Adaptive Multi-Scale Texture Analysis

With Application to

Automated Cytology

Ross Francis Walker

B.E. (Hons I, Electrical and Computer Engineering), Queensland University of Technology

Cytometrics Group, CSSIP,
Department of Electrical & Computer Engineering,
The University of Queensland,
St. Lucia, Brisbane, Queensland 4072, Australia.
Email: walker@elec.uq.edu.au
walker_ross@hotmail.com

THE UNIVERSITY
OF QUEENSLAND

Submitted as a requirement for the degree of
Doctor of Philosophy (Ph.D.),
The University of Queensland,
1997.

Statement of Originality

The work contained in this PhD thesis is, to the best of my knowledge and belief, original and my own work, except as acknowledged in the text. This material has not been submitted, either in whole or in part, for a degree at this or any other university.

Ross F. Walker

Key words

- Artificial Neural Networks
- Cervical Cancer
- Classification
- Co-occurrence
- Computer Vision
- Cross-Validation
- Image Analysis
- Image Processing
- Medical Diagnostics
- Multi-Scale
- Multi-Resolution
- Object Recognition
- Pattern Recognition
- Texture Analysis

Abstract

In this thesis we investigate the application of image analysis and self-adaptive algorithms to the detection of cell abnormalities in cervical smears. Cervical cancer is a preventable disease and, unlike most cancers, can be easily detected by a routine screening test. Current manual screening methods are costly and sometimes result in inaccurate diagnosis due to human error. The introduction of machine-assisted screening will therefore bring significant benefits to the community, by reducing financial costs and increasing screening accuracy.

One of the fundamental weaknesses of research efforts over the last 30 years has been in identifying a robust set of cell descriptors to allow accurate classification of cytological samples. Continuing advances in imaging technology and computing power have provided incremental gains in the diagnostic accuracy of automated cytology systems, however, there is a need for further improvement. The quantitative analysis of cell nuclear texture has shown the most promise in the past, and continues to be a major focus of research efforts around the world. Our motivation for the work in this thesis is a belief that greater effort in texture analysis research will yield further advances of significant benefit.

We investigate the history of texture analysis as applied to automated cytology, and identify Markovian techniques as being powerful methods of analysis. By Markovian methods, we mean those techniques which model the joint or conditional statistical dependence between neighbouring image pixels (Markov chains, Gibbs/Markov random fields, co-occurrence matrices etc.). In this thesis we concentrate on second-order co-occurrence-based techniques—a powerful and computationally light subset of higher-order Markovian approaches. We identify a weakness common to co-occurrence and many other methods of analysis. This weakness is the commonly used technique of applying a set of fixed functions to extract discriminant features. Such functions may provide good general performance across a wide range of texture types, but often fail to capture texture information which is specific to the subset of types being analysed. That is, the methods are globally applicable but are not locally optimised.

We present a theoretical approach to texture classification which is applicable to all texture types but which 'adapts' to the specific characteristics of the texture being analysed. The self-adaptive multi-scale techniques based on this approach allow the simultaneous capture of texture characteristics which exist at, and across, several spatial resolutions. We show by a critical appraisal of the presented methods that this approach can provide significant improvements in cell classification accuracy. We also show how the captured characteristics can be used to identify image locations where differences between texture classes occur—something which is generally not possible with other analysis methods. Finally, we demonstrate the broad applicability of our methods by classifying a wide range of texture images from natural, industrial and biological origins.

Contents

List of Figures

List of Tables

Abbreviations and symbols

Abbreviations and acronyms

AMSGLCM	Adaptive Multi-Scale Grey Level Co-occurrence Matrix.
B.E.	Bachelor of Engineering.
CDF	Cumulative Distribution Function.
CSSIP	Cooperative Research Centre for Sensor Signal and Information Processing.
GAoGLCM	Genetic Algorithm optimised Grey Level Co-occurrence Matrix.
GCM	Generalised Co-occurrence Matrix.
GLCM	Grey Level Co-occurrence Matrix.
GLEM	Grey Level Entropy Matrix.
GLRLM	Grey Level Run Length Matrix.
GLVM	Grey Level Variance Matrix.
GRF	Gibbs Random Field.
i.i.d.	independent and identically distributed (random variables).
LCPDF	Local Conditional Probability Density Function.
MRF	Markov Random Field.
NGLDM	Neighbouring Grey Level Dependence Matrix.
PDF	Probability Density Function.
Qld.	Queensland.
SGF	Statistical Geometric Features.
SFM	Statistical Feature Matrix.

Symbols

A, B	point sets in Euclidean space,
c, v, s, g	indices,
i, j, d	array indices,
k, l, m, n	points in $\mathbb{Z}^2$,
(x, y)	point in Euclidean two-space,
ρ	the covariance matrix,
σ	standard deviation,
$\mathbb{R}$	the real numbers,
$\mathbb{Z}$	the set of integers, $\{-\infty, \ldots, -1, 0, 1, \ldots, \infty\}$
$\mathbb{R}^k$	k-dimensional Euclidean space,
$\mathbb{Z}^k$	k-dimensional discrete space,
$\boldsymbol{D}, \boldsymbol{G}$	subsets of Euclidean space; domain of a function,
$\oplus$	morphological dilation; Minkowski set addition,
$\ominus$	morphological erosion; Minkowski set subtraction,
$\circ$	morphological opening,
$\bullet$	morphological closing,
$\mathcal{F}, f$	function; signal,
I	image,
$\#$	number of,
$\lvert A\rvert$	cardinality of set A; magnitude,
$\lVert . \rVert$	Euclidean norm,
$\forall x$	for all x,
$\exists x$	there exists an x such that,
$\Rightarrow$	implies,
$A \subset B$	A is included in B,
$A \supset B$	A contains B,
$a \in A$	point a belongs to set A,
$a \notin A$	point a does not belong to set A,
$\blacksquare$	end of proof mark; end of definition mark.

Publications

These are the publications which have been produced by, or in conjunction with, the author, during his Ph.D. candidacy:

Refereed Journal Papers

The following papers were to be submitted to the indicated journals. However, due to the commercially sensitive nature of our work, submission has been delayed pending patent applications.

1. Walker, R. F., Jackway, P. T., Longstaff, I. D. (1997b), 'Image Texture Analysis Via Co-occurrence Methods - Review and Extensions', *IEE Proceedings: Vision, Image and Signal Processing* (**to be submitted**).

2. Walker, R. F., Jackway, P. T. (1997), 'Adaptive Multi-Scale GLCM', *Pattern Recognition*, (**to be submitted**).

3. Walker, R. F., Jackway, P. T., Longstaff, I. D. (1997a), 'Genetic Algorithm Optimisation of Adaptive Multi-Scale GLCM Features', *IEE Proceedings: Vision, Image and Signal Processing*, The Institute of Electrical Engineers (**to be submitted**).

Refereed Conference Papers

5. Walker, R. F., Jackway, P. T., Longstaff, I. D. (1997c), 'Recent Developments in the Use of the Co-occurrence Matrix for Texture Recognition', in *Proceedings of DSP'97, the 13th International Conference on Digital Signal Processing*, Santorini, Greece, 2–4 Jul, 1997, pp. 63–65.

6. Walker, R. F., Jackway, P. T. (1996), 'Statistical Geometric Features—Extensions for Cytological Cell Analysis', in *Proceedings of ICPR, the 13th International Con-*

ference on Pattern Recognition, Technical University, Vienna, Austria, 25–29 Aug, 1996, pp. 790–794.

7. Walker, R. F., Jackway, P. T., Lovell, B. (1995), 'Cervical Cell Classification Via Co-Occurrence and Markov Random Field Features', in *Proceedings of DICTA-95, Digital Image Computing: Techniques and Applications*, Brisbane, Australia, 6–8 Dec, 1995, pp. 294–299.

8. Walker, R. F., Jackway, P. T., Longstaff, I. D. (1995), 'Improving Co-Occurrence Matrix Feature Discrimination', in *Proceedings of DICTA-95, Digital Image Computing: Techniques and Applications*, Brisbane, Australia, 6–8 Dec, 1995, pp. 643–647.

9. Jackway, P. T., Walker, R. F., (1995), 'Scaled Gradient Watersheds and Cell Feature Extraction', in *Proceedings of DICTA-95, Digital Image Computing: Techniques and Applications*, Brisbane, Australia, 6–8 Dec, 1995, pp. 68–73.

10. Walker, R. F., Jackway, P. T., Lovell, B., Longstaff, I. D. (1994), 'Classification of Cervical Cell Nuclei Using Morphological Segmentation and Textural Feature Extraction', in *Proceedings Second Australian and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 30 Nov – 2 Dec, 1994, pp. 297–301.

Internal Technical Reports

12. Walker, R. F. (1997b), 'Genetic Algorithm Optimisation of Adaptive Multi-Scale GLCM Features', CSSIP Technical Report, Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide, Australia, TR2/97.

13. Walker, R. F. (1997a), 'Adaptive Multi-Scale GLCM', CSSIP Technical Report, Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide, Australia, TR1/97.

14. Walker, R. F. (1996), 'Statistical Geometric Features - Refinements for Cytological Cell Analysis', Internal Technical Report, Dept. of Electrical and Computer Engineering, University of Queensland, St. Lucia, Brisbane, Australia.

15. Walker, R. F. (1995), 'Improving Co-occurrence Matrix Feature Discrimination', Internal Technical Report, Dept. of Electrical and Computer Engineering, University of Queensland, St. Lucia, Brisbane, Australia.

Original Contributions

I believe the material contained in this thesis contributes the following items to current research in image processing and texture analysis:

1. **The extension of SGF**
   A contribution of this thesis is our extension of the Statistical Geometric Features (SGF) algorithm to the analysis of cell nuclear chromatin. We define new feature functions which measure specific cell chromatin properties. We show the benefits of such 'manual adaptation' of feature functions to suit specific texture properties in terms of increases in classification performance and decreases in feature set dimensionality. Of greater significance is the fact that defining feature functions which measure specific image properties allows far better understanding of the properties which manifest discrimination between classes. This cannot be said for most other analysis methods.

2. **The Discrimination Matrix**
   The most significant contribution of this thesis is our presentation of several related methodologies for optimised, multi-scale, self-adaptive feature extraction. All are based on locating areas of discriminatory power among texture descriptors extracted across a range of spatial resolutions. In Chapter 4 we introduce the *discrimination matrix*—a two-dimensional matrix which, for the first time, provides a direct indication of the potential 'worth' of each co-occurrence matrix element for classification purposes. The spatial arrangement of discriminatory information within the matrix also suggests new approaches to extracting optimised features, and for enhancing the discriminatory power of currently defined feature functions. We demonstrate the success of our approaches by significant decreases in classification error of over 70%.

3. **Adaptive Multi-Scale GLCM**
   In Chapter 5 we introduce the first self-adaptive multi-scale feature functions for use with co-occurrence-based methods of texture analysis. Our method places no

reliance on pre-defined fixed feature functions, unlike all other co-occurrence-based methods published in the literature. In fact, feature definition is based solely on the specific statistical differences between texture classes. Furthermore, our technique is possibly the first co-occurrence-based method to provide simultaneous analysis of texture across several spatial resolutions. Once again, the technique attains significant increases in classification performance across a wide range of texture types.

4. **GA-optimised GLCM**
   Chapter 6 introduces what we consider to be the first application of optimisation techniques for extracting co-occurrence matrix features, without using neural network-based methods. While neural networks have been used in the past to classify texture data, this 'black-box' approach often provides little theoretical guide as to the image properties that are producing the classification result. The GAoGLCM method allows direct analysis of the optimised feature functions and, using the remapping technique described in Appendix B, the locating of areas within images which produce discriminatory information between classes. Moreover, by careful objective function design, we can optimise feature functions under a number of criteria, including correlation considerations, first-order or joint discriminatory power, etc.

5. **Wide applicability**
   We should point out that, while we apply the adaptive methods of Chapters 4, 5 and 6 to GLCM matrix data, their applicability is not restricted to this method of analysis. They can equally be applied to *any* analysis method where a series of feature vectors or matrices can be extracted via suitable constraint parameters. Furthermore, the methodologies we develop in this thesis are 'globally applicable' (they can be used to analyse any type of texture), yet their adaptive nature means that they are also 'locally optimised' for the application problem.

6. **Locating discriminatory areas in images**
   In Appendix B, we demonstrate another significant application of the discrimination matrix for image analysis. Discriminatory features derived from conventional analysis methods can provide only *qualitative* cues as to the characteristics of image classes which are statistically different, such as image contrast, entropy, or energy. By using the information contained in the discrimination matrix, we can directly locate actual *areas* within an image which provide such discriminatory information. We demonstrate this in Appendix B by 'remapping' discrimination matrix co-ordinates to actual image pixels. This allows a far better understand-

ing of the processes or physical attributes of image objects which differ between classes. We believe this to be the first demonstration of this capability. The technique should prove to be a valuable tool for use in the areas of image analysis and understanding.

Acknowledgements

This thesis was completed in the Department of Electrical and Computer Engineering at The University of Queensland, as part of a research programme in the Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP). From the Department I would sincerely like to thank my supervisors Professor Dennis Longstaff and Dr Brian Lovell for their valuable encouragement and support throughout my research programme. I would especially like to thank my project manager Dr Paul Jackway, also known as the 'Fountain of Knowledge', a person I admire and am indebted to for his guidance, encouragement and patience. I would also like to thank the other members of the Cytometrics group: Pascal Bamford, Andrew Bradley, Jennifer Hallinan, Damian Jones, Andrew Mehnert and Guy Smith, for their comradeship, academic input, and the stimulating intellectual discussions which arose in our Lab from time to time. Thanks also goes to Professor Bill Moran of Flinders University for his technical input to the works of Chapter 6 of this thesis.

I would like to acknowledge the bodies that provided financial assistance for this project, and the Department of Electrical Engineering and CSSIP for providing the necessary infrastructure and scholarship support.

I would also like to thank my dear family for not disowning me, even though I rarely visited during the course of this PhD, for being amazed at the most trivial of advances in my research work, and most importantly, for the love and strong family values which have allowed me to achieve throughout my life.

This thesis has been typeset using the LaTeX[1] package. The text font used is 12pt Computer Modern Roman[2], other fonts from the Computer Modern Roman Family are used where required with some special symbols from the AMS-TeX[3] macro package.

[1] LaTeX (Lamport 1994) is a "descriptive markup" document processing system based on the TeX (Knuth 1986) typesetting system. TeX and LaTeX are publicly available via FTP on the network at address: `labrea@stanford.edu.au`.

[2] The Computer Modern (CM) family of fonts were designed by Donald Knuth especially for the TeX system.

[3] AMS-TeX (AMS 1991) is a macro package for TeX which defines a wide range of specialist mathematical symbols.

Chapter 1

Automated Cytology and Image Processing

1.1 Introduction

Mass screening of the Australian population to identify seemingly healthy individuals who harbour undetected illnesses has been a growing trend for over 30 years. It is generally accepted that population screening should be undertaken if:

- the disease represents a substantial public health burden;
- an inexpensive screening test is available with reasonable sensitivity, excellent specificity, and low risk;
- the curative potential is better in early, compared to advanced stages of disease; and
- the treatment of screen-detected patients improves their prognosis.

Some diseases meeting the above criteria and which are currently screened in Australia include tuberculosis, and breast cancer. One of the earliest mass screening programmes in Australia was initiated in 1965 for the detection of cervical cancer. Cervical cancer is the seventh most common form of cancer among Australian women (Australian Bureau of Statistics 1994) affecting predominantly post-menopausal, but occasionally young, females. The whole-of-life probability of developing malignant abnormalities of the cervix is currently 1%, and each year there are over 1000 new cases reported (Department of Health, Housing, Local Government and Community Services 1993). Fortunately, the successful implementation of a nation-wide screening programme now prevents approximately 750 cases of cervical cancer each year (Australian Institute of Health 1991), and has substantially lowered mortality. Despite this, the incidence of death due to this preventable disease continues to remain far too high, with 336 deaths recorded in 1994 (Australian Bureau of Statistics 1994). Cervical cancer is rarely seen in women under 25 years of age. Its incidence increases until the late 30's (25 cases per 100 000 women), after which there is a steady decline until age 50 (17 cases per 100 000 women). There is a further increase in incidence until age 65, where the number of cases remains close to 35 per 100 000 women (Department of Health, Housing, Local Government and Community Services 1991).

Cervical cancer usually begins as *pre-cancer* in the cells which line the cervix. The affected cells, often called *neoplastic* cells, are so called because they show signs which suggest a potential to become cancerous. Cervical cancer usually takes more than a decade to develop from the initial signs of *neoplasia* (pre-cancer) to its invasive form. While the exact causes of cervical cancer are unknown, there appear to be a number of pre-disposing risk factors, including sexual behaviour, age of first pregnancy, smoking, occupation and even social class (Chomet & Chomet 1989). The Human Papilloma Virus (HPV) is strongly associated with the incidence of cervical cancer, as it is present in approximately 80% of smears which show signs of cervical neoplasia (Chomet & Chomet 1989, Young, Bevan & Johnson 1989, Hallouche 1993).

Neoplasia covers a wide range of abnormalities which occur in the skin of the cervix, but are *confined* in the skin. The general condition of cervical abnormality is called Cervical Intraepithelial Neoplasia (CIN), and the relative degree of abnormality is specified by a numbered suffix, i.e., CIN1, CIN2, CIN3, in order of increasing severity. The three CIN grades replace the older histological terms of mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma in situ (CIS), with the last two terms representing CIN3. It is interesting to note that up to 50% of CIN1 and 2 cases spontaneously regress without treatment (Chanen 1990). Some CIN3 cases have also been known to spontaneously regress. However, women with CIN do have a significantly increased risk of developing invasive cancer.

Cervical neoplasia most commonly appears in the *transformation zone* of the cervix—see Figure 1.1. The transformation zone is an area of the cervix which undergoes continual pathologic activity. Columnar cells originating in the endocervical canal undergo metaplasia to squamo-columnar cells, and then subsequently to squamous cells. This activity often gives rise to pathological changes, including cell neoplasia. It is therefore vitally important that a sample of cells are taken from the transformation zone when taking a Pap smear test, because it is this region that is most likely to undergo pre-malignant or malignant change.

Neoplastic cells are both chemically, functionally, and structurally different from healthy cells, and they are capable of inducing similar changes in neighbouring healthy cells (Chomet & Chomet 1989). As the severity of the neoplasia increases, so does the uncontrollable growth in neoplastic cell numbers, and the depth of penetration of these cells into healthy tissue—see Figure 1.2. If left untreated, the condition may progress to carcinoma in situ (a pre-cancerous condition involving the full thickness of the epithelium), and then invasive cancer. When invasive, affected cells form tumours and invade other body tissue and organs such as lymphatic channels and the blood stream. This spread of cancer around the body from the initial site is known as *metastasis*.

The long latency of cervical cancer, along with the relative ease with which a sample of cells can be collected from the cervix, makes this disease particularly amenable to mass screening. The test, generally known as the 'Pap test', involves the collection of cells from several areas of the cervix, using a wooden spatula or plastic brush. The cervical scrape specimen is then smeared on a glass slide and fixed with an alcohol

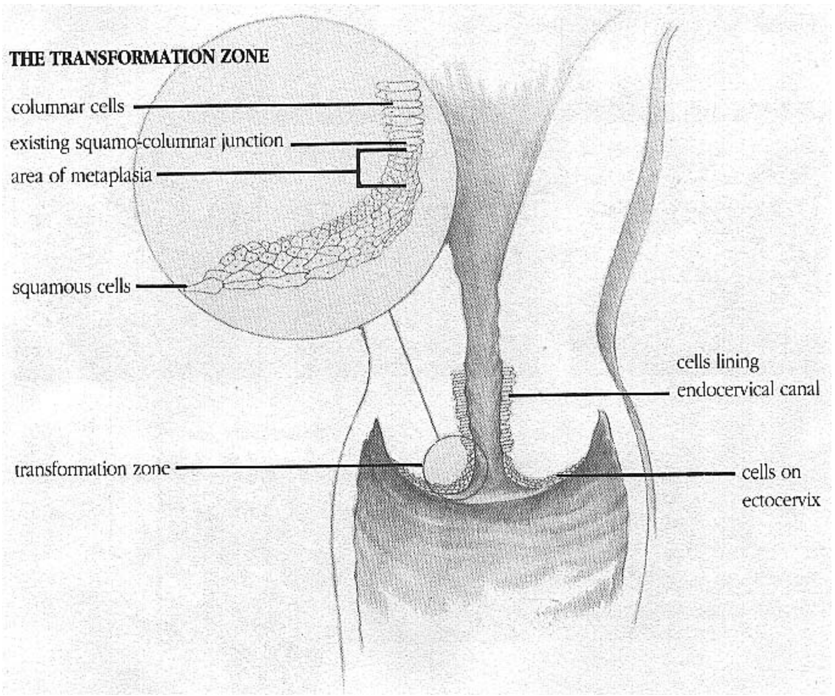
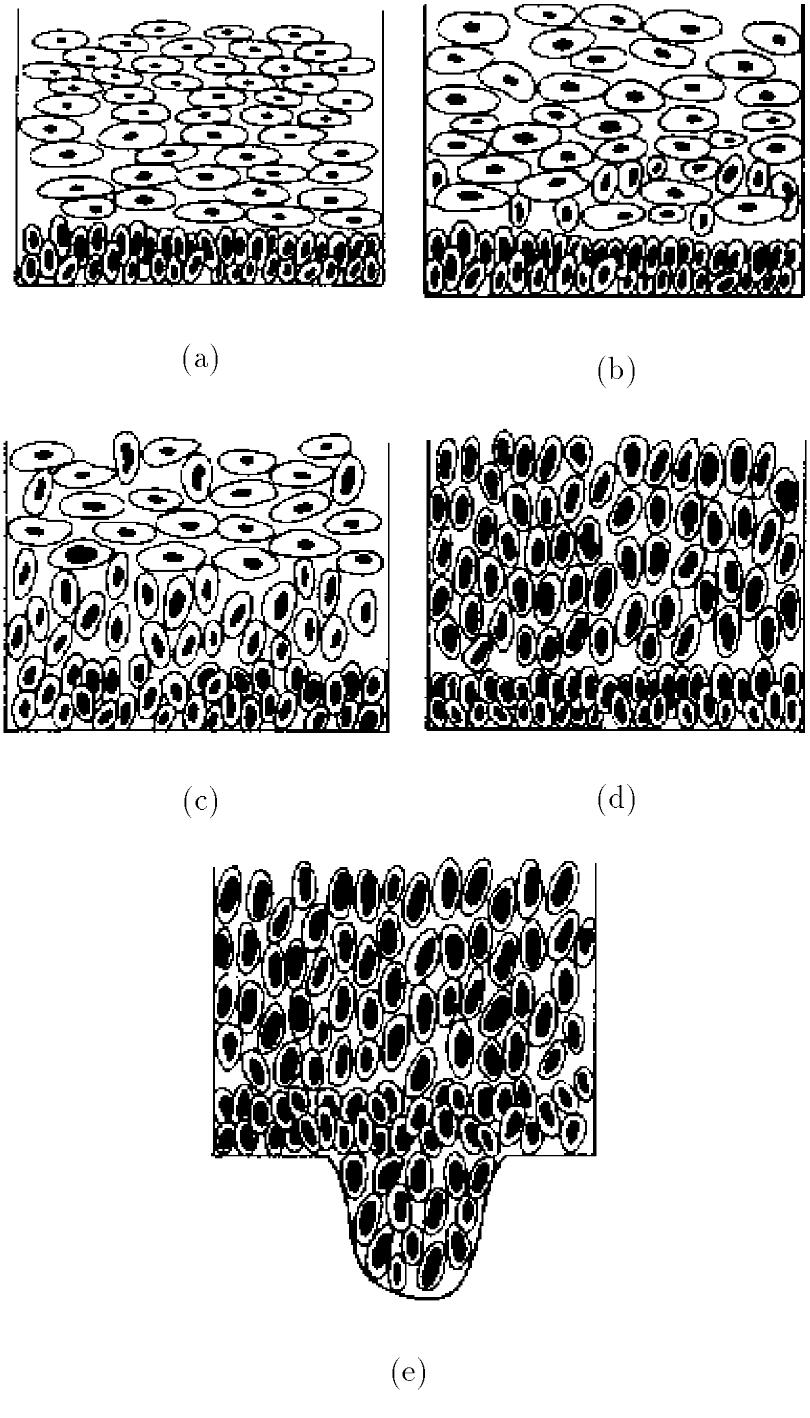


Figure 1.1: Location of the cervix and *transformation zone*. Around 85% of cervical cancers arise in the transformation zone. Reprinted from Department of Health, Housing, Local Government and Community Services (1991).

solution. Because the cells are translucent, they must undergo a process of staining to highlight cytological structure and to aid in their microscopic examination. The staining material binds to the cell nuclei, membranes, and other debris and artifacts on the slide. The most widely used staining procedure is the Papanicolaou staining method (commonly known as Pap stain), first introduced by George Papanicolaou (Papanicolaou & Traut 1943, Papanicolaou 1945). By deeply staining nuclei, while lightly staining cytoplasms, the Pap stain provides important colour differences and increased contrast between nuclei and cytoplasm. In Figures 1.3–1.5 we show several images of cervical smears stained using the Pap staining process.

1.1.1 Cervical Cancer Screening in Australia

The Australian Department of Health's national policy on cervical cancer screening encourages sexually active women to undergo a Pap smear test on a two-yearly basis (Department of Health, Housing, Local Government and Community Services 1991). Women are also encouraged to continue testing for cervical cancer until age 70. Two-



(a) (b) (c) (d) (e)

Figure 1.2: Grading of cervical cell neoplasia: (a) Normal; (b) CIN1; (c) CIN2; (d) CIN3; (e) CIS with micro invasive carcinoma.

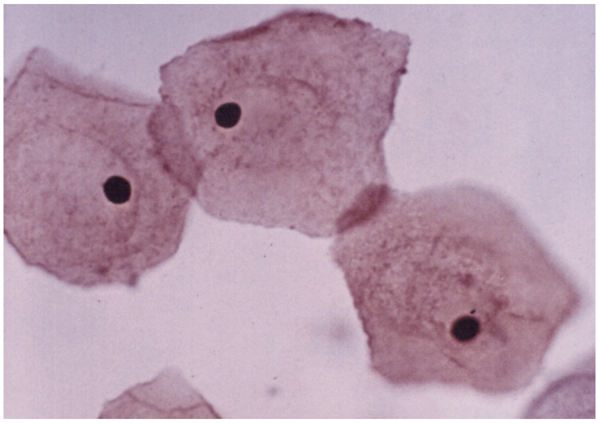


Figure 1.3: An example of squamous epithelial cells, captured at ×400 magnification. These cells are diagnostically normal, and are characterised by large cytoplasmic area and small, dense nuclei.

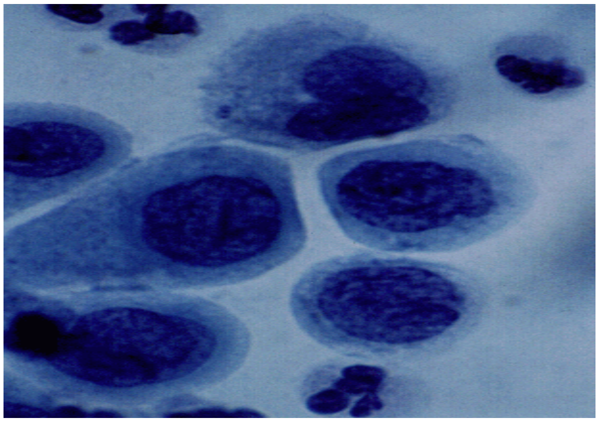


Figure 1.4: An example of CIN, captured at ×1000 magnification. These neoplastic cells are characterised by irregular nuclear shape, much larger nuclei, and large nuclear-to-cytoplasmic (NC) area ratio. Large NC ratio is generally a strong indicator of malignant abnormality.

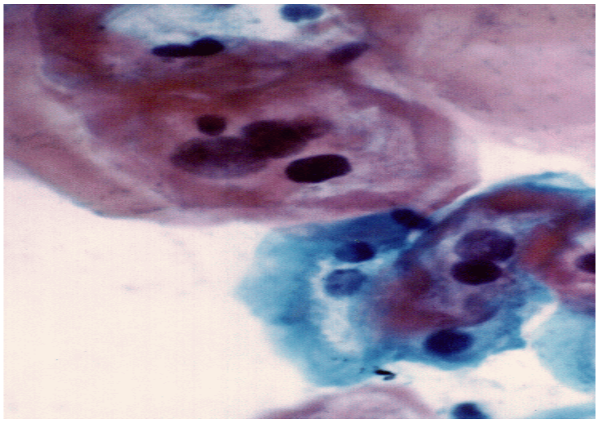


Figure 1.5: An example of cells infected by HPV, indicated by the 'halo effect' around the cell nuclei. This is a benign condition, however, there is a strong link between the presence of this virus and the occurrence of cell neoplasia.

yearly Pap smear tests reduce the cumulative incidence of cervical cancer by over 92% (Department of Health, Housing, Local Government and Community Services 1991). This programme raises two very important issues, relating to the financial cost to society, and to its effectiveness in significantly reducing the occurrence of cervical cancer in the community. Firstly, bi-annual screening has resulted in a costly and labour intensive programme of manually viewing millions of Pap smear slides each year. The current annual cost to Government and health insurance bodies is in the vicinity of 125 million dollars (Australian Institute of Health 1991). Secondly, the Pap test is explicitly a *screening* test, as opposed to a diagnostic procedure. As such, the huge number of Pap slides requiring analysis each year (currently 2 million annually in Australia), necessitates minimising the total time spent on each slide by cytologists to around 10 minutes. As each slide contains up to 400 000 cells, a thorough examination of every cell is impossible, resulting in an elevated risk of slide misclassification. It is generally accepted that 15–30% of Pap smears are mis-diagnosed in some way—a result of poor sampling of the cervix, poor staining of cells, poor cytologist training, fatigue, etc.

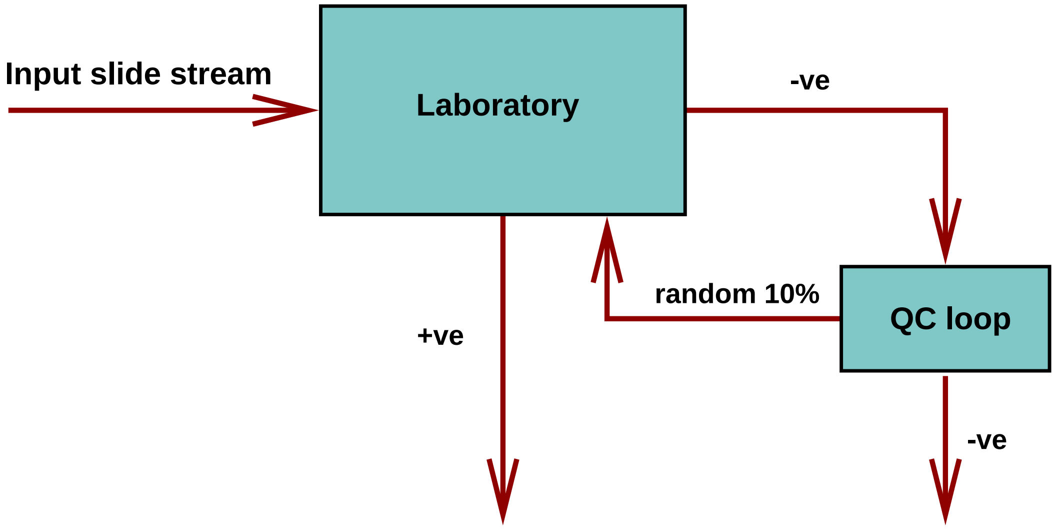


Figure 1.6: Manual screening procedures at cytology laboratories. As a quality control check, a random 10% of slides classified as negative (containing no detected abnormalities) are returned for further manual screening.

Figure 1.6 shows the operational procedures associated with manual screening undertaken at cytology laboratories. Manual screening of Pap slides is a two-stage process—screening and (if necessary) diagnosis. Each slide in the input stream is initially screened by a trained cytologist. This process involves microscopic examination of ‘fields’ of cells at low magnification. Suspicious cells are examined more closely using higher magnifications. The locations of suspicious cells are marked on the slide cover slip to allow their relocation. Slides which contain even a single abnormal cell, or which appear suspicious in any way, are further examined by a cytotechnologist or cytopathologist. These include slides which contain benign abnormalities such as HPV, candida, or the herpes virus etc. The grading of a slide is based on the grading of the slide's most abnormal cell—so-called *extreme value* grading.

Usually, a smear is reported as belonging to one of four groups:

- Negative;
- Abnormal;
- Positive;
- Unsatisfactory.

A smear may be reported as "unsatisfactory for evaluation" for several reasons. These include a slide with an insufficient number of squamous epithelial or endocervical cells (indicating poor sampling of the cervix and in particular, the transformation zone), obscuring blood or other artifacts, or a broken slide (Kurman & Solomon 1994). Such smears do not allow the reliable detection of cervical abnormalities. In such a situation a patient is requested to undergo a second smear test.

A "Negative" smear indicates a smear which was satisfactory for evaluation purposes, and upon which no abnormalities were detected. Patients receiving this result need not undergo a further examination for a period of two years.

An "Abnormal" smear report results when any changes which differ from a negative slide are detected. These changes include benign cellular abnormalities, as well as inflammatory changes which occur as a result of bacterial, viral and fungal infections (Chomet & Chomet 1989). Patients with abnormal smears are requested to have the cause of the abnormality treated, followed by a second smear test.

A "Positive" smear is reported when a pre-cancerous or cancerous abnormality is detected on the slide, even if only a single cell exhibits the abnormality. Because the Pap test is simply a screening test, it does not provide vital diagnostic information such as the position, spread or depth of the abnormality on the cervix. This information is necessary for determining which of the various treatments a patient will undergo—usually cone biopsy or laser for pre-cancer, and hysterectomy or radiotherapy for invasive cancer. Patients with positive smears are requested to under colposcopic examination, where the cervix is directly examined under magnification.

Smear examination is a qualitative and somewhat subjective process, where characteristics of the cells are examined under light microscope. The most common cues to possible cell abnormality include:

- Nuclear size. Pre-cancerous cells generally have larger nuclear area;
- Cytoplasm and nuclear shape. Pre-cancerous cells generally have cytoplasms and nuclei of irregular shape;
- The ratio of nuclear to cytoplasm area (NC ratio). For abnormal cells, the NC ratio is generally larger;
- Cell spatial context. Abnormal cells are often found in clumps;
- Nuclear chromatin density. Pre-cancerous cells tend to have denser chromatin.

It is common for several of these features to be present on an abnormal slide.

Of the slides which are initially diagnosed as 'negative' (i.e., those on which no abnormalities are found), a small sample—usually 10%—are selected for rescreening by a senior cytologist. This involves manually screening the selected slides a second time. This gives a measure of quality control for the screening process and the laboratory as a whole.

1.2 Automated Analysis of Cervical Smears

Because of the huge workload and financial cost involved in population screening, and the high incidence of slide misclassification due to human error, the mass-screening programme would benefit greatly from the introduction of automated analysis of cervical smears. Machine-assisted analysis may not only reduce the huge financial cost of slide processing, but more importantly may result in more accurate diagnosis. Computerised cervical screening has been the subject of research for over 30 years. Early automated systems included

- TICAS (Wied, Bartels, Barh & Oldfield 1968),
- CERVIFIP (Tucker & Shippey 1983),
- Cybest (Tanaka, Ikeda, Ueno, Watanabe, Imasato, Tsunekawa, Okamoto, Kashida & Mukawa 1979, Tanaka, Ueno, Ishikawa, Konoike, Shimaoka, Yamauchi, Hosoi, Okamoto & Tsunekawa 1980),

- Leytas (Ploem, Werwoerd, Bonnet & Koper 1979, Ploem, Goyarts-Veldstra, van Driel-Kulker, Zaaner & Meyer 1983, van Driel-Kulker & Ploem 1982),
- BioPEPR (Zahniser, Oud, Raaijmakers, Vooys & van de Walle 1979), and
- SAMBA (Brugal, Garbay, Giroud & Adelh 1979).

Many of these systems failed due to software and hardware deficiencies, and loss of financial backing. To be successful, an automated system needs to analyse over 10 000 cells per minute, and requires very specialised and expensive computer architecture (Anderson 1994). Recent automated systems utilise image processing techniques to measure various slide, cell, and nuclear characteristics called *features*. These include contextual features such as cell distributions (Garcia 1986), morphometric or photometric features of the cells including cell size and shape (Zahniser 1994, Knesel Jr, Geyer, Gahm, Nguyen, Fischer & Dorrer 1994), nuclear features including shape and texture, and Malignancy Associated Change (MAC) measurements (Garner, Ferguson & Palcic 1994). These features are known to change when cells become neoplastic. By measuring these features and applying Pattern Recognition (PR) techniques, an automated system can locate the 'most abnormal' cells on a slide, and flag them for review by a trained cytologist. This substantially reduces the task of manually locating abnormal cells among the 400 000 on a typical slide.

The most successful automated systems have now gained US FDA approval for use as quality control *rescreeners* or *adjunct screeners*. These include the AUTOPAP 300® (Anderson 1994) by NeoPath, Inc., and PapNet® (Mango & Herriman 1994), by Neuromedical Systems, Inc. A quality control rescreener replaces the *random* selection of negative slides for rescreening—see Figure 1.7. Rescreeners locate possibly-abnormal cells on slides which were previously classified as normal by human experts. By returning the 'most abnormal' negative slides for further manual analysis, rather than just a random sample, false-negative rates due to slide misclassification can be further reduced.

Recent negative publicity regarding the accuracy of manual screening has lead to the creation of new opportunities for machine assisted screening. In many countries, including the United States and Australia, women now have the choice of having their Pap slide rescreened by an adjunct screener. As Figure 1.8 shows, such slides are screened twice, both manually and (if negative) by machine. During rescreening, the most abnormal cells on a slide are located, and their images displayed on a high-resolution monitor for viewing by a cytologist. Adjunct screening currently incurs an extra charge (approximately $30), and is paid directly by the patient.

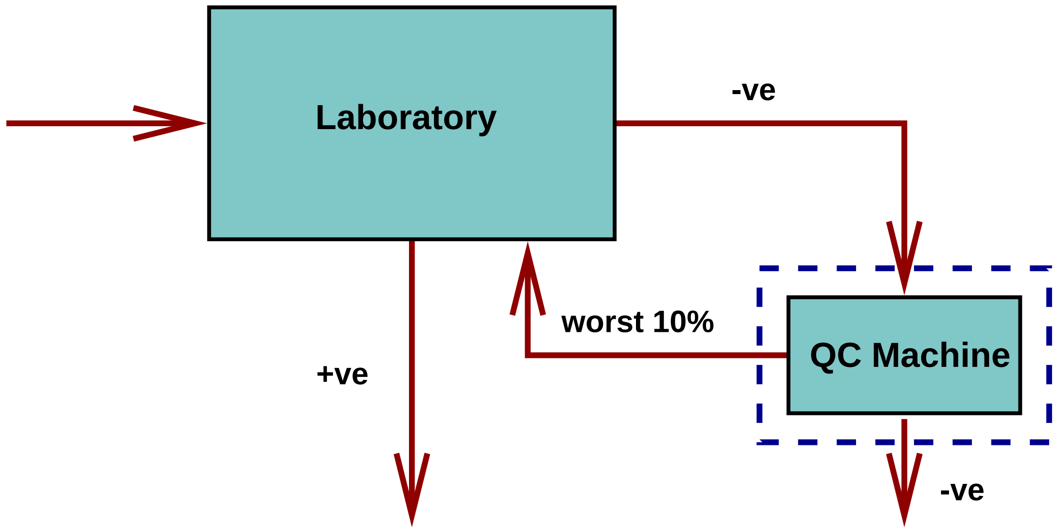


Figure 1.7: Using an automated system for quality control rescreening. All slides manually classified as normal are passed through a quality control check. The worst 10% of these normal slides are returned for further manual screening.

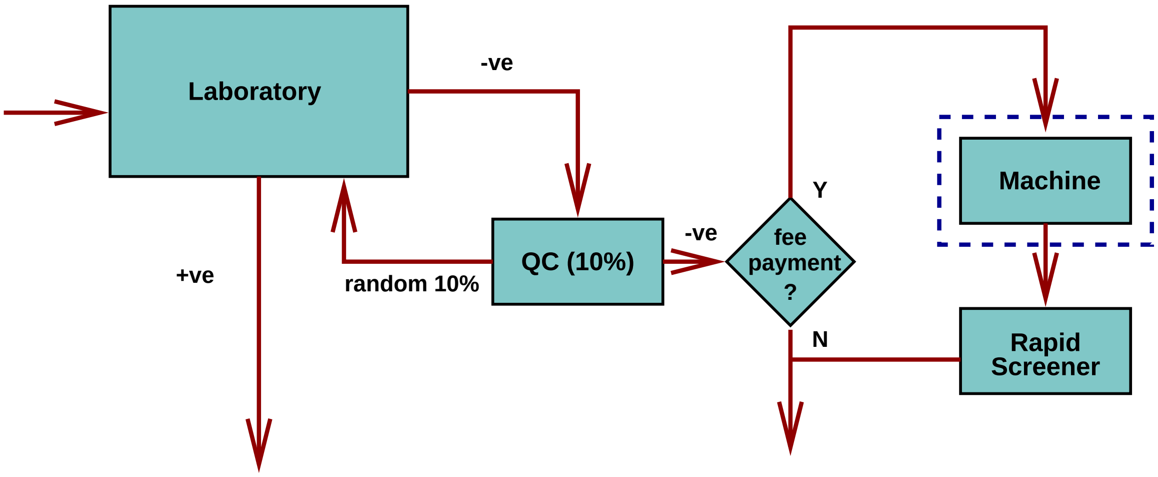


Figure 1.8: Using an automated system as an adjunct screener. Slides classified as normal by the QC manual screeners are further rescreened by an automated system. The 'most abnormal' cells on these slides are then reviewed by a cytologist.

Using *pre-screeners* can yield financial benefits for cytology labs and has the potential for improving slide diagnosis accuracy. A pre-screener reduces the volume of slides manually screened by removing the 'most normal' slides from the input slide stream. This reduced volume allows the use of fewer trained screeners and an equivalent reduc-

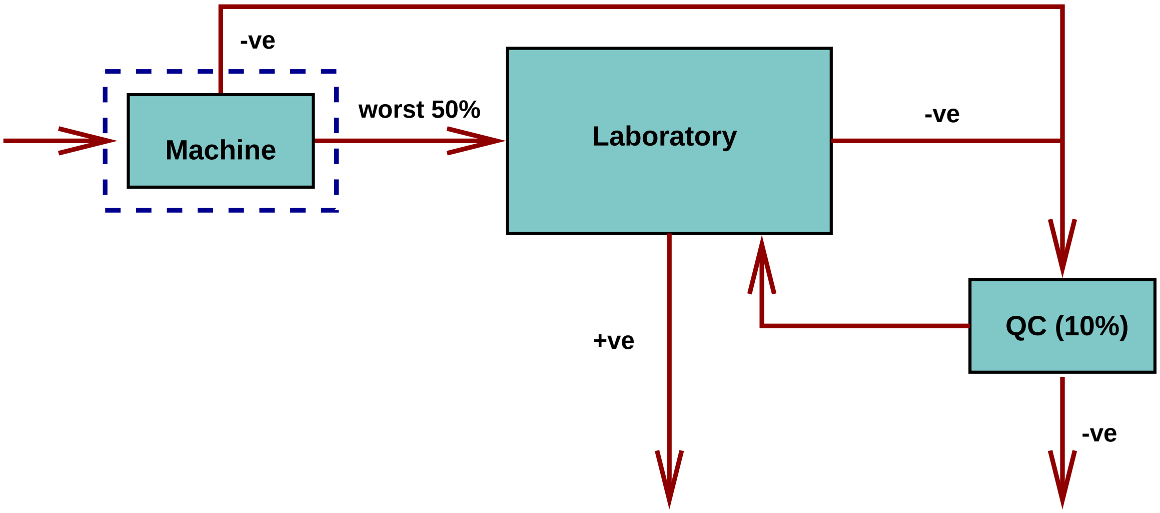


Figure 1.9: A pre-screener enriches the flow of abnormal slides to human screeners by removing the 'least abnormal' slides from manual screening.

tion in infrastructure and equipment. Alternatively, the time spend manually examining a slide can be increased. Potential increases in diagnosis accuracy can be achieved because many slides are screened twice—by both human and machine screening. Also, removing the 'most normal' slides enriches the stream of possibly abnormal slides, making the task of manual screening easier. Figure 1.9 shows such a system, where the 'most normal' 50% of slides are removed from manual screening. Because only 10% of the machine-classified normal slides are QC reviewed, the automated pre-screener *must* be accurate!

Of greatest benefit are automated systems for use as *primary screeners*, where all slides are diagnosed by machine. Several countries are already using the PapNet® automated screener as a primary screener, including Hong Kong, the Netherlands, Belgium, and Switzerland.

The success or failure of these systems rests with their acceptance by members of the medical profession. However, wide acceptance may not be as far away as we may initially believe. In a recent survey (Wied 1994) of a randomly selected 180 from a group of 1,344 internationally prominent cytologists from 56 countries, 166 responses were received to the question:

> "Do you believe that **interactive cytomorphometry automation** (i.e., that a computerised system will pre-scan the slide and show "alarms" for human intervention and decision) will become a useful routine addition to your laboratory in the foreseeable future?"

Over 83% of respondents answered "yes". Of these, 93% believed that such a system would be "economical, diagnostically accurate, and commercially available" within five years or less.

1.3 Malignancy Associated Changes

One weakness of both manual screening and current automated systems is their inability to detect abnormalities on slides from patients with cervical abnormalities, when the slide contains no diagnostic cells because of *sampling error*. Sampling error occurs when cells from an area of the cervix containing abnormality are not collected during a cervical examination. This is usually due to poor technique on the part of the physician collecting the smear. It has been conservatively estimated that around 15% of Pap smears test falsely-negative, because of the absence of any diagnostic cells on the slide—a result of poor sampling of the cervix (Gay, Donaldson & Goellner 1985, Department of Health, Housing, Local Government and Community Services 1993). Current automated or manual screening techniques cannot detect patients with cytological abnormalities when such abnormalities do not appear on the cervical smear. Recently, a Canadian Research team (Palcic & MacAulay 1994, Garner et al. 1994) demonstrated an automated screening technique using a completely different principle called *Malignancy Associated Changes* (MACs), which may overcome this sampling problem. MACs is thought to be a response by normal cells to a 'field effect' induced by malignant lesions or even neoplastic cells, and was first reported over 75 years ago (Gruner 1916). This response by apparently normal cells is in the form of very subtle but statistically measurable changes in cell characteristics, including

- slight increases in nuclear-to-cytoplasmic area ratio,
- orderly structure of heteropyknotic chromatin,
- numerous chromocentres,
- curved chromatin bands in crowded but orderly arrangements,

- clear spherical nuclear areas of uniform size surrounded by curved pyknotic chromatin bands, and
- an absence of nucleoli or multi-nucleated cells (Neiburgs 1968).

It is interesting to note that virtually all the above indicators relate to DNA organisation and distribution in the cell nucleus. We will discuss this further in Chapter 3. Hanselaar, MacAulay, Palcic, Garner & LeRiche (1992) and Payne, Lam, LeRiche, MacAulay, Ikeda & Palcic (1994) have also suggested that MACs may play a role in discriminating between progressive and regressive lesions, thus providing important prognostic information and possible reductions in treatment costs. That is, by careful monitoring of a *MACs score* or index[1], it may be possible to avoid expensive and somewhat traumatic procedures such as cone biopsy or laser evaporation. Furthermore, sampling error is no longer of concern, because we are no longer looking for diagnostically abnormal cells which may only number less than ten among 400 000 normal cells (so-called rare event detection). The MACs affect is expressed by apparently normal cells, so we only need a representative sample of normal cells from the cervix. But perhaps the biggest advantage of an automated system operating on the MACs principle is the fact that we no longer need to analyse every cell on a slide for possible abnormalities. With MACs, we only need to analyse a small sample of the best presented and most visible cells on a slide. By analysing far fewer cells, we can increase the analysis time per cell manyfold, thus increasing the accuracy of the screening process without increasing the overall time for screening each slide. Moreover, the highly expensive, custom-made, high-speed computer vision components of current automated systems, necessitated by the need to analyse 400 000 cells in minimal time, can now be replaced by more cost-effective components. We therefore have a much greater potential for introducing low-cost MACs-based automated screening systems into cytology laboratories.

The decreased photo-mechanical requirements of a MACs system are offset by increased computational requirements. Because the MACs affect is expressed in *very* subtle changes in the characteristics of normal cells, we are now faced with the challenge of defining new ways to measure such minute changes. Unfortunately, many of the cell descriptors published over the last 30 years and currently used in commercial systems, are of little use in a MACs system because they are based on discriminating between normal and truly-abnormal (neoplastic) cells which have vastly different characteristics to MACs-affected normal cells. However, they do provide us with a foundation from which to begin our search.

[1] a number indicating the severity of cervical abnormality.

This concludes our introduction to the history of, and current trends in, cervical cancer mass-screening in Australia and around the world. We have also covered the topic of automated systems which use computer vision as a basis for reducing costs and increasing the reliability of the screening process. We have not yet discussed the fundamental building blocks of such systems, or the computational processes which allow autonomous detection of pre-malignant abnormalities. In the next section we discuss the processes typically found in automated systems which utilise computer vision techniques.

1.4 Pattern Recognition

Pattern recognition is the application of mathematical and statistical techniques to the identification and classification of objects from differing classes. Generally speaking, PR is a science concerning the description and recognition of objects using measurements. The four engineering approaches to PR are statistical, structural, syntactic, and more recently, Neural Network (NN). This section concerns the former approach (statistically-based PR), because it is by far the most widely applied. PR techniques are generally developed in two stages—a *training* stage where the PR system is first designed, and a *testing* stage where the performance of the system is evaluated. By *classes*, we mean groups of objects whose properties intrinsically differ between groups, but which are similar for objects within groups. Cytologically normal and abnormal cells are one example of differing classes, as are species of animals or plants etc. We can characterise classes by the qualitative or quantitative measurement of object properties, known as *features*. Often, to facilitate automated analysis, these properties are measured from images of the object, rather than directly from the object. Prior to such measurement, it is often necessary to isolate the object of interest from its surrounding environment—a cell nucleus from its surrounding cytoplasm etc.—in a process called *segmentation*. The measurement of object features is known as *feature extraction*. We usually do not know *a priori* which features or object characteristics will best discriminate between classes, so it is common to extract a large quantity during the training stage of system design. This large set of features is then reduced by a process known as *feature selection*. Feature selection involve measuring statistical parameters (such as mean and variance) from class-conditioned distributions of features. Features are said to possess *discriminatory power* when their estimated statistics vary between classes in some manner.

