1993] technique is strictly a measure of how well a histogram consisting of two superimposed Gaussian distributions can be split into its constituent distributions by a single threshold. This is a somewhat limited evaluation and should ideally be used in conjunction with other methods. Overall, for any particular application, a subjective evaluation still appears to be the most consistently reliable.

It is proposed that future research should extend the work covered in this thesis to more general segmentation methods and development of a quantitative process of evaluation which agrees more closely with our own subjective evaluations of results and methods.

In conclusion, it appears that our problem does not lie with entropy: the reasons for its use, its strength as a prior distribution selector for Bayesian methods of inference and the quality of the results obtainable when it is correctly implemented can be and have been proved, although the arguments will doubtless rage for decades to come. The real problem is this question of "correct" implementation. Entropy, like other methodologies such as neural networks, is **not** a cure-all. It is not magic (to answer the question posed by Press *et al* [Press *et al*, 1992]), it is simply a sensible method for solving problems of inference. It is up to us to pose the problem correctly and to supply the data and information required in a useful form. We should also be certain about the distinction between information and belief.

This application has shown that we can obtain excellent thresholding results for very different images if we supply enough information of the right kind (e.g. context-related information about local uncertainty) in the right way and in the right place, e.g. putting variances in as the prior measure in the monkey model as opposed to using contextual information of the wrong kind (pixel coordinates) in the wrong place (the actual probabilities in the entropy expression). If we do it wrong, we get it wrong.

References

Abutaleb, A S (1989). "Automatic thresholding of gray-level pictures using two-dimensional entropy", *Computer Vision, Graphics, and Image Processing* 47, 22–32.

Albregtsen, F (1993). "Non-parametric histogram thresholding methods - error versus relative object area", *Proc. 8th Scandinavian Conference on Image Analysis*, Tromsø, Norway, 273–280.

Bretthorst, G L (1990). "An introduction to parameter estimation using Bayesian probability theory", in: *Maxim Entropy and Bayesian Methods*, P F Fougère (ed.), Kluwer Academic Publishers, Netherlands, 53–79.

Brink, A D (1987). *The Selection and Evaluation of Grey-level Thresholds Applied to Digital Images*, M Sc dissertation, Rhodes University, Grahamstown, South Africa.

Brink, A D (1989). "Grey-level thresholding of images using a correlation criterion", *Pattern Recognition Letters* 9, 335–341.

Brink, A D (1990). "Comments on grey-level thresholding of images using a correlation criterion", *Pattern Recognition Letters* 12, 91–92.

Brink, A D (1991a). "Edge-based dynamic thresholding of digital images", *Proc. IEEE COMSIG 91 S A Symposium on Communications and Signal Processing*, Pretoria, South Africa, 63.

Brink, A D (1991b). "Thresholding of digital images using the entropy of the histogram", *Proc. Second South African Workshop on Pattern Recognition*, Stellenbosch, South Africa, 48–53.

Brink, A D (1992a). "Thresholding of digital images using two-dimensional entropies", *Pattern Recognition* **25** (8), 803–808.

Brink, A D (1992b). "Maximum entropy threshold selection for image segmentation", *Proc. Third South African Workshop on Pattern Recognition*, Pretoria, South Africa, 97–103.

Brink, A D (1994). "Maximum entropy threshold selection based on the autocorrelation function of the image histogram", accepted for publication in *J. Computing and Information Technology*.

Brink, A D and G J De Jager (1987). "Evaluation of image grey-level thresholding methods", *Proc. Fifth South African Symposium on Digital Signal Processing*, Grahamstown, South Africa, 3.3.1–3.3.7.

Brink, A D and N E Pendock (1993). "Minimum cross-entropy threshold selection", submitted to *Pattern Recognition*.

Burg, J P (1967). "Maximum entropy spectral analysis", *Proc. 37th Meeting of the Society of Exploration Geophysicists*, Oklahoma City, USA. Reprinted (1978) in: *Modern Spectrum Analysis*, D G Childers (ed.), John Wiley and Sons, New York, USA, 34–41.

Cseke, I and Z Fazekas (1990). "Comments on grey level thresholding of images using a correlation criterion", *Pattern Recognition Letters* **11**, 709–710.

Dainty, J C and R Shaw (1974). *Image Science*, Academic Press, London, UK.

Donoho, D L, I M Johnstone, J C Hoch and A S Stern (1992). "Maximum entropy and the nearly black object", *J. Royal Statistical Society B* **54** (1), 41–81.

Edwards, B (1979). *Drawing on the Right Side of the Brain*, Fontana Books, UK.

Frieden, B R (1972). "Restoring with maximum likelihood and maximum entropy", *J. Optical Society of America* **62** (4), 511–518.

Frieden, B R (1973). "Statistical estimates of bounded optical scenes by the method of 'prior probabilities' ", *IEEE Trans. Information Theory* **IT–19** (1), 118–119.

Frieden, B R (1980). "Statistical models for the image restoration problem", *Computer Graphics and Image Processing* **12**, 40–59.