The aim of any pattern recognition system is to enable the allocation of an object, whose class membership is unknown, into one of several classes. Classifying a cell sample from a patient, into either the normal or abnormal class, is such an example. The allocation of objects into classes, based on its features, is called *classification*. Classifier design is based on *discriminant analysis*—a technique for separating and classifying two groups of data. Discriminant analysis has two parts to it. The first is discriminating between two groups of multivariate data from known sources. These two data groups are the control groups, known as the *training set*. The second step is classifying data of unknown origin into one of the two groups by applying the discriminating function to the unknown group.

A knowledge base plays an important role in PR systems. It encodes prior knowledge about the problem domain, and helps to guide each of the PR sub-blocks during system training and testing. The knowledge base can also accept feedback information from sub-blocks and thus control the interaction between them. An example of this would be error information from an image segmenter being used to control the image acquisition and pre-processing blocks for better image quality. It is commonly believed that the greater the amount of encoded prior knowledge, the greater the performance of the PR system. Figure 1.10 details each of the fundamental components in a typical pattern recognition system used for image processing. We will now discuss each of these components in more detail.

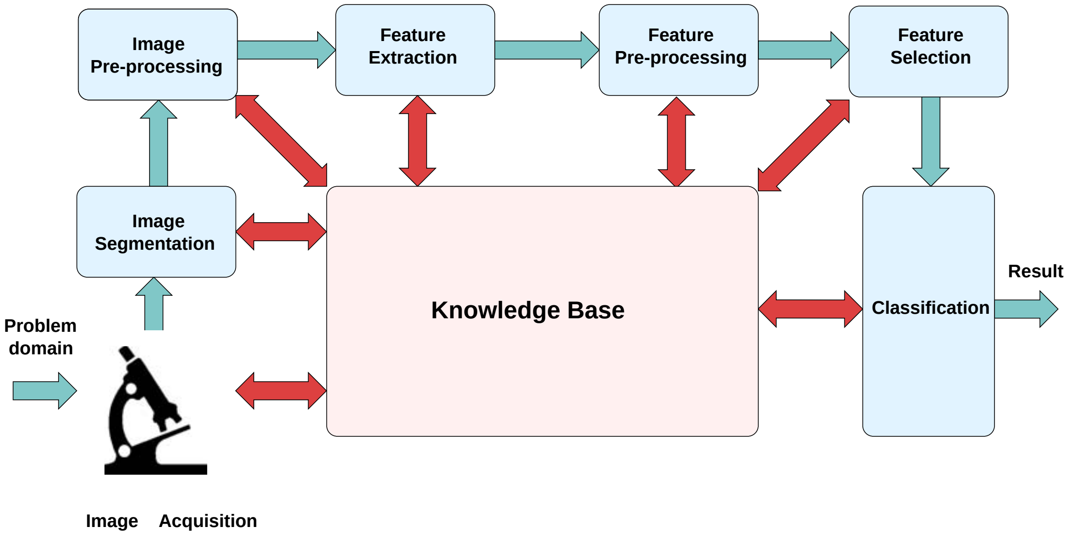


Figure 1.10: The fundamental components of a typical pattern recognition system used for image processing. Notice the central role of a knowledge base in guiding the operation of each PR sub-system.

1.4.1 Segmentation

Segmentation is the process of dividing an image into its constituent parts for further analysis. We commonly refer to such parts as *regions of interest.* For example, computer analysis of cervical cells commonly entails analysing cell cytoplasm, nucleus, and chromatin—see Figure 1.11. We need to isolate each region of interest from other regions which have no diagnostic importance, such as image background or artifacts. As shown in Figure 1.11, the level of segmentation required, and the algorithms used to achieve such segmentation, are influenced by the characteristics of each region of interest, i.e., the problem domain. As such, the success of a segmentation algorithm is inextricably linked to encoding as much prior knowledge as possible into the PR system. The segmentation process is complete when all regions of interest have been isolated.

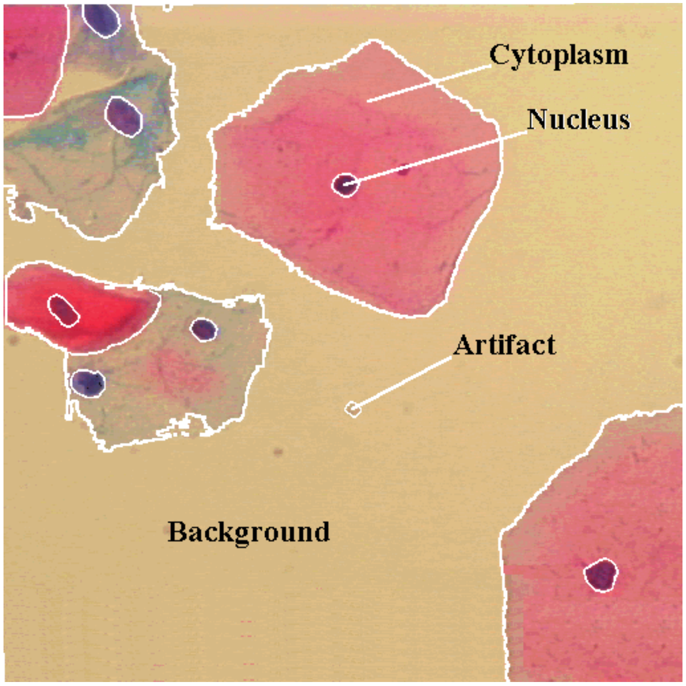


Figure 1.11: An example image of cervical cells, detailing typical regions of interest which may be of diagnostic importance—cell nuclei and cytoplasms. Other regions which may be of no diagnostic importance include artifacts and image background. The white lines represent segmentation boundaries which isolate the constituent parts of the image. Reprinted from Bradley (1996).

Segmentation algorithms are as numerous as the applications in which they are applied. Similarly, there are several dichotomies to which these algorithms can be grouped. One example is similarity and discontinuity methods such as *region growing* and *edge* *detection* (Gonzalez & Woods 1993). These two methods apply complementary approaches to image segmentation. In region growing, image components are detected based on pixel similarity or region homogeneity. Edge detection, however, segments an image based on dissimilarity or inhomogeneity between image pixels or regions. A second dichotomy is between global and local methods of segmentation. Both region growing and edge detection can be considered as local methods, because the classification of an image pixel as either boundary or object is based on properties which are spatially local to the pixel. An example of a global method of segmentation is *global thresholding*, where an image is segmented based on the statistics of the entire image. Global thresholding is particularly useful when the grey levels of the image components and background form two or more dominant modes. We demonstrate this method in Figure 1.12 by segmenting a grey-scale image of a cell from its surrounding background. A histogram of the image I reveals a bi-modal distribution of grey levels, representing the cell and background. Choosing a threshold level T in the valley between these two distributions ($T = 210$), and setting all pixel intensities below this threshold to 0, successfully segments this image. We can express this operation thus:

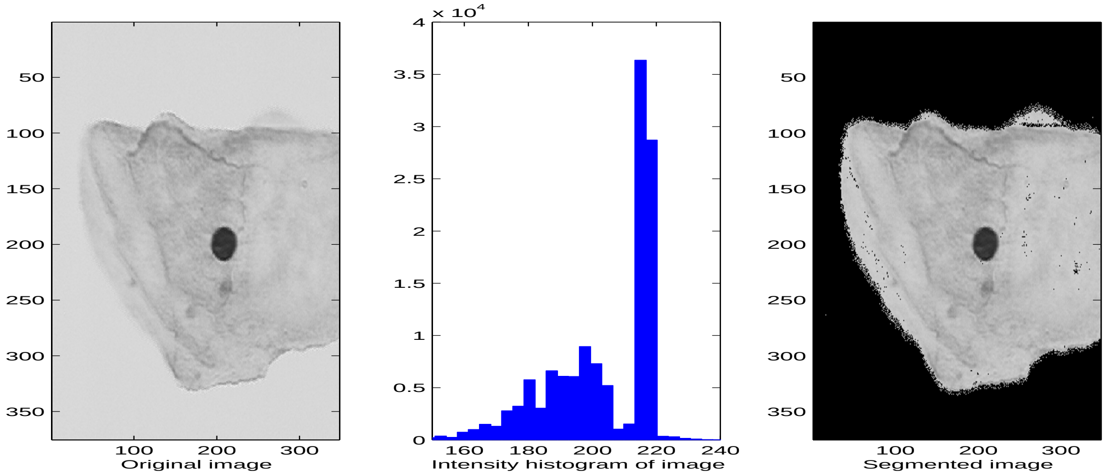


Figure 1.12: An example of image segmentation using a global threshold. Notice that the histogram is bi-modal, with two peaks at $I = 200$ and $I = 220$, representing foreground (cell) and background (slide) image components. By choosing a suitable threshold between these intensities, i.e., $T = 210$, we can successfully segment the image components.

$$I'(x,y) = \begin{cases} I(x,y) & \text{if } I(x,y) \geq T; \\ 0 & \text{if } I(x,y) < T, \end{cases} \tag{1.1}$$

where the co-ordinate pair (x, y) is a valid image pixel.

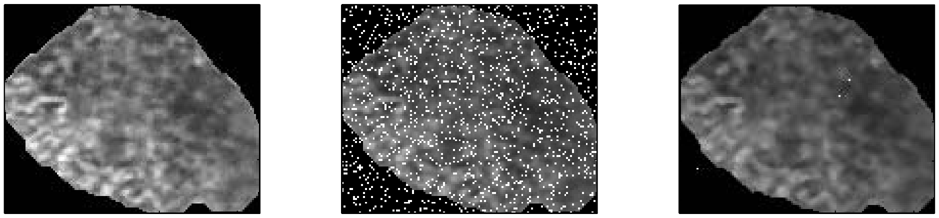
Recent techniques such as morphological (Vincent & Beucher 1989) and multi-resolution segmentation (Spann & Wilson 1985, Hong & Rosenfeld 1984), which work on images both globally and locally, have gained significant interest as they overcome the inherent deficiencies of the traditional methods mentioned above. We will discuss in detail our methodology for image segmentation, using a morphological process, in Section 3.4.2.

1.4.2 Image Pre-processing

Photometric pre-processing of digital images is a requirement of many image analysis systems. We apply pre-processing prior to the application of subsequent image operations for several reasons, including:

1. to restore images that have been corrupted in some way;
2. to enhance image attributes which are of particular interest; and
3. to requantise or normalise a set of images to ensure common photometric or statistical properties.

For example, images corrupted by 'impulse' noise can be pre-processed using a median filter to remove such noise. Impulse noise is a stochastic process which results in a random selection of image pixels taking on extreme intensity values. A median filter replaces the grey level of a pixel by the median value of the grey levels in a neighbourhood surrounding the pixel. It is a form of filtering which removes spike-like components from an image, yet preserves edge sharpness. Removing noise from images is an example of *image restoration*, and we show the results of such a process in Figure 1.13.



Original image Corrupted image Restored image

Figure 1.13: The process of restoring an image corrupted by impulse noise. The restored image is the result of median filtering the corrupted image, using a 3×3 neighbourhood.

A second type of pre-processing, called *image enhancement*, involves applying operators to enhance, or highlight, image properties. Enhancing such properties aids in the visual examination, analysis, and understanding of images. One example of enhancement, called *contrast stretching*, is often applied to images with low contrast, to increase the dynamic range of intensity levels in the image. Figure 1.14 shows the results of a simple linear contrast stretching operation defined as

$$\forall(x, y), \quad I'(x, y) = \left\{ \frac{I(x, y) - I_{\min}}{I_{\max} - I_{\min}} \times \mathrm{I_D}_{\max} \right\}, \tag{1.2}$$

where the co-ordinate pair (x, y) is a valid image pixel, $I'(x, y)$ is the requantised value for pixel intensity $I(x, y)$, and $\mathrm{I_D}_{\max}$ is the maximum photometric intensity of the device displaying the image.

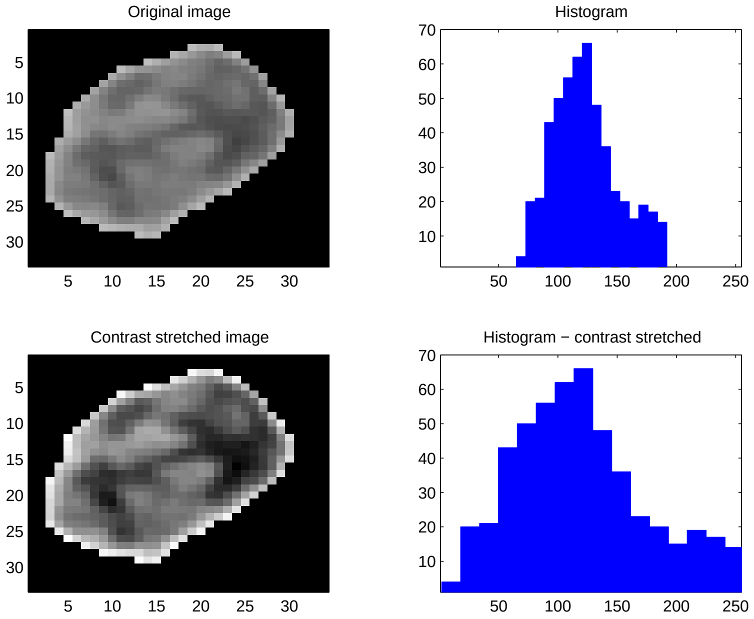
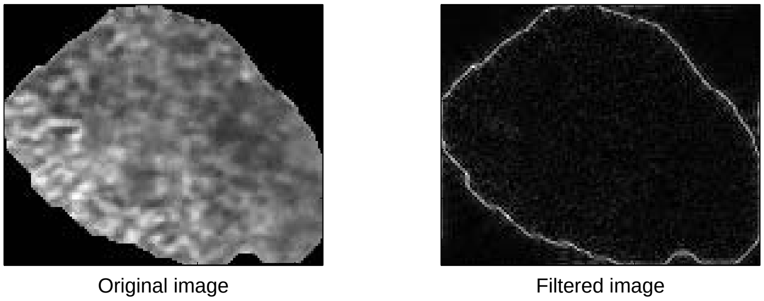


Figure 1.14: Increasing the dynamic range of pixel intensities by a process called *contrast stretching*. A comparison of the original and contrast-stretched images reveals the usefulness of this technique in enhancing the visibility of chromatin structure in cell nuclei. The resulting histogram of pixel intensities on the right shows that we have attained full use of all pixel intensities.

A further example of image enhancement involves applying a high-pass filter to enhance high gradients or edges in an image, as shown in Figure 1.15. This operation may be applied to an image to aid in its segmentation, or it may be used to highlight



Original image

Filtered image

Figure 1.15: Edge detection of an image by high-pass filtering. The resulting image can now be segmented using a simple global thresholding technique.

fine structure in an image that cannot be enhanced by standard contrast stretching operations.

Another example of image pre-processing is the *normalisation* of images captured from non-stationary photographic processes, such as x-ray images or microscopy images. Photographic processes are often prone to variation in photometric properties across a sequence of images, such as mean intensity, contrast, or photometric linearity. Variations of this type are a form of introduced artifact, or noise, in the image. It is essential that we remove such noise, prior to measuring image properties, to allow their most accurate estimation. Image normalisation can help remove the effects of non-stationary photographic processes, allowing the PR system as a whole to be independent of such variation. By pre-processing images using image normalisation algorithms (re-quantisation, histogram equalisation, etc.), we also ensure that all images are members of a common photometric domain. Pre-processing operations such as requantisation also fulfill a data-reduction role, by reducing a large, but variable, number of image intensities to a smaller, but fixed, number. In subsequent chapters, we will be comparing the performance of several algorithms by the classification of 8-bit grey-scale images. The computational burden and storage requirements imposed by these algorithms are related to the number of image intensities N_g, or the number of image intensities squared N_g^2. By requantising the original $8-$bit images from $N_g = 256$ to $N_g = 16$ grey-levels, we ensure not only common photometric domains for all images, but also a corresponding reduction in computational burden and storage requirements to $\frac{1}{16}$th or $\frac{1}{16^2}$th of the original. We will discuss image normalisation and histogram equalisation further in Section 3.4.3.

1.4.3 Feature Extraction

Perhaps the heart of all pattern recognition systems is measuring descriptive properties of image components. We call such measurement *features extraction*, and it involves measuring properties which in some way characterise the object of interest. An example of feature extraction would be measuring a cell's area by counting the number of image pixels within the confines of the cell boundary. In many cases, feature types are directly related to properties that are qualitatively measured by human vision. It is the ability of computer vision systems to accurately, rapidly, and repeatedly measure object properties *quantitatively*, without subjective bias, which provides one of their greatest benefits.

The types of image properties measured in cytometry generally fall into several categories:

- morphometric features – size and shape properties of image objects: cell nucleus or cytoplasm area, boundary properties such as circularity etc.;
- densitometric features – optical density of image components: cell DNA quantification;
- multi-spectral features – colour and frequency-domain properties: cell dye hue and saturation etc.;
- texture features – spatial relationships between image intensities: cell chromatin or cytoplasm texture.

While all of the above feature examples relate to cell properties, we can equally apply these measurement techniques to other areas of image analysis such as remote sensing, industrial, and other medical applications.

The cell features we have listed above are just a small subset of the many hundreds that have been evaluated in the literature over the last 30 years. Even in the early seventies, Prewitt (1972) listed about 100 features which were trialed in the CYTOSCAN system. Unfortunately, features which prove to be of worth in a particular cell analysis application may not work well in similar applications due to the inherent differences in the equipment and methodology used. These differences include the quality of the microscope, the resolution (magnification) of image capture, focus algorithms, the cell staining process, pre-processing methods etc. It is also generally accepted that a consistent staining process and using a *stoichiometric* stain[2] is vital to the success of any

[2] a dye whose staining density is proportional to DNA density.

automated cell analysis system. In fact, almost all cell descriptors in the published literature have been trialed on cells stained using stoichiometric stains. This represents a major hurdle to the commercial acceptance of automated systems, because the currently accepted stain (the Papanicolaou stain—commonly referred to as the Pap stain), is not stoichiometric. It is also unlikely that stoichiometric stains will be accepted by cytology labs in the near future because:

- The Pap stain is so widely accepted; and
- Positive slides will still need to be manually screened after automated screening. The Pap stain provides significant visual enhancement of cell properties that manual screening requires.

For these reasons we have concentrated on searching for features which prove to be robust descriptors of cell properties, when using Papanicolaou stained slides.

Size and density descriptors have been widely reported as being powerful discriminators of normal and abnormal cells. This is not surprising because it is well known that cell neoplasia results in morphometric and densitometric changes, including increased nuclear DNA content, enlarged nuclei, and reduced cytoplasm area (Bibbo, Bartels, Dutch & Wied 1984, Komitowski & Zinser 1985, Tanaka, Ikeda, Ueno, Mukawa, Watanabe, Okamoto, Hosoi & Tsunekawa 1987, Tucker 1979). The majority of these descriptors, including nuclear area, nuclear/cytoplasm ratio, mean nuclear density, variance, skewness, kurtosis etc., are easily and rapidly measured by simple pixel counting or extraction from image histograms.

Multi-spectral features for the analysis of cells have not gained wide support in the literature. This is particularly true for colour-based feature extraction, and may relate to the fact that cell colour is an artificial property—a result of the staining process. However, Harms, Gunzer, Baumann & Serbouti (1993) used a combination of colour and texture-based features to subtype MACs-affected monocyte and lymphocyte cells from histologically normal tissue. This in turn allowed automatically subtyping acute leukaemic conditions without the need for human screeners to search for rare 'blast' cells. Garcia (1986) and Poulsen (1973) found that features based on the average intensities of red and green filtered images provided a measure of discrimination between normal and abnormal cell nuclei. Nguyen, Poulsen & Louis (1983) and Noguchi (1985) have also successfully used colour-based descriptors for automated cytology. Fourier-based methods[3] such as the Power Spectral Method (PSM) (Lendaris & Stanley 1970)[4] have

[3] Fourier or spatial-frequency-based methods are sometimes considered to be texture methods

[4] also known as the Fourier Power Spectrum (FPS) method.

been applied to other medical applications (Kruger, Thompson & Turner 1974, Hall, Kruger & Turner 1974), but are generally considered to be inferior techniques (Conners & Harlow 1980, Wu, Chen & Hsieh 1992). These will be discussed further in Chapter 2.

Arguably the most important and widely-used image descriptors for automated cytology are textural features. The nuclei and cytoplasm of cervical cells are known to be rich in texture information which cytologists regularly use for specimen screening purposes. Texture analysis is such an important technique that we have chosen to discuss it in-depth in Section 1.5.

1.4.4 Feature Pre-processing

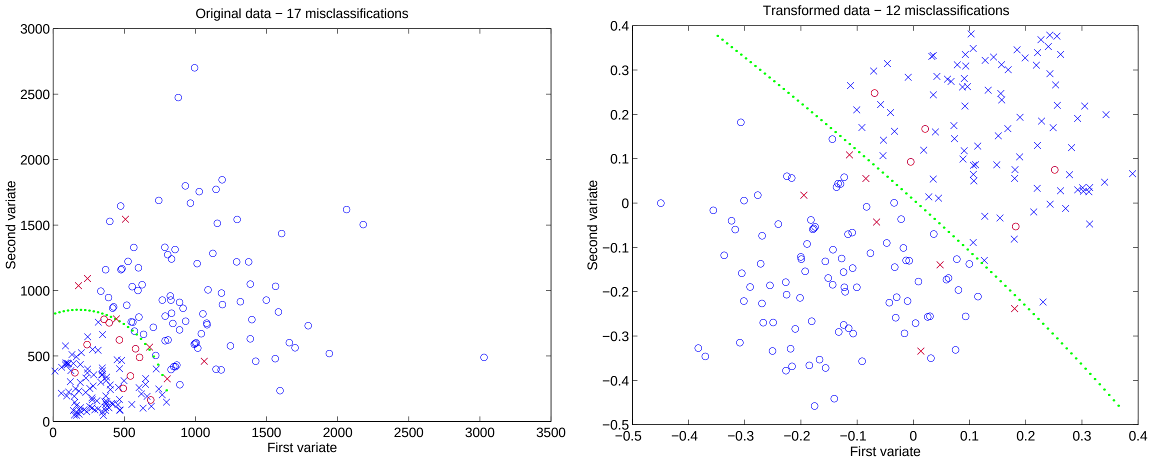
Feature pre-processing is the technique of transforming feature data so that it can be better utilised by subsequent processes. Neural network classifiers, for example, often require unbounded input feature data $x \in \mathbb{R}$ to be within numeric limits $x \in \mathbb{R}' \subset \mathbb{R}$ (i.e., between -1 and 1), or to be integer-valued, $x \in \mathbb{Z}$. Other classifiers may require binary-valued data, $x \in \{0, 1\}$. Other methods of pre-processing modify distribution statistics of the feature data, such as normalising the mean or variance of all features. While the exact method of pre-processing is not important, it is generally necessary to ensure that order-relationships between data are maintained, i.e.,

$$x_1 < x_2 \quad \Rightarrow \quad x'_1 < x'_2, \tag{1.3}$$

where x' is the numerical value of x after pre-processing.

As we will explain later, our feature selection and classification model relied on feature data whose distributions were, ideally, of a Gaussian form. These *parametric* models are based on normal-theory where assumptions are made about the distribution of the class-conditioned feature distributions. For methods which assume normality, or benefit from normality, it is often advantageous to transform feature data using normality transforms. While it is known that minor departures from normality are generally not detrimental to classification processes derived from normal-theory-based decision rules (McLachlan 1992), we could not guarantee that our feature distributions used throughout this thesis did not deviate significantly from Gaussian. To make our features more Gaussian, we applied preprocessing via a normality transform before discriminant analysis and classification. We transformed all features using a technique called the *Ladder of Powers* (Velleman & Hoaglin 1981). We will discuss this technique in detail in Section 3.4.4.

To demonstrate the benefits of using near-Gaussian data for normality-based classifiers, we show the classification results of highly non-Gaussian and Gaussian feature data in Figure 1.16. Classification errors for highly non-Gaussian feature data (a), were reduced from 17 to 12 after the data was made more Gaussian (b). We attained this decrease in error because the modelled PDFs of the feature data now closely matched those of the actual PDFs.



(a) (b)

Figure 1.16: Scatter plots for feature data, before and after normality transformation via equation (3.25) in Chapter 3: (a) A total of 17 misclassification errors resulted when using non-Gaussian data; (b) After normality transformation of the data, misclassifications were reduced to 12.

1.4.5 Feature Selection and Discrimination Measures

Given the pattern recognition system's requirement of classifying an object into one of several classes, it is necessary for us to capture information about the object, by extracting descriptive features. Generally, there are few limits on the number of feature measures which can be extracted from an object. However, it is common for many of these features to contain little or no *discriminatory* information. By the term 'discriminatory' information, we mean information which is class-dependent. That is, information which is similar for objects of the same class membership, but which differs for objects of alternate classes. Discriminatory information is essential to the success of classification processes. It is of vital importance in order to accurately define the different class distributions for each feature, and to minimise the overlap or common areas between these distributions. By removing non-discriminatory features from the feature set, we can achieve two very important advantages. Firstly, the resulting subset of features can be be processed in less time, simply because there is less data to process. Also, data storage requirements are similarly reduced. But more importantly, the reduction in feature set size can increase the estimation accuracy of the class-conditioned feature distributions. This is related to the 'curse of dimensionality' (Bellman 1961), and the 'peaking phenomenon', and will be discussed further on page 85 of Chapter 4.

The objective of feature selection is to select a minimal set of features which provide

- high discrimination between object classes, and
- accurate estimation of the real class-conditioned distributions from which the features were drawn.

The first problem we face when choosing an optimal subset of features is the *number* of possible subsets which we should (ideally) investigate. For example, for an N_V-dimensional feature set, there are $2^{N_V} - 1$ possible feature subsets, of cardinality $1, \ldots, N_V$. Assuming an initial population of 20 features, we have over 1 million subsets from which to choose!! Even if we know *a priori* a suitable subset cardinality, say v features, there are still $N_V!/((N_V - v)!\ v!)$ feature subsets of cardinality v. We can see that, even for small values of N_V, the problem of determining the worth of each feature subset becomes considerable. The method of determining the worth of feature sets presents us with a second problem. While it is best to optimise the worth of a feature subset based on minimising classification error rate, the computational cost of explicitly calculating such error rates is prohibitive. Fortunately, a number of attractive statistical methods are available which can address both of the aforementioned problems.

Discrimination measures

As an alternative to the direct determination of error rate, we can use more computationally tractable methods, called *class-separability measures*, to estimate it. These methods include error probability measures, probabilistic dependence measures, and interclass distance measures, among others (Hand 1981). While the exact form of measurement for each method differs, they all attempt to assign a figure-of-merit to a feature, based on how 'dissimilar' the class-conditioned distributions of the feature are. One measure of dissimilarity is the amount of overlap between the class-conditioned distributions, with less overlap meaning more dissimilarity. Other measures quantify the separation between the distributions, normalised by the spread or variance of each distribution.

Probabilistic distance measures are some of the most commonly applied methods of discriminatory power measurement, and attempt to quantify the *probabilistic distance* between two density functions. The greater this distance, the less overlap between the densities, and the smaller the probability of classification error. Of the many probabilistic distances measures defined in the literature (Chernoff, Divergence, Mahalanobis, etc), one of the most widely used is the *Bhattacharyya* metric (Bhattacharyya 1943), which takes the non-parametric form,

$$J_B = -\log \int [p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)]^{\frac{1}{2}} d\mathbf{x}. \tag{1.4}$$

where random variable $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_{N_v}]$ represents a set of candidate feature vectors and $p(\mathbf{x}|\omega_c)$ is the class-conditioned probability density of $\mathbf{x}$ for class ω_c. In simple terms, this is a measure of the amount of overlap between two class-conditioned PDFs. The square root of the product of the overlapping areas is integrated over the domain of the two PDFs. The larger the area of overlap, the smaller the discrimination measure.

When the class-conditional distributions of the features are known and of Gaussian form, we can express equation (1.4) in closed form as,

$$J_B = \frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}[\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\log\left[\frac{|\frac{1}{2}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)|}{\sqrt{|\mathbf{\Sigma}_1||\mathbf{\Sigma}_2|}}\right]. \tag{1.5}$$

where $\boldsymbol{\mu}_c$ and $\mathbf{\Sigma}_c$ are the mean vector and covariance matrix of $\mathbf{x}$ for class c, and $|\Sigma|$ represents the determinant of Σ. We have used the Bhattacharyya discrimination measure extensively throughout the course of our research. It is a parametric method, and provides the most robust estimates when the feature distributions are of Gaussian form. With this in mind, we pre-processed our features prior to feature selection and classification, using the ladder of powers normality transform introduced in Section 3.4.4.

Feature-set search algorithms

Having reviewed suitable, computationally light methods of evaluating feature-set worth, we will now discuss the considerable problem of choosing an optimal subset of features from a pool of possible candidate features. We should point out that, for the special case where all features are normally distributed, independent, and have equal covariance matrices, we can select the optimal feature set of cardinality v by choosing v features with the highest univariate discriminatory power (Hand 1981). However, features meeting the above restrictive criteria are extremely uncommon. In general, we cannot select feature sets based on their univariate discriminatory power.

As we showed previously, an exhaustive search of all possible feature subsets becomes computationally prohibitive when the cardinality of the feature pool is much above 10.

An alternate search method is the *branch-and-bound* algorithm (Hand 1981, Hand 1982, Kittler 1986, Yu & Yuan 1993). The branch-and-bound method yields the globally optimal feature subset *without the need to explicitly evaluate all possible subsets.* However, even this method of evaluation is computationally prohibitive for large feature sets. Fortunately, several other methods exist which produce an acceptable, albeit sub-optimal solution, by searching only a subset of all possible feature subsets. Examples of these are:

- *sequential forward selection* (SFS), which, starting with a subset containing only the best individual feature, successively adds one feature at a time to this subset (Hand 1981, Kittler 1986). One weakness of this method is that, once selected for inclusion, a feature cannot be removed;

- *sequential backward elimination* (SBE), which, starting with a set containing all features, successively removes one feature at a time from this set (Hand 1981, Kittler 1986). A weakness of this method is that, once removed, a feature cannot be re-introduced to the subset;

- *Min-Max feature selection*, which augments the current subset of features using individual and second-order discriminatory power considerations (Kittler 1986).

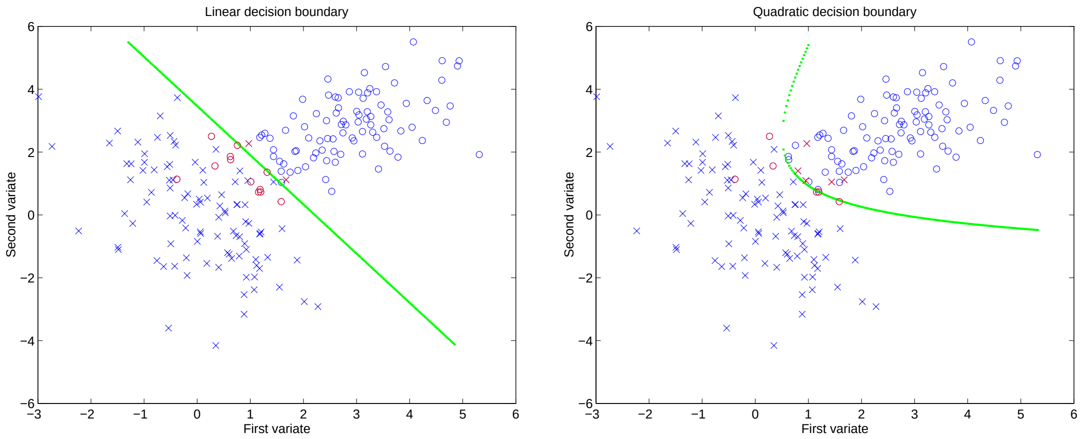
Arguably one of the best sub-optimal feature-set search methods is the *plus l–take away r* algorithm (Kittler 1978, Kittler 1986). This algorithm overcomes inherent weaknesses of other sub-optimal methods mentioned above (i.e., the 'nesting' problem) by alternately augmenting and depleting the current subset. Each augmentation adds l features to the current subset by applying sequential forward selection l times, while each depletion removes r features by applying sequential backward elimination r times. While being a sub-optimal method, Kittler (1978) reported that methods which forward-select and backward-eliminate several variables simultaneously were better than methods such as SFS and SBE which add or remove only one variable at a time. He also concluded that the combination of forward-selection and back-elimination almost always gave optimal results and computationally was comparable to less optimal approaches. For these reasons, we have chosen to use this method of feature selection throughout our work. More recent methods, such at *floating search* by Pudil, Novovicova & Kittler (1994) would be equally as applicable. Our method differs slightly from the above, in that we forward-select 2 new features using *joint* discriminatory power criteria, and backward-eliminates one feature. That is, rather than applying SFS twice, we add the *pair* of features which maximises the discriminatory power of the resulting set. This facilitates capturing feature subsets with exhibit higher-order discriminatory power. We begin with an exhaustive search for the best feature pair, and then augment this subset using the add-2/subtract 1 selection method. The discrimination measure used to determine feature set discriminatory power was the parametric Bhattacharyya distance measure defined in equation (1.5).

1.4.6 Classification

After selecting a suitable set of features whose class-conditioned statistics differ between the different classes of objects, we now need to design a suitable classifier which can be used to classify new, previously unseen, objects. We also need to determine the generalisation performance of this classifier by estimating the probability of classification error, i.e., the probability that an unknown object is allocated to the wrong object class. An example of such a *misclassification* would be the allocation of a cell containing pre-cancerous abnormalities to the normal class. The ramifications of such a dangerous error are obvious, and highlight the importance of minimising such errors through appropriate classifier design. We previously mentioned our use of normal-based feature pre-processing and classification throughout this thesis. Our preference for parametric methods is based on their wide-spread acceptance and use throughout the published literature. Moreover, even when the underlying assumptions of these methods are not strictly correct, we generally see minimal detrimental effect on their performance (James 1985).

Discriminant analysis and discriminant functions

The main goal of discriminant analysis is to derive a mathematical rule, known as a *discriminant function*, which can be used to separate the different groups of objects, e.g., the groups of normal and abnormal cells. The discriminant function is governed by statistical parameters drawn from populations of object features from known classes. It takes as an input, a vector of feature values extracted from a yet-to-be-classified object. Its output is usually a scalar value which can be used to determine the likely class membership of the object. In effect, a discriminant function defines an n-dimensional decision surface which separates the class-conditioned distributions of features in this n-dimensional feature-space. The two most common types of discriminant functions are *Linear Discriminant* and *Quadratic Discriminant* functions. Both of these are *Bayes* classifiers and assume that the class-conditioned feature data are multi-variate normally distributed. However, linear discriminant functions impose a restriction of equal class-conditioned covariance matrices, while this restriction is relaxed for quadratic discriminant functions. In Figure 1.17, we show typical examples of a linear and a quadratic decision boundary, calculated from simulated bivariate Gaussian-distributed feature data with unequal variance-covariance matrices.



(a) (b)

Figure 1.17: Two examples of decision boundaries in bivariate feature-space. (a) Linear discriminant boundary; (b) Quadratic discriminant boundary. We can see that the quadratic discriminant better models the true line of equal class-conditioned probability.

A quadratic decision function for assumed multivariate-normal feature distributions with unequal variance-covariance matrices can be expressed as (Gonzalez & Woods 1993):

$$d_c(\mathbf{x}) = \log P_{\omega_c} - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_c| - \tfrac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_\mathbf{c})\,\boldsymbol{\Sigma}_\mathbf{c}^{-\mathbf{1}}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{c})^\mathbf{T}\right], \quad c=\{1,\ldots,N_c\}, \tag{1.6}$$

where d_c is the discriminant measure for class c, P_{ω_c} is the class-conditioned prior probability, $\boldsymbol{\Sigma}_c$ is the covariance matrix, $\boldsymbol{\mu}_c = [\mu_1,\ldots,\mu_{N_v}]$ is the multivariate mean feature vector, and $\mathbf{x} = [x_1,\ldots,x_{N_v}]$ is the pattern vector to be classified. The resulting decision boundary $d_1 = d_2$ is of hyperquadric form, because no terms in $\mathbf{x}$ higher than the second power appear in the equation. By constraining the class-conditioned covariance matrices to be equal, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}, \quad c = 1,\ldots,N_c$, equation (1.6) reduces to

$$d_c(\mathbf{x}) = \log P_{\omega_c} - \tfrac{1}{2}\boldsymbol{\mu}_c\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c^\mathbf{T} + \mathbf{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c^\mathbf{T}. \tag{1.7}$$

The term $\frac{1}{2}\log|\boldsymbol{\Sigma}_c|$ in equation (1.6) is now class-independent and represents a constant offset for all classes. As such, it is not required in equation (1.7). The highest power of $\mathbf{x}$ is now 1, representing a hyperplane in $N-$dimensional feature-space, as suggested in Figure 1.17(a).

The quadratic discriminant function of equation (1.6) represents an optimal classifier for feature data which are truly multi-variate Gaussian. For real data, which is rarely multi-variate Gaussian, there remains the question as to what degree the classification accuracy is affected when the above assumptions are not met. According to the literature (Lachenbruch 1975, Seber 1984), quadratic discriminant functions are more sensitive to departures from multi-variate Gaussian in high-dimensional feature spaces, than linear discriminant functions. However, it is known that *minor* departures from normality are generally not detrimental to classification processes (McLachlan 1992, Hjort 1986). For non-normal feature data, normality transforms can be helpful in obtaining near-normality of feature's class-conditional densities. Moreover, to quote McLachlan (1992),

> *Even if a transform does not induce ... near normality, it will have played a useful role if it has been able to produce good symmetry in the data.*

Of further consideration is the fact that linear discriminant functions are sensitive to inequalities in class-conditioned covariance matrices (Lachenbruch 1975, Seber 1984). Quadratic discriminant functions exploit such inequalities and allow more optimal definition of decision surfaces, leading to the possibility of lower classification error.

Based on the above considerations, we chose the quadratic discriminant function of equation (1.6) as our classifier throughout this thesis. While there is a marginal increase in computational expense for quadratic functions and associated normality transforms, we feel the benefits of lower error probability are of much greater importance.

1.4.7 Evaluation of classifier performance

In the previous sections we overviewed the necessary sub-blocks of a typical pattern recognition system, and introduced further detail and justification for our selection of specific sub-block algorithms. In this section we will consider methodologies for assessing the classification *performance* of our system.

Measuring classifier performance generally entails quantifying how well the system classifies objects, and is usually expressed in terms of *misclassification rate* or *error rate*, i.e., the rate at which an object is allocated to the wrong object class. Classifier performance is normally quantified in terms of *apparent* error, or estimates of the *real* or *true* error. Both measures have an associated variance which indicates how close the estimates are to the true classification performance. Measuring apparent error usually results in an optimistically biased estimate, because the classified data used is the same as that which was used to design the system. This *training set* data is only a subset of the real population, and cannot possibly represent the total variability of the real population. Thus, statistical parameters extracted from training set data are only estimates, and contain inherent 'noise' or estimate errors. Designing a system based on minimising training set classification error runs the risk of *over-training*. The classifier learns characteristics which may only be present in the training set, and in so doing, loses its generalisation ability on unseen data. We usually prefer to estimate real error, because it provides an indication of how well the system will perform in practice, i.e., its *generalisation performance* using real, previously unseen data.

Estimating the real classification performance can be achieved by a number of evaluation strategies:

- Holdout (Weiss & Kulikowski 1991).
- Cross-validation (Efron 1982).
- Leave-one-out and jackknife methods (Quenouille 1949).

These and other methods have been extensively reviewed in the literature (Bradley 1996, Hand 1981, Kittler 1978, Lachenbruch 1975, Seber 1984, James 1985, Weiss & Kulikowski 1991). We will now give a short overview of these methods, as detailed in Bradley (1996) and Weiss & Kulikowski (1991).

Holdout

Holdout is particularly suitable for large datasets, e.g., for dataset size $N_s > 2000$, where the computational burden of other methods becomes prohibitive. It involves partitioning a dataset containing measures from N_s objects, into a training set of n measures and a test set of $N_s - n$ measures. We train the classifier system using the training set, and then estimate the real error rate by classifying the test set. Common training:test partitions are 67%:33% or 50%:50%. According to Weiss & Kulikowski (1991), the real error estimate obtained using this method is pessimistically biased. However, the estimate will converge to the real error for large dataset sizes, e.g., when there are over 1000 examples in the test set.

Cross-Validation

The n-fold cross-validation method (Efron 1982, Weiss & Kulikowski 1991) is particularly suitable for dataset sizes of $100 < N_s < 2000$. We randomise the dataset and partition it into n approximately equal test sets, resulting in each test set having approximately N_s/n members. For each test set, we form a training set consisting of the other $n-1$ test sets. We then use this training set to train the classifier system. The performance of the resulting classifier is then evaluated on the test set. We repeat this process n times, once for each of the n test sets. As a result, each and every example in the dataset is used only once to test the classifier. Also, an example is never tested on a system that it helped train. We obtain estimates of the real error rate by averaging the error rates of all n trials. Also, we can obtain an estimate of the error variance from these n trials. According to Breiman, Friedman, Olshen & Stone (1984), 10-fold cross-validation gave a good trade-off between error estimation and computational complexity. It provides an almost unbiased estimate of the error rate, but with high variance when used on small datasets. Fortunately, error estimate variance can be reduced by repeating the cross-validation process several times, each time using a different randomised partitioning (i.e., re-sampling).

Leave-One-Out

Leave-one-out strategies, including jackknife, are usually applied to datasets containing less than 100 members. These methods are reviewed in-depth in other published literature (Weiss & Kulikowski 1991), so we will limit our explanation to the basic principles of leave-one-out. Given a dataset of N_s samples, we train the classifier using N_s-1 samples, and test the system using the remaining sample. We repeat this process a total of N_s times until all examples have been tested once. Note that this is an extreme case of n-fold cross-validation, where $n = N_s$ (the number of samples in the dataset). Once again, the error estimate is almost unbiased but it may contain high variance for small datasets.

In Table 1.1 we show how the computational complexity increases from the holdout method to the leave-one-out method. We also detail how the training and test set partition sizes vary for each method.

	Holdout	n-fold CV	Leave-one-out
Training set	n	$N_s - \frac{N_s}{n}$	$N_s - 1$
Test set	$N_s - n$	$\frac{N_s}{n}$	1
Iterations	1	n	N_s

Table 1.1: Comparison of dataset partition sizes and number of classification trials for holdout, cross-validation, and leave-one-out estimators of real error rate.

1.5 Texture Analysis in Automated Cytology

Texture is a characteristic that is present in almost all images, and is considered to be one of the most important properties used in the identification or classification of image objects or regions. While the meaning of the term *texture* is difficult to concisely define, it has been described in several qualitative ways. Haralick, Shanmugam & Dinstein (1973) and Davis, Johns & Aggarwal (1979) described texture as the coarseness, homogeneity, and orientation of image structure. Julesz (1962) suggested that texture characterises the spatial relationships between image intensities or tones. Pressman (1976) proposed that texture is based on the variation of grey levels in a neighbourhood of a pixel, where the size of the neighbourhood depends on the size of the fundamental textural element, known as a *texton* (Julesz 1981, Julesz & Bergen 1983) or *texel*. According to the Collins English Dictionary, texture is *the structure, appearance, and feel of a woven fabric; the surface of a material, especially as perceived by the sense of touch; the general structure and disposition of the constituent parts of something...* (Makins 1992). Texture analysis is the *quantification* and use of such image texture properties. It is the basis of many image processing operations such as classification, segmentation, and synthesis of textured images.

Texture analysis methods can be loosely grouped into two classes—structural or statistical. Structural approaches generally model a texture as the deterministic or stochastic placement of texture primitives (textons), with the emphasis on texton characterisation such as size and shape—both local properties. This approach can fail where texton primitives are not readily identifiable, which is generally so. Statistical approaches focus on the global spatial relationships between intensity variations, and often fail to capture local properties of the texture. Texture characterisation requires both local (texton primitive) and global (spatial organisation) description. Although neither structural nor statistical methods satisfy this requirement fully, both classes of method, particularly statistical methods, have been widely accepted over the past two decades.

The statistical methods can be grouped into five general categories:

- gradient-based methods which define texture properties in terms of absolute differences in grey levels between neighbouring image pixels;
- frequency-based methods which model texture in terms of power spectra or autocorrelation criteria;
- mathematical morphology and granulometry-based methods which characterise

texture by the relationship between an image's connected components following morphological filtering operations and/or iterative thresholding;

- co-occurrence-based methods which model the statistical relationships between pairs of spatially separated image pixels.
- Markovian methods which which model higher-order statistical relationships between all elements in a defined neighbourhood.

Over the last 15 years, the most prolific and promising works in the area of automated cytology have been in the area of texture analysis of the nucleus. This is not surprising. As we mentioned earlier, pre-cancerous abnormalities are manifested in visual and subvisual changes in cell characteristics, including changes in chromatin content of the nucleus. In fact, it is generally believed that the initial signs of cell neoplasia first appear in the nucleus. Because nuclear chromatin and its spatial arrangement can be viewed as a type of texture, the use of texture analysis for detecting pre-malignant abnormalities in cells has seen widespread application.

Very little published research exists for the use of gradient-based texture analysis for automated cytology, possible due to the method's inherent susceptibility to noise in the image. One published work known to us was by Abmayr, Burger & Soost (1979), who used descriptors extracted from Laplacian filtered images to classify Papanicolaou stained cervical cells into four classes—basal, metaplastic, dyskaryotic, and carcinoma in situ. More specifically, they computed histograms of the filtered images, and extracted statistical descriptors such as mean and variance to discriminate between normal and abnormal cells. According to Abmayr et al. (1979), classification accuracy of their "TUDAB" system improved when these features were added to their existing feature database. Using cross-validation, they achieved a correct classification rate of 69.6%, which improved to 88.6% for a two-class case.

Kopp, Lisa, Mendelsohn, Pernick, Stone & Wohlers (1976) used features derived from optical Fourier transforms to detect cervical cell abnormalities. They attained false positive and false negative rates of 10-15% using a small database of 339 cells. Fourier descriptors were also used by Wang & Abmayr (1982) to classify mouse L fibroblasts into three classes – G1, S, and G2. Using a total of 28 Fourier features, they attained a correct classification rate of over 84%.

We have seen a marked increase in morphology-based image processing over the last 15 years, due to the growing popularity of this area among researchers. Salembier, Gasull, Marques & Sayrol (1992) applied image morphology to the automated detection of spermatozoa. They used a sequence of morphological filters to remove image artifacts and enhance image contrast. This facilitated the successful application of simple and rapid techniques such as thresholding for detecting individual spermatozoa. Young, Verbeek & Mayall (1986) used granulometry-based features to characterise the amount and arrangement of chromatin near the nuclear-cytoplasm interface. The four defined features quantified heterochromatin homogeneity, condensation, margination, and clump size distribution. In Walker & Jackway (1996) we used a related technique to extract features from images of regularly stained cervical cells. Our features were statistical descriptors from distributions of nuclear heterochromatin and euchromatin clump morphology (area, circularity, etc.), extracted from sequentially thresholded grey-scale images. Tanaka et al. (1987) thresholded grey-scale images of cervical cells to produce binary images containing connected components (heterochromatin clumps). The area of the sum of the individual components was then used as a feature in their CYBEST automated cytology system. One of the most prominent researchers in the area of cell analysis by image morphology is Fernand Meyer. He has published a large quantity of research discussing morphological approaches to all facets of automated cytology, including cell finding, segmentation, pre-processing and feature extraction. Such papers include Meyer (1980*a*), Meyer (1980*b*), Meyer & van Driel (1980), Meyer (1981), Meyer (1982), and Meyer & Beucher (1990).

Of all the texture methods, it is perhaps the Markovian methods which have received the widest application to cell texture analysis and have demonstrated the most promise. Markovian methods are those which model a texture, or extract texture descriptors, based on the interdependence among image pixels in localised spatial neighbourhoods. This interdependence is commonly expressed as transition probabilities (the conditional probability of occurrence of grey level g_1, given that grey level g_2 has occurred), or as co-occurrence probabilities (the joint probability of occurrence of pixel intensities). Examples of Markovian methods include Markov chains, Markov Random Fields (MRFs—multi-dimensional generalisations of Markov chains) and the closely related Gibbs Random Fields (GRFs), and co-occurrence-based methods.

MRF and GRF theory has been extensively reviewed in the literature, and we refer the reader to works by Besag (1974), Besag (1986), Chen (1988), Cross & Jain (1983), and Dubes & Jain (1989). Of particular interest is new research by Elfadel & Picard (1995), whose *aura measure* provides a generalised model framework which links GRF/MRF, grey-level co-occurrence, and the correlation matrix. The success of random field models generally relies on large image domains for accurate modelling—something which is commonly not available for cell images, due to the limited spatial resolution of existing imaging devices, and the small physical size of cell nuclei. For this reason, the number of published works using MRF techniques for cytological analysis are limited.

We have, however, trialed a combination of co-occurrence and MRF features for classifying regularly stained cervical cells (Walker, Jackway & Lovell 1995). We found that MRF features alone were poor discriminators of normal and abnormal cells. However, they appeared to possess higher-order discriminatory power when combined with co-occurrence descriptors. This may be because the MRF captured third and higher-order statistics, which complemented the second-order statistics captured by the co-occurrence method. Cross-validated classification of a small database of cells attained a correct classification rate of 88% using three MRF and two co-occurrence features. We believe this study is the first to use MRF features for classifying regularly stained cervical cells.

In this thesis we identify adaptive methods of analysis based on an assumed MRF model for the texture. For MRF and its related GRF, a complete statistical model within a defined window or neighbourhood is sufficient to represent the texture. This has been recently demonstrated by Paget & Longstaff (1996), who synthesised highly realistic textures, visually identical to textures from which the model estimates were derived. They achieved this by using an MRF model to capture the third-order local conditional PDF (LCPDF) of grey levels. This suggests that, for texture *synthesis*, a third-order LCPDF model is sufficient to characterise a large proportion of textures.

Paget, Longstaff & Lovell (1997) also showed that a second-order MRF model was sufficient for accurately *classifying* textures. Their model estimated the LCPDF of a pixel's grey level I_s, based on its eight nearest neighbours $\mathcal{N}_s = \{r \mid ||(s-r)||^2 \leq 2\}$, where s and r are valid image co-ordinate pairs. They determined the LCPDF $P(I_s \mid \mathcal{N}_s)$ directly from estimates of the local joint PDF $P(I_s, \mathcal{N}_s)$. As an example, the LCPDF $P(I_s \mid I_{s+(0,1)})$ under the neighbourhood constraint $\mathcal{N}_s = s + (0,1)$ shown in Figure 1.18(a) is determined from the joint PDF $P(I_s, I_{s+(0,1)})$ shown in Figure 1.18(b) by the equality

$$P(I_s \mid I_{s+(0,1)}) = \frac{P(I_s, I_{s+(0,1)})}{\sum_{i \in N_g} P(i, I_{s+(0,1)})}. \tag{1.8}$$

It is interesting to note that the numerator in equation (1.8) is simply the grey-level co-occurrence estimate for the texture. This establishes a link between this subset of MRF models and the GLCM method of Haralick et al. (1973). Because of the simpler GLCM neighbourhood definition (pair-wise joint-probability modelling), the co-occurrence method is an incomplete characterisation of texture. However, in many cases this is sufficient for texture classification. Moreover, co-occurrence methods have considerable advantages over MRF modelling, including computational efficiency, and better model estimate accuracy. This is especially so for images of small spatial domain, such as cell images, where it is difficult to accurately estimate higher-order spatial relationships necessary for most MRF/GRF models.

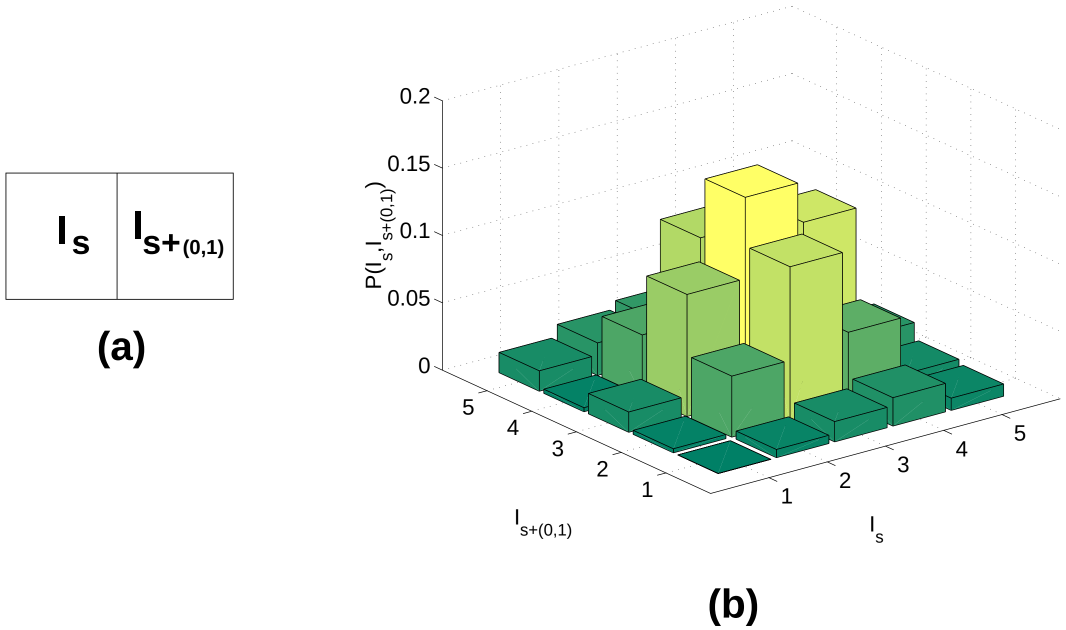


Figure 1.18: A neighbourhood (a), and its discrete local joint PDF (b).

Co-occurrence-based methods such as GLCM are second-order statistical methods which model the joint probability of occurrence of pairs of image properties. Such image properties include pixel intensities, intensity variance and entropy, and gradient descriptors. Co-occurrence-based methods are fundamental to the work contained in this thesis, and for this reason we provide an in-depth discussion of these methods in Chapter 2.

1.6 Thesis Background and Outline

It may be helpful to the reader for us to explain the motivations for the work we will present in this thesis. This work has been conducted within the Cytometrics Group—a multi-disciplinary team of CSSIP researchers with the task of implementing an automated analysis system for screening cervical smears for pre-cancerous abnormalities. The motivations for this research are many-fold, but are mainly based on the huge benefits such a system can bring to society. These include significant cost savings to the community, allowing funding of other important public health initiatives, and a reduction in mortality due to earlier detection and better prognostic leads.

The Cytometrics Group focuses on applying so-called 'smart algorithms' to detecting neoplasia and MACs in regularly stained cervical smears. Such algorithms are not only of benefit to new analysis systems, but can also enhance the performance of existing screeners. The group's recent change in focus to MACs detection is an important one of great medical and social significance. This is because MACs facilitates the possible application of our research to detecting other forms of cancer that exhibit the 'field effect' described in Section 1.3. In the future, it may be possible to detect pre-cancerous abnormalities in areas of the body, simply by the MACs analysis of a sample of saliva, or a scrape of cell tissue from the mouth.

Applying our algorithms to Pap-stained cervical smears is a relatively new area of research. It was generally considered that the variability of staining density produced by the Pap stain would mask any existing, often minute, differences in chromatin arrangement in pre-malignant cell nuclei (Zahniser et al. 1979, Wittekind, Hilgarth, Kretschmer, Seiffert & Zipfel 1983). For this reason, the vast majority of published literature on cell texture analysis used Feulgen and other stoichiometric staining chemicals to facilitate quantitative measurement of chromatin properties. By doing so, the direct application of the results of such research to cervical cancer screening programmes which use the widely-accepted Papanicolaou staining process is severely limited. We have therefore investigated the application of image cytometry to the problem of cervical cancer detection in Pap stained smears. Our initial investigation (Walker, Jackway, Lovell & Longstaff 1994) into the usefulness of the Pap stain produced encouraging results, which we reproduce in Figure 1.19. We can see from this figure that both normal and abnormal cell classes are, to a large degree, well clustered and separated.

One of the major impediments to our research effort has been the lack of an available database of classified cells for algorithm training. Compiling a database of cells at suitable spatial and photometric resolution is an arduous task involving hundreds of hours of work by researchers and trained cytologists. The resulting database is of such commercial value that its free availability to other researchers becomes unlikely. We have therefore compiled our own database of cytologically normal and abnormal cells from regularly stained cervical smears over the three and a half year course of our research. The selection of cell specimens was made by a trained cytologist from the Royal Womens Hospital, Brisbane. Towards the end of our research, a second, much larger database of normal and diagnostically abnormal cells was made available by Oncometrics Imaging Corporation[5]. These databases were unsuitable for investigating MACs, and therefore the work contained in this thesis is based on detecting diagnostically abnormal cells (so-called *rare event detection*). However, as we will show in subsequent chapters, the

[5] Oncometrics Imaging Corporation, Vancouver, B.C., Canada.

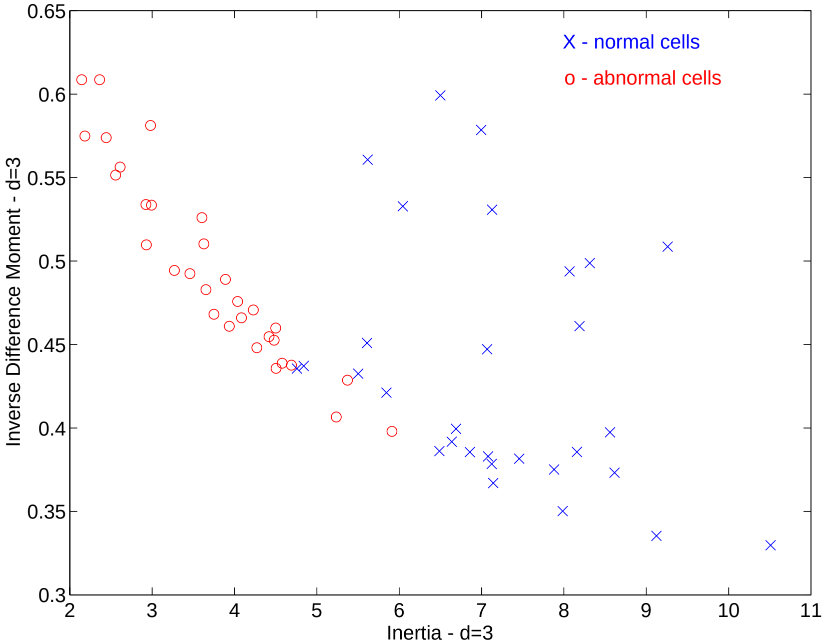


Figure 1.19: Scatter plot of two GLCM texture descriptors, extracted from images of Pap stained cervical cells.

techniques developed here are applicable to *any* class of texture problem, including MACs detection.

The focus of the research contained in this thesis is on *self-adaptive* texture analysis techniques. By self-adaptive techniques, we mean algorithms which do not rely on the explicit definition of fixed, problem-independent feature functions, but which use higher-level knowledge to form self-adaptive feature functions. These functions effectively 'tune' themselves to the specific characteristics of the texture classes being analysed. As we will show in the following chapters, such techniques provide significant benefits to the problem of image analysis including increased classification accuracy and lower computational burden. In short, self-adaptive techniques have the potential to extract more 'useful' image information using less features. They have the unique qualities of being an analysis method with general applicability to any texture classification problem (globally applicable), yet provide performance which only methods specifically 'tuned' or tailored to a particular texture problem can attain (locally optimised).

In the following chapters we introduce several new texture analysis techniques which extract texture descriptors with enhanced discriminatory power, by either manual or self-adaptation to specific texture characteristics. In Chapter 2, we begin with a comprehensive review of co-occurrence-based texture methods. As our review will show, they are generally considered to be the most powerful statistical techniques for analysing texture. Co-occurrence-based methods have received wide support from the research community since their introduction by Haralick et al. in 1973. We discuss the motivations for these various methods and highlight salient features and limitations as identified by the originating authors and others.

We introduce our first example of 'adaptive' texture analysis in Chapter 3. Here, we investigate a recently published method of texture analysis called Statistical Geometric Features (SGF). We review the method, discuss its salient features, and identify deficiencies in terms of its application to texture analysis. One such deficiency is the fact that the original authors defined only two texture descriptors, severely restricting the amount of information that can be extracted from texture images. We augment this pair of features with a set of features specifically defined for the task of cell nuclear chromatin texture analysis. We can say that these feature definitions have been 'tuned' or 'manually adapted' to the specific characteristics of cell nuclear chromatin. While this method is not self-adaptive, we introduce it to highlight the significant benefits that adaptive analysis can provide.

Chapter 4 discusses a method of improving the discriminatory power of Grey Level Co-occurrence Matrix (GLCM) features, using *discrimination matrices*. We investigate where co-occurrence matrix features derive their discriminatory power, and provide a theoretical basis for improving the GLCM method, using self-adaptive weighting functions which modify standard GLCM features. We then quantitatively evaluate the performance of our method by comparing the results of cross-validated classification of cervical cell texture images, using the unmodified and modified GLCM features. The proposed methodology provided up to a 70% decrease in classification error.

In Chapter 5 we continue the theoretical analysis of discriminatory power manifestation in co-occurrence matrices introduced in Chapter 4. We bypass the use of fixed feature functions as the basis for forming new features, because it is the *fixed* nature of these features which is the inherent weakness of all co-occurrence methods. We present a methodology for defining truly self-adaptive feature functions from co-occurrence matrices, which places no reliance on previously published feature definitions. Our proposed method, named Adaptive Multi-Scale GLCM (AMSGLCM), allows a natural extension of feature definition to include information across several spatial scales, thus capturing the benefits of multi-resolution analysis.

We present an alternate approach to self-adaptive multi-scale feature extraction in Chapter 6. While this technique is similar to that presented in Chapter 5, it does not rely on explicitly calculating discrimination matrices, which is essential to the successful operation of the AMSGLCM algorithm. Instead, we use a Genetic Algorithm (GA), and the *knowledge* that such discriminatory information exists, to form self-adaptive feature functions. Using a GA allows us to optimise feature definitions under specific criteria, such as minimising feature correlation while maximising feature discriminatory power. We include a complete review of GA principles and conclude with a critical appraisal of our method.

Finally, Chapter 7 provides a general summary of the main aspects of the work described throughout this thesis, and a discussion of our most significant achievements. We also offer some comments and suggestions for further research.

Chapter 2

Co-occurrence-Based Texture Algorithms

In Section 1.5 we introduced several qualitative definitions for the term *texture*, and reviewed many of the more popular texture analysis methods found in the literature. The subsequent chapters will deal intimately with the quantitative analysis of texture, due to its strong applicability to cell image analysis. Here, we review a number of second-order texture analysis methods, generally known as co-occurrence methods, for the analysis and classification of image texture. We discuss the motivations for these various methods, and highlight salient features and limitations as identified by the originating authors and others.

2.1 A Review of Co-occurrence-Based Texture Methods

As we identified in Section 1.5, the Markovian methods of texture analysis are considered by many to be the most powerful method for extracting texture information from images. Little theoretical evidence has been presented which attempts to explain why Markovian methods are intrinsically more powerful than, say, structural approaches or other statistical methods. However, if empirical evidence is considered as a valid measure of algorithm worth, then we need to look no further—the literature is ripe with quantitative comparisons of texture methods which conclude in favour of Markovian methods.

Among the Markovian methods, the most widely used are the co-occurrence-based methods which attempt to characterise second-order properties of an image. Co-occurrence is a measure of the relative frequency, or joint probability, of two image properties occurring, under predefined constraints, across the domain of the image. Image properties are pixel intensities, variance of intensities, gradient measures etc. These properties can be measured under constraints such as intersample spacing (both magnitude and orientation) and other higher-order neighbourhood definitions (windowing). For example, GLCM measures the probability of co-occurrence of image pixel intensities i and j, under the spatial constraint of d pixels separation between the pixels. Another method called the Gray Level Run Length Matrix (GLRLM) estimates the probability of image pixels with intensity i occurring in a co-linear sequence of length j. Prior to applying a texture operator such as GLCM or GLRLM, images are generally preprocessed photometrically using requantisation and/or histogram equalisation (Section 3.4.3). This normalisation ensures consistent co-occurrence domains independent of first-order properties such as an image's average, maximum, or minimum pixel intensity.

In the following sections we will review seven commonly-used co-occurrence-based texture operators: GLCM by Haralick et al. (1973), GLRLM by Galloway (1975), Grey

Level Entropy and Grey Level Variance Matrices (GLEM, GLVM) by Yogesan, Jorgensen, Albregtsen, Tveter & Danielsen (1996) and Yogesan, Albregtsen & Danielsen (1994), Statistical Feature Matrix (SFM) by Wu & Chen (1992), Neighbouring Grey Level Dependence Matrix (NGLDM) by Sun & Wee (1983), and Generalised Co-occurrence Matrices (GCM) by Davis et al. (1979)[1].

2.1.1 Grey Level Co-occurrence Matrix

The second-order statistical technique, GLCM, was first introduced by Haralick et al. (1973). They were among the first to characterise texture as an overall or average spatial relationship between grey tones in an image. The roots of this proposition can be found in earlier work by Julesz (1962) who conjectured that second-order probabilities were sufficient for human discrimination of texture.

The grey level co-occurrence matrix is determined as follows. We model a discrete grey-scale image on a domain $\boldsymbol{D} \subset \mathbb{Z}^2$ of N_g grey levels as a 2D function $I : \boldsymbol{D} \to \boldsymbol{G}$, where $\boldsymbol{G} = \{1, \ldots, N_g\}$. The GLCM $P(i, j|d, \theta)$ is an estimate of the second-order joint probability density function of grey-level pairs within the image. Each matrix element is an estimate of the probability that two image pixels, separated by the intersample displacement d, θ, have intensities i and j, where $i, j \in \boldsymbol{G}$:

$$P(i, j|d, \theta) = \frac{\# \{k, l \in \boldsymbol{D} \quad | \quad I(k) = i, I(l) = j, ||k - l|| = d, \angle(k - l) = \theta\}}{\# \{m, n \in \boldsymbol{D} \quad | \quad ||m - n|| = d, \angle(m - n) = \theta\}}, \tag{2.1}$$

where k, l, m, n are valid image pixel locations. Because of the discrete nature of digital image intensities, P is, in fact, a discrete density rather than a continuous one. Being a probability density, for any given d, θ,

$$\sum_{i,j} P(i, j|d, \theta) = 1, \tag{2.2}$$

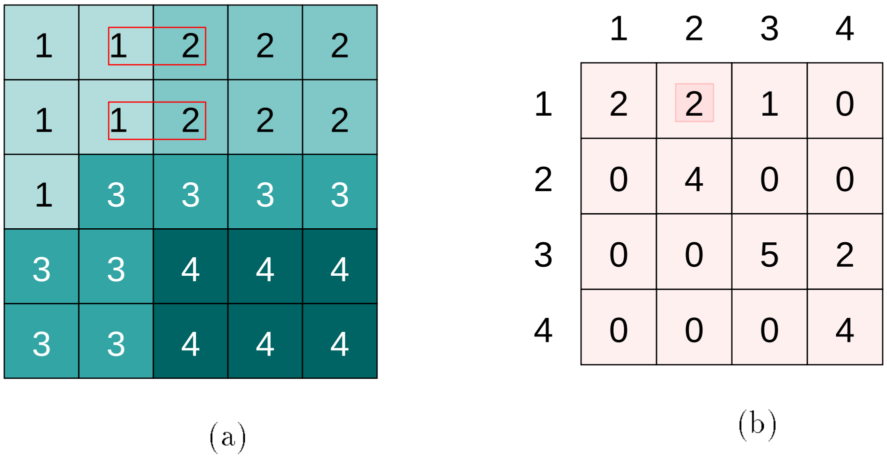
and

$$\forall i, j \in \boldsymbol{G}, \quad 0 \leq P(i, j|d, \theta) \leq 1. \tag{2.3}$$

As an example, Figure 2.1(a) shows a grey-scale image with a range of intensities from 1 to 4. The corresponding GLCM, calculated for pixel-pairs horizontally displaced by 1 pixel ($d = 1, \theta = 0^\circ$), is shown in Figure 2.1(b). For simplicity, we show the pixel-pair counts (the numerator of equation (2.1)), rather than probability estimates which would

[1] Another unrelated method called "Generalised Co-occurrence Matrix" was introduced by Hauta-Kasari, Parkkinen, Jaaskelainen & Lenz (1996).

follow matrix normalisation.



1	1	2	2	2
1	1	2	2	2
1	3	3	3	3
3	3	4	4	4
3	3	4	4	4

(a)

	1	2	3	4
1	2	2	1	0
2	0	4	0	0
3	0	0	5	2
4	0	0	0	4

(b)

Figure 2.1: An example of calculating a co-occurrence matrix from a grey scale image. (a) An image containing four intensities; (b) the GLCM for (a), under the intersample displacement constraint $d = 1, \theta = 0^o$. The co-occurrence matrix element $P(1,2)$ is determined by firstly scanning the image for all pixels that have an intensity of 1. Of these, we then count all which have a neighbouring pixel with intensity 2, displaced 1 pixel to its right ($d = 1, \theta = 0^o$). There are two such pixel pairs in this image, therefore, $P(1,2) = 2$.

Computational considerations necessitate limiting the number of image grey levels N_g, matrix intersample displacements d, and angular displacements θ. Images are generally requantised to 16 or 32 discrete levels using the techniques to be discussed in Section 3.4.3. The constraint θ is usually limited to four angles at 45^o intervals, with no distinction between opposite angles, i.e., $P(i,j|d,\theta) = P(i,j|d,\theta+\pi)$ and, therefore, $P(i,j|d,\theta) = P(j,i|d,\theta)$. Often, isometric matrices are formed by averaging these directional matrices. Displacement d is chosen according to the coarseness of the images to be analysed, and is generally varied from 1 to the maximum texture texton size in the image (Gool, Dewaele & Oosterlinck 1985).

Classical GLCM, as defined by Haralick et al. (1973), involves extracting scalar *secondary* features from the co-occurrence matrix. Haralick et al. defined 14 feature functions, and these were extended by Conners, Trivedi & Harlow (1984) and others. A listing of 19 such feature functions can be found in Pressman (1986). We also include a comprehensive listing of features in Appendix E. We detail 8 of the most commonly-used GLCM secondary features in Table 2.1, as defined in Haralick et al. (1973), Conners et al. (1984), and Trivedi, Harlow, Conners & Goh (1984).

Table 2.1: Commonly-used GLCM features.

Features	Equations
Energy:	$\sum_{i,j} P(i,j)^2$;
Entropy:	$-\sum_{i,j} P(i,j) \log P(i,j)$;
Homogeneity:	$\sum_{i,j} \frac{1}{1+(i-j)^2} P(i,j)$;
Inertia:	$\sum_{i,j} (i-j)^2 P(i,j)$;
Correlation:	$-\sum_{i,j} \frac{(i-\mu)(j-\mu)}{\sigma^2} P(i,j)$;
Shade:	$\sum_{i,j} (i+j-2\mu)^3 P(i,j)$;
Prominence:	$\sum_{i,j} (i+j-2\mu)^4 P(i,j)$;
Variance:	$\sum_{i,j} (i-\mu)^2 P(i,j)$
$\mu = \mu_x = \mu_y = \sum_i i \sum_j P(i,j) = \sum_j j \sum_i P(i,j)$;	
$\sigma = \sum_i (i-\mu_x)^2 \sum_j P(i,j) = \sum_j (j-\mu_y)^2 \sum_i P(i,j)$;	

GLCM features have been extensively used for texture classification in a diverse range of fields. Conners et al. (1984) used GLCM features for segmenting high-resolution satellite images of urban areas, while Weszka, Dyer & Rosenfeld (1976) classified terrain satellite images. In medical diagnostics, Pitts, Premkumar, Houston, Babaian & Troncoso (1993) used GLCM features to identify benign and malignant regions in prostate images; Yogesan, Albregtsen, Reith & Danielsen (1993) classified mice liver carcinoma; and we have applied GLCM features to the classification of normal and pre-malignant cervical cells (Walker et al. 1994, Walker, Jackway & Lovell 1995). Siew, Hodgson & Wood (1988) assessed carpet wear via the application of GLCM features.

Studies comparing the performance of GLCM features with other texture analysis techniques show that GLCM is one of the most powerful methods for general texture classification (Conners et al. 1984, Ohanian & Dubes 1992, Chen, Nixon & Thomas 1995, Augusteijn, Clemens & Shaw 1995, Gotlieb & Kreyszig 1990). Unser (1986) and Ojala, Pietikainen & Harwood (1996) concluded that while other methods, such as Law's Texture Energy Measures (TEM – Laws (1979)) and Markov Random Field models, may be superior to GLCM in some restricted cases (Stallings 1975, Kashyap, Chellappa & Khotanzad 1982), most fail when applied to other more general texture problems.

Conners & Harlow (1980) have shown, both theoretically and experimentally, that GLCM is a more powerful technique than Grey Level Difference Matrix (GLDM – Weszka et al. (1976)), Grey Level Run Length Method, and the Power Spectral Method (PSM – Lendaris & Stanley (1970))[2]. They showed that the group of Markov texture

[2] also known as the Fourier Power Spectrum (FPS) method.

pairs successfully discriminated by the other three methods were all subsets of the group of Markov texture pairs that could be discriminated by GLCM. However, they also presented examples of second-order textures which could not be discriminated by GLCM features, particularly texture pairs that were 180^o rotations of each other. The other texture operators were similarly limited. However, this is not strictly a limitation as such, because the algorithm used was *designed* to treat orientations of θ and $\theta + 180^o$ as equivalent, but can be designed to treat them as distinct.

Weszka, Rosenfeld, Carton, Kirby & Mohr (1975) had previously trialed the same texture operators on aerial images, with similar results. The only exception was that they found GLDM to be as powerful as GLCM. However, Conners & Harlow (1980) later showed this was due to the limited number of GLCM features used.

Ohanian & Dubes (1992) demonstrated that GLCM features perform better than fractal, MRF, and Gabor filter features in classifying a wide range of texture images, including fractal, Gaussian MRF, and natural images. Error rates were 9.3% (GLCM), 15% (fractal), 28% (Gabor) and 34% (MRF). They also showed that further decreases in classification error to 4.6% could be attained by using features from all four texture methods. This reveals GLCM's inability to capture all texture information, due to limitations in its feature definitions, and the fact that it only captures second-order texture information.

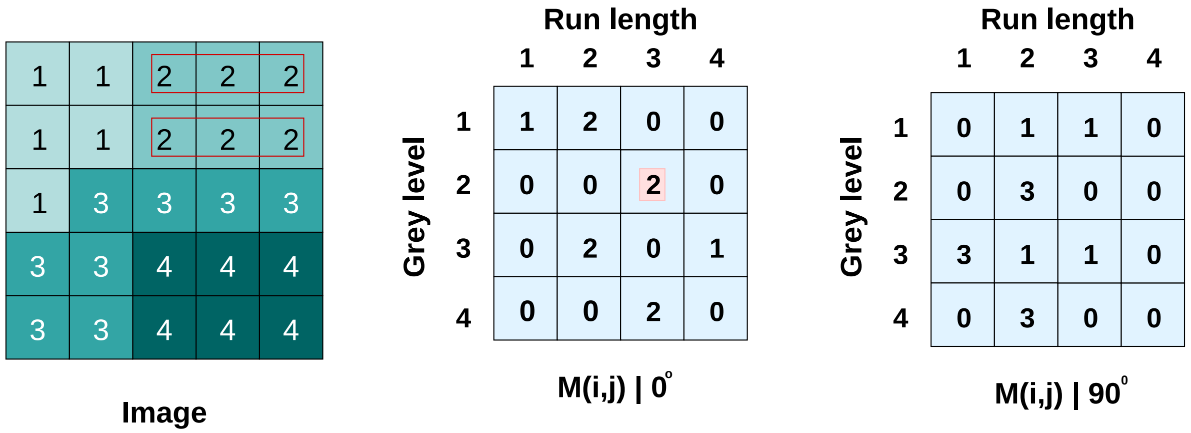
Chen & Dubes (1990) found that MRF performed better than GLCM. However, Ohanian & Dubes (1992) speculated that the success of MRF was dependent on large image sizes for adequate modelling—a very restrictive requirement in many image processing applications. Indeed, we have found MRF features to be very poor texture discriminators when used independently, for cell texture classification where small images are the norm (Walker, Jackway & Lovell 1995).

According to Ohanian and Dubes, GLCM has two drawbacks—the large number of potential features which can be extracted, and the lack of any theoretical guide to which features to extract for a particular problem. We address both these issues in Chapters 4 and 5.

A study by Wu et al. (1992) found GLCM (16% error rate) more effectively discriminated ultrasonic liver images than PSM (20% error rate), GLDM (25% error rate) and TEM (29% error rate). Wu et al. introduced a fractal-based approach, claiming its performance (11% error rate) was superior to that attained with GLCM. However, we find their comparison methodology flawed. They extracted GLCM features at only one spatial displacement, limiting analysis to one resolution, but for fractal analysis, they extracted features at four resolutions. They also claimed that the computational burden of GLCM was intolerable. This was because they did not requantise the 8-bit grey-scale images, which meant they had to calculate excessively large 256 × 256 co-occurrence matrices. Moreover, the resulting sparse matrices would not have allowed an accurate estimate of the true underlying joint PDF of grey-level pairs[3].

2.1.2 Grey Level Run Length Matrix

Another popular method for extracting co-occurrence-based texture descriptors is the GLRLM method introduced by Galloway (1975). Matrix calculation is computationally efficient, with the number of calculations required being directly proportional to the number of image pixels. A grey level *run* is a group of consecutive, collinear image pixels having the same grey level. The matrix element $M(i,j)$ quantifies the number of times a run of j pixels length occurs with an intensity of i. A series of matrices can be determined under angular constraints which define the direction of the run. Figure 2.2 shows an example of the GLRLM method.



Image

1	1	2	2	2
1	1	2	2	2
1	3	3	3	3
3	3	4	4	4
3	3	4	4	4

M(i,j) | 0°

Grey level \ Run length	1	2	3	4
1	1	2	0	0
2	0	0	2	0
3	0	2	0	1
4	0	0	2	0

M(i,j) | 90°

Grey level \ Run length	1	2	3	4
1	0	1	1	0
2	0	3	0	0
3	3	1	1	0
4	0	3	0	0

Figure 2.2: Calculating Grey Level Run Length Matrix element values for 0° and 90° orientations. In the image, there are two horizontal runs of intensity 2 with a length of 3 pixels. Thus, matrix element $M(2,3)|0^o = 2$.

Galloway extracted texture descriptors from the matrix using five weighting functions analogous to those proposed by Haralick et al. (1973). When extracted at four orientations (0, 45, 90, and 135 degrees), these features attained approximately 83% correct classification on the same database of terrain images used by Haralick et al.

[3] Wu et al. used 32 × 32 image tiles, giving an average of $32^2/256^2 = 0.0156$ pixel pairs per matrix element.

(1973). This result compares favourably with the 82.3% result of Haralick et al. using GLCM features.

A theoretical evaluation by Conners & Harlow (1980) concluded that GLRLM was less powerful than the GLCM method. They also identified two significant weaknesses—its susceptibility to noise, and its inability to capture important second-order grey level transition statistics of the form $P(i,j), \;\; i \neq j$. Weszka et al. (1976) found that GLRLM features performed the poorest when compared to GLCM, Fourier, and GLDM features. They suggested this was due to the method's sensitivity to image noise. Yogesan et al. (1993) classified four grades of cell pre-cancer using a combination of GLRLM and GLCM descriptors. They concluded that combining features from both texture methods provided lower classification error than GLRLM or GLCM features alone. Christen, Xiao, Minimo, Gibbons, Fitzpatrick, Galera-Dividson, Bartels & Bibbo (1993) also classified four classes of cell abnormality using a variety of morphological and statistical cell descriptors. Of the twenty features trialed, a run length feature (run length non-uniformity) was among the best three at discriminating the four classes of cells.

Bengtsson & Nordin (1994) suggested that "GLRLM is the most popular higher-order statistic used for texture analysis". While this may be true, the literature indicates that this technique is best used to complement other, more powerful methods of texture analysis.

2.1.3 Grey Level Entropy and Grey Level Variance Matrices

The GLEM and GLVM were introduced by Yogesan et al. (1996) and Yogesan et al. (1994) for the analysis of cell chromatin and Brodatz textures. Their motivation was based on the premise that existing statistical techniques, such as GLCM and GLRLM, are unable to measure the scale differences and grey level variation of minute structure. They proposed co-occurrence measures based on an image pixel's intensity, and the intensity entropy and variance in its neighbourhood. By varying the size of this neighbourhood, it should be possible to capture the size differences of the texture elements.

Formally expressing these matrices for a neighbourhood size of $w \times w$, the GLEM element $e(i,j|w)$ represents an estimate of the probability of grey level i occurring with neighbourhood entropy j. The entropy value j is defined as

$$j = -\sum_{g=1}^{G} P(g) \times \log[P(g)], \;\; P(g) \neq 0, \tag{2.4}$$

where G is the number of image grey levels and $P(g)$ is the probability of grey level g

occurring within a local neighbourhood of size $w \times w$, centred on a pixel with grey level i.

Yogesan et al. defined nine features which extract texture information from the GLEM. To trial this method, they extracted features from 2000 prostate cells from 20 patients, and used them to classify the patients into two classes—those who were resistant to hormone therapy and those who were hormone-sensitive. This trial was of considerable importance because the two prognostic groups could not be distinguished by histopathology or any other known means. Features from GLCM, GLRLM, and Local Intensity Transform (LIT-SNN) (Albregtsen, Kanagasingam, Farrants & Danielsen 1992) were also trialed. The highest correct classification rate of 95% was achieved by combining three GLEM features and one GLCM feature, showing not only the power of the GLEM method in this application, but also highlighting the potential of increased discriminatory power when combining different texture analysis methods.

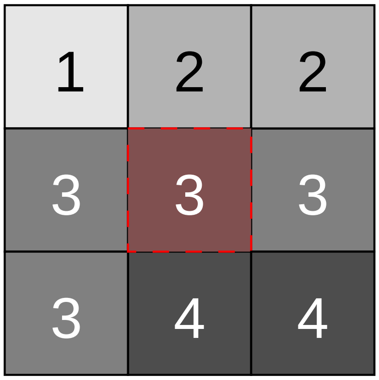
The GLVM element $p(i, j|w)$ represents an estimate of the probability of grey level j occurring with neighbourhood intensity variance i, where variance is defined in the usual manner. The size of the GLVM variance dimension is determined by the maximum variance in an image. Because each image will generally have a different maximum variance, Yogesan et al. suggest the following normalisation procedure:

$$Var' = \frac{Var - Var_{\min}}{Var_{\max} - Var_{\min}} * N_i, \tag{2.5}$$

where Var and Var' are the original and normalised variance values, and N_i is the size of the matrix dimension representing variance. This normalisation ensures that the range of variance values for each image are mapped to the full range of the matrix variance index i. For example, for a 3×3 window, the D_8 neighbourhood shown in Figure 2.3 has a variance of 0.94. Therefore, after normalising the variance via equation (2.5), $0.94 \rightarrow Var'$, this would contribute to the estimate of GLVM element $(Var', 3)$.

Yogesan et al. defined eight secondary features to extract the image structure's size and grey level variation information from the GLVM. Using just one feature (Grey Level Variance Ratio), they were able to classify, with no errors, 200 cell nuclei from 10 groups (5 normal, 5 cancer) into 2 classes (normal/abnormal) on a group-by-group basis. As a comparison, the best individual GLRLM feature provided 90% correct classification (i.e., 1 misclassified group).

By directly incorporating size information via neighbourhood constraints and measuring variance or entropy, Yogesan et al. redressed one of the deficiencies of GLCM—that features are global averages across entire images rather than direct measurement of texton properties. The spatial displacement parameter d in GLCM can only indirectly



$$
\begin{aligned}
i &= \mathcal{F}\left(\mathrm{var}\left(\begin{bmatrix} 1 & 2 & 2 \\ 3 & 3 & 3 \\ 3 & 4 & 4 \end{bmatrix}\right)\right) \\
j &= 3
\end{aligned}
$$

Figure 2.3: Calculating variance values for GLVM, using a 3×3 neighbourhood. The function $\mathcal{F}$ is a normalisation operation described in equation (2.5).

capture some of this information. A more in-depth comparative study is required to determine the general applicability of the methods to other types of texture, however, the results as presented are encouraging.

2.1.4 Statistical Feature Matrix

The SFM is a novel attempt by Wu & Chen (1992) to bypass the intermediate step of calculating co-occurrence matrices at various displacements, followed by feature extraction. Each element $M(i,j)$ of the $(L_r + 1) \times (2L_c + 1)$ matrix represents a *feature* extracted directly from the image. More specifically, the matrix *index* (i,j) represents an intersample displacement, $\mathbf{d} = (j - L_c, i)$, at which a particular feature will be extracted, while the matrix *value* $M(i,j)$ contains the feature value. The intersample displacement vector for each matrix element is shown in Figure 2.4, for the case of $L_r = L_c = 3$. As shown, a single matrix contains feature values extracted at several intersample lengths and orientations.

Only one matrix need be calculated for each type of statistical feature, as opposed to the necessity to calculate one matrix for each spatial displacement in GLCM. Wu and Chen defined three features—contrast, covariance and dissimilarity—which directly measure properties of the image. In contrast, in GLCM, it can be said that the features measure properties of the co-occurrence matrix directly, and of the image indirectly.

The choice of L_r and L_c, which represent the maximum intersample displacement in the image's row and column directions, is critical. Wu and Chen state that, for computational considerations, values of $L_r = L_c < 5$ are necessary. Larger values of L_r and L_c extract texture information across more scales and thus lead to matrices

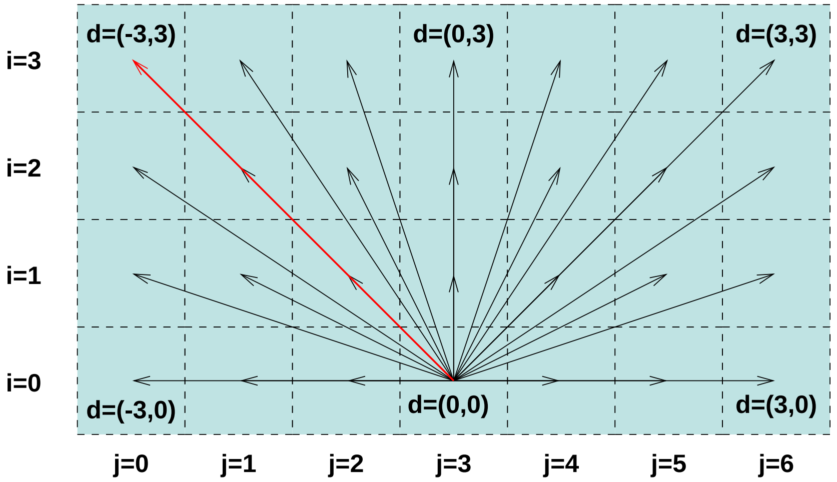


Figure 2.4: Magnitude and orientation of intersample displacement vector $\mathbf{d} = (j - L_c, i)$ for each SFM element $M(i, j)$, for $L_r = L_c = 3$. For example, the matrix element $M(3, 0)$ represents a feature extracted at intersample displacement $\mathbf{d} = (j - L_c, i) = \mathbf{d}(0 - 3, 3) = \mathbf{d}(-3, 3)$.

with more captured texture information. However, matrix storage and computational overhead quickly become prohibitive.

Because the matrix itself contains feature values, no feature extraction, via weighted sums of matrix elements, is performed. The matrix is used directly for classification. Wu and Chen measured the distance between two matrices as

$$D_{1,2} = \left[\sum_{i,j} |\mathbf{M}_1(i,j) - \mathbf{M}_2(i,j)|^2 \right]^{\frac{1}{2}}. \tag{2.6}$$

They performed two classification trials, the first using 16 Brodatz textures, and the second using ultrasonic liver images. Classification results were compared against those of GLCM using 12 features, and Liu and Jernigam's spatial frequency-based method using 8 features (Liu & Jernigam 1990). For Brodatz textures, SFM achieved 87.5% correct classification using only one feature, GLCM achieved 78.9%, while Liu and Jernigam's features achieved 76.6%. Similar results were recorded for liver texture classification, except that Liu and Jernigam's features performed considerably worse (50.7% classification). One could again question the fairness of the classification methodology. SFM features were calculated across intersample displacements of 1 to 4 pixels, while GLCM features were only calculated at effectively 1 pixel displacement.

The same three methods were also compared for noise performance. SFM features were surprisingly robust, even under strong additive noise up to 10dB signal-to-noise ratio. At this level of noise, SFM achieved 80% correct classification, compared to 19% for GLCM and 23% for Liu and Jernigam's features.

2.1.5 Neighbouring Grey Level Dependence Matrix

Sun & Wee (1983) introduced the NGLDM which models the co-occurrence of a pixel's intensity and properties of its surrounding pixels, under a neighbourhood constraint. The motivation for this approach was to produce angularly-independent features without explicitly calculating and averaging of co-occurrence matrices at several orientations. Other properties of this method, as reported by Sun and Wee, are computational simplicity, and invariance under linear grey level transformation—both properties of GLCM using 'normalised' images. This method is a generalisation of the Grey Level Difference Matrix of Weszka et al. (1976).

The NGLDM, $Q(k, s)$, is of dimensions $K \times S$, where K is the maximum grey level contained in the image, and $S + 1$ is the number of pixels within a neighbourhood constraint. The value of the matrix element (k, s) is the number of times the co-occurrence relationship exists between a pixel of intensity k and the number of its neighbours having intensity $\leq a$ being s. More formally,

$$\begin{aligned} Q(k,s) = \quad & \#\Big\{(i,j) \mid M(i,j) = k \text{ and } \#\Big[(q,r) \mid \rho((i,j),(q,r)) \leq d \\ & \text{and } |M(i,j) - M(q,r)| \ \leq a\Big] = s\Big\}, \end{aligned} \tag{2.7}$$

where $(i, j), (q, r)$ are valid image co-ordinates and $M(i, j)$ is the image grey level at (i, j); $\rho(.,.)$ is the D_8 distance measure, and d and a are positive integer constraints under which the matrix is determined. The constraint d, representing neighbourhood size, determines the scale or resolution at which the image is analysed. The threshold a allows the measurement of homogeneity within the neighbourhood defined by d.

Feature extraction from the matrix is analogous to that for GLCM, i.e., by a weighted sum of element values. Sun and Wee proposed 5 features, and suggested that these features measured properties of the image more directly than GLCM features. These features were then used to classify terrain images, and the results were compared to the classification rates of previously-published texture algorithms which used similar image data. NGLDM achieved 85% correct classification, compared to 83% by Haralick et al. (1973) using GLCM and terrain images, and 80% by Davis et al. (1979) using GCM and Brodatz textures. However, using the same data set as Sun and Wee, Weszka et al. (1976) achieved 92% correct classification using difference statistics and 91% using

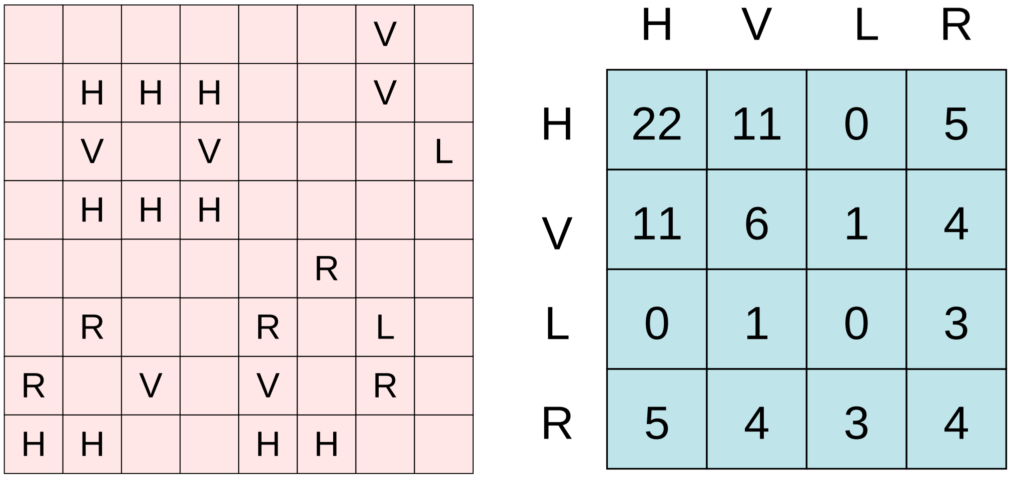
GLCM features. In a comparison of texture algorithms for analysing carpet wear, Siew et al. (1988) found that NGLDM had strong classification power compared to GLCM, GLDM and GLRLM features, but said no strong conclusions could be drawn because not all GLCM features had been trialed.

2.1.6 Generalised Co-occurrence Matrix

Davis et al.'s motivation for developing the GCM technique (Davis et al. 1979)[4] was based on the premise that macrotextures cannot be adequately defined by information contained in the GLCM. They suggested that, because of the large texton size of macrotextures, the GLCM predominantly captures statistics of grey level variation *within* the texton, and is unable to capture spatial arrangement statistics between textons. Their proposal was to capture spatial properties of texture edge information, via local maxima of the gradient image of the texture. They proposed two properties—magnitude and orientation of gradient local maxima—and extracted co-occurrence counts of these properties under several spatial constraint predicates. As an example, Figure 2.5(a) shows the location and corresponding orientation of a texture's local maxima of gradient, where H,V,L,R mean horizontal, vertical, left, and right, respectively. The corresponding GCM under the spatial constraint of co-occurrence within Euclidean distance 2 is shown in Figure 2.5(b).

Davis et al. suggested extracting features similar to those defined by Haralick et al. (1973). The extraction process is also the same, i.e., via a weighted sum of GCM elements. Davis et al. evaluated the performance of GCM features against those of standard GLCM. In the classification trial of 30 samples from 5 classes of texture, 8 GLCM features were extracted (4 feature functions at 2 spatial constraints), compared to 16 features for GCM (4 feature functions at 4 spatial constraints). Using pairs of features extracted under the same spatial constraint, the GCM method achieved greater than 80% correct classification, compared to GLCM which achieved only 57%. This is a significant difference in performance, however, once again it would appear that the classification methodology was somewhat biased. The GLCM features were rotation-invariant, being averaged across 45^o or 90^o intervals. The GCM features were, however, extracted at specific orientations. With the majority of the chosen textures having strongly oriented features, it is, therefore, not surprising that there was such a great discrepancy between the classification performance of the two methods. Nonetheless, the results presented by Davis et al. are encouraging enough to warrant a more rigorous trial.

[4] a similar method called Feature Frequency Matrix (FFM) was introduced by Shen & Bie (1992).



						V	
	H	H	H			V	
	V		V				L
	H	H	H				
					R		
	R			R		L	
R		V		V		R	
H	H			H	H		

(a)

	H	V	L	R
H	22	11	0	5
V	11	6	1	4
L	0	1	0	3
R	5	4	3	4

(b)

Figure 2.5: Calculating GCM element values: (a) Location and orientation of gradient local maxima; (b) GCM for Euclidean distance 2 neighbours. From Davis et al. (1979).

2.2 Discussion and Conclusions

Many co-occurrence-based techniques for analysing texture have been developed to address perceived weaknesses of the classic GLCM method, such as the inability to capture localised texture information. However, no comparative studies have shown conclusively that any one method provides superior performance to GLCM for all classes of texture. Unfortunately, there has been no unified framework for classification methodologies to permit fair comparison of algorithms. In some comparative studies, the chosen method of evaluation obviously favoured a particular analysis technique, such as extracting features at an unequal number of scales, or the extraction of rotation-invariant features for one technique, while using rotation-variant features for another. Only by eliminating these discrepancies and using a standardised image database of textures, can strong conclusions be drawn about the inherent power of the various techniques.

Several investigators attained improved classification performance by combining features from competing algorithms. This indicates the inability of any one method to capture all texture information. Some investigators attained improvement by extracting features across several scales. As we will clearly demonstrate in later chapters, information important to discriminating between texture classes can indeed exist at, and across, several spatial scales or resolutions.

Based on our review of published literature, there is strong evidence to suggest that the GLCM method of Haralick et al. is not only the most popular and powerful second-order texture method, but is perhaps one of the best among both structural and statistical approaches to texture analysis. However, a number of minor weaknesses were identified by various researchers. One such weakness is its reported inability to specifically capture local image properties mentioned previously. Another possibly related weakness is GLCM's inability to extract all texture information from co-occurrence matrices, as mentioned in the previous paragraph. This appears to be due to GLCM's fixed feature functions, and is the reason why many investigators report improvements in classification performance when features from other texture techniques are used to complement existing GLCM features. To ensure maximal capture of texture information, it is clearly an advantage to use several 'complementary' texture algorithms which extract both rotation-variant and rotation-invariant features at multiple scales. However, such an approach may prove to be computationally prohibitive. Can we address this problem of 'maximal capture of texture information' without resorting to using complementary algorithms? We will discuss this question in the following chapter.

Chapter 3

Statistical Geometric Features and Manual Feature Adaptation

In this chapter we begin our search for so-called 'smart algorithms' which use higher-level knowledge to extract a greater amount of information from texture images. We begin by investigating a recently published method of texture analysis called Statistical Geometric Features (SGF). After an initial review of the method, we discuss its salient features, and identify deficiencies in terms of its application to texture analysis. One of these deficiencies, common to most other analysis methods, is the definition of fixed, problem-independent feature functions which may be unsuitable for analysing specific texture types. As our first example of a method which uses higher-level knowledge, we define new SGF features more appropriate to cervical cell texture analysis. Those features showing discriminatory power are further investigated to determine the cytological properties manifesting the discrimination. While this method is not 'self adaptive', we will use it to demonstrate the significant benefits even manually-adaptive methods can yield over more conventional analysis methods such as GLCM. We will show that, by defining features specific to each analysis task, we can better target image properties which may contain important texture information. Using cell image analysis as an example, we will show that defining features in this way provides a far better understanding of textural changes within the cell nucleus upon neoplasia, than GLCM features.

3.1 Introduction

Recently, Chen, Nixon & Thomas (1995) proposed a novel set of 16 features for texture classification called Statistical Geometric Features. This work is of immediate interest since a thorough test by Chen et al. on *all* "Brodatz" textures (Brodatz 1966) has shown that:

- the SGF method exhibits a "substantially higher" correct classification rate than three other current methods—the Grey Level Co-occurrence Matrix (Haralick et al. 1973, Conners & Harlow 1980), the Fourier Power Spectral Method (Liu & Jernigam 1990), and the Statistical Feature Matrix (Wu & Chen 1992);
- the reduction in SGF classification performance, due to increasing the number of texture classes, is slower than the other methods; and
- the performance of SGF under additive noise conditions is good.

The SGF approach is to decompose a grey-scale texture image into a stack of binary images by threshold decomposition. Certain geometric properties of the connected regions (foreground and background) in each binary image are then measured, and a number of statistical parameters based on these geometric properties are computed.

These parameters then become the extracted texture features for the purpose of texture classification.

The results of Chen et al. are based on the seemingly ad hoc selection of *number of connected regions* and *irregularity* (compactness) as the two geometric properties to be measured from regions. Note, the 16 reported features are obtained by multiplying these properties by 2 (foreground/background regions) and then by 4 (statistical parameters). There are many other possible geometric properties of connected regions in binary images, therefore there seems ample scope for extending the SGF method. Further, it should be possible to tailor the geometric properties used to maximise performance in the particular texture classification problem at hand.

In this chapter we explore such extensions to the SGF method for discriminating normal and abnormal cervical cell images by classifying their nuclear texture. In particular, we seek to identify and define new SGF features which capture discriminatory information in cell texture images. Note, this problem may be somewhat more difficult than the Brodatz texture problem, because although there are only two classes (normal/abnormal), at the magnifications used in cytology, the discrimination between the two texture classes is not usually possible for humans without considerable training and other contextual information in the image.

As we mentioned in Section 2.1.1, we previously demonstrated in Walker et al. (1994) and Walker, Jackway & Lovell (1995) that the GLCM method performs quite well when classifying cervical cell texture images. We therefore use the GLCM method as our benchmark to assess this new method.

We review the SGF method in the next section before introducing our proposed extensions in Section 3.3. We then evaluate our new features, as well as those of Chen et al. (1995), in a feature selection and cell classification methodology which we describe in Section 3.4. The results are presented and discussed in Section 3.5, and our conclusions appear in Section 3.6.