Gablik, S (1970). *Magritte*, Thames and Hudson, UK.

Geman, S and D Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Analysis and Machine Intelligence* **6 (6)**, 721–741.

Gonzalez, R C and R E Woods (1992). *Digital Image Processing*, Addison–Wesley Publishing Co., Reading, Mass., USA.

Grimmett, G R and D R Stirzaker (1982). *Probability and Random Processes*, Oxford Science Publications, Oxford, UK.

Gull, S F (1991). "Some misconceptions about entropy", in: *Maximum Entropy in Action*, B Buck and V A Macaulay (eds.), Oxford University Press, Oxford, UK, 171–186.

Gull, S F and G J Daniell (1978). "Image reconstruction from incomplete and noisy data", *Nature* **272**, 686–690.

Gull, S F and J Skilling (1984). "Maximum entropy method in image processing", *IEE Proceedings* **131:F (6)**, 646–659.

Gull, S F and J Skilling (1985). "The entropy of an image", in: *Maximum Entropy and Bayesian Methods in Inverse Problems*, C R Smith and W T Grandy, Jr. (eds.), D Reidel Publishing Company, Netherlands, 287–301.

Haralick, R M and L G Shapiro (1985). "Survey: Image segmentation techniques", *Computer Vision, Graphics, and Image Processing* **29**, 100–132.

Hofstadter, D R (1979). *Gödel, Escher, Bach: an Eternal Golden Braid*, Penguin Books, UK.

Jain, A K and E Angel (1974). "Image restoration, modelling, and reduction of dimensionality", *IEEE Trans. Computers* **23 (5)**, 470–476.

Jaynes, E T (1965). "Gibbs vs Boltzmann entropies", *American J. Physics* **33**, 391–398.

Jaynes, E T (1968). "Prior probabilities", *IEEE Trans. Systems Science and Cybernetics* **SSC–4 (3)**, 227–241.

Jaynes, E T (1986). "Monkeys, kangaroos and N", in: *Maximum Entropy and Bayesian Methods in Applied Statistics*, J H Justice (ed.), Cambridge University Press, UK, 26–58.

Jaynes, E T (1988). "The relation of Bayesian and maximum entropy methods", in: *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, G J Erickson and C R Smith (eds.), Kluwer Academic Publishers, Netherlands, 25–29.

Jones, L K (1989). "Approximation–theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge", *SIAM J. Applied Mathematics* **49 (2)**, 650–661.

Journel, A G and C V Deutsch (1993). "Entropy and spatial disorder", *Mathematical Geology* **25 (3)**, 329–355.

Kapur, J N and H K Kesavan (1990). "The inverse MaxEnt and MinxEnt principles and their applications", in: *Maximum Entropy and Bayesian Methods*, P F Fougère (ed.), Kluwer Academic Publishers, Netherlands, 433–450.

Kapur, J N, P K Sahoo and A K C Wong (1985). "A new method for gray–level picture thresholding using the entropy of the histogram", *Computer Vision, Graphics, and Image Processing* **29**, 273–285.

Kesavan, H K and J N Kapur (1989). "The generalized maximum entropy principle", *IEEE Trans. on Systems, Man, and Cybernetics* **19 (5)**, 1042–1052.

Kikuchi, R and B H Soffer (1977). "Maximum entropy image restoration. I. The entropy expression", *J. Optical Society of America* **67 (12)**, 1656–1665.

Kirby, R L and A Rosenfeld (1979). "A note on the use of (gray level, local average gray level) space as an aid in threshold selection", *IEEE Trans. Systems, Man, and Cybernetics* **SMC-9 (12)**, 860–864.

Kittler, J and J Illingworth (1986). "Minimum error thresholding", *Pattern Recognition* **19 (1)**, 41–47.

Kullback, S (1959). *Information Theory and Statistics*, John Wiley, New York.

Liang, Z (1988). "Statistical models of *a priori* information for image processing: neighbouring correlation constraints", *J. Optical Society of America A* **5 (12)**, 2026–2031.

Loredo, T J (1989). "From Laplace to Supernova SN 1987A: Bayesian inference in astrophysics", in: *Maximum Entropy and Bayesian Methods*, P F Fougère (ed.), Kluwer Academic Publishers, Netherlands, 81–142.

Mardia, K V (1989). "Markov models and Bayesian methods in image analysis", *J. Applied Statistics* **16 (2)**, 125–130.

Marqués, F and A Gasull (1993). "Stochastic image model for segmentation. Application to image coding", *Proc. 8th Scandinavian Conference on Image Analysis*, Tromsø, Norway, 265–272.

Mood, A M, F A Graybill and D C Boes (1974). *Introduction to the Theory of Statistics, 3rd edition*, McGraw-Hill, Singapore.

Narayan, R and R Nityananda (1986). "Maximum entropy image restoration in astronomy", *Annual Review of Astronomy and Astrophysics* **24**, 127–170.