3.2 Statistical Geometric Feature Algorithm

Once again we model a discrete grey-scale image on a domain $\boldsymbol{D} \subset \mathbb{Z}^2$, of N_g grey levels, as a 2D function $I : \boldsymbol{D} \rightarrow \boldsymbol{G}$, where $\boldsymbol{G} = \{1, \ldots, N_g\}$. The statistical geometric feature algorithm as given in Chen et al. (1995) is:

Step 1 A stack of binary images $I_b(x, y; \tau)$ is produced from $I(x, y)$ by thresholding at each discrete intensity level $\tau \in \{1, 2, \ldots, N_g\}$. Each binary image $I_b(x, y; \tau)$ is obtained by:

$$I_b(x,y;\tau) = \begin{cases} 1 & \text{if } I(x,y) \geq \tau; \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

Note: The mapping of the space of discrete grey-scale images to the space of binary image stacks is bijective (one-to-one and onto). The term *one-to-one* means that each distinct point in the image space of I (an image) maps to a distinct stack of binary images $I_b(x,y,\tau)$. The term *onto* means every possible point in the space of $I_b(x,y,\tau)$ (a binary image stack) is an image of a point in I. No loss of information occurs when representing a grey-scale image as a binary stack because we can always recover the image without loss:

$$I(x,y) = \sum_{\tau=1}^{N_g} I_b(x,y) \quad \text{for all} \quad (x,y) \in \boldsymbol{D}. \tag{3.2}$$

For each binary image $I_b(x,y,\tau)$, a group of '1'-valued pixels is defined as being a *4-connected region* if, for all pixels in the group, each pixel has at least one 4-connected neighbour within the group. Groups of '0'-valued pixels are similarly defined[1].

Step 2 A geometric property is measured for each 4-connected region in each binary image. These measurements are then summed or averaged across all the '1'-valued regions and all the '0'-valued regions at each threshold to give a pair of geometric properties $g_1(\tau), g_0(\tau)$ as functions of threshold, τ.

Step 3 Several statistics which characterise the distributions of $g(\tau)$ across τ are then computed. These statistics are then used as texture features for classification.

■

Chen et al. (1995) used two sets of geometric properties. The first is a simple count of the number of connected regions:

$$NC_1(\tau) \;=\; \text{the number of 4-connected '1'-valued regions} \tag{3.3}$$
$$NC_0(\tau) \;=\; \text{the number of 4-connected '0'-valued regions.} \tag{3.4}$$

[1] We discuss connectivity definitions in Appendix C.

The second, an average measure weighted by region size, of the irregularity or non-circularity of the regions, is defined as:

$$\overline{IRGL}_1(\tau) = \frac{\sum_{j=1}^{NC_1(\tau)} IRGL_j(\tau).NOP_j(\tau)}{\sum_{j=1}^{NC_1(\tau)} NOP_j(\tau)}; \quad (3.5)$$

$$\overline{IRGL}_0(\tau) = \frac{\sum_{j=1}^{NC_0(\tau)} IRGL_j(\tau).NOP_j(\tau)}{\sum_{j=1}^{NC_0(\tau)} NOP_j(\tau)}, \quad (3.6)$$

where index j is the jth 4-connected region, $NOP_j(\tau)$ is the number of pixels in the jth region at level τ, and $IRGL_j(\tau)$ is the irregularity or non-circularity of each region, given by:

$$IRGL = \frac{1 + \sqrt{\pi}.\max_{i \in I} \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}}{\sqrt{|R|}} - 1, \quad (3.7)$$

where,

$$\bar{x} = \frac{\sum_{i \in R} x_i}{|R|}, \quad \bar{y} = \frac{\sum_{i \in R} y_i}{|R|}. \quad (3.8)$$

R is the set of all indices to pixels in the region, and $|R|$ is the cardinality, or number of indices in the set R. We discuss the characteristics of this measure as defined above in Appendix D.

The four feature functions of threshold level τ defined above ($NC_1(\tau)$, $NC_0(\tau)$, $\overline{IRGL}_1(\tau)$, $\overline{IRGL}_0(\tau)$) represent statistical distributions. Chen et al. defines four statistics based on these feature functions, namely,

$$max\ value = \max_{\tau} g(\tau), \quad (3.9)$$

$$average\ value = \frac{1}{N_g - 1} \sum_{\tau} g(\tau), \quad (3.10)$$

$$sample\ mean = \frac{1}{\sum_{\tau} g(\tau)} \sum_{\tau} \tau.g(\tau), \quad (3.11)$$

$$sample\ S.D. = \sqrt{\frac{1}{\sum_{\tau} g(\tau)} \sum_{\tau} (\tau - sample\ mean)^2.g(\tau)}, \quad (3.12)$$

where $g(\tau)$ is one of the four feature functions. This gives a total of 16 features based on the statistics of the geometric properties of the image.

We note that many other statistics could be used here, including higher-order moments and rank order statistics, but in this chapter we have remained with the above four as proposed by Chen et al.

3.3 Analysis of the Proposed Method

3.3.1 Cytological Interpretation of SGF Regions

Intensity images of cervical cell nuclei represent chromatin *density* within the nuclei. As discussed in Section 1.1 (page 4), the cells undergo a staining process to allow imaging of the normally opaque cells. During this cell staining process, chromatin is stained proportional to its density. Areas of condensed chromatin known as *heterochromatin* absorb larger quantities of stain than the more sparse *euchromatin*. Thus, low intensity areas of a nuclear image represent predominantly heterochromatin, while high intensity areas represent euchromatin. The use of threshold level τ effectively segments the nuclear image based on chromatin density. Figure 3.1 details a series of thresholded images of a single nucleus.

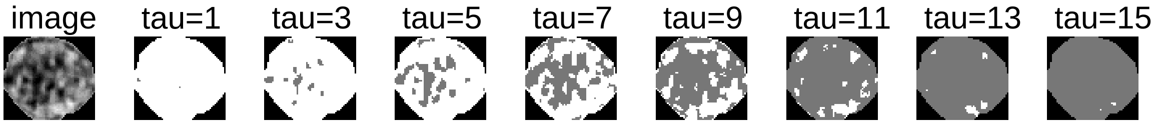


Figure 3.1: A series of thresholded images of a single nucleus.

Features based on '1'-valued pixels are measures of nuclear regions containing predominantly euchromatin. Features based on '0'-valued pixels measure characteristics of nuclear regions containing predominantly heterochromatin. For example, in cytological terms, the feature NC_1 represents the number of euchromatin clumps, while NC_0 represents the number of heterochromatin clumps. This representation is somewhat weaker at the extreme τ levels (at $\tau = 1$ for '1'-valued regions and $\tau = N_g$ for '0'-valued regions. For example, at low threshold levels, '1'-valued clumps represent areas containing not only euchromatin, but also low density heterochromatin. At high threshold levels, '0'-valued regions represent areas containing all heterochromatin, plus an amount of the higher density euchromatin. This presents a problem when analysing the results of subsequent feature classification in terms of heterochromatin or euchromatin properties. To minimise this problem, we calculate '1'-valued features at threshold levels $\tau = 4, \ldots, N_g$ and '0'-valued features at threshold levels $\tau = 1, \ldots, N_g - 3$. This removes the more 'contaminated' clumps from the analysis, allowing stronger conclusions to be drawn.

3.3.2 Refinements to Features

The definition of the feature NC (number of connected regions) presents us with an immediate problem in terms of the assumptions that are generally implicitly made about texture. An example of such an assumption is that the texture image is stationary (in a statistical sense) or homogeneous. Or alternatively, if the image is comprised of several textures, it is assumed that each texture area is homogeneous. More importantly, texture measures should ideally be independent of the amount or area of texture analysed, i.e., of window size. This is particularly important for the purpose of cell texture analysis, where even for the highest magnifications, the number of sample points (image pixels) is not only limited but dependent on cell size. It is difficult to choose a fixed window size to analyse cell texture, due to the variability of cell/nuclei shapes. In general, the entire cell texture image is analysed as a whole.

The feature NC is unfortunately linearly dependent on image size. That is, doubling the image area will double the number of connected regions in the image. Because it is well known that the size of abnormal cell nuclei are generally larger than those of normal cells, a feature such as NC will be highly discriminatory irrespective of whether the *texture* in the normal and abnormal nuclei are the same or different. Thus, the feature as proposed by Chen et al. is not specifically a texture measure, but a measure of both texture *and* morphological (area) characteristics. We propose re-expressing this feature in a form that is independent of image size, by normalising the measure based on image area:

$$NCA_1(\tau) = \frac{NC_1(\tau)}{|R_I|} \tag{3.13}$$

where NCA_1 is the number of connected '1'-valued regions normalised by the image area, R_I is the set of pixel indices in the image I, and $|R|$ is the cardinality of R. The feature NCA_0 is similarly defined.

3.3.3 New Features Tailored to Cytologic Applications

Defining new features specific to the problem at hand represents a significant advantage of adaptive methods over other methods such as GLCM, where feature definition is arbitrary and independent of the analysis task. As we will demonstrate, using tailored features not only allows better targeting of possible discriminatory properties within a texture, but also allows much stronger conclusions to be drawn from subsequent classification results. Prior to feature definition, an analysis of texture properties and an understanding of the process which generated the texture is warranted. Such properties dictate the kind of geometric features to be measured. As an example, we detail an analysis for the case of cell texture classification, but similar procedures can be adopted for many texture classification problems (Brodatz textures, crop identification, land use, etc.).

As discussed in Section 1.5 (page 35), the initial signs of cell neoplasia appear in the nuclei of affected cells. Various cell enzymes are responsible for gene regulation, and hence nuclear DNA and chromatin production. Changes in chromatin structure are known to be a result of changes in this gene regulation. It is also known that carcinogenesis can severely affect normal gene regulation (Danielsen, Farrants & Ruth 1989), resulting in not only increased quantity but also structural differences in the chromatin within the nucleus (Tucker 1979, Komitowski & Zinser 1985). It would be reasonable to assume that cell carcinogenesis may be detectable by way of measuring chromatin properties.

The following features attempt to measure specific cytological properties of the heterochromatin and euchromatin clumps in cell nuclei. These measures are based on properties of the cell such as the centre of gravity of the nucleus, chromatin clump size, and clump position within the nucleus (contextual information).

Firstly, we define the centre of gravity of the jth clump at some threshold τ as $(\bar{x}_j, \bar{y}_j)$, as in equation (3.8). We can define the centre of gravity of the entire nucleus as

$$x_{COG} = \frac{\sum\limits_{i \in R_I} x_i}{|R_I|}, \quad y_{COG} = \frac{\sum\limits_{i \in R_I} y_i}{|R_I|}, \tag{3.14}$$

where R_I is the set of all indices to pixels in the entire nuclear image, and $|R_I|$ is the cardinality of R_I.

We can define the normalised clump displacement of the jth clump from the centre of gravity of the nucleus as

$$D_j = \sqrt{\pi}\frac{\sqrt{(\bar{x}_j - x_{COG})^2 + (\bar{y}_j - y_{COG})^2}}{\sqrt{|R_I|}}. \tag{3.15}$$

For circular regions, we express D_j as a proportion of the radius of the region.

- **Average Clump Displacement**

$$\overline{DISP}_1(\tau) = \frac{\sum\limits_j D_{1,j}}{NC_1(\tau)} \tag{3.16}$$

This feature measures the average displacement of '1'-valued regions from the centre of gravity of the nucleus (normalised for nuclear area). The feature $\overline{DISP}_0(\tau)$ is similarly defined. We are attempting to measure whether cell neoplasia results in euchromatin or heterochromatin clumps whose displacements from the centre of gravity of the nucleus are, on average, greater or less than that of normal cells. Such changes have been noted in (Danielsen et al. 1989).

- **Average Clump Inertia**

$$\overline{INERTIA}_1(\tau) = \frac{\sum_j D_{1,j}.NOP_{1,j}(\tau)}{NC_1(\tau)} \tag{3.17}$$

This feature measures the average inertia of '1'-valued regions, where *inertia* is defined as the product of region area times region displacement from centre of gravity. We are attempting to determine whether cell neoplasia results in contextual changes in chromatin clump distribution. That is, whether larger chromatin clumps are displaced further from or closer to the nucleus centre of gravity. Such changes have been noted in Danielsen et al. (1989). The feature $\overline{INERTIA}_1(\tau)$ is similarly defined.

- **Total Clump Area**

$$TAREA_1(\tau) = \frac{\sum_j NOP_{1,j}(\tau)}{|R_I|} \tag{3.18}$$

This feature measures the total area of '1'-valued regions relative to the area of the nucleus. This feature will determine whether cell neoplasia results in more/less chromatin as a proportion of cell area (Danielsen et al. 1989). The feature $TAREA_0(\tau)$ is similarly defined.

- **Average Clump Area**

$$\overline{CAREA}_1(\tau) = \frac{\sum_j NOP_{1,j}(\tau)}{NC_1(\tau)} \tag{3.19}$$

Measures the mean area of '1'-valued clumps. Any correlation between cell abnormality and increased/decreased chromatin clump size will be detected by this feature (Danielsen et al. 1989, Komitowski & Janson 1990). The feature $\overline{CAREA}_0(\tau)$ is similarly defined.

3.4 Feature Evaluation

In the preceding sections, we defined a total of twelve feature functions,

1. $NCA_0(\tau)$,
2. $\overline{IRGL}_0(\tau)$,
3. $\overline{DISP}_0(\tau)$,
4. $\overline{INERTIA}_0(\tau)$,
5. $TAREA_0(\tau)$
6. $\overline{CAREA}_0(\tau)$
7. $NCA_1(\tau)$,
8. $\overline{IRGL}_1(\tau)$,
9. $\overline{DISP}_1(\tau)$,
10. $\overline{INERTIA}_1(\tau)$,
11. $TAREA_1(\tau)$
12. $\overline{CAREA}_1(\tau)$

of which eight have been defined to target specific properties of the types of texture we are analysing. From these twelve functions, we extract the four statistical features defined in equations (3.9) to (3.12), giving a total of 48 feature measures. Each of these feature measures can be considered as an *average* measure across all threshold levels τ. This averaging process may in fact hide discriminatory power at specific threshold levels. To determine whether discriminatory power is present at particular threshold levels, we also choose to analyse each of the twelve feature functions at each of the threshold levels τ.

3.4.1 Cell Database

Our data consists of a small set of 117 cells captured from 12 cervical slides, processed using ThinPrep® slide preparation[2] and regular Papanicolaou staining. Image capture was at a magnification of ×100, giving a spatial resolution of 0.12μm per pixel. We captured a total of 59 cells, with abnormalities ranging from mild neoplasia (CIN1) to Carcinoma in situ (CIS), from 11 abnormal slides, while 58 normal cells were captured from both the 11 abnormal and 1 normal slides. We note that some of the normal cells from abnormal slides may be MACs affected. However, MACs results in such minute changes that it can only usually be detected in *population* statistics, and not in individual cells. Moreover, we suspect that most cell databases used throughout the world contain MACs-affect normal cells, due to the problem of sampling error discussed in Section 1.3 on page 13.

We photometrically recalibrated the imaging system between each capture session, and ensured that the imaging of normal and abnormal cells were randomly interspersed.

[2] Cytyc Corporation, Massachusetts, U.S.A.

All cells were classified through the microscope before capture, by a cytologist. We show examples of typical normal and abnormal cell nuclei in Figure 3.2. We can see from these images that it is quite difficult for the untrained observer to distinguish visual differences between normal and abnormal cell nuclei in isolation.

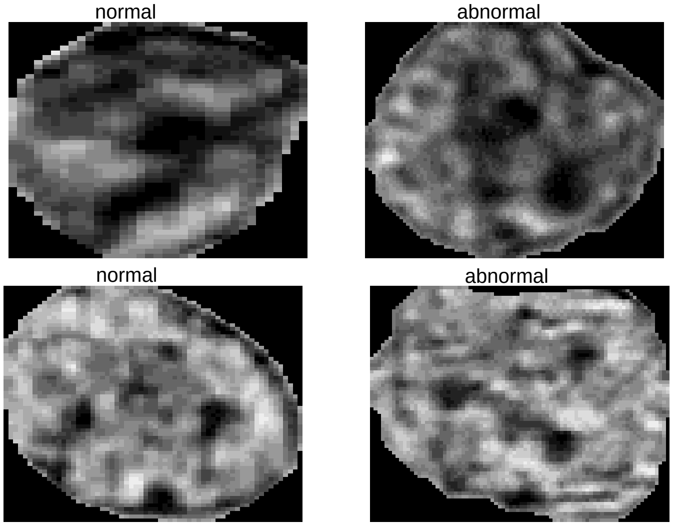


Figure 3.2: Typical examples of both normal and abnormal cervical cell nuclei, segmented using the technique described in Subsection 3.4.2.

3.4.2 Nuclear Segmentation

We now introduce the reader to our methodology for cell image segmentation, used in this and subsequent chapters. Our image analysis focused on the chromatin texture within the nucleus, so it was necessary for us to segment the nuclear image from its surrounding cytoplasm and slide background. Following image capture, we segmented each nucleus by a series of automated fast morphological transforms using octagonal structuring elements. The coding of each operation was based on work by Lee (1985), and proved to be very computationally efficient. The roots of image morphology can be found in set theory. For binary images I_b, each image component (groups of '1' or '0'-valued pixels) represents a set. Often, the image is acted upon by a *structuring element* B, also represented by a set. We will now explain in detail, two operations used to segment our cell images: the morphological *opening* and the morphological *closing*. A more in-depth discussion of image morphology can be found in (Vincent & Beucher 1989) and (Gonzalez & Woods 1993).

The opening of a set I_b by a structuring element B can be expressed as:

$$I_b \circ B = \bigcup \left\{ (B)_l \mid (B)_l \subset I_b \right\}, \tag{3.20}$$

where I_b (a binary image) and B (the structuring element) are sets in $\mathbb{Z}^2$, $(B)_l$ represents B translated by $l \in \mathbb{Z}^2$, and $\circ$ represents the opening operation. The opening can be thought of as the union of all translates of B that fit inside I_b. It effectively removes any image components which cannot completely 'hold' the structuring element. Figure 3.3 illustrates this concept.

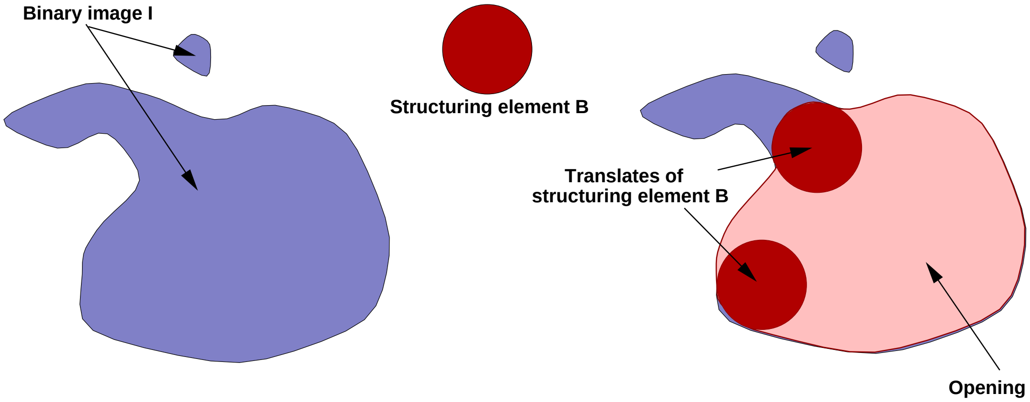


Figure 3.3: The morphological *opening* of binary image I, using a circular structuring element B.

The morphological closing of a set I_b by a structuring element B can be expressed as:

$$I_b \bullet B = \bigcup \left\{ k \mid \forall l \text{ s.t. } k \in B_l,\ B_l \cap I_b \neq \emptyset \right\}, \tag{3.21}$$

where $k \in \mathbb{Z}^2$ is a valid coordinate pair, and $\bullet$ represents morphological closing. A closing operation effectively 'fills in' any missing image components that are smaller than the structuring element. We show a simplified closing operation in Figure 3.4.

Our procedure for segmenting cell images using the opening and closing operations described above is as follows. Firstly, we globally thresholded each grey-scale cell image I, resulting in an incomplete segmentation of the nucleus in binary form I_b. The thresholding process can be expressed as:

$$\forall (x,y) \quad I_b(x,y) = \begin{cases} 0 & \text{if } I(x,y) \geq \tau; \\ 1 & \text{otherwise,} \end{cases} \tag{3.22}$$

where the co-ordinate pairs (x,y) are valid image pixels and τ is the threshold level.

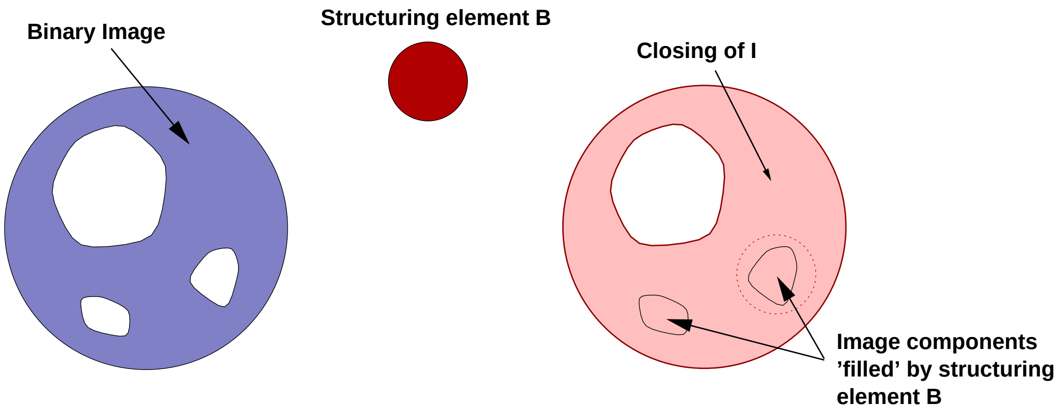


Figure 3.4: The morphological *closing* of binary image I, using a circular structuring element B.

An example of this thresholding operation is shown in Figure 3.5(a) and (b). The initial global threshold sometimes resulted in an incomplete nuclear image, particularly in areas of the nucleus which contained the least-dense *euchromatin*. Euchromatin areas represent high image intensities within the nucleus, and such intensities were occasionally above the threshold intensity, as we can see in Figure 3.5(b). We corrected this nuclear inhomogeneity by applying a morphological *closing* operation, using an octagonal structuring element slightly smaller than the smallest nucleus we expect to find on a slide.

Figure 3.5(c) shows the results of this operation. We then removed cytoplasmic artifacts, such as blood cells and leucocytes, by performing a morphological *opening* of the image. By using an octagonal structuring element slightly smaller than the *smallest* nucleus, we were able to ensure only artifacts, and not valid nuclei, were removed. Comparing Figures 3.5(c) and 3.5(d), we can clearly see the removal of such artifacts. We then used the resulting binary image as a mask for extracting the nucleus from the grey-scale image. Figure 3.5(d) shows an example of a binary mask, while Figure 3.5(e) shows the resulting segmented image.

We can express the entire morphological process as:

$$I_S = I * \Big((I_b \bullet B_1) \circ B_2 \Big) , \tag{3.23}$$

where I_S is the resulting segmented image, I_b is the thresholded image, and B_1 and B_2 are structuring elements. The symbols $\bullet$ and $\circ$ denote morphological closing and opening, respectively. Our method proved quite robust, and the majority of cell images were successfully segmented without human intervention. However, on some darkly

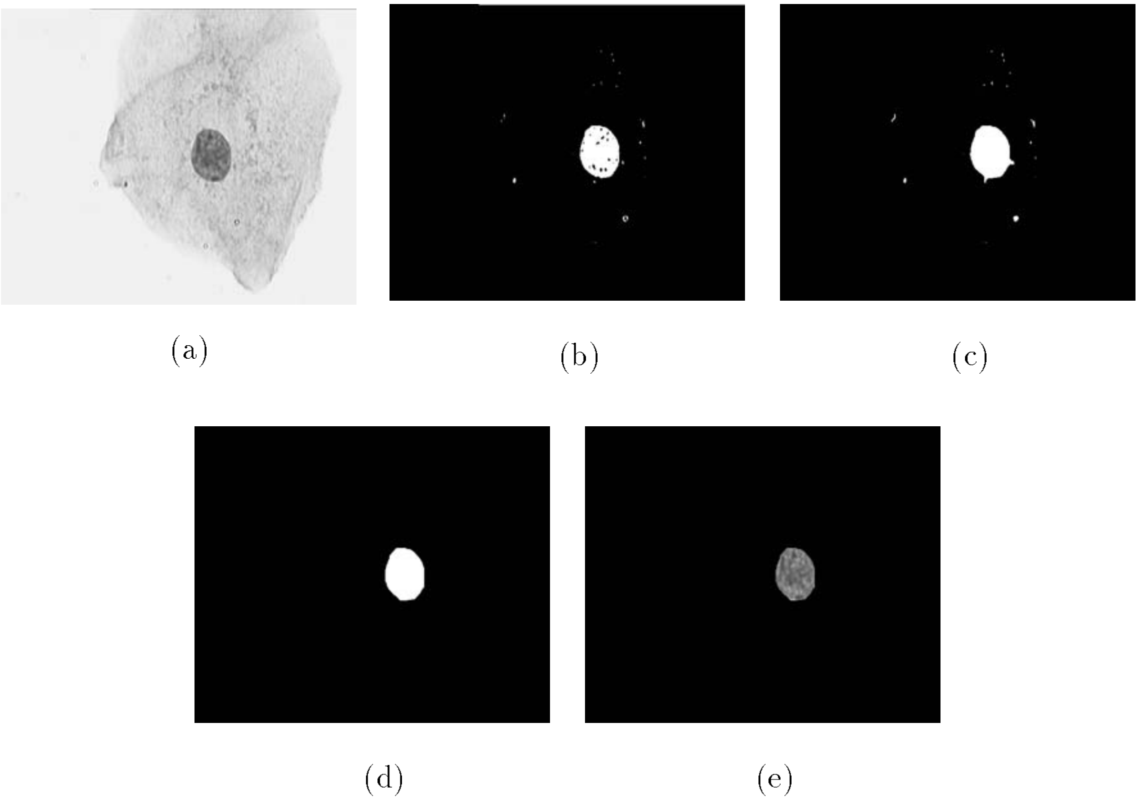


Figure 3.5: The process of segmenting a nucleus from the surrounding cytoplasm and image background, using image morphology: (a) the original greyscale image; (b) thresholded binary image; (c) removing nuclear inhomogeneity by applying an closing operation; (d) removing image artifacts by an opening operation. The resulting image is used as a mask for segmentation; (e) the segmented greyscale nucleus.

stained cells, we found it necessary to interactively adjust the initial global threshold.

3.4.3 Image Pre-Processing

Following segmentation, we pre-processed each nuclear image prior to feature extraction. See Section 1.4.2 for a discussion of image pre-processing. We requantised all images to 16 grey levels to reduce the computational expense of subsequent PR operations, and provide a common photometric domain for all images.

Our requantisation operation can be expressed thus:

$$\forall(x,y), \quad I'(x,y) = \text{floor}\left\{\frac{I(x,y) - I_{\min}}{I_{\max} - I_{\min}} \times 15.999\right\} + 1, \tag{3.24}$$

where the co-ordinate pair (x, y) is a valid image pixel, $I'(x, y)$ is the requantised value of pixel intensity $I(x, y)$, and the function floor(.) reduces a real-valued quantity to the largest integer lower than the argument. The addition of 1 in equation (3.24) results in a requantised intensity range of $I'(x, y) \in 1, \ldots, 16$. We have used intensity level '0' to indicate background pixels which were not to be processed.

Figure 3.6 shows the effectiveness of our requantisation method. The two images were chosen because of their gross difference in mean intensity—a result of a variation in imaging conditions and inherent differences in the density of each cell's nuclear chromatin. We can see that requantisation resulted in a reduction in image intensities; a normalised photometric domain (grey levels of intensity 1 to 16); enhanced image contrast; and the maintenance of histogram shape. Maintaining histogram shape is an important quality, because it ensures that second-order probability statistics, which are constrained by first-order statistics, are not modified. An alternate form of requantisation, called histogram equalisation or equiprobability quantising (Conners & Harlow 1978), could also have been used. Histogram equalisation is often used where the imaging process introduces photometric non-linearities (commonly found in X-ray imaging). We chose not to use this form of requantisation for two important reasons:

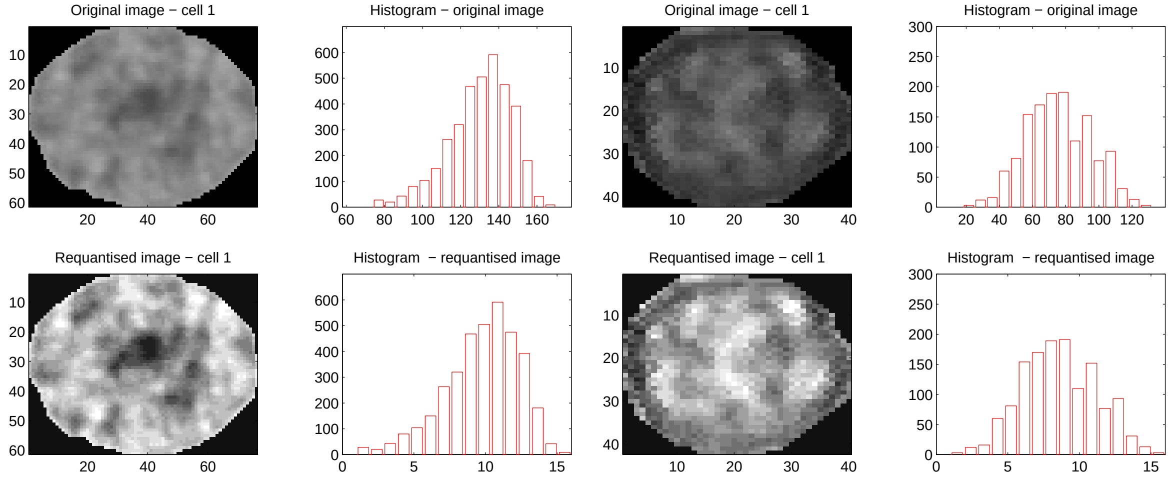


Figure 3.6: Requantisation of nuclear images. Notice that histogram shape is maintained—an important property if second-order statistics are to be preserved.

1. we have measured the photometric linearity of our imaging system by fitting a linear regression line to intensity data at varying exposures. The recorded coefficient of regression R^2 was greater than 0.999, and we therefore consider our system to be photometrically linear;

2. histogram equalisation is a non-linear operation and results in all images having the same, flat, histogram. That is, all images processed using histogram equalisation have the same first-order distribution—a uniform distribution. Because second-order joint probabilities are constrained by first-order marginal probabilities, histogram equalisation can result in the space of possible second-order probabilities being *reduced*. The effect is to make second-order properties measured from our images more *similar*, compared to when using our linear requantising technique of equation (3.24). This also has the effect of making the classification task more difficult than it need be.

3.4.4 Feature Pre-Processing

We pre-processed each of the 48 SGF and 40 GLCM features prior to feature selection and classification, using the Ladder-of-Powers technique of Velleman & Hoaglin (1981) previously mentioned in Section 1.4.4. This power transform re-expressed feature data to minimise any departures from normality, and allowed our normality-based parametric classifier to better profit from that normality. The Ladder-of-Powers technique raises the data by a power $\gamma \in \mathbb{R}$. For each feature $\underline{x}$, both classes $\underline{x}_1$ and $\underline{x}_2$ were re-expressed using the same power γ. For each power, the empirical CDF (or *ogive*) was compared to a Gaussian CDF Φ, defined by the first two moments of the re-expressed data. We measured the goodness of fit of these two CDFs by an error metric based on the total squared area between the two CDFs. Thus,

$$Error(\gamma) = \sum_{c=1,2} \left(\int_x \left(\left[cdf \left\{ \underline{x}_c^\gamma \right\} - \Phi \left\{ \text{mean} \left(\underline{x}_c^\gamma \right), \text{var} \left(\underline{x}_c^\gamma \right) \right\} \right]^2 \right) \right), \tag{3.25}$$

where $\Phi \left\{ \mu, \sigma^2 \right\}$ represents the CDF of a Gaussian distribution of mean μ and variance σ^2. The power which minimised the sum of the errors for the two classes, γ_{opt}, was used as the power by which the data were transformed:

$$\gamma_{opt} = \min_\gamma \left\{ Error(\gamma) \right\}. \tag{3.26}$$

Figure 3.7 shows a typical example of the histograms of a class-conditioned feature, before and after normality transformation. The data used was a GLCM feature from Chapter 6 whose class-conditioned distributions were highly non-Gaussian. The two lower plots in this figure show the distributions after transformation via equation (3.25). We can see the resulting distributions are now near-Gaussian, and represent a much closer match to a true Gaussian distribution then the unmodified data.

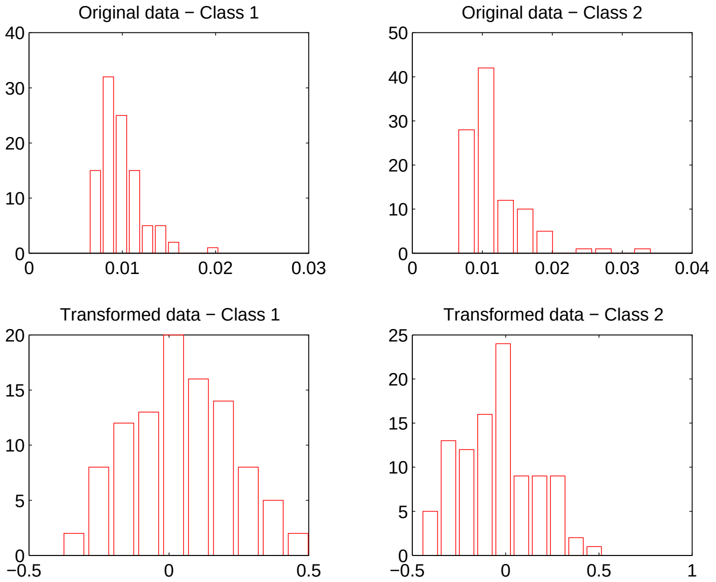


Figure 3.7: Histograms of class-conditioned feature data, before and after normality transformation using equation (3.25).

The ladder-of-powers method is an example of a *univariate* normality transform. The optimisation of γ for each feature is independent of all other feature variates. That is, the transformation optimises the marginal normality of each feature variate. Other, more optimal methods of normality transforms exist, such as those which enhance the joint-normality of the feature data. We refer the reader to McLachlan (1992) for a discussion of these transforms, including a multi-variate extension to the Box-Cox univariate normality transform (Box & Cox 1964). An extensive review of the Box-Cox power transform can also be found in Sakia (1992). Multi-variate transforms are generally computationally expensive and may not yield further improvement in classification performance. For these reasons, we have chosen to apply the uni-variate ladder of powers method throughout this thesis.

3.4.5 Feature Selection and Classification

We used the discriminant analysis methodology discussed in Section 1.4.5 (page 26) to reduce the high dimensionality of the feature space to a lower dimension. This allowed a more robust estimation of the class-conditioned distributions within this new space. We reduced the 48 dimensional feature space to 20 dimensions using Kittler's *plus 2-take away 1* feature set search algorithm (Kittler 1986) and the Bhattacharyya discrimination measure (Hand 1981). To critically appraise the SGF method, we compared the classification performance of SGF features to those of GLCM. A total of 40 GLCM features were trialed, derived from the 8 standard feature functions defined in Section 2.1 (*Energy, Entropy, IDM, Correlation, Inertia, Variance, Shade, Promenance*). Each of these features were calculated at spatial displacements of 1,2,4,8, and 16 pixels. The GLCM features were similarly reduced to 20 dimensions.

To highlight the benefits of manual feature adaptation, we used the Bhattacharyya discrimination measure detailed in Section 1.4.5 to evaluated the discriminatory power of each of the SGF features. Those with high discriminatory power were further investigated to determine the cytological properties which were manifesting the discrimination. This was an easy step, because the features were manually chosen to measure specific properties of the cell chromatin.

We applied leave-one-out classification (detailed in Section 1.4.7) to the optimised sets of SGF and GLCM features to obtain accurate estimation of the real classification error. For each trial, we trained the quadratic classifier of Section 1.4.6 on all but one sample, and evaluated the performance of the resulting classifier on this sample. We repeated this process until all 117 samples had been classified once. The sum of the misclassifications represented the real (as opposed to the apparent) misclassification rate, and is a minimally biased estimate.

3.5 Results and Discussion

3.5.1 Discriminant Analysis Results

Figure 3.8 details the real misclassification rates produced by leave-one-out classification. We can see that SGF features provide stronger discriminatory power and better classification performance than GLCM features at lower dimensions. We feel this is due to the tailoring of SGF features to measure specific texture properties, as opposed to the more ad hoc application of pre-defined GLCM features. We notice that most of the discriminatory power exhibited by the SGF features defined in this chapter is contained in two features. A feature space of two dimensions provides good discrimination with low computational burden.

3.5.2 Feature evaluation

We determined the discriminatory power of all SGF features and further analysed those features which expressed discriminatory ability between normal and abnormal cells.

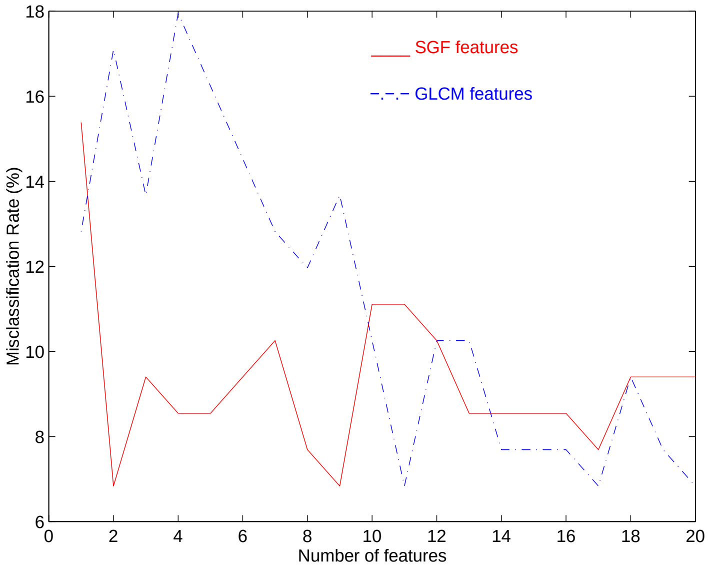


Figure 3.8: Comparison of SGF and GLCM misclassification rates for optimal feature sets.

Each of the discriminatory features held valuable information directly relating to euchromatin and heterochromatin changes upon cell neoplasia. The following figures detail several discriminatory features.

The feature NCA_1 measures the average number of euchromatin clumps per unit nuclear area. We can see from Feature 3.9 that there is great variability in the average number of euchromatin clumps per unit area for normal cells, however, abnormal cell euchromatin appears to be more stable. Albregtsen et al. (1992) also noted tighter clustering of certain features from abnormal cells. Furthermore, we can see that the number of clumps per unit area is considerably smaller for abnormal cells, suggesting larger clump sizes. There is much published literature to support this hypothesis.

The feature $max\ value(CAREA_0)$ shown in Figure 3.10 exhibits further discriminatory information. This feature measures the average size of heterochromatin clumps within the nucleus. We can clearly see that abnormal cells not only have greater variability in average clump size, but also have much larger clumps than normal cells.

The distribution of the feature $average\ value(IRGL_0)$ shows greater variability for normal cells than abnormal cells. This feature measures the shape of heterochromatin clumps in cell nuclei, with low regularity indicating more circular shape. The distribution of this feature suggests that the shapes of abnormal cell heterochromatin clumps are more uniform than that of normal cells.

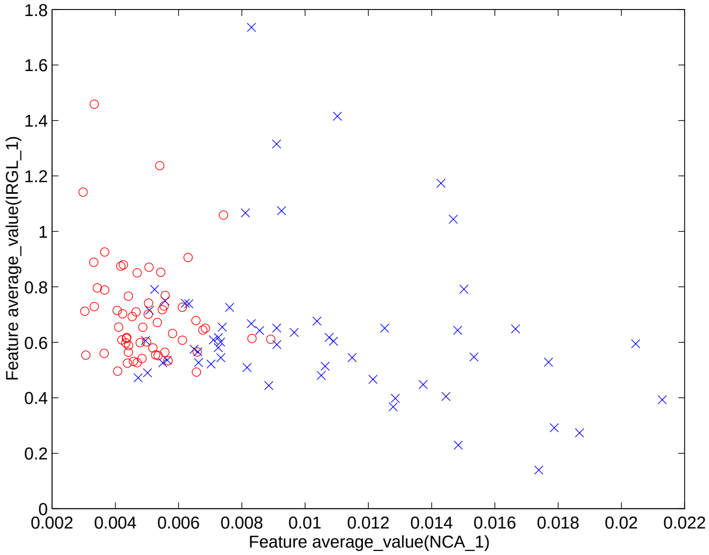


Figure 3.9: An SGF feature showing clear disparity between normal and abnormal cell nuclei. The 'x' marks represent normal cells while 'o' marks represent abnormal cells. The feature $average\ value(NCA_1)$ indicates abnormal cells have less euchromatin clumps per unit nuclear area. Also, the variability of this average is far less for abnormal cells than normal cells.

Heterochromatin changes during cell neoplasia

Analysing the feature $\overline{INERTIA_0}$ at each τ level revealed a prominent discrimination peak at threshold level $\tau = 6$. This feature is sensitive to large heterochromatin clumps displaced from the centre of the nucleus. The graph on the left in Figure 3.11 details the discriminatory power of the feature $\overline{INERTIA_0}$ at each τ level. The plot on the right shows a scatter plot of this feature at $\tau = 6$. This figure suggests that pre-cancerous cells have larger chromatin clumps nearer to the nuclear-cytoplasmic membrane, than normal cells.

3.6 Conclusions

In this chapter, we have reviewed the method of Statistical Geometric Feature texture analysis, and demonstrated, by example, its extension to other texture analysis problems. The flexibility of this method, by tailoring specific features to the problem at hand, is an important advantage over other methods which use pre-defined, problem-independent features. For example, we defined a further four feature functions which attempt to measure specific properties of chromatin texture distribution in cell nuclei

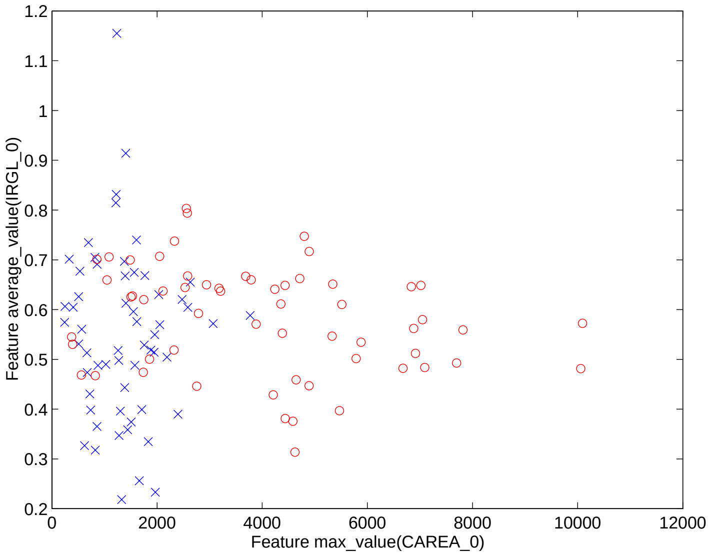


Figure 3.10: Another two features showing clear disparity between normal and abnormal cell nuclei. The 'x' marks represent normal cells while 'o' marks represent abnormal cells. From the feature *max value*$(\overline{CAREA}_0)$ we can see that abnormal cells have larger heterochromatin clumps and greater clump size variability. From the feature *average value*$(IRGL_0)$ we find that abnormal cells have less variability in heterochromatin clump shape.

images. These features provided insights into the cytological properties of abnormal cells, and more importantly, highlighted strong differences between chromatin structure in normal and abnormal cells.

We modified a feature proposed by the original authors to provide invariance to texture area and thus image window size. This allowed the feature to be applied to problems where a fixed window size is inappropriate or impossible to use (such as cell texture analysis).

To highlight the advantages of this adaptive method, we evaluated a total of 48 features, in the form of statistics of the six feature functions, on a data set of high-resolution cell nucleus images. After applying discriminant analysis, feature set reduction and leave-one-out classification, we found that a misclassification rate of less than 7% could be attained using only two features. This compared favourably to GLCM, which required eleven features for the same error rate.

Many of the SGF features provided insights into the cytological properties of neoplastic cells, and more importantly, highlighted strong differences between chromatin structure in normal and neoplastic cells. By defining feature functions which measure

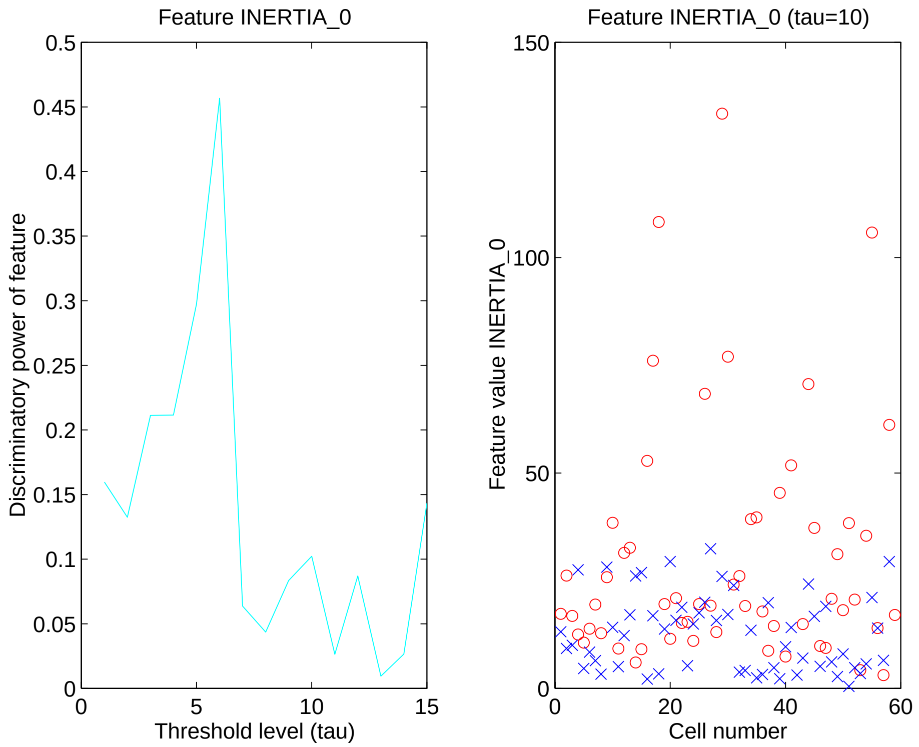


Figure 3.11: The plot on the left shows the discriminatory power of feature $\overline{INERTIA}_0$ at different thresholds. The peak at $\tau = 6$ indicates differences in the locations of large heterochromatin clumps between normal and abnormal cell nuclei. The figure on the right clearly shows abnormal cells having higher 'inertia', suggesting larger clumps nearer to the nuclear-cytoplasmic membrane, than normal cells.

specific properties of cell texture, we found that:

- many neoplastic cells appear to contain heterochromatin clumps with greater average area than those of normal cells (Figure 3.10),
- the average number of euchromatin clumps per unit nucleus area in normal cell nuclei is greater than that of neoplastic cells (Figure 3.9),
- normal cells have a greater variability in the number of euchromatin clumps per unit area, whereas this quantity is far more stable in neoplastic cells (Figure 3.9),
- neoplastic cell nuclei may have larger heterochromatin clumps at greater distances from the centre of gravity of the nucleus, compared to normal cell nuclei (Figure 3.11).

To conclude, the method of SGF texture analysis using the features defined in this work, provided good discriminatory power when detecting textual changes in high-resolution cervical cell images. Preliminary results indicate that the method may be as powerful as the Grey Level Co-occurrence Matrix method. Moreover, using feature functions derived specifically for the purpose of cell chromatin analysis allowed quantitative as well as qualitative descriptions of chromatin texture changes in abnormal cell nuclei. The process of manually adapting SGF features tailored to the geometric properties of the textures allows far stronger conclusions to be drawn from the feature distributions and classification results than is possible with many other texture methods, thus making this technique a powerful analysis tool.

Manual adaptation of feature functions means that we need to define new features each time a new type of texture is analysed. It would be more convenient if the method could 'self-adapt' its feature functions without the need for human intervention. In the next chapter we will begin our search for self-adaptive texture analysis techniques, which will greatly enhance their general applicability to a wider range of texture types.

Chapter 4

Improving Co-occurrence Feature Discriminatory Power

In this chapter we discuss a method of improving the discriminatory power of co-occurrence matrix features. We investigate where co-occurrence matrix features derive their discriminatory power, and provide a theoretical basis for improving this discrimination. The new method of texture analysis we present here is self-adaptive and requires no human intervention. We critically appraise our method in classification trials against the benchmark GLCM method, and present examples of discrimination improvement using real-world data. Cross-validation results indicate remarkable increases in feature discriminatory power for almost all features trialed.

4.1 Introduction to GLCM Feature Extraction

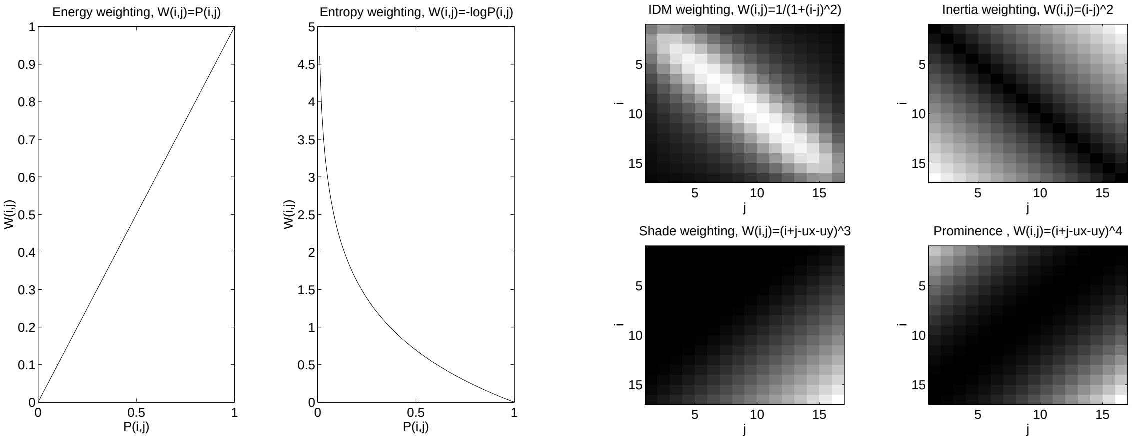
We will begin with a summary of the GLCM method reviewed in Section 2.1.1, and focus more on the process of extracting texture information from its matrices. As readers will remember, the co-occurrence matrix is an estimate of the joint PDF of grey-level pairs in an image. The matrix is generally symmetric and, when normalised, element values are bounded by [0,1], and the sum of all element values equals 1. Information is extracted from the matrices by applying *secondary* feature functions. Approximately 20 such secondary features appear in the literature (Conners et al. 1984, Haralick et al. 1973, Haralick 1979, Conners & Harlow 1980), and they measure four main types of texture information:

1. measures of an image's statistical properties,
2. measures of an image's visual characteristics,
3. measures based on information theory, and
4. measures of information based on correlation.

Many of these secondary features are derived by weighting each of the co-occurrence matrix element values, and then summing these weighted values to form the feature value. The weighting applied to each element is based on a feature weighting function, so by varying this function, different texture information can be extracted from the matrix. These weighting functions fall into two general classes:

1. co-occurrence matrix element weighting based on the element's value, and
2. co-occurrence matrix element weighting based on the spatial position of the element.

Table 4.1 lists eight of the most popular feature weightings used in the literature. In this chapter we will use these eight features specifically (applied to isotropic co-occurrence matrixes), but the method can be applied to any feature calculated by the weighted sum of co-occurrence matrix elements. Figures 4.1(a) and (b) give examples of some of the more common weighting functions applied to a co-occurrence matrix. In Figure 4.1(b), lighter shades indicate larger element weighting.



(a) (b)

Figure 4.1: Two classes of GLCM weighting functions: (a) Weighting dependent on element value: $W(i,j) = \mathcal{F}\left(P(i,j)\right)$; (b) Weighting dependent on an element's spatial position: $W(i,j) = \mathcal{F}(i,j)$.

Notation and terminology common to Chapters 4, 5, and 6

Indices	
Class index	$c = 1, \ldots, N_c$
Variate index	$v = 1, \ldots, N_v$
Pattern index	$s = 1, \ldots, N_s$
Grey level index	$g = 1, \ldots, N_g$
GLCM matrix indices	$i, j = 1, \ldots, N_g$
GLCM displacement index	$d = 1, \ldots, N_d$
GLCM angle index	$\theta = \frac{\pi}{4}\{0, \ldots, N_\theta - 1\}$

We represent a single measure drawn from a pattern as x, and a single measure drawn from a set of N_s patterns as the column vector $\underline{x} = [x_1, \ldots, x_{N_s}]^{\mathbf{T}}$. The row vector $\mathbf{x} = [x_1, \ldots, x_{N_v}]$ denotes N_v measures or variates drawn from a single pattern,

Type 1 - Weighting dependent on matrix element value, i.e., $W(i,j) = \mathcal{F}\left(P(i,j)\right)$

Function	**Weighting**
• $Energy = \sum_{i,j} P(i,j)^2$;	$W(i,j) = P(i,j)$
• $Entropy = -\sum_{i,j} P(i,j) \log P(i,j)$;	$W(i,j) = \log P(i,j)$

Type 2 - Weighting dependent on spatial position, i.e., $W(i,j) = \mathcal{F}\left(i,j\right)$

Function	**Weighting**
• $IDM = \sum_{i,j} \frac{1}{1+(i-j)^2} P(i,j)$;	$W(i,j) = \frac{1}{1+(i-j)^2}$
• $Inertia = \sum_{i,j} (i-j)^2 P(i,j)$;	$W(i,j) = (i-j)^2$
• $Correlation = -\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)}{\sqrt{(\sigma_x \sigma_y)}} P(i,j)$;	$W(i,j) = \frac{(i-\mu_x)(j-\mu_y)}{\sqrt{(\sigma_x \sigma_y)}}$
• $Shade = \sum_{i,j} (i+j-\mu_x-\mu_y)^3 P(i,j)$;	$W(i,j) = (i+j-\mu_x-\mu_y)^3$
• $Prominence = \sum_{i,j} (i+j-\mu_x-\mu_y)^4 P(i,j)$;	$W(i,j) = (i+j-\mu_x-\mu_y)^4$
• $Variance = \sum_{i,j} (i-\mu)^2 P(i,j)$;	$W(i,j) = (i-\mu)^2$

GLCM Notation

$P(i,j)$ is the (i,j)th element of a normalised co-occurrence matrix .

$P_x(i) = \sum_j P(i,j)$;	$P_y(j) = \sum_i P(i,j)$
$\mu_x = \sum_i i \sum_j P(i,j) = \sum_i i P_x(i) = E\{i\}$	$\mu_y = \sum_j j \sum_i P(i,j) = \sum_j j P_y(j) = E\{j\}$
$\sigma_x = \sum_i (i-\mu_x)^2 \sum_j P(i,j)$	$\sigma_y = \sum_j (j-\mu_y)^2 \sum_i P(i,j)$

$\mu = \mu_x = \mu_y$ for symmetric matrices.
$W(i,j)$ is the weighting applied to the (i,j)th element of a normalised co-occurrence matrix .

Table 4.1: GLCM feature functions which are widely used in the literature. These can be classed into two types of weighting function: those dependent on the *value* of each co-occurrence matrix element (type 1), and those dependent on the *spatial position* of each element (type 2).

while $\underline{\mathbf{x}} = [\underline{x}_1, \ldots, \underline{x}_{N_v}]$ denotes a set of N_v different measures where each measure $\underline{x}_v$ is as defined above.

A co-occurrence matrix calculated at displacement $d\angle\theta$ is represented as $\mathbf{P}$ or $\mathbf{P}(i, j|d, \theta)$. When averaged across θ, it is represented as $\mathbf{P}(i, j|d)$ or its equivalent array form, $\mathbf{P}(i, j, d)$. By $P(i, j)$ we mean a single *element* of $\mathbf{P}$ measured from one pattern, while $\underline{P}(i, j)$ indicates a column vector of measures for this element, extracted from N_s patterns. The notation $\underline{Pw}(i, j)$ represents the vector of element values after $\underline{P}(i, j)$ has been weighted by the weighting function $W(i, j)$. Similarly, the discriminatory power of element $\underline{Pw}(i, j|d)$ is denoted as $J(i, j|d)$ or equivalently as $J(i, j, d)$.

4.1.1 Feature Extraction and the Curse of Dimensionality

By extracting these secondary features from co-occurrence matrices calculated at varying intersample spacings d, we are able to analyse texture content at different spatial resolutions. It is well known that texture information can exist at varying spatial resolutions (Conners & Harlow 1980, Rosenfeld & Kak 1982, Shen, Bie & Chiu 1993) and this *multi-scale* approach to feature extraction is an attempt to capture such information. However, extracting secondary features at several scales can result in a feature space of unsuitably large dimensions when compared to the number of training set examples available to define or characterise this space (Silverman 1986). This problem is commonly known as the *curse of dimensionality* (Hand 1981) and can be easily shown by a simple example.

Let $\mathbf{X}$ denote a two-dimensional normally distributed random vector. Let the two-dimensional measurement vector $\underline{\mathbf{x}}$, where $\underline{\mathbf{x}} = [\underline{x}_1, \underline{x}_2]$, represent 100 realisations of random variable $\mathbf{X}$. The histogram of the first variate $\underline{x}_1 = [x_{1,1}, \ldots, x_{1,100}]^{\mathbf{T}}$, representing the extraction of one feature from 100 exemplars, is shown in Figure 4.2. It should be noted that each histogram bin contains a number of entries (100/10 = 10 measures per bin, on average), and the histogram shape is an adequate approximation of the true underlying distribution from which the realisations were extracted.

The second variate $\underline{x}_2 = [x_{2,1}, \ldots, x_{2,100}]^{\mathbf{T}}$ represents the extraction of a second feature from the 100 exemplars. It is now necessary to estimate this new two-dimensional histogram containing 100 bins, with only the original number of (now 2-D) data points. This histogram, shown on the right of Figure 4.2, contains many empty bins, and all other bins have very low counts. On average there is only 100/100 = 1 measure per bin. Clearly, the more features we extract, the more inaccurate our estimate of the underlying multivariate distribution becomes.

For classification purposes, where an unknown texture is to be allocated to one of sev-

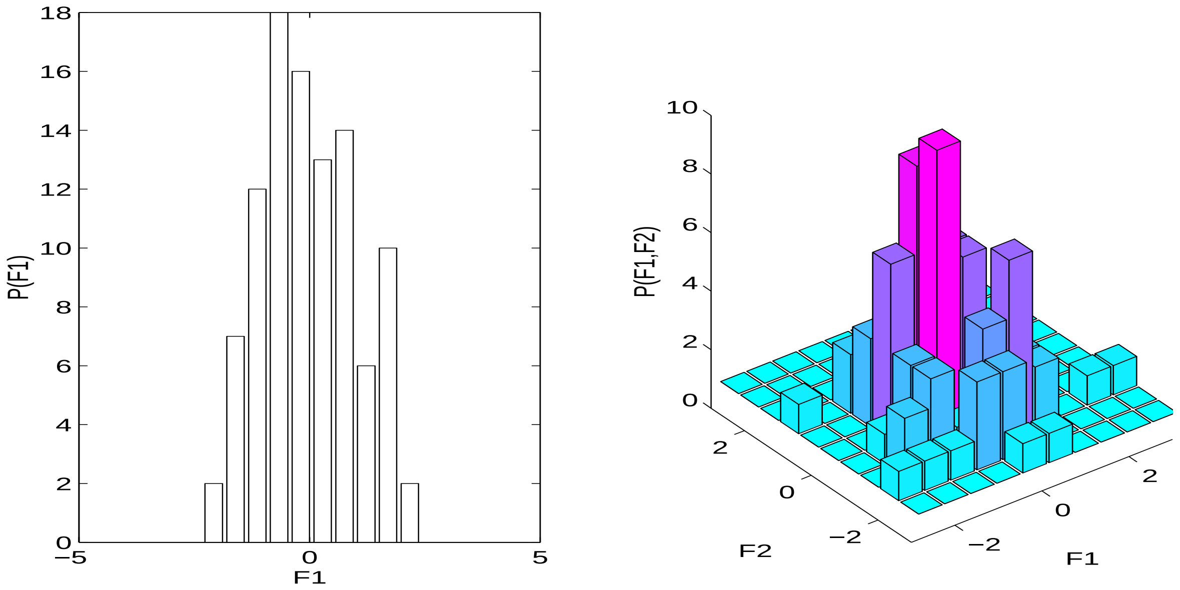


Figure 4.2: The ‘curse of dimensionality’ at work. An example showing the decrease in the accuracy of distribution estimation as dimensionality is increased from 1-D to 2-D. For each feature, 100 realisations were drawn from normally distributed i.i.d. data.

eral classes, it is of paramount importance to accurately estimate these underlying class-conditioned feature distributions. While it would be desirable to extract only a minimal number of ‘useful’ secondary features (where we define ‘useful’ as being those features whose class-conditioned distributions exhibit statistical differences between classes), it is usually not known *a priori* which features will be useful. It is common practice to extract a large number of secondary features, and subsequently reduce the resulting high-dimensional feature space, using the discriminant analysis and feature selection techniques reviewed in Section 1.4.5. This allows better estimation of the true distribution of features for each class based on the limited training data available. Reducing feature set dimensionality usually involves removing those features that are redundant to the classification process. That is, removing those features which provide little or no extra information to distinguish between texture classes. This process is accomplished by removing highly correlated features, which provide very little *extra* information, and by removing features whose discriminatory power is very low (whether correlated or not). We discussed discriminatory power measures in Section 1.4.5. Ideally, removing such features will not result in an increase in classification error.

For secondary features which prove to be discriminatory, the question arises as to where this discriminatory power is derived.

4.2 The Discrimination Matrix

For an $N_g \times N_g$ co-occurrence matrix, a univariate secondary feature vector $\underline{x} = [x_1 \ldots, x_{N_s}]^{\mathbf{T}}$ is generally a weighted sum of these N_g^2 elements, that is

$$\underline{x} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} W(i,j)\underline{P}(i,j) \tag{4.1}$$

$$= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \underline{Pw}(i,j), \tag{4.2}$$

where $W(i,j)$ is the weighting applied to the element $P(i,j)$, $Pw(i,j)$ is the *weighted* element, and i and j are the matrix row and column indices. The column vector $\underline{P}(i,j) = [P_1(i,j), \ldots, P_{N_s}(i,j)]^{\mathbf{T}}$ represents a measurement vector for matrix element $P(i,j)$, calculated from N_s exemplars. The feature $\underline{x}$ can be modelled as realisations of a random variable $\mathcal{X}$, whose distribution is characterised by its moments, μ_x, σ_{xx}, etc. These moments, when calculated for two or more classes of data, can be used to determine the discriminatory power of this feature, using a metric such as equation (1.5) on page 27. The feature $\underline{x}$ is comprised of the sum of weighted co-occurrence matrix elements $\underline{Pw}(i,j)$, and these individual weighted matrix elements can also be considered as random variables, each with individual discriminatory power. In the following work we call $\underline{Pw}(i,j) = W(i,j)\underline{P}(i,j)$ the *weighted elemental features*, for these weighted elements can themselves be used as features for classification purposes. The discriminatory power of the univariate secondary feature $\underline{x}$ will be related to how well each of the individual weighted elemental features $\underline{Pw}(i,j)$ discriminate between the classes. Ideally we would like all weighted elements to individually have high discrimination, but in general this is rarely the case. Some elements may provide good discrimination, while others will be poor. We can only hope that enough weighted elements possess discriminatory power to provide the secondary feature with adequate discriminatory power.

Estimates of the discriminatory power of each weighted co-occurrence matrix element $\underline{Pw}(i,j)$ can also be determined by using the class-separability measure J_B defined in equation (1.5). We apply this metric to the weighted elements derived from a training set of co-occurrence matrices, extracted from two or more classes of texture data. Denote the discrimination for weighted element $\underline{Pw}(i,j)$ as $J(i,j)$, and the set of such elemental discriminations for the entire co-occurrence matrix as $\mathbf{J}$. We call this matrix a *discrimination matrix*, first introduced in Walker, Jackway & Longstaff (1995). The discrimination matrix is rich in information, and quantitatively expresses a number of significant differences between the pair of textures being analysed:

- The discrimination matrix directly gives an indication of which elements or areas of the matrix are providing the most discrimination. This grants the potential to select only these elements to form secondary features.

- The intensity pairs to which these elements belong can be used to determine which areas of the original image provide the discrimination. That is, it localises areas within the image that differ between classes. This is demonstrated by example in Appendix B.

- Finally, because texture can be viewed from macroscopic to microscopic resolutions, the elemental discrimination information can be used to determine the resolution or *scale* at which to view the texture, to provide the most discrimination. We can achieve this by varying the co-occurrence matrix displacement parameter d, effectively using it as a scale parameter, and determining which scale provides the highest average or maximum elemental discrimination.

In Figure 4.3, we show in detail the steps involved in calculating a discrimination matrix from a training set of grey-scale images.

Figure 4.4 details the results of one such application of a discrimination measure to a set of 16×16 co-occurrence matrices from the study detailed in Walker et al. (1994). In this study, co-occurrence matrices were extracted from high-resolution images of cervical cell nuclei. For each of the two classes of nuclei (normal and abnormal), the images were requantised from 256 to 16 grey levels prior to co-occurrence matrix calculation. Using this data, the Bhattacharyya discrimination measure defined in equation (1.5) was applied to each elemental feature vector,

$$J(i,j) = J_B\left(\underline{Pw}_1(i,j), \underline{Pw}_2(i,j)\right), \quad i = 1, \ldots, 16, \; j = 1, \ldots, 16. \qquad (4.3)$$

The resulting matrix of discrimination values $\mathbf{J}$, derived from co-occurrence matrices calculated at a displacement of 3 pixel, is shown in Figure 4.4(a). In this figure, lighter intensities indicate elements which provide high class discrimination. It is interesting to note that elements with high discrimination tend to be grouped together. It is these areas which contribute most to a secondary feature's discrimination ability. It is no coincidence that in Walker et al. (1994), the feature *Inertia*, which provided the highest discrimination for this study, weighted these off-diagonal elements highly. The element weighting for the feature *Inertia* is shown in Figure 4.4(b). Note the general similarity in topography between *inertia* weighting and elemental discriminatory power shown in Figure 4.4(a).

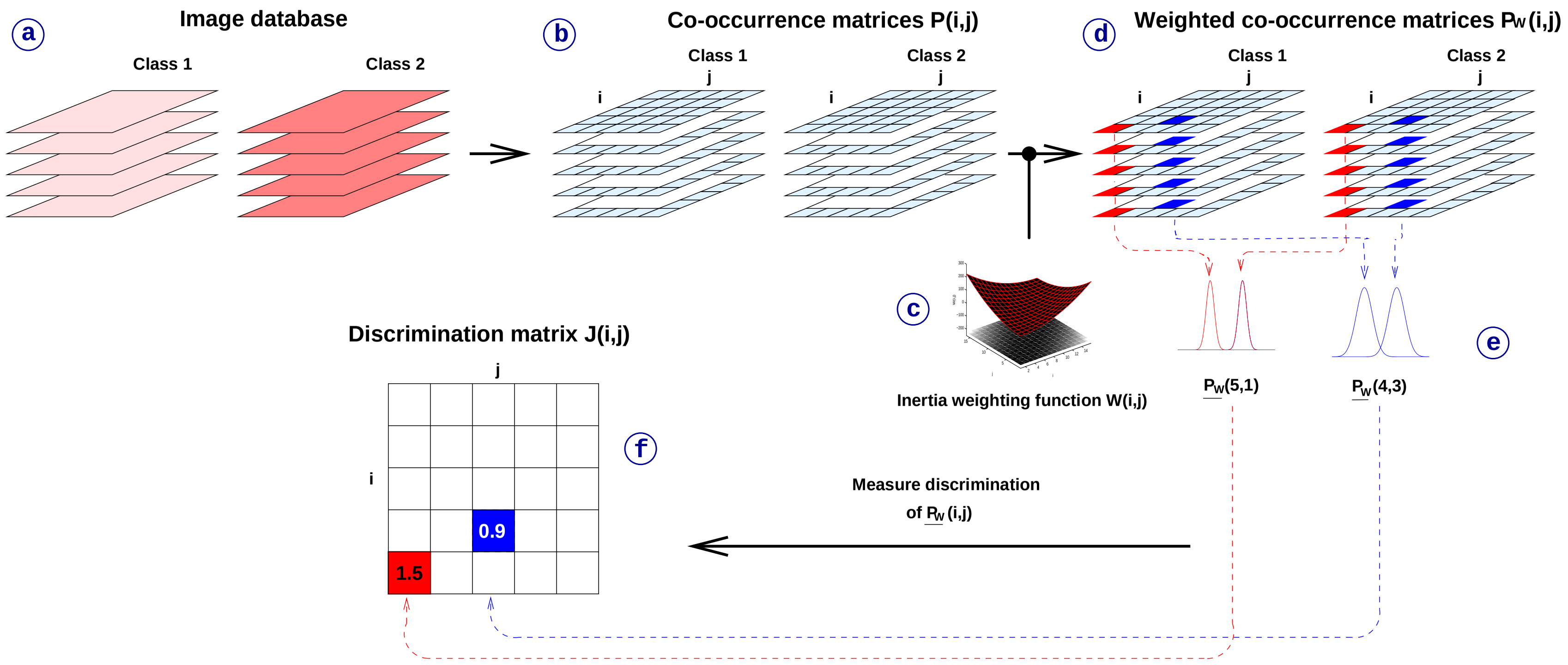
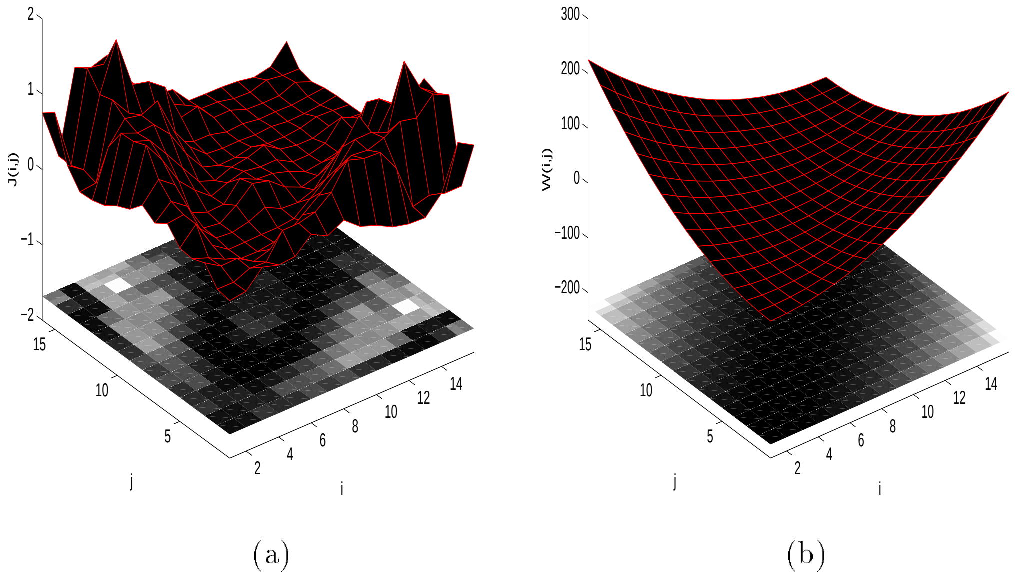

a
Image database
Class 1
Class 2
b
Co-occurrence matrices P(i,j)
Class 1
Class 2
i
j
d
Weighted co-occurrence matrices Pw(i,j)
Class 1
Class 2
i
j
c
Inertia weighting function W(i,j)
e
Pw(5,1)
Pw(4,3)
Measure discrimination
of Pw(i,j)
f
Discrimination matrix J(i,j)
i
j
0.9
1.5

Figure 4.3: Calculating a discrimination matrix from weighted co-occurrence matrices. From a set of normalised grey-scale images (a), calculate a set of co-occurrence matrices (b). Using a standard GLCM feature weighting function (c), weight each of the co-occurrence matrices, to from a set of weighted matrices (d). From the class-conditioned distributions of each weighted matrix element (e), calculate the discriminatory power using equation (1.5). The corresponding element in the discrimination matrix (f) is set to this value.



(a)

(b)

Figure 4.4: (a) A *discrimination matrix*, representing the estimated elemental discrimination $J(i,j)$ for co-occurrence matrices of cervical cells; (b) Element weighting for the feature *Inertia*. Lighter colours indicate higher weighting.

4.3 Class Separability Improvements Using Co-occurrence Element Discrimination Measures.

Based on the above discussion, there is strong evidence to suggest that discriminatory secondary features are a result of weighting highly those co-occurrence elements which provide high discrimination. This alludes to a simple process by which the discriminatory power of currently defined features can be enhanced. That is, by modifying the secondary feature's weighting function such that the weighting applied to elements with high discrimination is further increased. The discrimination matrix is ideal for this purpose, because it *directly* expresses the discriminatory power of each co-occurrence element.

In our proposed methodology, the elemental discrimination information $J(i,j)$ is used to increase a secondary feature's class separability, by modifying the weight $W(i,j)$ applied to each element $P(i,j)$. The discrimination matrix can be used directly,

$$W'(i,j) = J(i,j)W(i,j), \qquad (4.4)$$

or a smoothed version used to reduce the dependency on the training set,

$$W'(i,j) = \mathcal{F}\{J(i,j)\}W(i,j), \tag{4.5}$$

where $\mathcal{F}$ is a smoothing function and $W'(i,j)$ is now the modified weighting. We can now express a secondary feature as

$$\underline{x} = \sum_{i,j} \mathcal{F}\{J(i,j)\} \times W(i,j) \times \underline{P}(i,j). \tag{4.6}$$

We use this weighting modification to suppress the contribution to the secondary feature of elements with low discriminatory power, while increasing the contribution of elements with high discrimination. The net result should be a secondary feature with enhanced discriminatory power.

As we mentioned previously, the discrimination matrix was determined directly from a training set of co-occurrence data. To increase the generalisation ability of this technique to new, unseen data, we need to smooth this matrix before we use it as a second weighting. As Figure 4.4(a) shows, we can view a discrimination matrix as a topographic surface, and as such we can fit a smooth parametric surface to it. We can fit this surface by minimising an error metric (such as mean square error). We can also control the degree of fitting or 'generality' by varying the order of the function, say, a 2-D Gaussian or an nth-order polynomial. The number of parameters used in the modelling function varies the degrees of freedom by which the function can be fit to the surface. In the extreme, by using N_g^2 parameters in the modelling function, and thus N_g^2 degrees of freedom, the fitting function closely matches the original. Using a smaller number of parameters reduces the dependency of $\mathbf{J}$ on the training set by producing a smoother surface, and also reduces the need to store all N_g^2 discrimination matrix parameters. Usually, co-occurrence matrices are symmetric, and therefore, so is the discrimination matrix $\mathbf{J}$. Thus it is only necessary and indeed preferable to fit only part of the matrix. We factor $\mathbf{J}$ into a product of lower and upper triangular matrices $\mathbf{J} = \mathbf{LU}$. The surface is then fit to the non-zero elements of either the lower or upper triangular matrix. This surface is then duplicated in the alternate matrix.

4.4 Methodology

We now detail experiments using discrimination-matrix-dependent re-weighting to improve the class separability of co-occurrence features. Our data consisted of a set of 61 segmented nuclei of cervical cells from a previous trial (Walker et al. 1994). Of the 61 nuclei, 31 were from abnormal cells, while 30 were from normal cells. After requantising the images to 16 grey levels using the technique of Section 3.4.3, we calculated co-occurrence matrices for each of these nuclei over sixteen spatial displacements $d = 1, \ldots, 16$. We weighted each of the matrix elements using the eight feature weighting functions defined in Table 4.1, and then determined the discriminatory power of each weighted element. For each displacement d, this produced eight discrimination matrices (one for each feature function), which we used to further weight the co-occurrence matrix elements. We refer the reader to Figure 4.3 for details of our methodology for calculating the discrimination matrices.

4.4.1 Smooth Surface Fitting to Discrimination Matrix

To smooth the discrimination matrices, we fit a quadratic surface to the matrix values $J(i, j)$. We chose a quadratic surface because it was computationally simple and provided an adequate degree of smoothing. Higher-order surfaces or two-dimensional filters would be equally as valid. The general equation of a quadratic surface is (Philipp & Smadja 1994)

$$q(x, y, z) = Ax^2 + By^2 + Cz^2 + 2Dxy + 2Eyz + 2Fzx + 2Gx + 2Hy + 2Iz + J = 0, \quad (4.7)$$

and for surfaces whose major axis lies in the z-plane, i.e., biquadratic forms, the coefficients C, E, F in equation (4.7) are zero. We re-express the biquadratic form as

$$q(i, j) = Ai^2 + Bj^2 + 2Cij + 2Di + 2Ej + F \quad (4.8)$$

with $q(i, j)$ representing the fitted discrimination value at matrix element (i, j). We fit the surface by minimising the squared surface error S,

$$S = \sum_{i=1}^{N_g} \sum_{j=1}^{i} \Big(q(i, j) - J(i, j) \Big)^2, \quad (4.9)$$

where N_g is the number of grey levels in the image (the co-occurrence matrix dimension). We achieved the minimisation by differentiating equation (4.9) with respect to each of the coefficients A to F, and solving these 6 simultaneous equations for the 6 coefficients. We solved this symbolically which allowed a closed-form solution (see Appendix F). Following surface fitting, we multiplied the weighted feature value $Pw(i, j)$ by the surface weighting $q(i, j)$, effectively reducing the contribution to the secondary feature $\underline{x}$ of elemental features with low discrimination:

$$\underline{x} = \sum_{i,j} q(i,j)W(i,j)P(i,j) = \sum_{i,j} q(i,j)\underline{Pw}(i,j). \tag{4.10}$$

4.4.2 Feature Pre-processing

We used the *Ladder of Powers* normality transform (Velleman & Hoaglin 1981) detailed in Section 3.4.4, to ensure near-Gaussian class-conditioned feature distributions for each of the 128 features (eight feature functions calculated at sixteen spatial displacements). After normality transformation, we determined the new discriminatory power of the modified features, and compared it to the discriminatory power of the unmodified features.

4.4.3 Classification

To estimate the real classification error rate for this technique, we implemented a classification regime using 10-fold cross-validation (Weiss & Kulikowski 1991), as discussed in Section 1.4.7. For each of the 128 features, the data set was randomly partitioned to form ten mutually exclusive test sets with equal class proportions. For each test set, its corresponding training set consisted of the other nine test sets. Discrimination matrices were determined from only the training set co-occurrence matrices. We calculated new modified features from the training set, and then used the *Ladder of Powers* technique to transform the features, to make their distributions more Gaussian. A quadratic classifier was designed based on the transformed training set, and the transformed test set features classified. We repeated this process a total of ten times; the misclassification rate being equal to the sum of the test set misclassifications.

4.5 Results and Discussion

4.5.1 Smooth Surface Fitting to Discrimination Matrix

We detail the results of fitting a biquadratic surface to a typical discrimination matrix—that of the feature $Inertia_{d=3}$. Figure 4.5 shows the discrimination matrix to be fitted with a topographic surface, along with the resulting fitted quadratic surface. We can see that the resulting weighting function provides an adequate degree of smoothing.

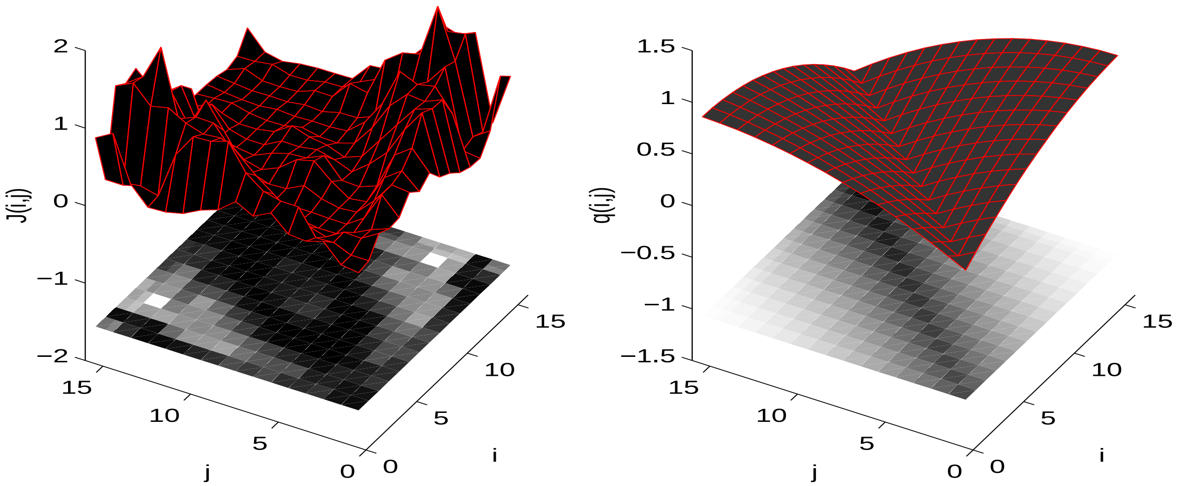


Figure 4.5: Fitting of a biquadratic surface to the discrimination matrix **J** for the feature $Inertia_{d=3}$. The figure on the left shows the discrimination matrix as a topographical surface. The figure on the right shows the resulting biquadratic surface fitted to the discrimination matrix.

4.5.2 Discrimination Matrix Weighting

Figure 4.6 shows the discrimination improvements as a result of discrimination-matrix-based re-weighting. The graphs detail class-discrimination over sixteen co-occurrence displacements $d = 1 \ldots 16$, for each of the eight features. We can see that significant increases in class separability were achieved for almost all features and displacements. The graph for the feature *variance* illustrates well the improvements in discriminatory power attainable using our method. At some displacements, discriminatory power has been increased by an order of power (from approximately 0.1 to 1). The most discriminatory feature for this dataset (*inertia*) as also attained greater class separability.

Discrimination improvements produced by biquadratic surface re-weighting are also shown in Figure 4.6. Once again we can observe that, for almost all features and displacements, significant increases in class separability have been achieved.

4.5.3 Cross-validation Results

Figure 4.7 details the classification results of our method. Each graph compares the cross-validated misclassification rate of standard GLCM features to that of the discrimination-

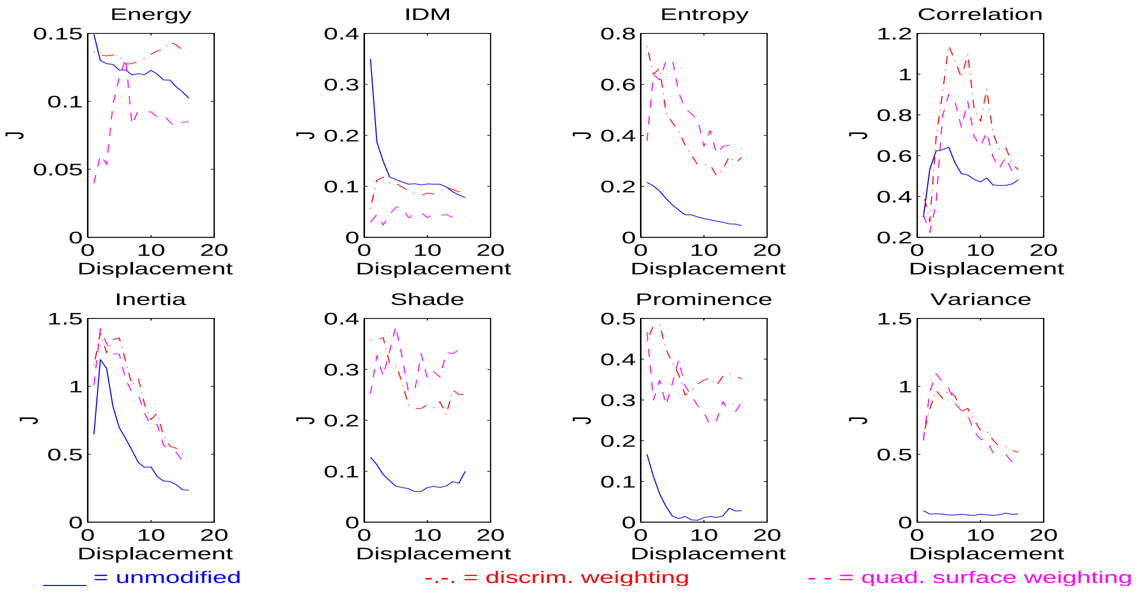


Figure 4.6: This figure compares class-discrimination J_B for unmodified, discrimination-matrix-modified, and quadratic-surface-modified features. Solid lines indicate the discriminatory power of unmodified features. Dot-dashed lines indicate the discriminatory power of features which were weighted by the discrimination matrix $J(i,j)$ without smoothing. The dashed lines indicate the discriminatory power of features which were weighted by a quadratic surface $q(i,j)$.

modified features, across the 16 spatial displacements. We can see that remarkable decreases in classification error were attained for 6 of the 8 features. By using the discrimination matrix as a further weighting function, we have successfully increased the contribution of discriminatory elemental features to the secondary feature. For the feature *variance*, an average 70% decrease in misclassification occurred following this modification. Results for quadratic surface-fitted reweighting were similar to those of discrimination matrix-reweighting, and are therefore not shown.

4.6 Conclusions

In this chapter, we presented a novel method of self-adaptive feature extraction which significantly improves the discriminatory power of many co-occurrence features. Our improvement is based on generating a discrimination matrix which indicates the discriminatory power of each element of a weighted co-occurrence matrix. Based on this discrimination matrix, we can suppress the influence of weighted elements with low

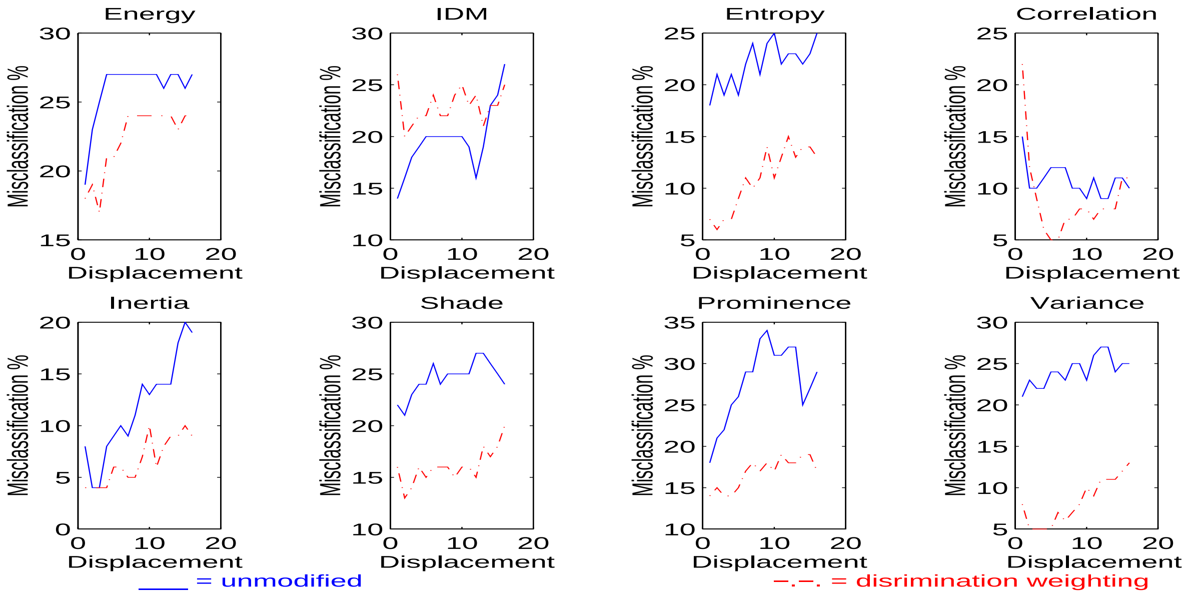


Figure 4.7: Cross-validation results for unmodified and discrimination modified features. For many features, we have attained remarkable decreases in classification error.

discriminatory power on the secondary feature. In the same way, the influence of elements with high discrimination is enhanced. We used the discrimination matrix (or a smoothed version of it) as an extra weighting function to be used in conjunction with known standard co-occurrence matrix feature weightings.

We critically appraised our method by classifying a small database of nuclear texture images. Cross-validation results indicate remarkable increases in feature discriminatory power for most features, when compared to the discriminatory power of standard GLCM features. For example, modifying the feature *variance* resulted in approximately 70% decrease in classification error. Although we used GLCM features as an example, our methodology can be applied to many co-occurrence-based method, such as those discussed in Chapter 2.

Our technique can be viewed as an *extension* to the co-occurrence method, because it used the standard feature functions as its basis for modification. As we will show in the next chapter, the fixed nature of these features is an area of weakness in the GLCM technique which is open to further improvement.

Chapter 5

Self-Adaptive Multi-Scale Feature Extraction

In this chapter we introduce a new second-order method of texture analysis called Adaptive Multi-Scale Grey Level Co-occurrence Matrix (AMSGLCM). We present the motivation for this new technique, based on the inherent limitations of standard GLCM. Our new method deviates significantly from the GLCM method in that features are extracted, not by a fixed weighting function of co-occurrence matrix elements, but by a variable summation of elements in neighbourhoods containing proven high discrimination. We critically appraised the performance of AMSGLCM and GLCM in pair-wise classification of images from visually similar texture classes, captured from natural, synthesised, and biologic origins. In these cross-validated classification trials, AMSGLCM demonstrated significant advantages over GLCM, including increased feature discriminatory power and decreased classification error.

5.1 Introduction

In the previous chapter we reviewed the current methods of extracting feature descriptors from co-occurrence matrices, and investigated the origins of discriminatory power manifestation in these secondary features. We showed that highly discriminatory secondary features were a result of summing weighted elemental features with high discrimination. However, the mechanism for extracting features from the GLCM has a number of drawbacks:

- the defined features do not extract all texture information (Trivedi et al. 1984);
- the large number of potential features which can be extracted;
- the lack of any theoretical guide to which features to extract for a particular problem (Ohanian & Dubes 1992);
- the defined features may not extract information from discriminatory areas of the co-occurrence matrix.

In this chapter we present a novel method of extracting co-occurrence matrix features *self-adaptively*. That is, the extracted features adapt to suit the specific characteristics of the classes of texture to be analysed, without human intervention. We will show that this Adaptive Multi-Scale GLCM (AMSGLCM) has a number of significant advantages over standard GLCM, including increased feature discriminatory power and decreased classification error.

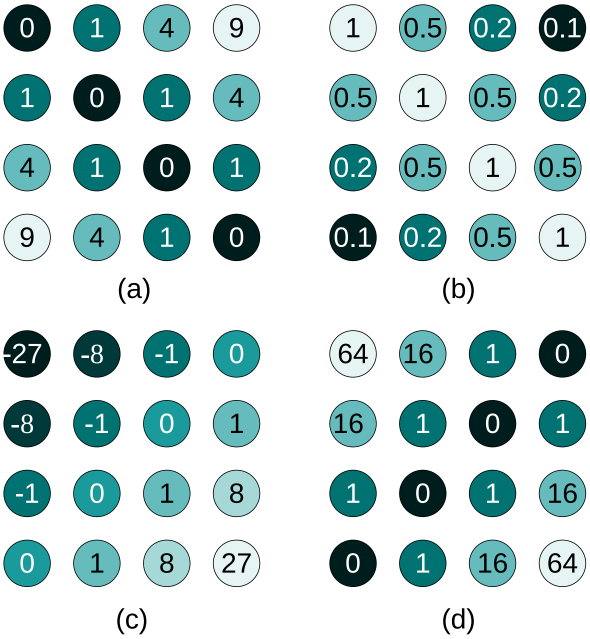
5.2 Adaptive Multi-Scale Texture Analysis

The use of co-occurrence matrices can be viewed as a form of data reduction, where images of arbitrary size and photometric resolution are transformed to a lower fixed dimensional data space of dimensions $N_g \times N_g$. It is assumed that this data reduction process captures all the relevant texture information contained within the image[1]. It is still advantageous to further reduce the order of this N_g^2-D data-space. Classical GLCM as defined by Haralick et al. (1973) involves extracting *secondary* features from the co-occurrence matrix, effectively reducing this N_g^2-dimensional space to 1 dimensional secondary features. As we mentioned previously, many secondary features are calculated by a weighted sum of co-occurrence matrix element values. For example, GLCM's feature *Inertia* weights each element $P(i, j)$ with the weighting of $(i - j)^2$. Figure 5.1 visually represents weightings for several of the most popular GLCM features. For the feature *Inertia*, we can see that only the off-diagonal elements of the matrix are weighted highly. In Chapter 4, we identified that it is the characteristics of these off-diagonal matrix elements which contribute most to the characteristics of the extracted feature *Inertia*. Thus, we can view the extraction of a feature from a co-occurrence matrix as being the summation of elements from a *localised area* or *subset* of the matrix.

In the previous chapter, we demonstrated that when classifying texture into one of two classes $\mathcal{T}_1, \mathcal{T}_2$, the *discriminatory power* of a co-occurrence matrix feature is related to the discriminatory power of the *individual matrix elements* (see our proof in Appendix A.0.3). That is, a feature formed by summing co-occurrence matrix elements with high discriminatory power will generally have high discriminatory power. We can also show that including elements with low discriminatory power in this weighted sum can severely reduce the feature's discriminatory power. We demonstrate this in Figure 5.2 for the case of two independent features $\underline{F}_{1,c}$, $\underline{F}_{2,c}$, $c = \{1, 2\}$, where c is the class index. The first variate $\underline{F}_{1,c}$ has high discriminatory power while the second variate $\underline{F}_{2,c}$ has low discriminatory power. We can see that feature 2's poor discrimination ($J = 0.03$) results in a summed feature having less discriminatory power ($J = 0.53$) than feature 1 ($J = 1.0$). Including such poor elemental features in a weighted sum can severely reduce the potential discriminatory power of a secondary feature. It would therefore be preferable to exclude features with poor discriminatory power from the summation process.

By defining various weighting functions which weight some elements or areas of the matrix more highly, Haralick and others have effectively provided a method of extracting

[1] Being a 2nd-order method, the co-occurrence matrix is unable to capture 3rd or higher-order texture information



(a) (b) (c) (d)

Figure 5.1: Feature weighting functions for GLCM. (a) Inertia; (b) Inverse Difference Moment; (c) Cluster Shade; (d) Cluster Prominence. Lighter shades indicate higher weighting.

features from more localised areas of the matrix. The weightings are, however, **independent of the type of texture being analysed**, and thus may not directly target the areas of the co-occurrence matrix that contain high discriminatory power. The fixed nature of these weighting functions results in the possibility that *no* secondary feature will possess high discriminatory power, despite the fact that certain elements may possess such power. Without some guide to which elements of the matrix to use, it is necessary to define a large number of ad hoc secondary feature functions, in the hope that one or more will possess high discriminatory power. This clarifies one weakness of GLCM and other co-occurrence-based methods—that their *fixed feature functions* extract information from a reduced subset of elements (those that are weighted highly) independent of the *usefulness* of these elements in discriminating between textures.

As we demonstrated in Chapter 4, the discriminatory power of any co-occurrence matrix elemental feature $\underline{P}(i,j)$ can be easily estimated by applying a class-separability measure, such as the Bhattacharyya or Mahalanobis distance measures (Hand 1981), to the elements derived from a training set of co-occurrence matrices, extracted from two or more classes of texture data $\mathcal{T}_1, \mathcal{T}_2$. We denoted the discrimination for matrix element $\underline{P}(i,j)$ as $J(i,j)$, and the set of such elemental discriminations for the entire co-occurrence matrix as $\mathbf{J}$. In Chapter 4, we determined the discriminatory power of

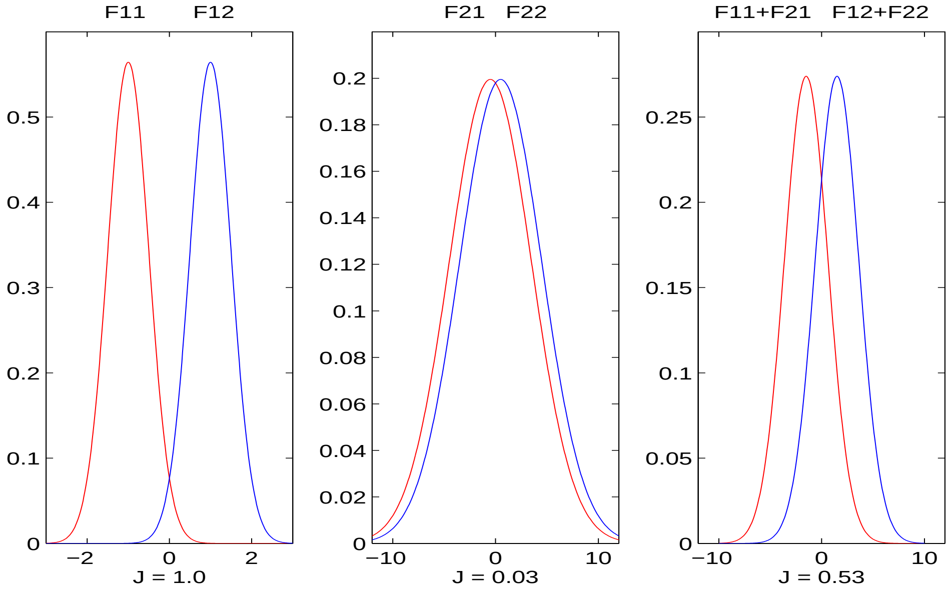


Figure 5.2: Including features with poor discriminatory power in the summation of co-occurrence matrix elemental features can result in an overall decrease in discriminatory power. The left and middle plots show the class-conditioned PDFs of two features, $\underline{F}_{v,c}$, where $v = 1, 2$ represents the variate number and $c = 1, 2$ represents the class number. The plot on the right shows the resulting class-conditioned PDFs when the two features are summed, $\underline{F}_{1,c} + \underline{F}_{2,c}$. The corresponding Bhattacharyya discrimination measure J is shown below each plot. The summed feature (representing a co-occurrence secondary feature) has poor discrimination due to feature two's low discriminatory power.

the elements after they had been weighted by a feature function, i.e., $\underline{Pw}(i, j)$. In this chapter, we wish to avoid any dependence of our technique on fixed feature functions. We therefore calculate J directly from the co-occurrence matrix elements $\underline{P}(i, j)$.

While the discrimination matrix may indicate which of the N_g^2 elemental features provide high discriminatory power, we need to further consider how to best use these elemental features to provide a new set of secondary features with dimensionality of less than N_g^2 while maintaining high discrimination[2]. While in theory it would be possible to use all the co-occurrence matrix elements individually as secondary features, the computational implications of such an approach become prohibitive. If N_d is the number of intersample displacements at which co-occurrence matrices are to be calculated, then the total number of elemental features becomes $N_d * N_g^2$. Feature set dimensionality

[2] In actuality, the use of symmetric co-occurrence matrices results in only $\frac{N_g^2}{2} + \frac{N_g}{2}$ distinct elements per matrix

reduction via discriminant analysis and even sub-optimal feature set search algorithms would be clearly computationally prohibitive, as we demonstrated in Section 1.4.5 on page 26. Take for example co-occurrence matrices extracted from sixteen grey-level images, calculated at four displacements. The resulting feature space is of $4 \times 16^2 = 1024$ dimensions!

The posed problem is:

> *How do we reduce this $N_d * N_g^2$ dimensional feature space, while still maintaining the highest possible feature set discriminatory power?*

One approach would be to simply use those elemental features whose discriminatory power is above a certain threshold T. That is,

$$\underline{\mathbf{x}} = \{\underline{P}(i,j,d) \mid J(i,j,d) > T\}, \quad i,j = 1,\ldots,N_g, \quad d = 1,\ldots,N_d \qquad (5.1)$$

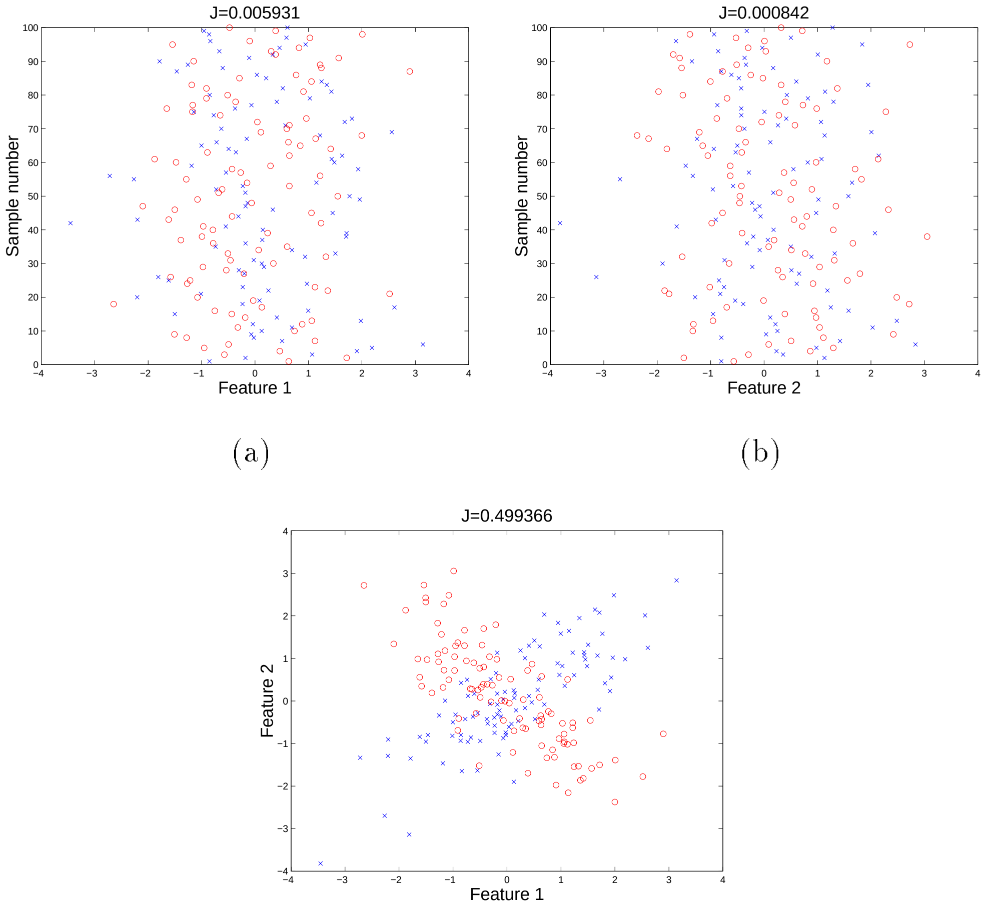
where $\underline{\mathbf{x}} = [\underline{x}_1, \ldots, \underline{x}_{N_v}]$ is now an N_v-dimensional multivariate set of elemental features containing only those elemental features with high discriminatory power, and $N_v = \#(J(i,j) > T)$. This approach has several pitfalls:

- The discrimination measure used to calculate the discrimination matrix $\mathbf{J}$ is a measure of first-order discriminatory power. It is known that features with low first-order discriminatory power may in fact possess higher-order discriminatory power when used in combination with other features. We illustrate this point in Figure 5.3 for two features which have very low first-order discriminatory power, but which possess high second-order discriminatory power. Thus, selecting features based on first-order discriminatory power and a threshold T may exclude such important elemental features.