Otsu, N (1979). "A threshold selection method from grey-level histograms", *IEEE Trans. on Systems, Man and Cybernetics* **9 (1)**, 62–66.

Pal, N R and S K Pal (1989a). "Entropic thresholding", *Signal Processing* **16**, 97–108.

Pal, N R and S K Pal (1989b). "Object-background segmentation using new definitions of entropy", *IEE Proceedings* **136:E (4)**, 284–295.

Pal, N R and S K Pal (1991a). "Image model, Poisson distribution and object extraction", *Int. J. of Pattern Recognition and Artificial Intelligence* **5 (3)**, 459–483.

Pal, N R and S K Pal (1991b). "Entropy: a new definition and its applications", *IEEE Trans. Systems, Man, and Cybernetics* **21 (5)**, 1260–1270.

Pan, G and D P Harris (1990). "Three nonparametric techniques for the optimum discretization of quantitative geological variables", *Mathematical Geology* **22 (6)**, 699–722.

Pendock, N E (1993a). "Bayesian source estimation from potential field data", in: *Maximum Entropy and Bayesian Methods*, A Mohammad-Djafari and G Demoment (eds.), Kluwer Academic Publishers, Netherlands, 287–293.

Pendock, N E (1993b). "Likelihood distributions for radio wave tomography", *Proc. IEEE COMSIG 93 S A Symposium on Communications and Signal Processing*, Johannesburg, South Africa, 35–39.

Pratt, W K (1991). *Digital Image Processing, 2nd edition*, John Wiley & Sons, Inc., USA.

Press, W H, S A Teukolsky, W T Vetterling and B P Flannery (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd edition*, Cambridge University Press, Cambridge, UK.

Pun, T (1980). "A new method for grey-level picture thresholding using the entropy of the histogram", *Signal Processing* **2**, 223–237.

Sahoo, P K, S Soltani, A K C Wong and Y C Chen (1988). "A survey of thresholding techniques", *Computer Vision, Graphics, and Image Processing* **41**, 233–260.

Sears, F W and G L Salinger (1975). *Thermodynamics, Kinetic Theory and Statistical Thermodynamics*, 3rd edn., Addison–Wesley Publishing Co., Reading, Mass., USA.

Seth, A K and J N Kapur (1990). "A comparative assessment of entropic and non–entropic methods of estimation", in: *Maximum Entropy and Bayesian Methods*, P F Fougère (ed.), Kluwer Academic Publishers, Netherlands, 451–462.

Shannon, C E and W Weaver (1949). *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, USA.

Shore, J E and R W Johnson (1980). "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross–entropy", *IEEE Trans. Information Theory* **IT–26** (1), 26–37.

Sieracki, M E, S E Reichenbach and K L Webb (1989). "Evaluation of automated threshold selection methods for accurately sizing microscopic fluorescent cells by image analysis", *Applied and Environmental Biology* **55** (11), 2762–2772.

Skilling, J (1986). "Theory of maximum entropy image reconstruction", in: *Maximum Entropy and Bayesian Methods in Applied Statistics*, J H Justice (ed.), Cambridge University Press, Cambridge, UK, 156–178.

Skilling, J (1988). "The axioms of maximum entropy", in: *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, G J Erickson and C R Smith (eds.), Kluwer Academic Publishers, Netherlands, 173–187.

Skilling, J (1989). "Classic maximum entropy", in: *Maximum Entropy and Bayesian Methods*, J Skilling (ed.), Kluwer Academic Publishers, Netherlands, 45–52.

Skilling, J (1990). "Quantified maximum entropy", in: *Maximum Entropy and Bayesian Methods*, P F Fougère (ed.), Kluwer Academic Publishers, Netherlands, 341–350.

Skilling, J (1991). "Fundamentals of MaxEnt in data analysis", in: *Maximum Entropy in Action*, B Buck and V A Macaulay (eds.), Oxford University Press, Oxford, UK, 19–40.

Titterington, D M (1984). "The maximum entropy method for data analysis", *Nature* **312**, 381–382.

Toda, M, R Kubo and N Saitō (1983). *Statistical Physics I: Equilibrium Statistical Mechanics, 2nd edition*, Springer–Verlag, Heidelberg, Germany.

Tolman, R C (1979). *The Principles of Statistical Mechanics*, Dover Publishers, New York, USA.

Tsai, W–H (1985). "Moment–preserving thresholding: a new approach", *Computer Vision, Graphics, and Image Processing* **29**, 377–393.

Wernecke, S J and L R D'Addario (1977). "Maximum entropy image reconstruction", *IEEE Trans. Computers* **26** (4), 351–364.

Weszka, J S (1978). "A survey of threshold selection techniques", *Computer Graphics and Image Processing* **7**, 259–265.

Weszka, J S and A Rosenfeld (1978). "Threshold evaluation techniques", *IEEE Trans. Systems, Man, and Cybernetics* **8** (8), 622–629.

Zellner, A (1988). "Optimal information processing and Bayes's Theorem", *The American Statistician* **42** (4), 278–284.