- Many elemental features are 'noisy' estimates, because of low pixel-pair counts. This is particularly so for off-diagonal elements. Using such individual features for classification may reduce the robustness of the classifier modelled on this data.

- Many of the elemental features with high discrimination may be highly correlated and therefore redundant, because they provide little additional discriminatory information.

To elaborate this last point, we would expect that neighbouring elements within a co-occurrence matrix are correlated, because they are measures of similar image qualities, and therefore tend to possess similar discriminatory power. For example, we would expect neighbouring elements $P(1,1)$ and $P(1,2)$ to contain similar pixel-pair counts,



(a)

(b)

(c)

Figure 5.3: An example of two features which possess higher-order discriminatory power, despite having very low first-order discriminatory power. (a) The first feature variate. ‘x’ and ‘o’ marks represent class 1 and class 2 data respectively; (b) The second feature variate. Both the first and second variates possess minimal first-order discriminatory power; (c) A scatter plot of the two features. Note the high second-order discriminatory power.

compared to, say, element pairs $P(1,1)$ and $P(1,16)$. This same assumption can be applied to elements with equivalent indices, from co-occurrence matrices calculated at ‘neighbouring’ intersample displacements, e.g. $P(i,j|d)$, $P(i,j|d{+}1)$. The spatial extent of high element-pair correlation can be easily determined from training set data, by plotting a graph of correlation ρ versus element-pair displacement d_e,

$$\begin{aligned} Corr(d_e) &= E\Big\{\rho\left(\underline{P}(i_1,j_1,d_1),\underline{P}(i_2,j_2,d_2)\right) \\ &\qquad \mid \left((i_1-i_2)^2+(j_1-j_2)^2+(d_1-d_2)^2\right)^{\frac{1}{2}} = d_e\Big\}, \\ &\forall\ i_1,i_2,j_1,j_2 \in \{1,\ldots,N_g\}, \quad d_1,d_2 \in \{1,\ldots,N_d\}, \end{aligned} \tag{5.2}$$

where E and ρ are the expectation and correlation operators respectively.

Our experience suggests this to be correct. Figure 5.4 details a graph showing average element-pair correlation versus element separation determined using equation (5.2). The data used were co-occurrence matrices from a previous study (Walker, Jackway & Lovell 1995), calculated at sixteen intersample displacements $d = 1,\ldots,16$.

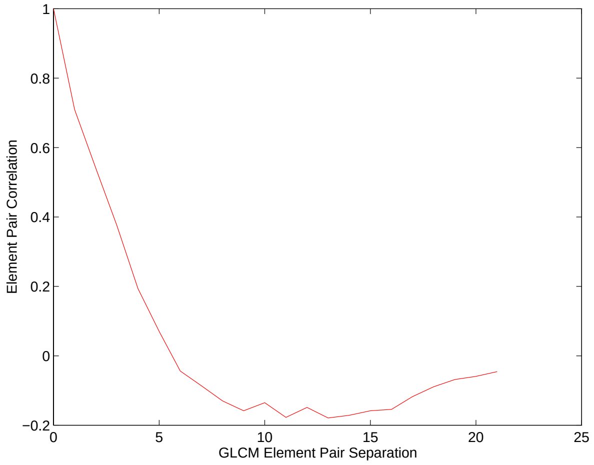


Figure 5.4: Average co-occurrence matrix element-pair correlation ρ versus element separation d_e. This figure clearly shows that neighbouring co-occurrence elements are highly correlated, while spatially distant neighbours are uncorrelated.

We can see that neighbouring elements are, on average, highly correlated, while very little correlation exists between more distant elements. More empirical evidence is suggested in Figure 5.5 where it can be seen that discriminatory power varies somewhat smoothly over the domain of the co-occurrence matrix. That is, high discriminatory power is exhibited in localised *areas* or neighbourhoods of elements, as opposed to individual isolated elements. If neighbouring elements were uncorrelated, they would be unlikely to exhibit such similar discriminatory powers.

This leads us to a second approach to dimensionality reduction. Highly correlated elemental features within the neighbourhoods of local discrimination maxima could be removed (because they provide minimal additional discriminatory power), while keeping those that are uncorrelated. A better approach would be to sum together highly correlated elements in a given neighbourhood of high discrimination. This summation also has the advantage of reducing any random noise that may be present in the feature estimates, increasing the robustness of the feature estimates.

Unfortunately, Appendix A.0.1 proves that such a summation will always result in a *decrease* in apparent discriminatory power, compared to if the elemental features were left separate as individual secondary features. However, for highly correlated features, the resulting decrease in discriminatory power following the summation is minimal. This is demonstrated in Appendix A.0.2. Moreover, if we can reduce feature set dimensionality by using this prior knowledge (that of GLCM spatial correlation) in such a way that the discriminatory power is hardly reduced, then it increases the chance of finding a robust discriminant function which is also good for the original population set, rather than just for the training set. The use of a lower-dimensional feature space results in a corresponding *increase* in class-conditioned PDF model accuracy and the possibility of increased classification performance by avoiding the 'curse of dimensionality' (Hand 1981).

Based on the above discussions, we propose a new approach to secondary feature formation by summing element values. We restrict the number of such summations for each secondary feature to those elemental features within a neighbourhood of a local maximum in discrimination. That is, each secondary feature is formed by summing only those elements within the neighbourhood of a maxima in discrimination. This helps to ensure that the secondary feature is not *contaminated* by including the remaining elements external to this neighbourhood which may contain low discriminatory power or are uncorrelated—a problem which occurs in classical GLCM. Being a fully self-adaptive weighting method, we can see that the range of possible element weightings is far greater than that of the 20 listed GLCM features defined in the literature. That is, if we consider the set of *all* possible weighted sums of co-occurrence matrix elements, our approach provides a 'more complete' set of weighting functions than the minimal set of 20 fixed GLCM feature functions.

It should be emphasised that we are now considering a three-dimensional *stack* of co-occurrence matrices for each texture image, calculated at the various intersample displacements $d = 1 \ldots N_d$, and a corresponding three-dimensional neighbourhood. The size of this 3-D neighbourhood is determined empirically using a graph of correlation versus element separation as shown in Figure 5.4. The calculation of co-occurrence matrices at N_d displacements, from grey-scale images requantised to N_g levels, results in a three-dimensional *discrimination space* of $N_d \times N_g \times N_g$ elements. We show an example of a stack of discrimination matrices in Figure 5.5, calculated from texture images detailed in Section 5.3.1. This figure demonstrates that discriminatory power can exist at not only several scales, but also at varying positions or areas within the co-occurrence matrices. It is also interesting to note that discriminatory power exists in localised *clusters* which extend across scale-space d as well as across matrix-space i, j.

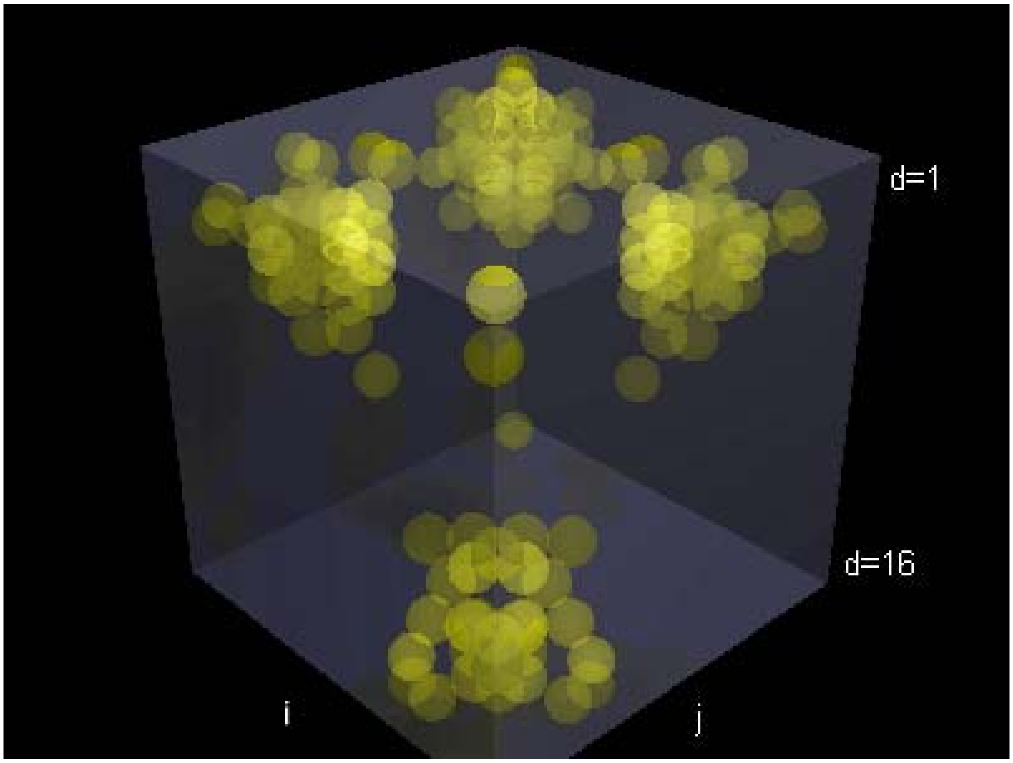


Figure 5.5: A vertical stack of 16 discrimination matrices for spatial displacements $d = 1, \ldots, 16$. Spheres with high intensity indicate corresponding co-occurrence matrix elements which possess high discriminatory power.

This three-dimensional representation of discriminatory power shows which areas of the co-occurrence matrix to use to form features and at what resolution to extract texture information. It also highlights the limitations of the fixed feature functions of standard GLCM, and other co-occurrence matrix-based methods, which often weight highly non-discriminatory areas of the matrix, or use an inappropriate resolution to extract the texture information.

In summary, forming secondary features by the weighted sum of elements from localised areas of high discriminatory power has a number of advantages, namely:

- because these elements are highly correlated, the summation results in minimal loss of texture information;

- the summation reduces feature estimate error caused by independent noise in the element estimates;

- being a summation of only elements with high discriminatory power, the resulting feature also has high discriminatory power;

- as can be seen in Figure 5.5, the number of discriminatory areas is generally limited. Hence, the majority of discriminatory information can be extracted in a minimal number of secondary features.

5.2.1 Adaptive Multi-Scale GLCM Algorithm

Our proposed AMSGLCM algorithm selects localised areas of co-occurrence matrices for summation based on *Seeded Region Growing*, a recently published methodology for segmenting images (Adams & Bischof 1994). Seeded region growing begins with the choice of a number of *seeds* (individual pixels or groups of pixels) from which the segmented regions are *grown*. The growth of these regions (that is, the inclusion of unclassified neighbouring pixels) is controlled by a criterion—usually homogeneity or similarity. After all pixels have been allocated to one (and only one) region, the resulting tessellated image is considered to be segmented.

The seeded region growing algorithm incorporated in AMSGLCM differs from the above in the following ways:

- The method was extended from 2-D to 3-D, by growing regions within the three-dimensional stack of discrimination matrices. Each region thus represents a corresponding group of co-occurrence matrix elements to be summed to form a secondary feature for classification purposes.

- Seed points are chosen iteratively (one at a time), rather than all being designated at the start of the growing process. This is because the seed criterion is the global maxima of discrimination for the current iteration, and only one global maxima exists in the 3-D discrimination space at any given iteration.

- Regions are constrained to occupy a volume no greater that the three-dimensional local neighbourhood determined by the correlation graph. This ensures that only highly correlated elements are summed to form secondary features.

The AMSGLCM algorithm is explained below, and in Figure 5.6.

Step 1: For each texture image, requantise the image to N_g intensity levels and calculate co-occurrence matrices at each of the N_d intersample displacements $d = 1 \dots N_d$.

Step 2: For each pair of texture classes, calculate an $N_g \times N_g$ discrimination matrix at each of the N_d intersample displacements. This produces a three-dimensional stack of discrimination matrices, showing the discriminatory power of each co-occurrence matrix element at each displacement.

Step 3: To determine a suitable neighbourhood size N for calculating secondary features, plot the graph of average element-pair correlation versus element displacement using equation (5.2), as shown in Figure 5.4. From the graph, choose a 3-D neighbourhood size which will contain elements whose correlation is above a threshold, say $\rho \geq 0.5$.

Step 4: From the stack of discrimination matrices, locate the global maxima of discrimination $J_{max}(i,j,d) = \max_{i,j,d} \{J(i,j,d)\}$. This element becomes a new seed for region growing, and the corresponding elemental feature $\underline{P}(i,j,d)$ becomes the basis for a new secondary feature. Call this new feature the *current sum* $\underline{CS} = \underline{P}(i,j,d)$. Consider elements $\underline{P}(i,j,d)$ such that $J(i,j,d) \in \mathrm{N}\left(J_{max}(i,j,d)\right)$, where N represents the 3-D neighbourhood determined by equation (5.2). Then,

```
if
  $J_B\{\underline{CS} + \underline{P}(i,j,d)\} \geq J_B\{\underline{CS}\}$
then
  $\underline{CS} = \underline{CS} + \underline{P}(i,j,d),$
  $J(i,j,d) = 0.$
end
```

where J_B is the discriminatory power measure of equation (1.5). That is, for each element within the three-dimensional neighbourhood of this maxima, add the corresponding elemental feature to the current sum if it does not reduce the resulting discrimination. The resulting summed feature becomes a secondary feature to be used for classification purposes. For all elements which were summed, reset the corresponding discrimination measure to zero. This ensures that the elemental feature is not summed again in another secondary feature, thus helping to make the resulting secondary features more uncorrelated and independent.

Step 5: Repeat from step 4 until a suitable number of features have been extracted.

■

Comment

While elemental features within the defined neighbourhood are, on average, highly correlated, the question of whether a pair of neighbouring elemental features are, in fact, correlated should be determined on an individual basis. This is why, in Step 4, we do not automatically sum *all* elemental features within the neighbourhood. We only sum elemental features if the discriminatory power of the resulting summed feature is not less than that of the previous summed feature. We use a discrimination measure, rather than a correlation measure, simply because it is easily calculated and directly indicates the discriminatory power of the resultant feature.

5.3 Evaluation of AMSGLCM Methodology

We now detail a methodology for comparing the classification performance of our proposed algorithm to that of classical GLCM, for the 2-class problem.

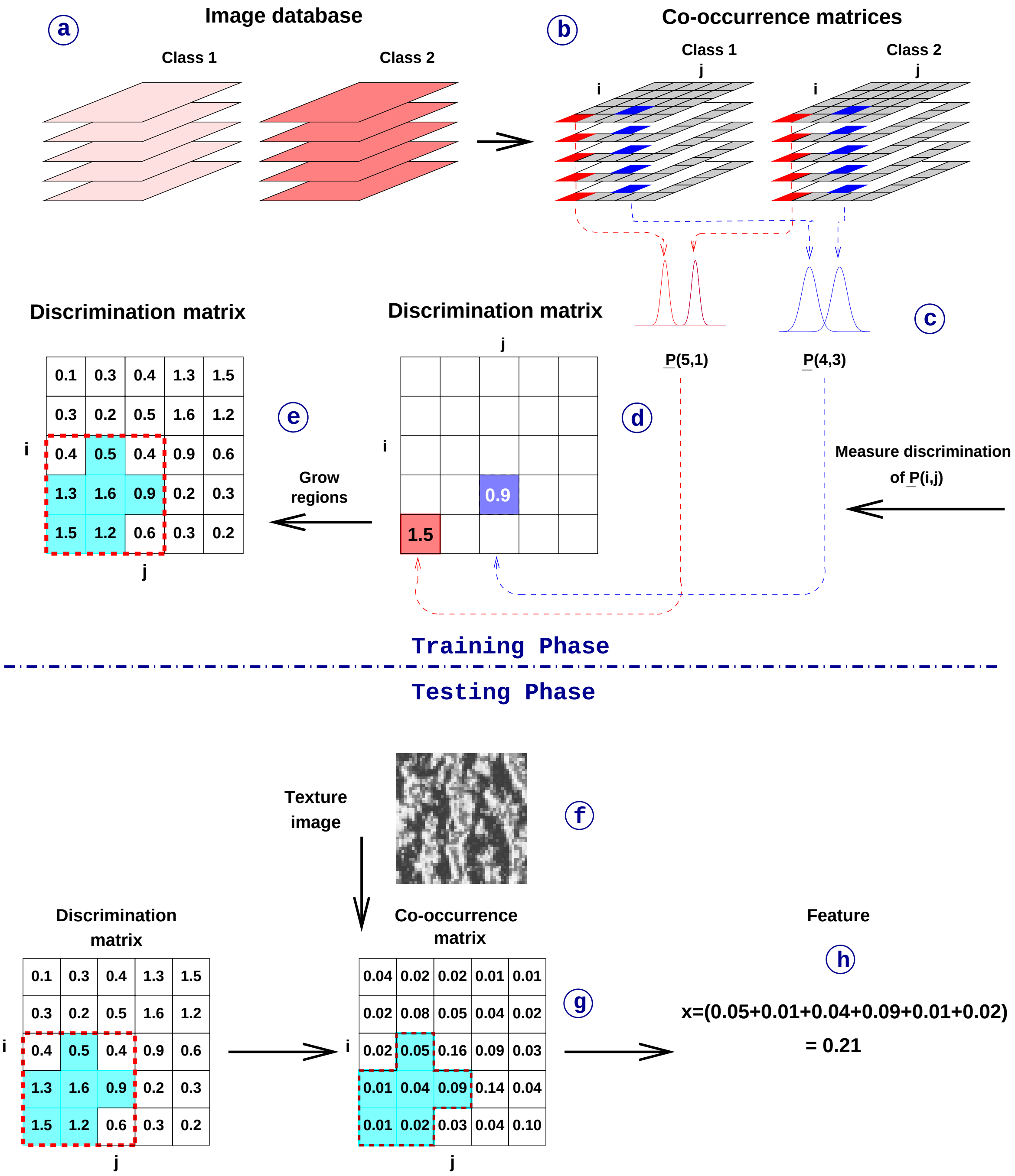
5.3.1 Texture Database

To demonstrate the power of our method, we used as test data pairs of grey-scale images containing visually similar textures from several sources. This is in contrast to much of the literature where visually distinct textures were often used in the classification experiments. Our chosen texture pairs are:

1. Brodatz images.
   Textures D4 and D9 from Brodatz's Photographic Album (Brodatz 1966). Grey-scale images of 8-bit photometric resolution, 700 × 700 pixels in size.

2. MIT's VISTEX database.
   Textures Sand.0000 and Sand.0002 from Massachusetts Institute of Technology's Vision Texture database (Picard, Graczyk, Mann, Wachman, Picard & Campbell 1995). Grey-scale images of 8-bit photometric resolution, 512 × 512 pixels in size.

3. Synthesised Brodatz texture images.
   Brodatz texture D22, synthesised using a non-parametric multi-scale non-causal Markov random field model. The images were generated by a multigrid technique using the Gibbs sampler and a novel pixel temperature function. Full details can be found in Paget & Longstaff (1996). Class 1 data uses an MRF neighbourhood size of 5 × 5 pixels, while class 2 uses a neighbourhood size of 7 × 7 pixels.



0.1	0.3	0.4	1.3	1.5
0.3	0.2	0.5	1.6	1.2
0.4	0.5	0.4	0.9	0.6
1.3	1.6	0.9	0.2	0.3
1.5	1.2	0.6	0.3	0.2

0.1	0.3	0.4	1.3	1.5
0.3	0.2	0.5	1.6	1.2
0.4	0.5	0.4	0.9	0.6
1.3	1.6	0.9	0.2	0.3
1.5	1.2	0.6	0.3	0.2

0.04	0.02	0.02	0.01	0.01
0.02	0.08	0.05	0.04	0.02
0.02	0.05	0.16	0.09	0.03
0.01	0.04	0.09	0.14	0.04
0.01	0.02	0.03	0.04	0.10

Figure 5.6: The AMSGLCM algorithm, shown in 2-D for clarity. For each image in the database (a), calculate co-occurrence matrices (b). From the pair of class-conditioned distributions of each co-occurrence matrix element (c), calculate the discrimination measure (d). Locate discrimination maxima in the resulting discrimination matrix, and grow regions within their neighbourhoods (e) as per step 4 on page 108. To extract a feature for a texture image (f), calculate its co-occurrence matrix (g), and sum the elements represented by the grown region (h).

4. Synthesised Brodatz texture images.
   Brodatz texture D9, synthesised as detailed above. Both class 1 and class 2 data use the same neighbourhood size, however, class 2 uses a texture image generated at a later iteration in the synthesis process than class 1. Photometric resolution was 8-bit and spatial resolution was 700 × 700.

5. Cervical cell nuclear chromatin texture images.
   High-resolution images of cervical cell nuclear chromatin, captured from cytologically normal and abnormal cells, at a spatial resolution of 0.12μm per pixel. Full details of this database can be found in Section 3.4.1

Figure 5.7 shows images of each of the 5 texture pairs.

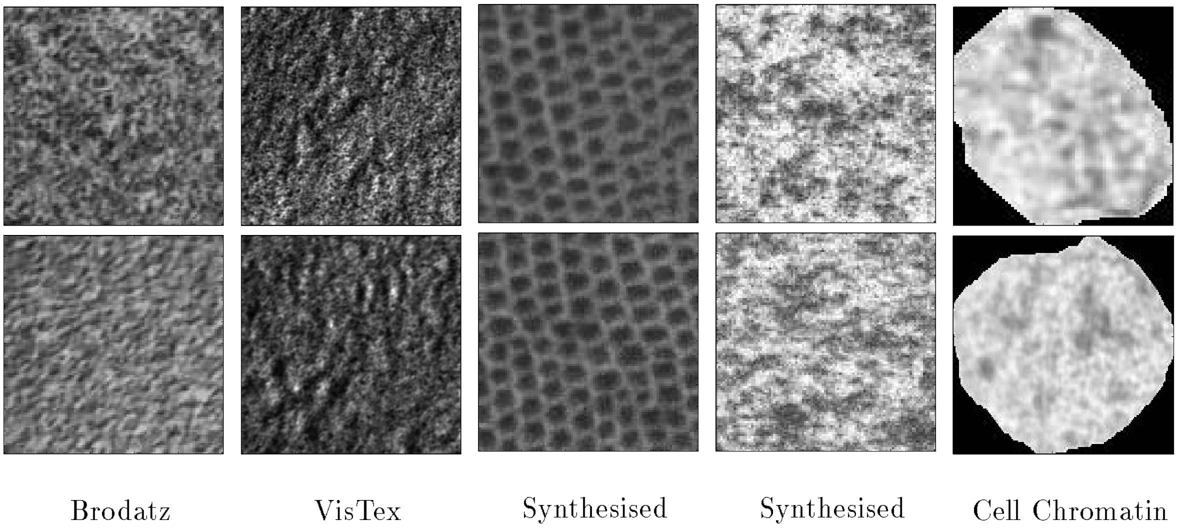


Figure 5.7: Pairs of textures used in the classification trials of GLCM and Adaptive Multi-Scale GLCM. The top row of images are class 1 textures, while the bottom row are class 2.

For texture pairs 1 to 4, we extracted a total of 100 image tiles for each class by randomly sampling the original image. To obtain enough data, tiles were allowed to overlap. For each texture pair, the tile size was varied until a measurable misclassification rate was achieved. This was necessary because using a fixed tile size for all four texture pairs resulted in some textures being classified perfectly using both standard and adaptive GLCM.

For texture pair 5, we used the same database of 58 normal and 59 abnormal chromatin texture images discussed in Section 3.4.1. Due to the small size of each image (as small as 20 × 20 pixels), the entire image was used in the feature extraction process.

5.3.2 Feature Extraction

For standard GLCM, we extracted the eight secondary features defined in Table 4.1 at sixteen intersample displacements $d = 1, \ldots, 16$, giving a total of 128 features.

For AMSGLCM, to ensure a fair classification comparison, we calculated discrimination matrices at each of the above sixteen displacements. Using the algorithm defined in Section 5.2.1, we extracted secondary features from neighbourhoods of the first 128 local maxima in discrimination, starting with the global maximum in discrimination, and repeatedly searching for the new global maxima after each iteration.

We used the normality transform introduced in Section 3.4.4 to pre-process the features prior to discriminant analysis and classification.

5.3.3 Feature Selection

Once again, we used the add-2/subtract-1 algorithm described in Section 1.4.5 to find an optimal set of features at reduced dimension. For texture pairs 1 to 4, we reduced the 128-dimensional feature space to fifteen dimensions, based on the rule of thumb of six exemplars per feature per class (Foley 1972). Using the same rule, we reduced the dimensionality of features for texture pair 5 (cell data) to ten dimensions, due to the reduced number of exemplars available. We used the parametric Bhattacharyya distance measure defined in equation (1.5) to determine feature set discriminatory power.

5.3.4 Classification

For each analysis method, we used pair-wise ten-fold cross-validated classification (Section 1.4.7) to provide a robust estimation of the real classification error. We randomised the data set and partitioned it into ten approximately-equal test sets. The complement of each test set was used to train the quadratic classifier, and the performance of the resulting classifier was evaluated on the test set. It was necessary to include the feature selection process in this step, i.e., a new feature set was selected based only on the training set data. We then classified the corresponding test set features, using the quadratic discriminant function defined in equation (1.6).

5.4 Results and Discussion

In all five classification trials, AMSGLCM outperformed GLCM with significantly lower classification error rates. This is clearly shown in Table 5.1, which details the minimum error rates achieved by each method and the corresponding number of features used to obtain those rates. The final column details the extent of the improvement in classification achieved by AMSGLCM—a reduction in errors ranging from 12 to 35%[3]. Figure 5.8 provides a visual comparison of the performance attained by the two methods.

	GLCM		AMSGLCM		
Texture Pair	Minimum Error	Number of Features	Minimum Error	Number of Features	Error Decrease
1	5.0%	10	4.0%	5	20%
2	3.5%	6	3.0%	7	14%
3	7.0%	4	4.5%	6	35%
4	12.0%	6	10.5%	15	12%
5	12.0%	1	8.0%	2	33%

Table 5.1: Comparison of classification performance of the AMSGLCM algorithm to that of standard GLCM.

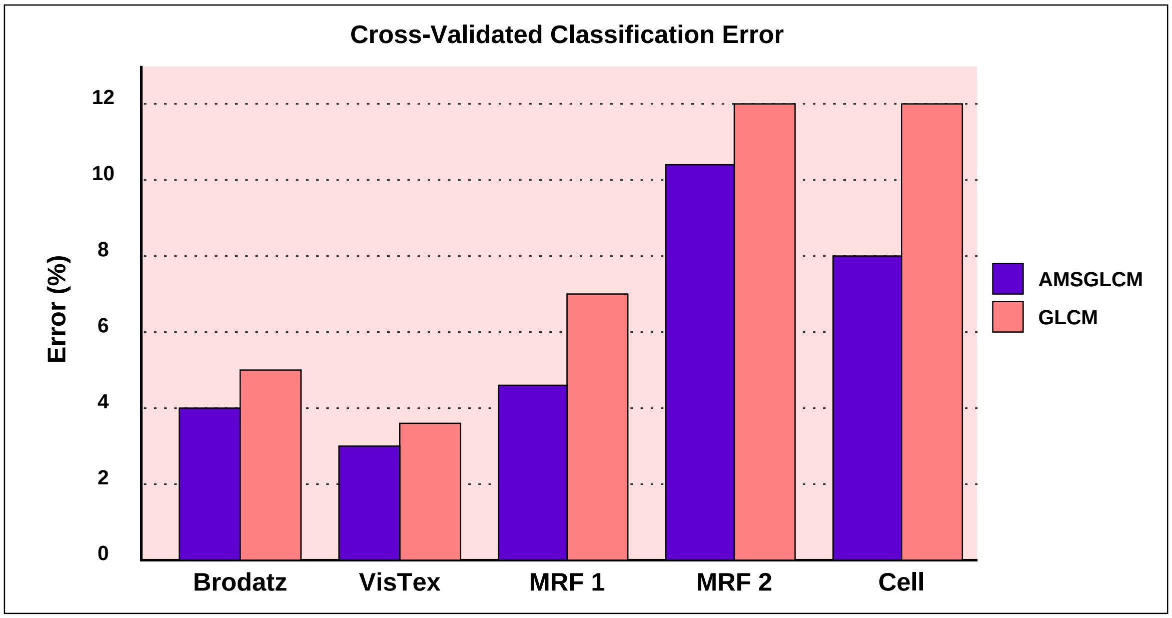


Figure 5.8: Graphical comparison of classification performance of the Adaptive Multi-Scale GLCM algorithm to that of standard GLCM.

In most real-world systems using texture classification, it is usual to include features from several analysis methods. Usually, only the first few highly discriminatory features

[3] relative error, defined as $\frac{\mathrm{Error}_{\mathrm{GLCM}} - \mathrm{Error}_{\mathrm{AMSGLCM}}}{\mathrm{Error}_{\mathrm{GLCM}}}$.

from each method would be included in the final classification system. Thus, it is important to consider not only the minimum error rates, but also the number of features required to achieve this rate. The graphs of Figure 5.9(a-e) show that AMSGLCM attained, on average, lower misclassification rates for all feature sets from 1 to 10 dimensions. We summarise this information in Table 5.2.

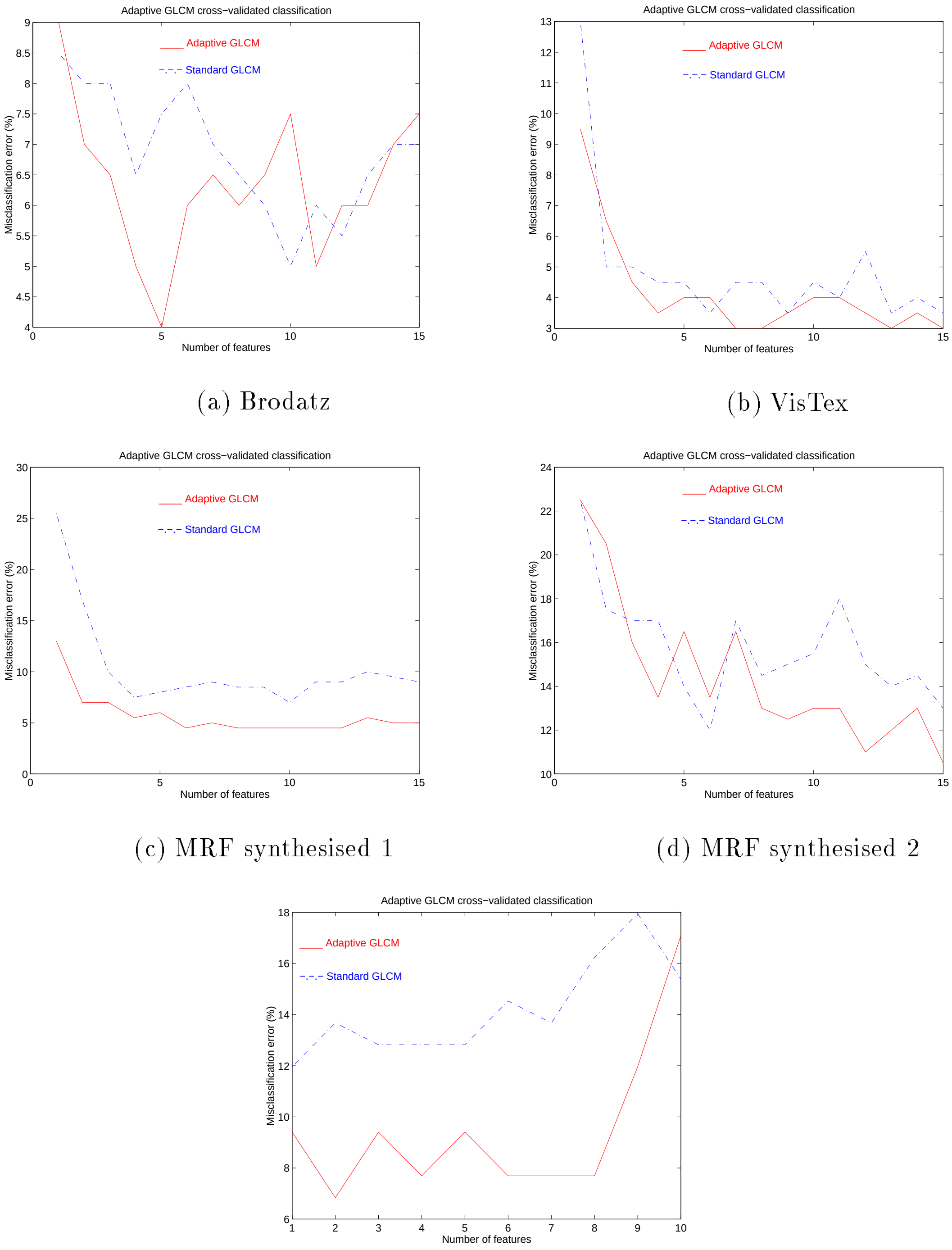
Texture Pair	Feature set dimensions									
	1	2	3	4	5	6	7	8	9	10
1	L	W	W	W	W	W	W	W	L	L
2	W	L	W	W	W	L	W	W	T	W
3	W	W	W	W	W	W	W	W	W	W
4	T	L	W	W	L	L	W	W	W	W
5	W	W	W	W	W	W	W	W	W	W
Total Win/Lose	3W/1L	3W/2L	5W/0L	5W/0L	4W/1L	3W/2L	5W/0L	5W/0L	3W/1L	4W/1L

Table 5.2: Win-Lose comparison of classification performance of AMSGLCM versus standard GLCM, for feature sets of 1 to 10 dimensions (W=win, L=lose, T=tie).

In this table, we awarded a 'win' to the AMSGLCM algorithm if, at the particular feature set dimensionality, its misclassification rate was lower than that of GLCM, or a *lose* if its misclassification rate was greater. It can be seen that AMSGLCM won three or more of the five classification trials, for all feature sets up to at least 10 dimensions. Therefore, in a classification system comprising of several composite analysis methods, the inclusion of AMSGLCM should be considered over GLCM.

We found that AMSGLCM provided secondary features with, on average, higher first-order discriminatory power than standard GLCM. This is represented by histograms in Figures 5.10(a-e), a comparison of the first-order discriminatory power of all 128 AMSGLCM and GLCM features for the five classification trials. The 'x' axis represents the first-order Bhattacharyya discrimination measure as defined in equation (1.5), and the 'y' axis represents the number of features which possessed this discrimination. Notice the general shift to the right in the distributions of AMSGLCM discriminatory power (compared to GLCM), indicating features with higher average discriminatory power. Also notice the vastly reduced number of AMSGLCM features which possess *no* discriminatory power.

AMSGLCM was able to provide secondary features with higher discriminatory power because it exploited higher-level knowledge (the information in the discrimination matrices). This allowed direct targeting of localised areas of the co-occurrence matrices with proven discriminatory power for feature formation, and provided the inherent ability of the method to *adapt* to the type of texture being analysed. The method of localised summation excludes elemental features with low discriminatory power, or those which are uncorrelated (and therefore possibly more *independent* and better used as separate features).



(a) Brodatz

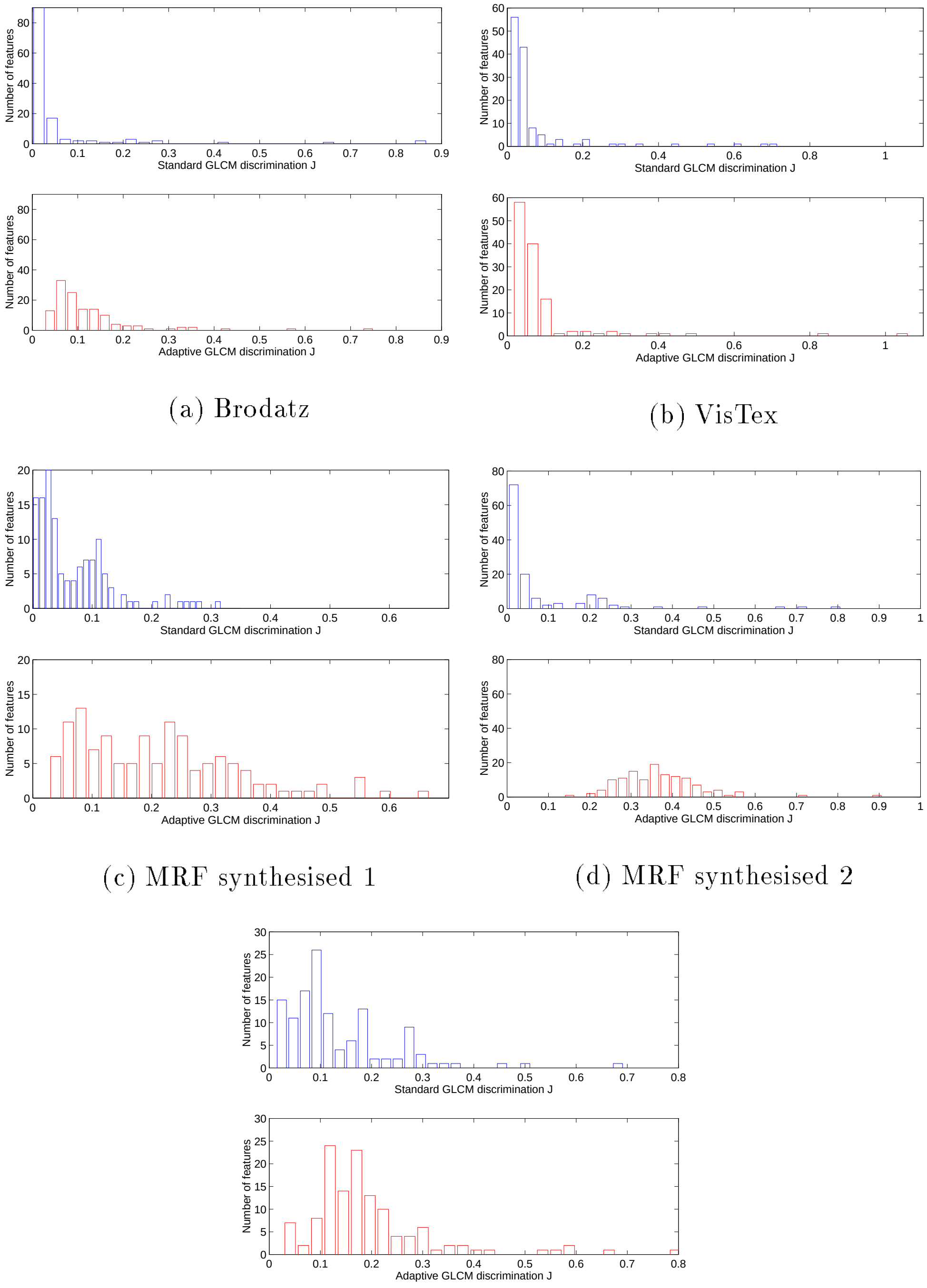
(b) VisTex

(c) MRF synthesised 1

(d) MRF synthesised 2

(e) Cell

Figure 5.9: Cross-validated error rate versus feature set size for the five classification trials (a-e).



(a) Brodatz

(b) VisTex

(c) MRF synthesised 1

(d) MRF synthesised 2

(e) Cell

Figure 5.10: Histograms of GLCM and AMSGLCM feature discriminatory power: (a) Brodatz textures D4 and D9; (b) VISTEX textures Sand.0000 and Sand.0002; (c) MRF synthesised texture D22; (d) MRF synthesised texture D9; (e) Cell chromatin texture. Notice the general shift to the right in the distribution of AMSGLCM discriminatory power (compared to GLCM), indicating features with higher average discriminatory power. Also notice the vastly reduced number of AMSGLCM features which possess *no* discriminatory power.

Computationally, the AMSGLCM and GLCM algorithms are similar and have many common steps. For AMSGLCM, a number of further steps are necessary, namely:

- the calculation of discriminatory power for each co-occurrence matrix element,
- the determination of 3-D neighbourhood size via a correlation graph, and
- the formation of secondary features via a decision metric.

These additional steps do increase the computational burden. However, after training and evaluation, classification times for new unseen data are similar to existing techniques. Moreover, the capacity of the AMSGLCM method to isolate neighbourhoods of high discriminatory power and form features from just these areas, means it extracts more discriminatory information in fewer features than GLCM. This, in turn, markedly reduces feature selection/discriminant analysis computation time, offsetting most of the increase due to additional steps.

5.5 Conclusions

We have introduced a new second-order method of texture analysis called Adaptive Multi-Scale Grey Level Co-occurrence Matrix, based on the well-known GLCM technique of Haralick et al. (1973). Our method deviates significantly from the original in that features are extracted, not via a fixed two-dimensional weighting function of co-occurrence matrix elements, but by variably summing elements in three-dimensional neighbourhoods containing proven high discrimination. The ability to detect such potentially useful regions within the co-occurrence matrix, by using a discrimination matrix, results in a number of significant advantages over the traditional GLCM method, namely:

- the features extracted using AMSGLCM have, on average, higher discriminatory power than the standard GLCM features defined in published literature (Figure 5.10);
- using AMSGLCM features, on average, provides lower misclassification error than that of standard GLCM (Figure 5.8);
- for a given misclassification error, the number of AMSGLCM features required is generally less than that of standard GLCM features (Figures 5.9);

We demonstrated these advantages in trials comparing the performance of AMSGLCM and GLCM in classifying a range of visually similar images captured from natural, synthetic and biologic origins. AMSGLCM achieved significantly lower classification error rates and increased feature discriminatory power. The multi-resolution technique is fully self-adaptive and requires no human intervention. It self-adapts to the specific types of texture being analysed (locally optimised) yet it can be applied to a wide range of textures (globally adaptive).

Once again, it is important to mention the general applicability of our method. Although the technique was trialed on *grey level* co-occurrence matrices, it is not only limited to GLCM. AMSGLCM can be extended to *any* analysis method where a series of matrices are determined via constraint parameters, such as NGLDM (Sun & Wee 1983), Yogesan's GLEM and GLVM methods (Yogesan 1995) or GCM (Davis et al. 1979).

Chapter 6

Optimised, Self-Adaptive Feature Extraction Using Genetic Algorithms

In the previous chapter we introduced our first fully self-adaptive multi-resolution technique for analysing texture in images. Here we present a slightly different approach which attempts to extract an optimised set of highly discriminant features without the need for explicitly calculating a discrimination matrix. We use Genetic Algorithm optimisation to produce a set of features whose worth is evaluated by first-order discriminatory power and feature correlation considerations. Once again, we critically appraised the performance of our Genetic Algorithm optimised GLCM (GAoGLCM) method and GLCM in pair-wise classification of images from visually similar texture classes, captured from synthesised and biologic origins. In these cross-validated classification trials, our method demonstrated significant advantages over GLCM, including increased feature discriminatory power and decreased classification error.

6.1 Introduction

The Adaptive Multi-Scale GLCM algorithm of Chapter 5 provided a multi-resolution approach to texture analysis which extracted highly discriminant features with low correlation compared to standard GLCM features. We used a set of discrimination matrices to locate discriminatory co-occurrence matrix elements. We then used localised groups of these elements to form discriminant secondary features, by summing these elements. In fact, we can consider the formation of AMSGLCM features as being a *weighted* sum of *all* elements, just like standard GLCM. In the case of AMSGLCM, however, this weighting is *binary*, i.e., '0' or '1'. That is, elements to be summed were given the weighting of '1', while all other elements were given a weighting of '0'. This parallelled the approach of standard GLCM where secondary features are also weighted sums of all co-occurrence elements. However, our method allowed the *location* and *size* of high weighting ('1's) to be adapted to the types of texture being analysed. In GLCM, these locations are fixed.

These binary weighting functions are somewhat restrictive, in that the weighting can only be either '1' or '0', and '1' weightings are restricted to locations within a fixed neighbourhood of high element correlation. Both of these restrictions can be removed by using a parameter-controlled continuous weighting function. Such a function could allow precise control of weight values by allowing each weight to take on a real value, $W(i,j,d) > 0$. Also, greater flexibility in weighting localisation could be attained by adjusting the parameters upon which the function is dependent. But how do we control the values of these parameters without the need for human intervention? Also, is it possible to control these weighting functions in such a way that the extracted features are optimised for high discriminatory power and low correlation?

The Genetic Algorithm is ideally suited to such parameter-based optimisation of a highly complex solution space. In this chapter we will use a GA to optimise the locations and shapes of a set of Gaussian weighting functions. Each function will be used to form one secondary feature, by a weighted sum of all co-occurrence matrix elements. Gaussian weighting functions allow precise yet simple weighting control using a minimum of parameters—mean and variance-covariance. We will once again use a three-dimensional stack of co-occurrence matrices, calculated across several spatial displacements, for feature extraction. Therefore, each Gaussian weighting is also three-dimensional. We will now review genetic algorithm optimisation in detail.

6.2 Genetic Algorithm Optimisation

The Genetic Algorithm is a recent and novel optimisation algorithm whose mechanisms mirror processes observable in natural evolution (Holland 1975, Bethke 1981). In its most basic form, the GA consists of three functions:

1. genetic selection,
2. genetic operation,
    - crossover
    - mutation
3. genetic replacement.

These functions work on a population of candidate solutions called *chromosomes*. Figure 6.1 details the *life-cycle* of a GA in terms of these operations.

The life-cycle of a GA begins with a population of chromosomes upon which we apply the above operations. Each chromosome consists of a string of *genes* which encode input variables to the problem for which a solution is sought. An individual gene usually takes a binary or real value. In this work, we only consider binary-valued genes, as they are the most widely used in the literature. Given that each chromosome represents a trial solution to a problem, we need to quantitatively measure how optimal each solution is. We achieve this by assigning a measure of chromosome *worth* to each chromosome. The worth of a chromosome is expressed using an *objective* function O, or alternatively via a related function called *fitness*. A chromosome's worth directly determines its chance of being represented in a new generation.

We randomly generate an initial population of chromosomes, and evaluate their corresponding worth. From this initial population pool, we select a subset population of

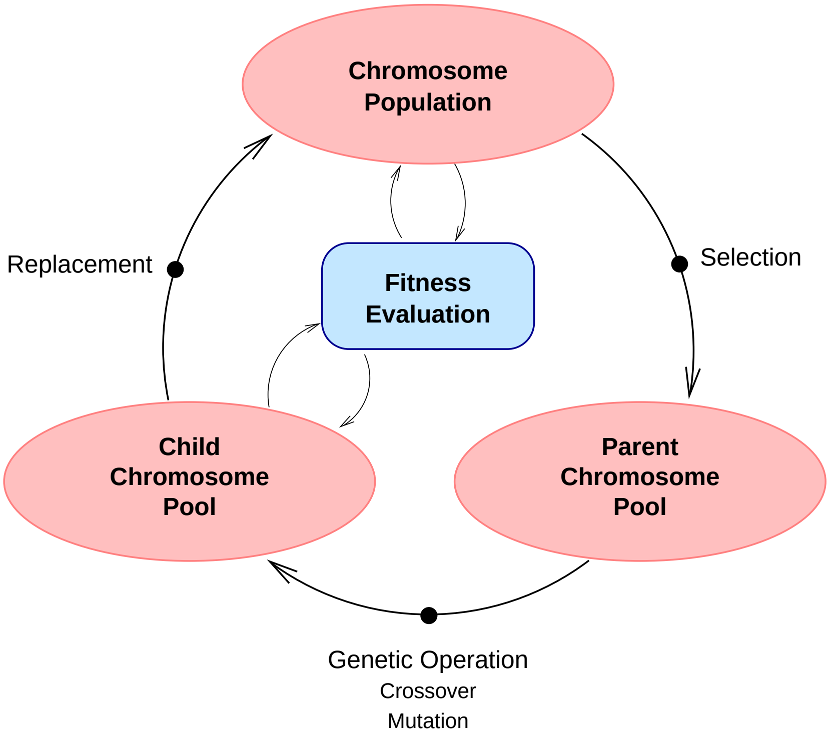


Figure 6.1: The life-cycle of a Genetic Algorithm.

parent chromosomes with high chromosome worth. It is from this subset of parent chromosomes that offspring or *child* chromosomes are generated, via the genetic operations of crossover and mutation. We evaluate the worth of each offspring, and use it to form a new population by replacing 'weaker' parent chromosomes, based on a replacement strategy. GA operations involve random processes, and many are weighted random processes where the probability of an outcome is weighted by an input quantity such as chromosome worth. Such a weighting strategy helps to ensure the 'survival of the fittest' chromosomes, with the result that each successive cycle or generation of chromosomes has, ideally, increasing average and maximum worth.

We continue the above cycle of parent selection, child generation, and replacement, until a minimum acceptable fitness is attained by one of the chromosomes, or until a set number of cycles has concluded. The chromosome of highest worth in the terminating population represents a highly optimised solution to the problem.

In Figure 6.2, we show a simple application of a GA to the optimisation of an independent variable x, such that x maximises the objective function $O(x) = -x^2 + 7x$. In this example, we encode a population of four chromosomes with random initialisations of the input variable $x \in \{0, 0.5, 1, \ldots, 7.5\}$, using 4-bit binary encoding.

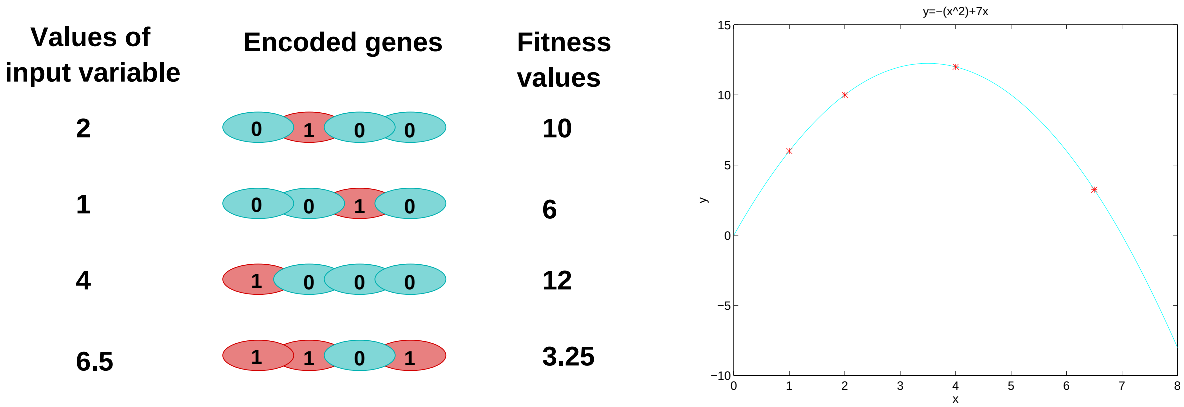
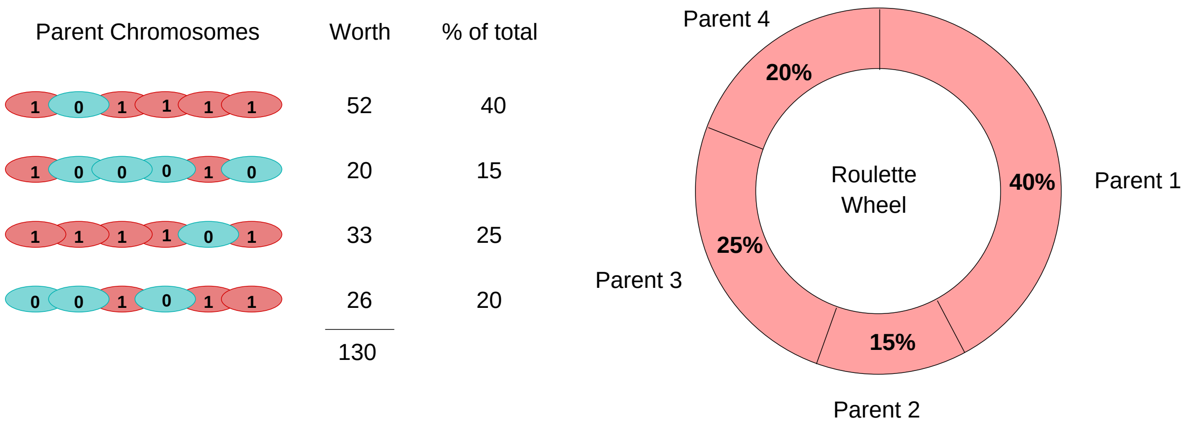


Figure 6.2: Chromosome encoding and fitness evaluation. Fitness is evaluated directly from the objective function $O(x) = -x^2 + 7x$.

From this initial pool, parent chromosomes are selected according to their worth or fitness. We assign to each chromosome a fitness value f determined directly from the objective function $O(x) = -x^2 + 7x$. Because the value encoded by the third chromosome ($x = 4$) is closest to the optimal value ($x_{\text{max}} = 3.5$), it achieves the highest fitness value and thus has the greatest chance of passing on genetic information to succeeding generations of chromosome populations. We discuss next the process of parent selection based on fitness.

6.2.1 Genetic Selection

Selection or *reproduction* is a mechanism where chromosomes from the current population are chosen as parents for possible 'mating', according to their fitness values f. The idea is to give parent chromosomes with high fitness a greater chance to pass on their genetic information to subsequent offspring. This mechanism simulates 'survival of the fittest' and natural selection in nature. The process of parent selection is most commonly achieved using *Roulette Wheel Selection* (Goldberg 1989), where a roulette wheel's slot sizes are proportional to each chromosome's fitness value. Each spin of the wheel selects a parent from the current population of chromosomes. The process is shown in Figure 6.3. The proportionate worth of each chromosome (expressed as a percentage) is determined by dividing each chromosome's worth by the total worth of all chromosomes. We can see from Figure 6.3 that the chance of a chromosome being selected as a parent from this biased wheel is directly proportional to its worth.



Parent Chromosomes	Worth	% of total
1 0 1 1 1 1	52	40
1 0 0 0 1 0	20	15
1 1 1 1 0 1	33	25
0 0 1 0 1 1	26	20
	130	

Figure 6.3: The process of genetic selection. The roulette wheel is partitioned in accordance with the worth of each chromosome. The chance of a chromosome being selected for reproduction is therefore directly proportional to its worth.

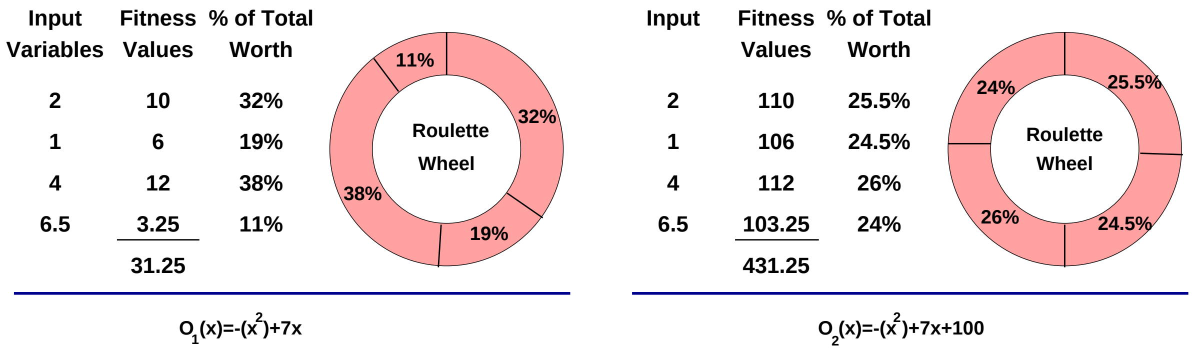
6.2.2 Fitness Techniques

In the previous example shown in Figure 6.2, we determined the fitness of a chromosome directly from an objective function—the function to be optimised. While such an approach is acceptable for this example, it can result in poor GA performance in other optimisation problems. This is best illustrated by example—see Figure 6.4. We take the case of optimising two similar functions, $O_1(x) = -x^2 + 7x$ and $O_2(x) = -x^2 + 7x + 100$. In Figure 6.4, we show the fitness values and resulting roulette wheel biases for both functions, where fitness is determined directly from the objective functions. For objective function O_2, we find that the small range of fitness values *as a proportion of total fitness*, results in all chromosomes having almost equal chance of selection as parents.

To increase the range of fitness values, we use *fitness techniques* to re-map objective function values. One of the simplest fitness techniques, called *windowing*, assigns a fitness value f to the ith chromosome $\mathcal{C}_i$ proportional to the difference between its objective value $O(\mathcal{C}_i)$ and that of the weakest chromosome $O_{\min} = \min_j O(\mathcal{C}_j)$:

$$f_i = O(\mathcal{C}_i) - O_{\min} + k, \tag{6.1}$$

where k is a suitable constant. This normalisation process ensures an appropriate range of fitness values are assigned, irrespective of the range of objective values. We demonstrate this in Figure 6.5.



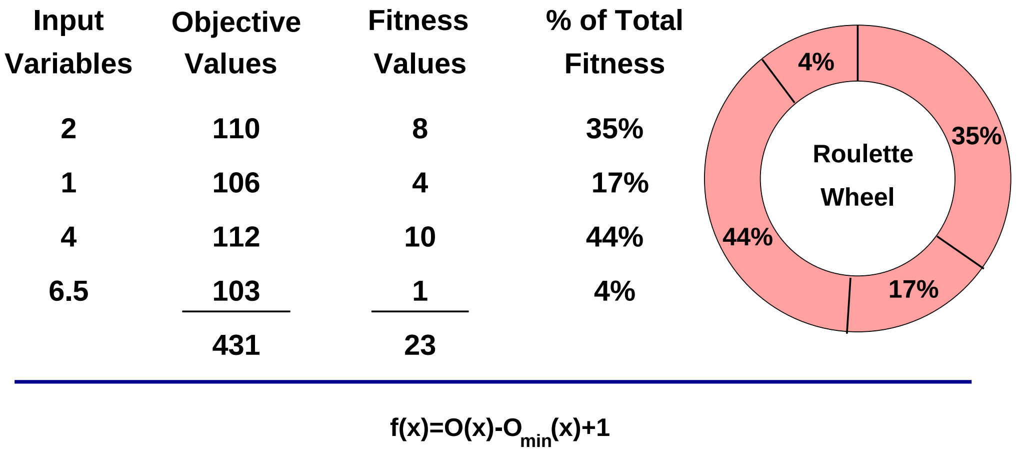
Input Variables	Fitness Values	% of Total Worth
2	10	32%
1	6	19%
4	12	38%
6.5	3.25	11%
	31.25	

$O_1(x)=-(x^2)+7x$

Input	Fitness Values	% of Total Worth
2	110	25.5%
1	106	24.5%
4	112	26%
6.5	103.25	24%
	431.25	

$O_2(x)=-(x^2)+7x+100$

Figure 6.4: Using objective values directly as fitness values can result in inappropriate biasing of the roulette wheel. For objective function O_2, the weakest and strongest chromosomes have approximately equal fitness values.



Input Variables	Objective Values	Fitness Values	% of Total Fitness
2	110	8	35%
1	106	4	17%
4	112	10	44%
6.5	103	1	4%
	431	23	

$f(x)=O(x)-O_{min}(x)+1$

Figure 6.5: An example of the *windowing* technique, applied to objective function O_2 of Figure 6.4. We choose k to be approximately 10% of the total range of objective values, i.e., $(112 - 103) * 10\%$. Compared to the roulette wheel for O_2 of Figure 6.4, we now have a larger spread of fitness values.

6.2.3 Genetic Algorithm Operations

Crossover

Crossover or *recombination* is a simple process where pairs of parent chromosomes ‘swap’ groups of their genes. This process enhances the probability that fitter parents pass on beneficial subsets of their genes, thereby increasing the chances that children of higher fitness are produced. From a solution-space search viewpoint, where each chromosome represents a single point in solution-space, the crossover operation ensures that new areas of solution-space are searched.

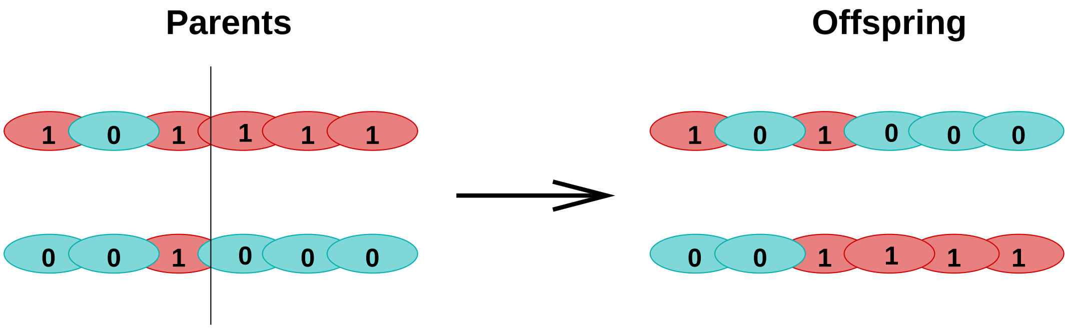


Figure 6.6: The inheritance of genetic information from both parents is achieved via crossover

The most common and simplest form of crossover operation is *single-point* crossover, shown in Figure 6.6. In this operation, we form a pair of new children by randomly selecting a single crossover point along a pair of randomly chosen parent genes, and then swapping the portions of the chromosomes beyond this point. The rate of occurrence of a crossover operation can be controlled by a probability term P_c, whose value typically lies in the range 0.6 to 1. If no crossover occurs, children become direct copies of their parents. The choice of P_c is a compromise between 'exploitation' of localised areas of solution-space, and 'exploration' of new areas. High values of P_c help to ensure new areas of the solution-space are searched. However, this means that fit parents may no longer be represented in the next generation. Lower values of P_c allow more children to be exact copies of their parents which, along with the mutation operation, allow more localised exploration of solution-space and thus exploitation of more optimal solutions. However, low P_c values can create too many offspring which are simply copies of their parents. This can result in slow GA convergence or stagnation due to the lower rate of generating unique offspring.

Mutation

In GAs, mutation is the occasional random alteration of a single gene, and is performed on a bit-by-bit (or gene-by-gene) basis. Its purpose is two-fold. Firstly, it helps successive generations to acquire new genetic information, which may not be added via simple crossover operations alone. For example, if the n-th bit of *all* chromosomes in the current population is the same value, say, '0', the crossover operation will never result in this bit being set to '1'. Applying the mutation operation will, given time, introduce this new information. Secondly, mutation helps to maximise near-optimal solutions via more localised searches.

For binary-encoded chromosomes, a mutation occurring with probability P_m complements the binary value of a single gene, as shown in Figure 6.7. Each bit of the chromosome undergoes mutation if its probability test is passed. Thus, for a chromosome of bit length $L_{\mathcal{C}}$, the average number of gene mutations N_m is

$$N_m = P_m \times L_{\mathcal{C}}. \tag{6.2}$$

A second form of mutation occurring with probability P_m replaces a bit with a randomly generated binary value of equal probability $p(0) = p(1) = 0.5$. We can see that this rate of mutation is effectively half the previous rate, as there is equal chance that the new bit value will be the same as the current value, thus,

$$\begin{aligned} N_m &= P_m \times p(\text{transition}) \times L_{\mathcal{C}} && (6.3) \\ &= 0.5 \times P_m \times L_{\mathcal{C}}, && (6.4) \end{aligned}$$

where $p(\text{transition})$ is the probability of the generated bit being the binary complement of the current bit. Typical ranges for P_m are from 0.001 to 0.01.

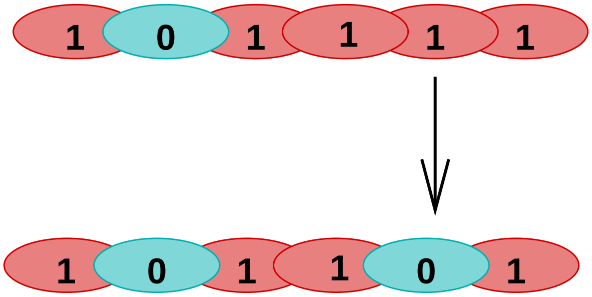


Figure 6.7: Introducing new genetic information is achieved by the mutation operation.

6.2.4 Generation Replacement Strategies

Each generation of parent chromosomes produces an equal number of child chromosomes or *offspring* via the aforementioned genetic operations. Forming a new generation by replacing *all* parents with children may result in the best parent failing to be represented in the new generation. Two similar forms of replacement strategy have been implemented to overcome this: *elitist* and *steady-state* replacement. The elitist strategy replaces all but the single best, or several best, parent chromosomes, while steady-state replacement replaces only the weakest parent chromosomes. The elitist strategy is the most popular and provides good GA performance, but can result in a single 'super chromosome' dominating the population. The steady-state replacement strategy can suffer from slow convergence, as the rate of introducing new genetic information is slower.

6.3 Adaptive Multi-Scale Texture Analysis Using Genetic Algorithm Optimisation

Our proposed algorithm extracts features by selecting localised areas of co-occurrence matrices for element summation. The method differs from classical GLCM feature formation in the following ways:

- The method is extended from 2-D to 3-D, by extracting secondary features from a *stack* of co-occurrence matrices calculated over a range of spatial displacements $d = 1, \ldots, N_d$.

- Rather than using GLCM's fixed element weighting functions for element summation, our algorithm uses adaptive 3-D Gaussian weighting functions $W(i, j, d)$ to form weighted sums of localised areas of the stack of co-occurrence matrices. We use a genetic algorithm to optimise the location and size of these Gaussian weights in co-occurrence-space $P(i, j, d)$.

- The appropriate choice of GA objective function allows the simultaneous optimisation of a feature's first-order discriminatory power, whilst minimising feature correlations.

The Gaussian weighting functions take the form

$$\begin{aligned} W(i,j,d) &= \exp^{-\left(\left[[i\ j\ d]-\boldsymbol{\mu}_W\right]\Sigma_W^{-1}\left[[i\ j\ d]-\boldsymbol{\mu}_W\right]^{\mathrm{T}}\right)}, \\ & \qquad i,j = 1,\ldots,N_g, \quad d = 1,\ldots,N_d, \end{aligned} \tag{6.5}$$

where $\boldsymbol{\mu}_W = [\mu_1\ \mu_2\ \mu_3]$ is the three-dimensional mean vector , and

$$\Sigma_W = \begin{bmatrix} \sigma_{ii} & \sigma_{ij} & \sigma_{id} \\ \sigma_{ji} & \sigma_{jj} & \sigma_{jd} \\ \sigma_{di} & \sigma_{dj} & \sigma_{dd} \end{bmatrix} \quad \Big| \quad \sigma_{ij} = \sigma_{ji},\ \sigma_{id} = \sigma_{di},\ \sigma_{jd} = \sigma_{dj}$$

is the variance-covariance matrix for Gaussian weighting function W. Thus, a secondary feature $\underline{x}$ is extracted by a Gaussian-weighted sum of co-occurrence matrix elements:

$$\underline{x} = \sum_{i,j,d} W(i,j,d)\underline{P}(i,j,d). \tag{6.6}$$

The genetic algorithm optimises the form of these weightings, W_v, via the encoding of three mean and six unique variance-covariance parameters[1] for each Gaussian weighting. Typical examples of Gaussian weighting functions are detailed in Figure 6.8. We show these in 2-D for clarity.

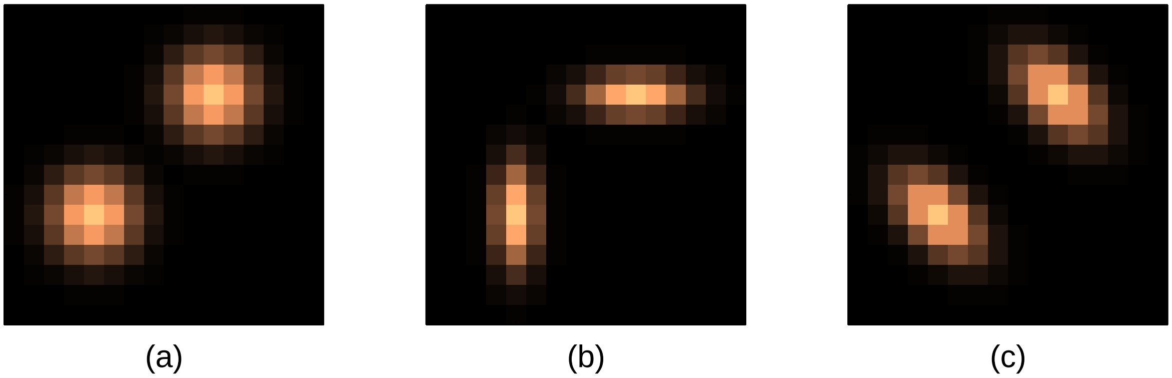


Figure 6.8: Typical examples of Gaussian weighting functions, shown in 2-D for clarity. Note that the weightings are symmetric, because we are using symmetric co-occurrence matrices: (a) Equal variances, no covariance; (b) Unequal variances, no covariance; (c) Unequal variances, non-zero covariance.

In this work, each GA chromosome $\mathcal{C}$ encodes binary information for eight 3-D Gaussian weightings $W_v, \quad v = 1, \ldots, 8$, representing eight features for classification purposes. Our choice of eight features is based on two considerations:

- Our experience suggests there are generally a limited number of discriminant areas in the 3-D discrimination space. Evidence for this can be seen in Figure 5.5.
- Extracting larger numbers of features can be counter-productive due to problems associated with the 'curse of dimensionality' (Hand 1981, Kittler 1978). That is, it is impossible to accurately estimate multi-dimensional feature distributions much above ten dimensions, without a prohibitively large number of training set data (Friedman 1994).

[1] Three of the nine variance-covariance parameters are repeated, e.g., $\sigma_{ij} = \sigma_{ji}$.

Using a $16 \times 16 \times 16$ co-occurrence matrix space $P(i, j, d)$, we encode each of the three components of the mean vector using four bits,

$$\boldsymbol{\mu}_{W_v} \in \{1, \ldots, 16\}^3, \tag{6.7}$$

and encode each of the six variance-covariance elements σ_{st} using three bits,

$$\begin{aligned} \sigma_{ss} &\in \{4, 4.5, \ldots, 7.5\}, \quad s = 1, \ldots, 3, & (6.8) \\ \sigma_{st} &\in \{-2, -1.5, \ldots, 1.5\}, \quad s = 1, 2, \quad t = 2, 3, \quad s \neq t. & (6.9) \end{aligned}$$

Note that there are a total of nine unique parameters for each Gaussian weight, requiring a total of 30 bits for encoding. These nine degrees of freedom allow precise control of weighting location and form in 3-D space. It is obvious that the approximately 2^{30} possible weighting functions form a far larger set than the 20 fixed feature functions defined in the literature for GLCM.

The choice of minimum variance in equation (6.8) is an important one needing clarification: it is a compromise between strong adaptation to training set data, and the algorithm's ability to 'generalise' to unseen test data. A low variance value allows formation of a highly localised weighting function, excessively 'tuned' to the training set. The resulting secondary feature is mainly comprised of a single co-occurrence element. Any estimate error, or noise, in this element equally affects the resulting secondary feature. Forming such features can result in algorithm *over-training*, which is characterised by good classification performance on training set data, but poor performance on test set data. By enforcing a minimum variance criteria, we ensure that a secondary feature comprises the weighted sum of a sufficient number of elements to avoid such over-training and reduce feature estimate noise. Our choice of a minimum variance of 4 in equation (6.8) was based on the size of our co-occurrence matrix space ($16 \times 16 \times 16$), and on empirical evidence obtained during the development of our algorithm.

We show the encoding of a chromosome $\mathcal{C}$ with mean and covariance parameter information in Figure 6.9. Each of the eight 30-bit binary words in this figure hold the parameters for one 3-D weighting function.

The requirement of the GA was to extract eight secondary features by weighted summation of co-occurrence elements, maximising the first-order discriminatory power of each feature whilst minimising the cross-correlations between features. Maximising the *joint* discriminatory power of the eight features may seem more appropriate, as it allows the capture of higher-order discriminatory information. However, such a maximisation would necessitate using *all* eight GAoGLCM features in the final classification

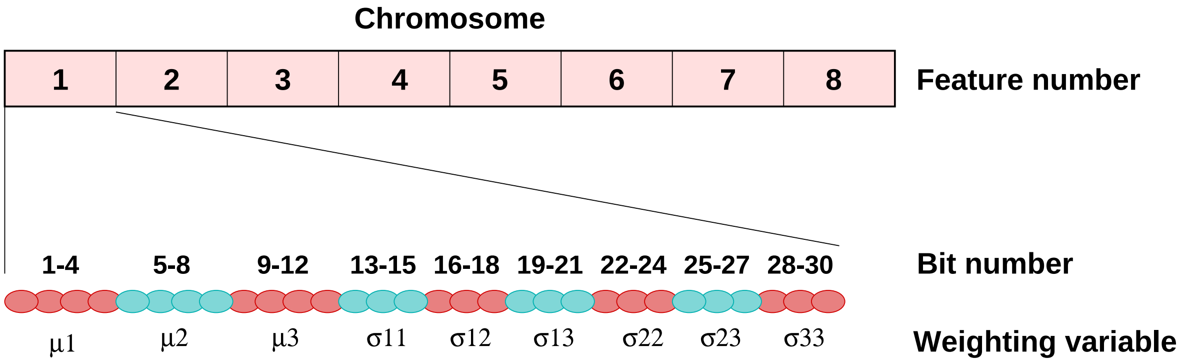


Figure 6.9: The encoding of mean and variance-covariance parameters in a chromosome. Each chromosome contains encodings for eight Gaussian feature weightings.

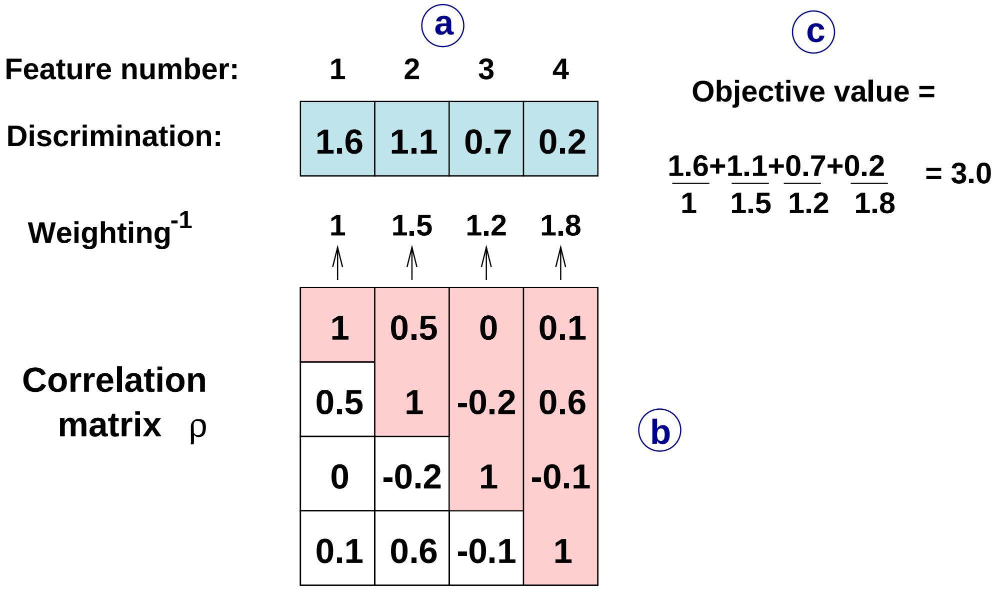
design. Usually, the overall classifier design will use features from other unrelated analysis methods, and we will want to apply a feature selection process after combining our eight features. It would be impossible to guarantee that all eight GAoGLCM features would be selected. Therefore, if the eight features are separately assessed (rather than jointly), then the process will be more effective. We acknowledge that the overall performance of the classifier, when based on texture alone, will not be as high, when compared to using joint discriminatory power as the optimisation criteria. By maximising first-order discriminatory power while minimising correlation, we are attempting to extract features which will possess discriminatory power no matter how many are selected in the final classification system. We can achieve this by expressing our objective function as

$$O(\mathcal{C}) = \sum_{a=1\ldots N_v} \frac{J_a}{\sum\limits_{b=1\ldots a} |\rho_{a,b}|}, \tag{6.10}$$

where $\underline{J} = \{J_a\}, \quad a = 1, \ldots, N_v$ is a vector whose N_v components are discrimination measures for the N_v secondary features $\underline{\mathbf{x}} = \{\underline{x}_v\}, \quad v = 1, \ldots, N_v$, encoded by chromosome $\mathcal{C}$. The components of $\underline{J}$ are numerically ordered from maximum to minimum discrimination,

$$\underline{J} = [J_{\max}, \ldots, J_{\min}]\,, \tag{6.11}$$

and $\rho_{a,b}$ is an element of the similarly ordered $v \times v$ correlation matrix for feature set $\underline{\mathbf{x}}$. Here, we express the worth of any feature set by a weighted sum of the first-order discriminatory powers of its N_v features. The weighting for each feature is equal to the inverse of the sum of its cross-correlation measures with its 'more discriminatory' counterparts, as shown in Figure 6.10. We are effectively reducing the contribution to the objective value O, of a feature's discrimination measure J, by dividing by the sum of its correlations with the other features. Thus, O is maximised by highly discriminatory, uncorrelated features.


a
c
b
Objective value =
$\frac{1.6}{1}+\frac{1.1}{1.5}+\frac{0.7}{1.2}+\frac{0.2}{1.8}$ = 3.0
Weighting^{-1}
Correlation matrix ρ

Feature number:	1	2	3	4
Discrimination:	1.6	1.1	0.7	0.2
Weighting^{-1}	1	1.5	1.2	1.8

1	0.5	0	0.1
0.5	1	-0.2	0.6
0	-0.2	1	-0.1
0.1	0.6	-0.1	1

Figure 6.10: An example of calculating the objective value for a feature set of four variates. Each element of the ordered discrimination vector (a) is divided by the sum of the corresponding feature's cross-correlations (b) with its 'more discriminatory' counterparts. The objective value for the entire feature set (c) is simply the summation of these weighted discriminations.

The GAoGLCM algorithm follows.

Step 1: For each texture image, requantise the image to N_g intensity levels and calculate co-occurrence matrices at each of the N_d intersample displacements $d = 1, \ldots, N_d$.

Step 2: Encode N_k chromosomes with random initialisations for each of the $N_v \times 3$ mean and $N_v \times 6$ unique variance-covariance parameters for the N_v features to be extracted.

Step 3: Develop a suitable objective function O, such as equation (6.10), to measure chromosome worth.

Step 4: Allow the GA to optimise the weighting functions over a suitable number of life cycles.

Step 5: Determine the chromosome with the highest worth among the N_k chromosomes of the terminating population. Use the N_v weighting functions encoded in this chromosome for future feature extraction via equation (6.6).

■

6.4 GAoGLCM Classification Evaluation

We now detail a methodology for comparing the classification performance of the proposed algorithm to that of classical GLCM, for the 2-class problem.

6.4.1 Texture Database

To evaluate the power of our method, we used as test data pairs of grey-scale images containing visually similar or indistinct textures from several sources:

- **Texture pair 1:** Synthesised Brodatz texture images.
  Brodatz texture D9, synthesised using a non-parametric multi-scale non-causal Markov random field model. This is the same database as set 4 of the previous chapter. The technique used to generate the images was based on a multigrid approach using the Gibbs sampler and a novel pixel temperature function. Full details can be found in Paget & Longstaff (1996). Both class 1 and class 2 data used the same neighbourhood size, however, class 2 used a texture image generated at a later iteration in the synthesis process than class 1. Photometric resolution was 8-bit and spatial resolution was 700 × 700.

- **Texture pair 2:** Cervical cell nuclear chromatin texture images.
  High-resolution images of cervical cell nuclear chromatin, captured from cytologically normal and abnormal cells. This is the same database as set 5 of the previous chapter. Full details of this database can be found in Section 3.4.1.

- **Texture pair 3:** Cervical cell nuclear chromatin texture images.
  Images of thionine-stained cervical cell nuclear chromatin, captured from cytologically normal and abnormal cells. This is a much larger database containing images of lower resolution than the previous set.

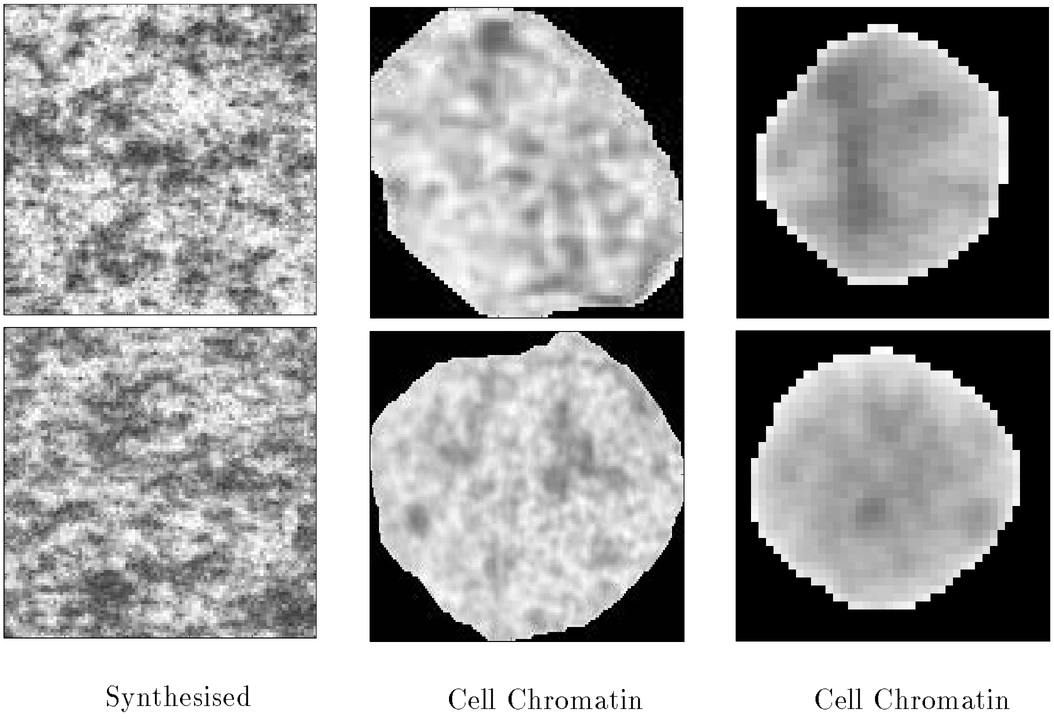


Figure 6.11: Examples of pairs of textures used in the classification trials of GLCM and GA-optimised GLCM. The top row of images are class 1 textures, while the bottom row are class 2.

For texture pair 1, we extracted a total of 100 image tiles for each class by randomly sampling the original image, and allowing tiles to overlap. A suitable tile size was chosen such that both texture methods produced a measurable misclassification rate.

For texture pair 2, we used the entire image within the nuclear boundary to extract co-occurrence matrices, due to the small size of each image.

For texture pair 3, we used a much larger database of 431 normal and 431 abnormal chromatin texture images[2], captured at ×63 optical magnification. Once again, we used the entire nuclear image in the feature extraction process.

6.4.2 Feature Extraction

For both the GLCM and GAoGLCM methods, we calculated co-occurrence matrices from the database images across sixteen spatial displacements, $d = 1, \ldots, 16$. For standard GLCM, we extracted the eight secondary features defined in Table 4.1 at each of the sixteen displacements, from the training and test sets, giving a total of 128 features.

[2] Supplied by Oncometrics Imaging Corporation, Vancouver, B.C., Canada.

Using the GAoGLCM algorithm defined in Section 6.3, we extracted eight GA optimised features using an initial population of $N_k = 50$ chromosomes. We allowed each population to evolve over a total of $N_i = 200$ cycles, using only training set data for optimisation. We set GA control parameters for cross-over rate P_c to 0.8, and mutation rate P_m to 0.01, and used an *elitist* replacement strategy. These choices were based on work by Grefenstette (1992) which recommended similar values following extensive control parameter optimisation trials.

We used the eight encoded Gaussian weightings $W_v, \quad v = 1, \ldots, 8$ from the fittest chromosome in the terminating population to extract features from the test set data.

6.4.3 Feature Preprocessing and Feature Selection

In line with our work of the previous chapters, we used the normality transform discussed in Section 3.4.4 to provide near-Gaussian class-conditioned feature distributions. From the 128-dimensional GLCM feature set, we selected eight optimal feature sets of 1 to 8 dimensions, using the add-2/subtract-1 method of feature selection described in Section 1.4.5. This was similarly done for the eight GAoGLCM features.

6.4.4 Classification

For each analysis method, we used pair-wise ten-fold cross-validated classification to provide a robust estimation of the real classification error. This procedure was outlined in Section 1.4.7. For completeness, it was necessary to perform feature extraction and selection on *each* of the 10 training sets, as explained below.

For GAoGLCM feature *extraction*, it was necessary to determine optimised weighting functions independently for each of the ten training sets. This is due to the adaptive nature of the features, i.e., the exact form of each feature weighting function is training set dependent. Because standard GLCM feature functions are fixed and independent of training and test set partitions, feature extraction need only be completed once on the entire data set.

For both GLCM and GAoGLCM feature *selection*, a new feature set was selected based only on the training set data. The corresponding features of the test set were then classified.

6.5 Results and Discussion

In all three classification trials, GAoGLCM outperformed GLCM with significantly lower classification error rates. This is clearly shown in Table 6.1, which details the minimum error rates achieved by each method, and the corresponding number of features used to obtain those rates. The final column details the extent of the improvement in classification achieved by GAoGLCM—reduction in errors by up to 47%[3]. Figure 6.12 provides a visual comparison of the performance achieved by the two methods.

	GLCM		GAoGLCM		
Texture Pair	Minimum Error	Number of Features	Minimum Error	Number of Features	Error Decrease
1	9.2%	7	6.8%	8	26%
2	10.6%	6	5.6%	4	47%
3	14.1%	2	13.5%	5	4%

Table 6.1: Comparison of classification performance of the GA-optimised GLCM algorithm and standard GLCM. The GAoGLCM algorithm has attained up to 47% decrease in classification error.

The classification results shown in Figure 6.13 are expressed as graphs of cross-validated error rate versus feature set size, for feature sets of 1 to 8 dimensions. Referring to these graphs and to Table 6.1, we can see that GA-adapted features improved classification performance, represented by a decrease in cross-validated classification error, in all three trials. Moreover, performance improved with all feature set sizes from 1 to 8 variates[4]. For texture pair 3, increasing GLCM feature set size *deteriorated* classification performance—see Figure 6.13(c). It would appear that the incremental increases in discriminatory power were not sufficient to overcome decreased class-conditioned PDF estimation accuracy for the feature variates—a classic example of the 'curse of dimensionality'.

Figure 6.14(a) shows an example of 3-D Gaussian weightings optimised by our genetic algorithm. We applied this optimisation to co-occurrence matrices extracted from texture pair 2 (Cytometrics cell data). The weightings are symmetrical because we used symmetric co-occurrence matrices. While not clearly visible in this diagram, each left-right half [5] of this 3-D 'weighting-space' contains eight Gaussian weightings. Analysis

[3] relative error, defined as $\frac{\text{Error}_{\text{GLCM}}-\text{Error}_{\text{GAoGLCM}}}{\text{Error}_{\text{GLCM}}}$.

[4] feature set 1 of texture pair 3 excepted.

[5] separated by the plane $i - j = 0$.

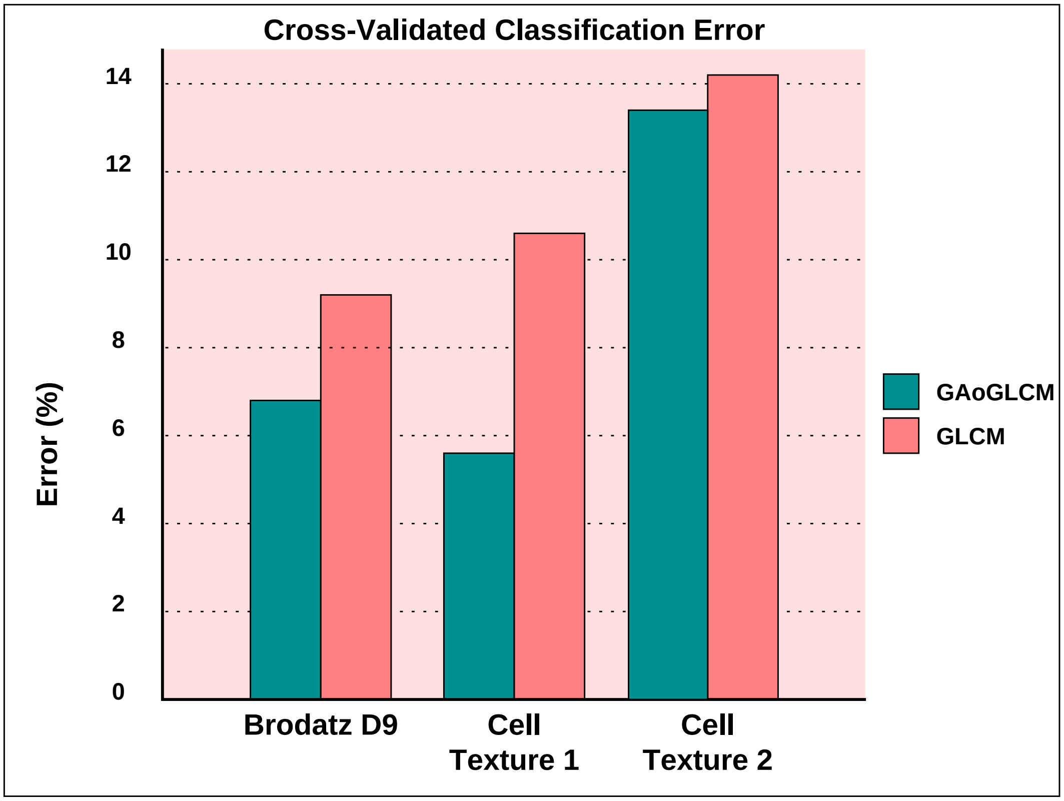
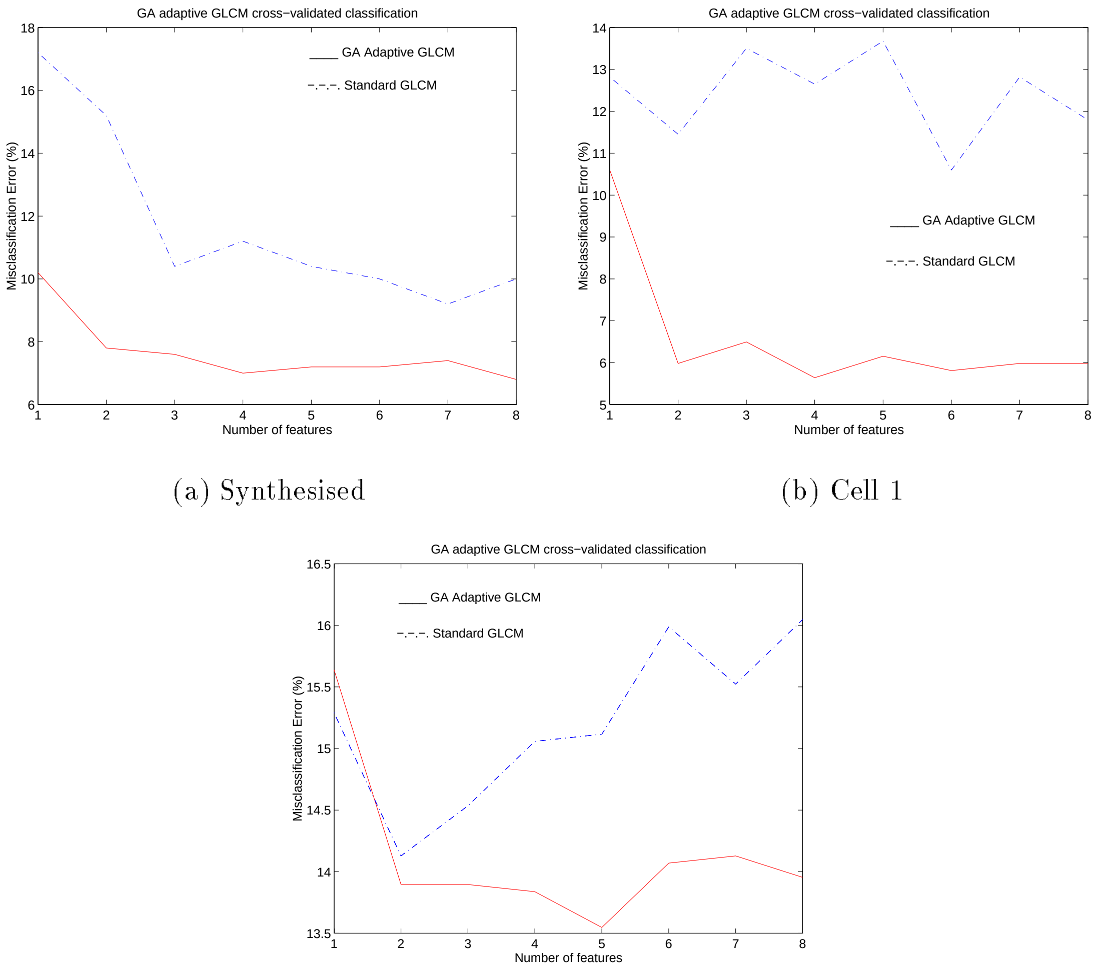


Figure 6.12: Comparison of classification performance of the GA-optimised GLCM algorithm and standard GLCM.

of the corresponding 'discrimination-space' for this data—see Figure 6.14(b)—reveals that the locations of each Gaussian weighting correlate highly with the locations of high discriminatory power. The algorithm has extracted features across several scales, adding credibility to the worth of this multi-resolution extension of standard GLCM. We can also see from the varying size and shape of each weighting that Gaussian weighting functions allow simple, yet precise, control of feature extraction localisation. Comparing Figures 6.14(a) and (b) raises two important questions:

1. **Why were some large regions of high discrimination represented by more than one weighting function? Why didn't the GA form one large weighting in these areas?**.

   Careful examination of Figures 6.14(a) and (b) show some large 'clumps' of discrimination being represented by up to three Gaussian weighting functions. We believe the reason for this is revealed in Figure 5.4 of the previous chapter. Co-occurrence elemental features are only highly correlated over *small* neighbour-



(a) Synthesised

(b) Cell 1

(c) Cell 2

Figure 6.13: Cross-validated error rate versus feature set size for the three classification trials. (a) Brodatz D9 texture, (b) Cytometrics cell data, (c) Oncometrics cell data.

hoods. Spatially distant elements of a large discrimination clump tend to be uncorrelated, and therefore provide higher discrimination contribution when used in *separate* secondary features. Thus, the GA partitions these large discrimination clumps in a way that maximises the extracted discriminatory information.

2. **Why is one Gaussian weighting localised in an area which appears to have no discriminatory power?**
   Our algorithm encodes each chromosome with eight Gaussian weighting functions. Even if there are a limited number of discriminatory areas, say seven, the GA will still place the remaining weighting *somewhere*, even if it is in an area of seemingly

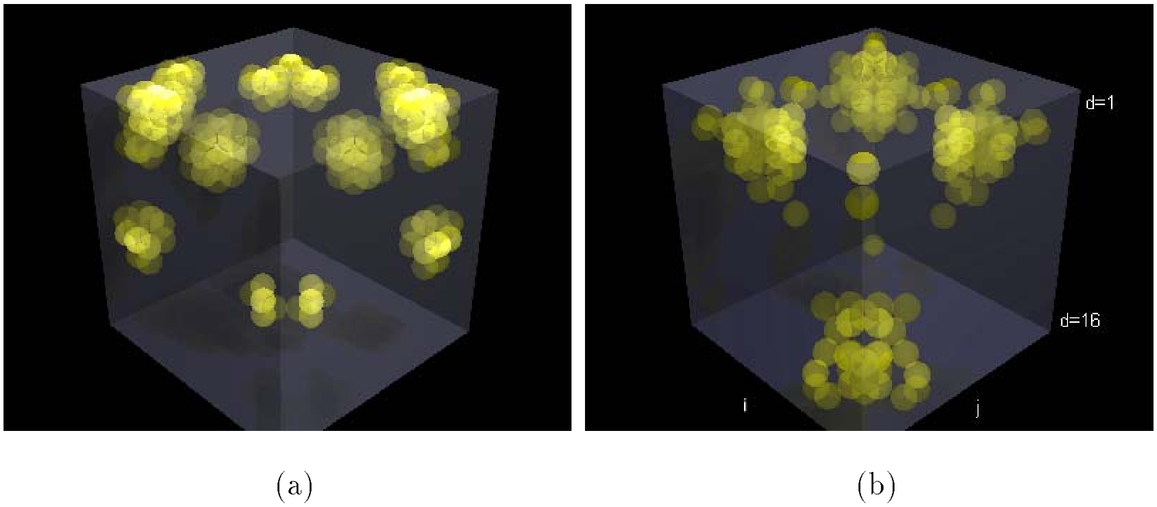


Figure 6.14: Comparing a 3-D discrimination space and the resulting optimised 3-D Gaussian weighting functions: (a) An example of optimised 3-D Gaussian weightings optimised for texture pair 2. Spheres with higher intensities indicate larger weightings; (b) The 3-D discrimination-space of texture pair 2 data. Spheres with higher intensities indicate co-occurrence matrix elements with higher discriminatory power. Note that the locations of each Gaussian weighting correlate well with the locations of high discriminatory power.

low discriminatory power! In future research, we may allow the GA to select an optimal number of feature weightings, based on training set data.

GAoGLCM was able to provide secondary features with higher discriminatory power because it exploited higher-level knowledge (the existence of localised areas of discriminatory co-occurrence elements). This allowed direct targeting of these areas for feature formation, and provided for the inherent ability of our method to *adapt* to the type of texture being analysed. The method of localised summation via Gaussian weighting functions excludes elemental features with low discriminatory power, and those which are uncorrelated (and therefore possibly more *independent* and better used as separate features).

The GAoGLCM algorithm is computationally expensive during the training phase, as it is an optimisation technique based on a randomised search of a multi-dimensional solution space. Such techniques are always computationally intensive, but the need to cross-validate the results exacerbates this drawback. Using N-fold cross-validation, it is necessary to train the GA N times. However, the existence of isolated neighbourhoods of high discriminatory power, and the formation of features from only these areas, allows the extraction of most discriminatory information contained within the co-occurrence matrices, in *fewer* features than for GLCM. This in turn allows a marked decrease in feature selection/discriminant analysis computation time, off-setting some of the increase mentioned above. Moreover, once a suitable set of optimised weighting functions is obtained, the classification times for new data are similar to other methods.

6.6 Conclusions

In Chapter 5 we introduced a new second-order method of texture analysis called Adaptive Multi-Scale GLCM. This method extracted features via a variable summation of elements in neighbourhoods containing proven high discrimination, as measured using a *discrimination matrix*. In this chapter, we have presented a new methodology to optimally extract adaptive secondary features which bypasses the need to explicitly calculate a discrimination matrix. Based on the knowledge that discriminatory areas exist across scales in localised neighbourhoods of co-occurrence matrices, we have proposed the extraction of features using adaptive 3-D Gaussian weightings. We use a genetic algorithm to determine the location and scale of these weightings. Optimisation of these weightings allows extraction of features with high discriminatory power and low correlation.

We have shown that this approach has a number of significant advantages over the traditional GLCM method, namely:

- The features extracted using GAoGLCM have, on average, higher discriminatory power and lower correlation than the standard GLCM features defined in the literature.

- As a result, the use of GAoGLCM features can provide significantly lower classification error rates than standard GLCM.

- The direct targeting of discriminatory areas of co-occurrence matrices allows most discriminatory texture information to be extracted in fewer features. Extracting only 8 GAoGLCM features proved to be far more effective than extracting 128 standard GLCM features!

We demonstrated these advantages in trials comparing the performance of GAoGLCM and GLCM in classifying a range of visually similar images captured from synthetic and biologic origins. GAoGLCM significantly lowered classification error rates and increased feature discriminatory power.

Once again, the applicability of this method is not only limited to GLCM. It can be extended to *any* analysis method where a series of matrices are determined via constraint parameters, such as Neighbouring Grey Level Dependence Matrix (Sun & Wee 1983), Grey Level Entropy Matrix and Grey Level Variance Matrix (Yogesan 1995), or Generalised Co-occurrence Matrices (Davis et al. 1979).

Due to severe time constraints we have been unable to evaluate our method on a more extensive database of texture types. Also, we need to further investigate GA design to determine more optimal settings for GA control parameters, such as number of chromosomes N_k and number of GA cycles. The values of minimum and maximum variance in equation (6.8) also needs further investigation to determine if these values are problem specific or generally applicable. However, based on the results obtained, we can say that the performance so far is very encouraging and worthy of consideration in any texture analysis application.

Chapter 7

Summary and Conclusions

7.1 Summary of the Thesis

In this thesis we have presented the results of research and experimental work aimed at enhancing the machine assisted analysis of cytological specimens. The main body of our work has concentrated on so-called ‘smart algorithms’ for analysing image texture, and we have attempted to maintain the general applicability of these methods to applications other than cell analysis. We have confirmed the efficacy of our methods by their application to a range of textures which exist in industry, medicine, and nature. This chapter contains a summary of the work contained within this thesis, general conclusions and accomplishments made, and suggestions on further research avenues.

In Chapter 1 we reviewed manual and automated cervical cancer screening, and identified avenues for enhancing the performance of current and future automated cytology systems. We introduced pattern recognition and discussed PR operations typically used for image analysis. We also provided some detail (extended in Chapter 3) of specific PR algorithms used to support our texture analysis work throughout this thesis. For example, in Section 1.4.5 we justified our choice of feature selection algorithm, and suggested a minor change which facilitates the capture of feature sets with higher-order discriminatory power. We also presented our preferred method of discriminatory power measurement (the Bhattacharyya discrimination measure), and quadratic classification. Finally, we identified texture analysis as an avenue for further exploitation and reviewed many of the methods widely used in the literature.

In Chapter 2 we provided a comprehensive review of co-occurrence-based methods of texture analysis. We discussed the motivations, advantages, and limitations of each method, and where available, presented the results of comparative studies found in the literature. We found that several comparison studies were limited due to poor methodologies which favoured one or another technique. We also found that inter-study comparisons were difficult, due to differences in evaluation methodology and databases. We felt that a critical appraisal based on a more unified framework of classification methodologies and a common, comprehensive, texture database was needed, if stronger conclusions were required as to the power of each method. However, we did conclude that the GLCM method of Haralick et al. (1973) was one of the most widely used method of texture analysis, and considered by many researchers as the most powerful method for general texture analysis. We therefore used the GLCM method as our benchmark for comparing the performance of the algorithms introduced in subsequent chapters.

Chapter 3 reviewed the SGF method of texture analysis. The original authors (Chen et al. 1995) suggested that the SGF method was more powerful than the GLCM method. However, we found it to be significantly weaker, due to the limited number of feature functions (two) defined by the original authors. We suggested new SGF features whose definitions were based on measuring specific properties of the texture types to be analysed—in our case, chromatin texture. We showed that such manual adaptation of features can yield significant advantages over other methods based on pre-defined feature functions, namely:

- it allowed better targeting of possible discriminatory texture properties;
- features which were found to be discriminatory provided a far better understanding of the properties of the texture which discriminated between classes;
- the tailored SGF features used in our classification trial provided classification performance equal to GLCM features, but with fewer features.

We also provided further details of our PR algorithms which supported the texture analysis work of this and subsequent chapters. For example, in Section 3.4.3 we presented our method of linear image requantisation, and justified our choice by identifying weakness of the widely used histogram equalisation method which has detrimental effects on second-order statistical analysis. In Section 3.4.4 we presented our method of feature normality transformation, and demonstrated the usefulness of this method when using normality-based parametric methods of feature analysis and classification.

Our success with manual adaptation of feature functions motivated further research into methods of *self-adaptive* feature extraction. In Chapter 4 we commenced a theoretical investigation into why discriminatory power was manifested in some GLCM features but not in others. Using a *discrimination matrix*, we were able to show that discriminatory power was expressed in only localised areas of co-occurrence matrices. GLCM feature functions which, by chance, weighted these discriminatory areas highly, also possessed high discrimination. The discrimination matrix provided us with our first avenue for self-adaptive feature formation. By using the matrix values as a second weighting function, we could increase the discriminatory power of the majority of GLCM's pre-defined features. We attained up to 70% decreases in classification error when applying our proposed method to classifying a database of regularly stained cervical cell nuclei.

In Chapter 5 we completely removed our method's reliance on the fixed feature functions of GLCM. Rather than using the discrimination matrix as a weighting to be applied in addition to the fixed weightings of GLCM, we used the matrix alone as a basis for defining new feature functions. We used the discrimination matrix and measures of GLCM element correlation to define neighbourhoods of GLCM elements which, when summed, formed new features for texture classification. Because we only targeted areas of high discriminatory power, we were able to capture most of the discriminatory power contained in the matrices in far fewer features. The discrimination matrices explicitly showed the existence of discriminatory texture information across several spatial scales or resolutions. This led to the extension of our method to simultaneous multi-scale analysis. We then critically appraised our Adaptive Multi-Scale GLCM by applying it to the classification of a range of textures from nature, industry and medicine. When compared to standard GLCM, our method attained significant decreases in misclassification error of between 12 and 35%.

We introduced a method of optimised, self-adaptive multi-scale feature extraction for texture analysis in Chapter 6. Based on AMSGLCM, we used a Genetic Algorithm to search a multi-dimensional Gaussian weighting space for extracting co-occurrence features. Using a suitable GA objective function, we were able to successfully extract a group of features which were optimised for high discrimination and low feature correlation. Our choice of Gaussian weights allowed precise control of weighting localisation and shape in multi-dimensional weighting-space, using a minimum of control parameters. An advantage of this method over AMSGLCM is that it does not require the explicit calculation of discrimination matrices at each of the spatial resolutions being analysed. Once again, we critically appraised our GA-optimised GLCM (GAoGLCM) by applying it to the classification of synthetic and biological textures. We attained significant decreases in misclassification error of up to 47% when compared to standard GLCM features.

7.2 Thesis Discussion

The techniques presented in this thesis represent the evolution of our research over the last three and a half years. The ordering of the chapters is approximately in line with our advances in adaptive analysis methodology, from basic manual feature adaptation in Chapter 3, to fully automated self-adaptive, multi-resolution, optimised texture analysis in Chapter 6. Our journey began by showing the benefits of adaptive feature extraction, using manually defining feature functions specific to cell chromatin analysis (Chapter 3). By doing so, the defined features yielded far greater understanding of the pathological processes which accompany cell carcinogenesis than many other current methods of texture analysis using problem-independent features. The weakness of this technique is in its lack of general applicability to other texture classes without human intervention.

The requirement to define and hand-code feature functions (i.e., manual adaptation) provided the motivation to further investigate a completely new approach which required no human interaction (i.e., self-adaptation). By theoretically analysing discriminatory power manifestation in co-occurrence matrix secondary features, we were able to implement a self-adaptive approach to enhancing the discriminatory power of currently used feature functions (Chapter 4). Here, we introduced the discrimination matrix, which quantitatively measures the discriminatory power of each co-occurrence matrix elemental feature, after it has been weighted by one of the standard feature weighting functions defined in the literature. We were able to show that the discriminatory power of each weighted feature directly influenced the discriminatory power of a secondary feature comprised of the sum of these weighted features (Appendix A.0.3). Thus, by using the discrimination matrix as a further weighting function, we were able to increase the contribution made by highly discriminatory elemental features, to the secondary feature. In this way, the discriminatory power of the secondary feature was increased.

At this point we should enter a word of caution about *over-training* and its relationship to the discrimination matrix and adaptive methods. All datasets contain estimate errors, a natural consequence of finite sample size. We are, of course, speaking here in a statistical sense, meaning that sampled distributions can not completely capture the true statistical nature of real data. Because the discrimination matrix is determined directly from sampled data, it too contains estimate errors in its discrimination measures. This error is exacerbated further by elemental features whose distributions are not Gaussian (we have used normality-based discriminatory power measures). Thus we need to ensure that the measured discriminatory powers are an accurate representation of those which exist for real data, and are not some artifact present only in the dataset—a result of over-training, or 'tuning' the algorithm too much to the dataset. To this end, we have always applied normality transformation to feature distributions prior to discriminatory power estimation. More importantly, we have ensured that discriminatory power measures (as well as feature selection and classifier design) were independent of test set data. That is, no test set data was used during algorithm training.

The method of self-adapting standard feature functions (Chapter 4) proved very successful, and resulted in truly outstanding decreases in classification error. The technique is computationally light, requiring only 136 discriminatory power calculations for each discrimination matrix (using 16 grey-level requantised images). It can be easily implemented into PR systems to enhance the discriminatory power of any GLCM features which may be used. Finally, our method is easily extended to other co-occurrence-based and more general analysis techniques.

Introducing the AMSGLCM method of Section 5 removed the reliance of our previous method on pre-defined feature functions, which we identified as being a fundamental weakness of standard analysis methods. Extending the discrimination matrix to the scale domain was a natural step which allowed the incorporation of more higher-level knowledge—that of multi-resolution discrimination variation. We now not only had a greater number of elemental features from which secondary features could be formed (thus reducing estimate errors), we also had far greater flexibility in choosing which elements to use to form summed secondary features. In effect, the discrimination space showed directly *how many* elements to use for feature formation, and showed *which* elements to use to ensure capturing the most texture information. This led to a fully self-adaptive method incorporating multi-scale texture information, which attempted to optimise the capture of texture information, yet which required no human intervention.

Using a three-dimensional discrimination space resulted in a small increase in computational expense. It was now necessary to calculate $(N_g^2 + N_g) * N_d/2$ discrimination measures—a total of 2176 in our implementation. Computational cost was further increased by the need to cross-validate the classification process. The above process needed to be repeated a total of ten times for ten-fold cross-validation. This cost is unavoidable if we are to maintain robust classifier training and performance evaluation methodologies. On a positive note, directly targeting discriminatory elemental features facilitates capturing discriminatory information in far fewer features than in traditional methods. As a result, the computational cost of feature selection can be reduced. Furthermore, computational cost is only high during classifier design. After training and evaluation, classification times on new, unseen data are similar to existing techniques.

GA-based feature extraction (Chapter 6) extends the work of the previous chapter, and represents the most recent evolution of our approach to self-adaptive, multi-scale texture analysis. This extension allows far greater flexibility in the size and shape of neighbourhoods from which elements are to be summed to form secondary features. In AMSGLCM, these neighbourhoods were confined to a three-dimensional cubic neighbourhood based on element correlation. For GAoGLCM, these neighbourhoods are defined by Gaussian weighting functions with full flexibility in all three domains of the co-occurrence space. The extension also allows optimising the extraction process based on explicit criteria—discriminatory power and feature correlation criteria in our implementation.

The biggest impediment to our method is its computational expense. On a DEC AlphaStation 255, a single GA training session takes around one day. As a result, full cross-validated classification evaluation takes over one week. However, as with AMSGLCM, after training and evaluation, classification times on new unseen data are similar to existing techniques.

At this point it seems appropriate to compare the performance of the AMSGLCM and GAoGLCM algorithms. Obviously the GAoGLCM approach is far more computationally expensive, both in terms of time and storage requirements, because it involves a parameter-based optimisation in a highly complex solution space. Computation time and storage requirements during the training phase are approximately an order of magnitude greater than for AMSGLCM. After training however, both techniques have computational cost which is similar to existing techniques. Classification performance can be determined by comparing the two texture types which were trialed for both techniques:

1. synthesised Brodatz texture D9—texture pair 4 for AMSGLCM and texture pair 1 for GAoGLCM.

2. Cytometrics cell database—texture pair 5 for AMSGLCM and texture pair 2 for GAoGLCM.

We compare these results in Figure 7.1 Based on the classification results, the GAoGLCM method has attained an average 30% decrease in classification error when compared to AMSGLCM. This is probably due to its ability to extract features which are optimised for high discriminatory power and low correlation. More comparison trials on a wider range of texture types will need to be run before stronger conclusions can be drawn. However, we would suggest that the GAoGLCM be used if computational cost is not of significant concern.

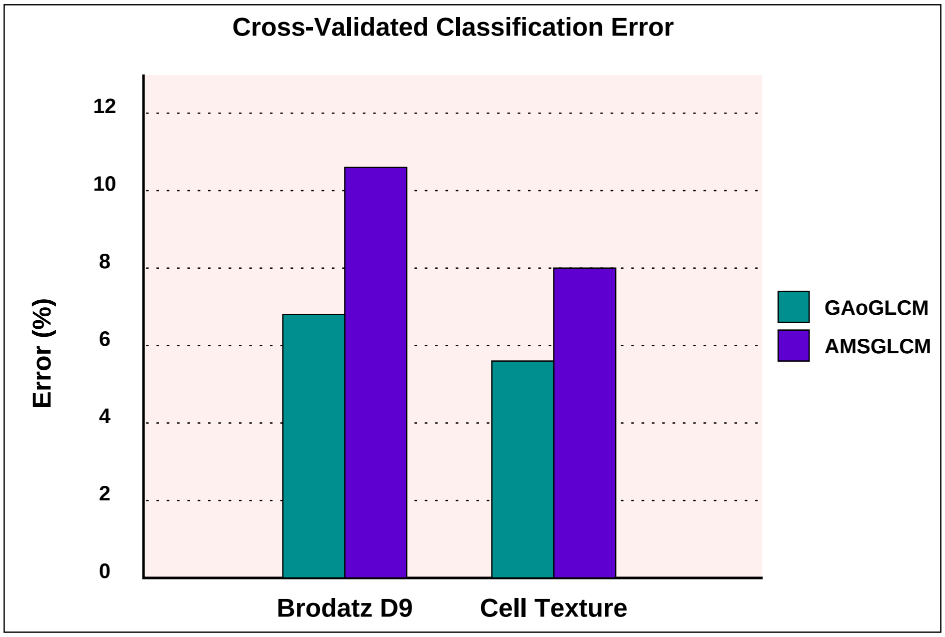


Figure 7.1: Comparing classification error rates for AMSGLCM and GAoGLCM. Classification errors are decreased by an average 30% when using the GAoGLCM technique.

7.3 Thesis Contribution

This thesis presents a comprehensive analysis of statistical texture analysis methodologies, and introduces several new approaches to optimised feature extraction. Here, we describe the main contributions made by this thesis, chapter by chapter.

A contribution of this thesis is our extension of the SGF algorithm to analysing cell nuclear chromatin. We have shown the benefits of manually adapting feature functions to suit specific texture properties by an increase in classification performance and decrease in feature set dimensionality. But perhaps of greater significance is the fact that defining feature functions which measure specific image properties has allowed a far better understanding of the cytological properties which manifested the discrimination between normal and abnormal classes. This cannot be said for most other analysis methods.

One of the main contribution of this thesis was our presentation of several related methodologies for optimised, multi-scale, self-adaptive feature extraction. All are based on locating areas of discriminatory power among texture descriptors extracted across a range of spatial resolutions. In Chapter 4 we introduced the *discrimination matrix*—a two-dimensional matrix which, for the first time, provided a direct indication of the potential 'worth' or usefulness of each co-occurrence matrix element for classification purposes. The spatial arrangement of discriminatory information within the matrix also suggested possible approaches to using such information for enhancing the discriminatory power of currently defined feature functions. The success of our approach was demonstrated by significant decreases in classification error of over 70%.

Chapter 5 introduced what we believe is the first self-adaptive multi-scale feature functions for use with co-occurrence-based methods of texture analysis. Our method placed no reliance on pre-defined fixed feature functions, unlike all other co-occurrence-based methods published in the literature. In fact, feature definition was based solely on the specific statistical differences between texture classes. Furthermore, our method is possibly the first co-occurrence-based technique to provide simultaneous analysis of texture across several spatial resolutions. Once again, the technique attained significant increases in classification performance across a wide range of texture types.

Chapter 6 introduced what we consider to be the first application of optimisation techniques to the extraction of co-occurrence matrix features, without the use of neural network-based methods. While neural networks have been used in the past to classify texture data, this 'black-box' approach often provides little theoretical guide as to the image properties that are producing the classification result. The GAoGLCM method allows direct analysis of the optimised feature functions and, using the remapping technique described in Appendix B, the location of areas within images which produce any discriminatory information between classes. Moreover, by suitable objective function design, we can choose to optimise feature functions under a number of criteria, including correlation considerations, first-order or joint discriminatory power, etc.

We should emphasise that, while the adaptive methods of Chapters 4, 5, and 6 were applied to GLCM co-occurrence matrices, their applicability is not restricted to this method of analysis. They can equally be applied to *any* analysis method where a series of feature vectors or matrices can be extracted via a suitable constraint parameter.

In Appendix B, we demonstrated another significant benefit of the discrimination matrix for image analysis. Discriminatory features derived from conventional analysis methods usually provide only *qualitative* cues as to the characteristics of images which differ statistically between classes, such as image contrast, entropy, or energy. By using the information contained in the discrimination matrix, we can directly locate actual *areas* within an image which provide such discriminatory information. We demonstrated this in Appendix B by 'remapping' discrimination matrix co-ordinates to actual image pixels. Examining actual areas of an image which provide discriminatory power allows a far better understanding of the processes or physical attributes of image objects which differ between classes. We believe this is the first demonstration of this capability.

7.4 Fundamental Limitations

Although this thesis claims an original and significant contribution to the field of pattern classification, there are some limitations to the outlined techniques, and several research questions which remain unanswered. We hope that this thesis is the genesis for such further research.

- The discrimination matrix represents the 'worth' of each co-occurrence matrix elemental feature, by measuring first-order discriminatory power. However, it is known that features which do not exhibit first-order discriminatory power may, in fact, exhibit higher-order discriminatory power, as demonstrated in Section 5.2 on page 102. It may be advantageous to measure the true worth of elemental features by considering both first and higher-order discriminatory power. Computing such higher-order discriminatory power quickly becomes prohibitive above powers of two. For example, for symmetric $N_g \times N_g$ co-occurrence matrices calculated at N_d displacements, there are a total of $(N_g^2 + N_g) * N_d/2$ unique elemental features $\underline{P}(i, j, d)$. Thus there are a total of approximately $((N_g^2 + N_g) * N_d/2)^2$ feature pairs for which second-order discriminatory power measurements are required. Clearly, third and higher-order calculations are infeasible.

- The self-adaptive methods of feature extraction introduced in Chapters 5 and 6 rely on the existence of localised *groups* of discriminatory elements. It is currently unknown to what degree performance will be affected if discrimination is only exhibited in isolated *individual* elements. However, because of the reasons discussed in Section 5.2 (neighbouring co-occurrence elements being measures of similar image properties, and therefore being highly correlated in general), it is unlikely that textures which exhibit such characteristics are common. Highly regular textures containing strong gradients (such as some fabricated structures), which produce highly structured co-occurrence matrices with vastly different statistics for neighbouring elements, may not produce significantly better performance than GLCM and other common techniques.

- The AMSGLCM algorithm of Chapter 5 used a three-dimensional neighbourhood constraint of equal dimensions in both the co-occurrence domain (i, j) and the scale domain (spatial displacement d). We determined this cubic neighbourhood by measuring the average pair-wise correlation between all element pairs in the 3-D stack of co-occurrence matrices. We realise that there is no direct relationship between the co-occurrence axes i, j and the scale axis. For this reason, it may be more appropriate to determine neighbourhood size independently in the scale domain by only measuring correlation between element pairs from differing spatial constraints d. That is:

$$Corr(d_d) = E\left\{\rho\left(\underline{P}(i,j,d_1), \underline{P}(i,j,d_2)\right) \mid (d_1 - d_2) = d_d\right\}, \tag{7.1}$$

  where d_d is the element pair displacement in the spatial displacement domain d.

- As mentioned in Chapter 6, time constraints have not allowed the GAoGLCM technique to be fully evolved. An area that needs further research is in the correspondence between the eight weighting functions across each of the ten cross-validation trials. We have noticed that, on some occasions, the locations of some of the eight feature weightings differed between trials. This may be due to two causes:

  1. The differences between the elemental feature statistics $\underline{P}(i, j, d)$ of each of the ten cross-validation sets. While these differences are a natural consequence of the random partitioning of the dataset, ideally the algorithm should be invariant to these differences. If this is indeed the problem, it suggests that

the GA is not generalising sufficiently, and is thus too highly tuned to the training set.

2. The GA is not finding a globally optimal solution, but is finding a different *locally optimal* solution for some of the cross-validation training partitions. Given that each solution is represented by a 30-bit binary chromosome, we have a solution space of cardinality 2^{30}! This solution space is searched by the GA a total of $N_k = 50 \times N_i = 200 = 10\ 000$ times in our implementation, which may not be sufficient.

We can test for (2) by trialing the GA multiple times on the *same* training set. If there is variation in any of the solutions, it indicates an optimal solution is not being found by the GA. This would suggest that we need to further investigate GA control parameters such as number of iterations N_i and number of chromosomes N_k, to ensure that a globally optimal result is attained.

If we find stable solutions after running this test, it would suggest that indeed, the GA is not generalising sufficiently. Therefore, we need to further increase the minimum variance criteria $\sigma_{\min}$ for the Gaussian weighting functions, as discussed on page 130. This will ensure that secondary features are comprised of a larger number of elemental features, which helps to average any estimate noise introduced by the partitioning of the data set.

However, based on our results, we can conclude that:

- even locally optimal solutions provide significant benefits over more traditional methods such as GLCM; and
- even further improvements in classification performance may be attained, because some of the results presented may represent sub-optimal solutions.

- We mentioned in Chapters 5 and 6, that the AMSGLCM and GAoGLCM algorithms are applicable to a wide variety of existing techniques, where a series of matrices are determined via constraint parameters. To confirm this general applicability, we expect to apply our adaptive methods to techniques such as NGLDM (Sun & Wee 1983), Yogesan's GLEM and GLVM methods (Yogesan 1995) or GCM (Davis et al. 1979) in future research.

Appendix A

Feature Summation and Discriminatory Power Considerations

A.0.1 Proof that the discriminatory power of a 2-D feature set is always greater than or equal to the discriminatory power of a 1-D feature formed by the summation to the two individual features.

As discussed in Section 5.2, we are faced with the question of when to sum elemental co-occurrence matrix features (to reduce feature set dimensionality) and when to leave them separate (to maintain or increase discriminatory power). We will consider the case of a feature containing two variates, and whether to sum together the individual variates to form a 1-D feature, from the point of view of maintaining overall discriminatory power. Extensions to the n-D case are straightforward.

For mathematical tractability we consider a simplified form of a discrimination measure, used as the basis for discrimination metrics such as the Bhattacharyya, Divergence, Mahalanobis and Matsushita metrics,

$$J = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}, \tag{A.1}$$

where the row vector $\boldsymbol{\mu}_c = [\mu_1, \ldots, \mu_{N_v}]$ is the class-conditioned mean vector of feature set $\underline{\mathbf{x}}$, and $\boldsymbol{\Sigma}_c$ is the variance-covariance matrix. Equation (A.1) is simply a measure of the distance between the means of the multivariate PDFs of the two classes, normalised by the class-conditioned variance-covariance matrices.

Notation

c is the class index, $c \in \{1, 2\}$.
v is the feature or variate index, $v \in \{1, 2\}$.
$\underline{x}_{c,v}$ is the vth feature vector for class c.
$\mu_{c,v}$ is the sample mean of feature v for class c.
σ_{c,v_1v_2} is the covariance of features v_1 and v_2 for class c.
$B = \mu_{1,1} - \mu_{2,1}$.
$C = \mu_{1,2} - \mu_{2,2}$.
$A_{v_1v_2} = \sigma_{1,v_1v_2} + \sigma_{2,v_1v_2}$, and $A_{12} = A_{21} = A$ for covariance matrices.
J_{1D} is the discriminatory power of the summed features $\underline{x}_{c,1} + \underline{x}_{c,2}$,

$$J_{\mathrm{1D}} = J(\underline{x}_{1,1} + \underline{x}_{1,2}\ ,\ \underline{x}_{2,1} + \underline{x}_{2,2}).$$

J_{2D} is the joint discriminatory power of the 2-D feature set $\underline{\mathbf{x}}_c = [\underline{x}_{c,1}, \underline{x}_{c,2}]$,

$$J_{\mathrm{2D}} = J\left([\underline{x}_{1,1}, \underline{x}_{1,2}] \ , [\ \underline{x}_{2,1}, \underline{x}_{2,2}] \right).$$

1-D case

When the two features are summed to form a 1-D feature, we can show from equation (A.1) that

$$\begin{aligned} J_{1D} &= \frac{(\mu_{1,1}+\mu_{1,2})^2 - 2(\mu_{1,1}+\mu_{1,2})(\mu_{2,1}+\mu_{2,2}) + (\mu_{2,1}+\mu_{2,2})^2}{\sigma_{1,11}+\sigma_{1,22}+2\sigma_{1,12}+\sigma_{2,11}+\sigma_{2,22}+2\sigma_{2,12}} && \text{(A.2)} \\ &= \frac{(B+C)^2}{A_{11}+A_{22}+2A_{12}} && \text{(A.3)} \\ &= \frac{(B+C)^2}{A_{11}+A_{22}+2A}. && \text{(A.4)} \end{aligned}$$

2-D case

When the two features are left as a 2-D feature, we can show from equation (A.1) that

$$\begin{aligned} J_{2D} &= \begin{bmatrix} \mu_{1,1}-\mu_{2,1} \\ \mu_{1,2}-\mu_{2,2} \end{bmatrix}^{\mathbf{T}} \begin{bmatrix} \sigma_{1,11}+\sigma_{2,11} & \sigma_{1,12}+\sigma_{2,12} \\ \sigma_{1,21}+\sigma_{2,21} & \sigma_{1,22}+\sigma_{2,22} \end{bmatrix}^{-1} \begin{bmatrix} \mu_{1,1}-\mu_{2,1} \\ \mu_{1,2}-\mu_{2,2} \end{bmatrix} && \text{(A.5)} \\ &= \begin{bmatrix} B \\ C \end{bmatrix}^{\mathbf{T}} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} B \\ C \end{bmatrix} && \text{(A.6)} \\ &= [B, C] \frac{\begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix}}{(A_{11}A_{22}) - (A_{12}A_{21})} \begin{bmatrix} B \\ C \end{bmatrix} && \text{(A.7)} \\ &= \frac{(B^2A_{22} - BCA_{21} - BCA_{12} + C^2A_{11})}{A_{11}A_{22} - A_{12}A_{21}} && \text{(A.8)} \\ &= \frac{(B^2A_{22} - 2BCA + C^2A_{11})}{A_{11}A_{22} - A^2}. && \text{(A.9)} \end{aligned}$$

Sum the features, or leave as separate variates?

We use the decision rule

$$J_{1D} \gtrless J_{2D} \Rightarrow \begin{cases} \text{sum} \\ \text{separate} \end{cases} \tag{A.10}$$

to determine whether the pair of features are summed to form a univariate feature, or left separate and used as two individual features. We can re-express equation (A.10) in terms of equations (A.4) and (A.9) thus:

$$\frac{(B+C)^2}{A_{11}+A_{22}+2A} \gtrless \frac{(B^2A_{22}-2BCA+C^2A_{11})}{A_{11}A_{22}-A^2} \Rightarrow \begin{cases} \text{sum} \\ \text{separate} \end{cases} \tag{A.11}$$

Expanding equation (A.11) gives

$$\begin{aligned} B^2A_{11}A_{22} &- B^2A^2+2BCA_{11}A_{22}-2BCA^2+C^2A_{11}A_{22}-C^2A^2 \\ &\gtrless B^2A_{11}A_{22}+B^2A_{22}^2+2B^2AA_{22}-2BCAA_{11}-2BCAA_{22} \\ &- 4BCA^2+C^2A_{11}^2+C^2A_{11}A_{22}+2C^2A_{11}A \Rightarrow \begin{cases} \text{sum} \\ \text{separate,} \end{cases} \end{aligned} \tag{A.12}$$

$$\begin{aligned} -B^2A^2 &+ 2BCA_{11}A_{22}-2BCA^2-C^2A^2 \\ &\gtrless B^2A_{22}^2+2B^2AA_{22}-2BCAA_{11}-2BCAA_{22} \\ &- 4BCA^2+C^2A_{11}^2+2C^2A_{11}A \Rightarrow \begin{cases} \text{sum} \\ \text{separate.} \end{cases} \end{aligned} \tag{A.13}$$

Reducing this further (and multiplying by -1) gives

$$B^2(A+A_{22})^2-2BC(A^2+AA_{11}+A_{11}A_{22}+AA_{22})+C^2(A+A_{11})^2 \lessgtr 0 \Rightarrow \begin{cases} \text{sum} \\ \text{separate.} \end{cases} \tag{A.14}$$

This can be reduced to

$$(\ B(A+A_{22})-C(A+A_{11})\ )^2 \lessgtr 0 \Rightarrow \begin{cases} \text{sum} \\ \text{separate} \end{cases} \tag{A.15}$$

Clearly, the left-hand term of equation (A.15) can never be less than 0. This proves that the summation of features will never result in an increase in discriminatory power above that of their joint discriminatory power.

A.0.2 Proof that summing features with similar distribution statistics results in minimal loss of discriminatory power.

The left hand term of equation (A.15) expresses the relative loss of discriminatory power which results when two feature variates are summed to form a single univariate feature. The larger the magnitude of this term, the greater the loss of discriminatory power:

$$\text{Loss}_J = |B(A + A_{22}) - C(A + A_{11})|, \tag{A.16}$$

where Loss_J represents the relative loss in discriminatory power resulting from the summation. Readers will remember our methodology for reducing the $N_g^2 \times N_d$ elemental co-occurrence features to a smaller number of dimensions with minimal loss of discriminatory power, was to sum together neighbouring features, because these were measures of similar image properties. As such, they had similar class-conditioned statistics and were thus highly correlated. Here, we wish to show in a more theoretical sense that such an approach does, indeed, minimise loss while reducing feature set dimensionality.

As the similarity between a pair of neighbouring elemental features increases, $B \approx C$ and $A_{11} \approx A_{22} \approx A$. In the limit as the feature statistics become identical, we can re-express equation (A.16) as

$$\begin{aligned} \text{Loss}_J &= B(A + A) - B(A + A) \\ &= 0. \end{aligned} \tag{A.17}$$

Clearly, the more similar the feature variates to be summed, the lower the loss of discriminatory power.

A.0.3 Proof that the discriminatory power of a summed feature can be expressed in terms of the discriminatory powers of the individual variates

From equation (A.1), we can express the discriminatory powers, J_1 and J_2, of the two feature variates $\underline{x}_{c,1}$, $\underline{x}_{c,2}$ as

$$J_1 = \frac{B^2}{A_{11}}, \tag{A.18}$$

and

$$J_2 = \frac{C^2}{A_{22}}. \tag{A.19}$$

From equations (A.18) and (A.19) we have

$$B^2 = J_1\, A_{11}, \tag{A.20}$$

$$C^2 = J_2\, A_{22}. \tag{A.21}$$

Re-expressing equation (A.4) in terms of equations (A.21) and (A.21) gives

$$J_{1D} = \frac{(\sqrt{J_1 A_{11}} + \sqrt{J_2 A_{22}})^2}{A_{11} + A_{22} + 2A} \tag{A.22}$$

$$J_{1D} = \frac{J_1 A_{11} + 2\sqrt{J_1 A_{11}}\sqrt{J_2 A_{22}} + J_2 A_{22}}{A_{11} + A_{22} + 2A}. \tag{A.23}$$

Clearly, the greater the discriminatory power of the univariate features J_1, J_2, the greater the discriminatory power of the resulting summed feature J_{1D}.

Appendix B

Locating Discriminatory Areas in Images

Locating areas of an image which provide discriminatory power, by using the discrimination matrix.

Section 4.2 indicated that it was possible to use the discrimination matrix to locate discriminatory areas within the images being analysed, thus providing a link between the *statistical* differences between two texture images and their corresponding *physical* or morphological differences. We can easily demonstrate this by the following example.

We take two copies of the same texture, representing two texture classes, and modify the second-order statistics of the second image. For this example, we choose the texture Tiles.0009 from the VISTEX Vision Texture Database (Picard et al. 1995)—see Figure B.1. Both images are requantised to 16 grey levels in order to calculate 16×16 co-occurrence matrices. The second-order statistics of the second image are modified in the following way:

- for each pair of pixels with intensities $i = 1$ and 7, separated by a displacement of $d = 1$, we change the second pixel of the pair to intensity $j = 9$:

  $$\forall\, k, l \in \boldsymbol{D},\ ||k - l|| = 1,\ \text{s.t. } I(k) = 1,\ I(l) = 7, \quad I(l) \leftarrow 9 \ \text{ iff } \ Rand < 0.2, \tag{B.1}$$

  where k and l are valid image co-ordinates, I is an image of domain $\boldsymbol{D} \subset \mathbb{Z}^2$, $||.||$ represents the norm of a vector in 2-space, and $Rand$ is a uniformly distributed random variable. The second-order statistics of the image are such that intensity pairs (1,7) are far more common than intensity pairs (1,9). The modification defined in this equation has the effect of grossly modifying the second-order grey-level joint-probability $P(1, 9)$ of the image, and any co-occurrence matrices calculated from the image. To maintain the first-order statistics of the image, we change a randomly chosen pixel with intensity 9 to intensity 7, for each $7 \rightarrow 9$ change. Such a simple modification method also results in slight changes to other second-order statistics, however, the dominant change occurs at $P(1, 9)$. The intensity pair $(1, 7)$ was chosen because it represented perforation edge pixels in the images (areas of high gradient), which are clearly visible.

As we will show, the resulting changes make negligible visual difference to the original image. That is, the texture images for class 1 and class 2 remain visually indistinct.

To calculate the discrimination matrix for this texture pair, we extract 100 texture tiles of 50×50 pixels from both texture classes. Co-occurrence matrices for each of these tiles are extracted, and the discriminatory power of each matrix element calculated as discussed in Section 4.2. The resulting discrimination matrix is shown in Figure B.2.

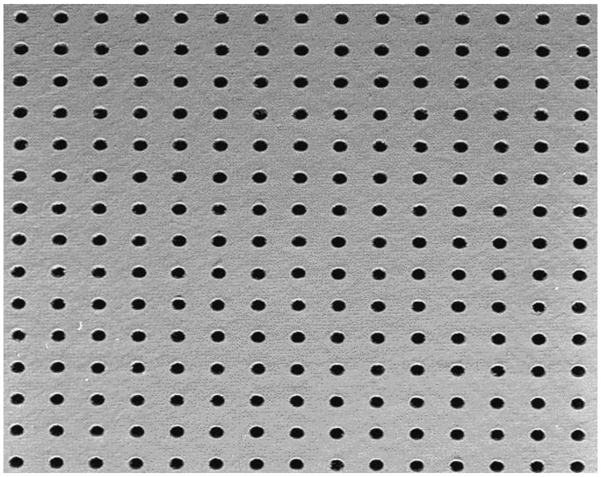


Figure B.1: Texture TILE.0009 from the VISTEX database.

We can see that the second-order statistic $P(1,9)$ clearly shows high discriminatory power, as expected. To link this statistic back to a physical property of the textures, it is simply a matter of searching the images for the occurrence of pixel pairs with intensities 1 and 9, separated by a displacement of 1 pixel. The results of this search are detailed in the right-hand image of Figure B.3, where the corresponding pixel pairs are marked. The mapping has correctly located the perforation edge pixels—the only areas of the image with significant gradient. We believe this to be the first demonstration

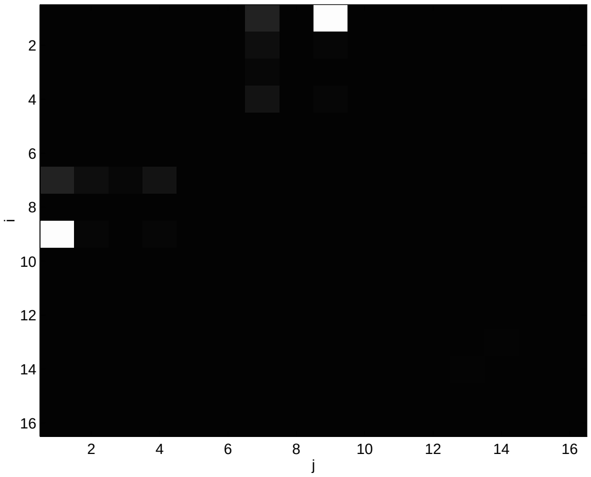


Figure B.2: The resulting discrimination matrix for the two texture image classes. We can clearly see that the second-order statistic $P(1,9)$ exhibits high discriminatory power.

of discriminatory power localisation in images. The technique should prove to be a valuable tool for use in the areas of image analysis and understanding.

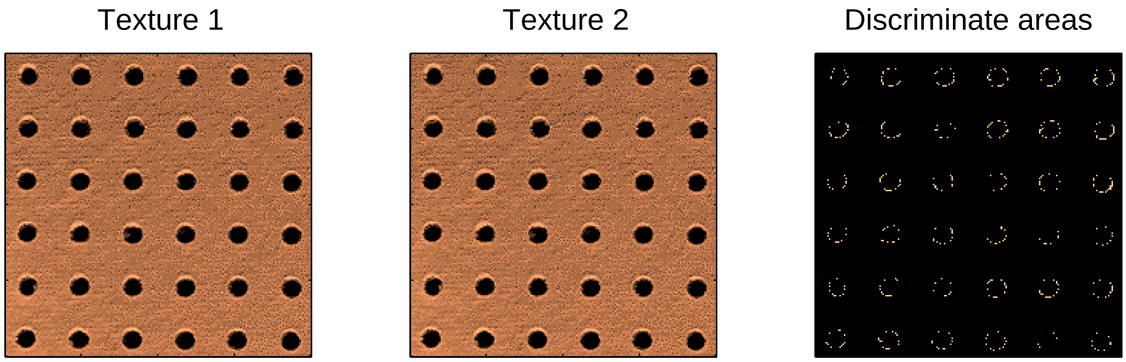


Figure B.3: Mapping a statistical difference back to physical areas of an image. The left and centre images show the two texture classes. The image on the right shows the physical areas of the two images which differ. For clarity, we show only a 200 × 200 pixel area of the original 512 × 512 images.

Appendix C

Connectivity Issues

Connectivity in digital images traces its roots to set theory and topology, and relies on an understanding of terms such as *paths*, *4-connected*, and *components* etc. Let S denote any subset of integer points (i, j) in an image lattice. A *4-path* is any N-tuple of points $(i_n, j_n),\ n = 1, \ldots, N$ where $N > 1$, and for all n the pair $(i_n, j_n), (i_{n+1}, j_{n+1})$ are horizontal or vertical neighbours. That is,

$$\Big|(i_n - i_{n+1}) + (j_n - j_{n+1})\Big| \leq 1,\ \forall n. \tag{C.1}$$

Similarly, an *8-path* is any N-tuple of points $(i_n, j_n),\ n = 1, \ldots, N$ where $N > 1$, and for all n the pair $(i_n, j_n), (i_{n+1}, j_{n+1})$ are horizontal, vertical, or diagonal neighbours. That is,

$$\max\Big\{|i_n - i_{n+1}|, |j_n - j_{n+1}|\Big\} \leq 1,\ \forall n. \tag{C.2}$$

Any pair of points $(i_1, j_1), (i_2, j_2)$ in S are said to be *4-connected* if there exists in S a 4-path having (i_1, j_1) as the first point and (i_2, j_2) as the last point. If there exists in S a 4-path containing all elements of S, then S is called *4-connected*, containing only one 4-connected component known as a *4-component*. 8-connected components are similarly defined.

In SGF terminology, a 4-component S in binary image I_b is simply a 4-connected region whose pixel intensities are either all '1'-valued or '0'-valued (for binary images I_b). That is,

$$I_b(i, j) = 1,\ \forall (i, j) \in S, \tag{C.3}$$

or

$$I_b(i, j) = 0,\ \forall (i, j) \in S. \tag{C.4}$$

Set theory dictates that, if 4-connectivity is used to describe a set S, then 8-connectivity is used for the set's complement S', and visa-versa. This is intuitively appealing, since if we consider a pair of diagonally adjacent neighbours in S are not connected and therefore not touching (i.e., 4-connectivity), then the complementary pair of neighbours belonging to S' should be regarded as touching (i.e., being 8-connected). This arrangement is shown in Figure C.1(a). In a practical sense, this is easy to visualise if we consider foreground regions being placed on a continuous background—foreground regions that are separated *must* be separated by background, as shown in Figure C.1(b). Rosenfeld (1970) points out that without such dual connectivity, Euler's theorem of polygonal networks breaks down. Further discussion can be found in Latecki (1993).

Applying complementary connectivity to digital images can lead to serious ramifications when applied to quantitative and qualitative measurement of image properties.

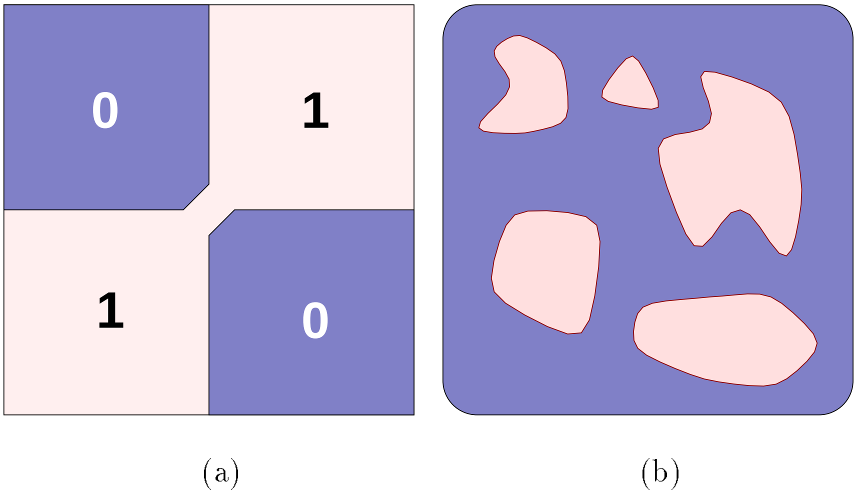


Figure C.1: An illustration of the complementary nature of connectivity in digital images. (a) If we consider a pair of diagonally adjacent neighbours are not connected and therefore not touching (i.e., 4-connectivity), then the complementary pair of neighbours should be regarded as touching (i.e., being 8-connected); (b) Separate foreground regions on a continuous background.

In terms of cervical cell analysis as proposed in this thesis, it means that the definition of a chromatin *clump* or region is different for euchromatin and heterochromatin. There is no intuitive reason to consider euchromatin clumps as being, say, 8-connected, yet heterochromatin clumps as being 4-connected. Moreover, such a situation results in feature measures which cannot be compared between the clump types (euchromatin and heterochromatin). For example, consider a nucleus with clump arrangement as shown in Figure C.2. If we consider 8-connectivity for euchromatin and 4-connectivity for heterochromatin, the nucleus would be considered as having 2 heterochromatin clumps of area A, yet having only one euchromatin clump with area $2A$. Features such as NCA and $\overline{CAREA}$ (Section 3) which measure the number of clumps per unit nuclear area, and average clump area, respectively, would provide vastly different measures for the two clump types. Clump irregularity measures on these regions would also yield wildly different results, despite both regions being equivalent.

It would be better to consider a nuclear image as being comprised of two types of regions which are both separate and equivalent entities, rather than to consider the image to consist of euchromatin foreground clumps over a background of heterochromatin, as Figure C.1(b) espouses. While this approach may mean that set theory rules and Euler's theorem are not completely satisfied, the benefits of such an approach, in terms of image understanding, far outweigh this disadvantage.

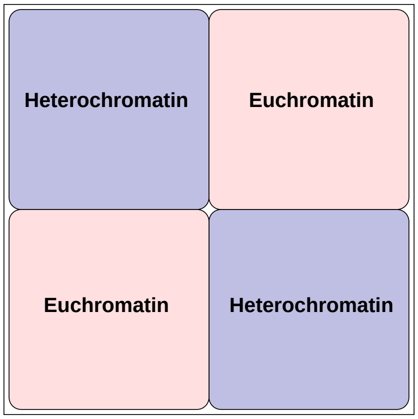


Figure C.2: The representation of euchromatin and heterochromatin clumps in digital images is best modelled using 4-connectivity for both clump types.

Appendix D

Irregularity Measure Characteristics

Several salient features of the measure $IRGL$ (page 62) require discussion. One measure of a region property closely related to irregularity is *circularity* or *compactness*, based on the perimeter P and area A of a region,

$$circularity = \frac{P}{\sqrt{4\pi A}}. \tag{D.1}$$

It can be seen that, for a circular region of radius r,

$$\begin{aligned} circularity &= \frac{2\pi r}{\sqrt{4\pi}\sqrt{A}} \\ &= \frac{2\pi r}{\sqrt{4\pi}\sqrt{\pi r^2}} \\ &= 1. \end{aligned} \tag{D.2}$$

A somewhat undesirable result of this definition is that larger measures of circularity indicate *less* circularity. Redefining and renaming this feature to *irregularity* is more appropriate. From equation (D.2), the irregularity of a circular region is now:

$$irregularity = \frac{2\pi r}{\sqrt{4\pi}.\sqrt{A}} - 1 \tag{D.3}$$

$$= \frac{\sqrt{\pi}r}{\sqrt{A}} - 1, \tag{D.4}$$

where r is now defined as the distance from the centre of gravity of the region to the farthest point on the perimeter,

$$r = \sup_{(x,y)\in A} \sqrt{(x-\bar{x})^2 + (y-\bar{y})^2}, \tag{D.5}$$

$$\bar{x} = \int_A x\, dx, \quad \bar{y} = \int_A y\, dy. \tag{D.6}$$

The additional -1 defines the irregularity of a circular area to be 0, which is intuitively appealing.

On a digital grid, equations (D.4) and (D.5) become

$$irregularity = \frac{\sqrt{\pi}\max_{i\in R}\sqrt{(x_i-\bar{x})^2 + (y_i-\bar{y})^2}}{\sqrt{A}} - 1 \tag{D.7}$$

where R is the set of indices of pixel co-ordinates in the region. We can define region area A as

$$A = |R|, \tag{D.8}$$

the cardinality of R (or the number of pixels in the region).

Using a digital grid introduces problems when dealing with regions containing only a single pixel:

$$\begin{aligned} irregularity &= \frac{\sqrt{\pi}\max\limits_{i\in R}\sqrt{(x_i-\bar{x})^2+(y_i-\bar{y})^2}}{\sqrt{|R|}}-1 \qquad \text{(D.9)} \\ &= \frac{\sqrt{\pi}.0}{1}-1 \\ &= -1. \end{aligned}$$

Chen et al. overcame this problem by the addition of a '1' term to the numerator,

$$IRGL = \frac{1+\sqrt{\pi}.\max\limits_{i\in R}\sqrt{(x_i-\bar{x})^2+(y_i-\bar{y})^2}}{\sqrt{|R|}}-1, \qquad \text{(D.10)}$$

where

$$\bar{x} = \frac{\sum\limits_{i\in R} x_i}{|R|}, \quad \bar{y} = \frac{\sum\limits_{i\in R} y_i}{|R|}. \qquad \text{(D.11)}$$

Defined in this way, the measure of irregularity has a number of salient characteristics:

1. For single pixel regions, the measure becomes,

$$\begin{aligned} IRGL &= \frac{1+\sqrt{\pi}.0}{1}-1 \qquad \text{(D.12)} \\ &= 0. \end{aligned}$$

2. As the radius of a circular region tends to infinity, or alternatively, as the sampling grid spacing ϵ approaches 0,

$$\begin{aligned} IRGL &= \lim_{\epsilon\to 0}\left(\frac{1+\sqrt{\pi}.\max\limits_{i\in R}\sqrt{(x_i-\bar{x})^2+(y_i-\bar{y})^2}}{\sqrt{|R|}}-1\right) \qquad \text{(D.13)} \\ &= \lim_{\epsilon\to 0}\left(\frac{1}{\sqrt{|R|}}+\sqrt{\pi}\frac{\max\limits_{i\in R}\sqrt{(x_i-\bar{x})^2+(y_i-\bar{y})^2}}{\sqrt{|R|}}-1\right) \\ &= 0+1-1 \\ &= 0. \end{aligned}$$

3. The measure $IRGL$ is invariant under rotation and translation.
   This can be easily seen, because $|R|$ and $\max_{i \in R} \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}$ remain constant.

Appendix E

Standard Grey Level Co-occurrence Features

Table E.1: Commonly-used GLCM features.

Features	Equations
Energy:	$F1 = \sum_{i,j} P(i,j)^2;$
Entropy:	$F2 = -\sum_{i,j} P(i,j) \log P(i,j);$
Homogeneity:	$F3 = \sum_{i,j} \frac{1}{1+(i-j)^2} P(i,j);$
Inertia:	$F4 = \sum_{i,j} (i-j)^2 P(i,j);$
Correlation:	$F5 = -\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y} P(i,j);$
Shade:	$F6 = \sum_{i,j} (i+j-\mu_x-\mu_y)^3 P(i,j);$
Prominence:	$F7 = \sum_{i,j} (i+j-\mu_x-\mu_y)^4 P(i,j);$
Variance:	$F8 = \sum_{i,j} (i-\mu_x)^2 P(i,j)$
Sum Average:	$F9 = \sum_{i=2}^{2N_g} i P_{x+y}(i)$
Sum Entropy:	$F10 = -\sum_{i=2}^{2N_g} P_{x+y}(i) \log P_{x+y}(i)$
Sum Variance:	$F11 = \sum_{i=2}^{2N_g} (i - F9)^2 P_{x+y}(i)$
Difference Average:	$F12 = \sum_{i=0}^{N_g-1} i P_{x-y}(i)$
Difference Entropy:	$F13 = \sum_{i=0}^{N_g-1} -P_{x-y}(i) \log P_{x-y}(i)$
Difference Variance:	$F14 = \sum_{i=0}^{N_g-1} (i - F12)^2 P_{x-y}(i)$
Information Measure:	$F15 = \frac{F2-HXY1}{\max(HX,HY)}$
Coefficient of Varaition	$F16 = \frac{\sigma(P(i,j)}{\mu(P(i,j))}$
Peak Transition Probability	$F17 = \max(P(i,j))$
Diagonal Variance	$F18 = \text{variance of } P(i,j)$
Diagonal Moment	$F19 = \sum_{i,j} (0.5\lvert i-j \rvert P(i,j))^{\frac{1}{2}}$
Second Diagonal Moment	$F20 = \sum_{i,j} (0.5\lvert i-j \rvert P(i,j))$
Triangular Symmetry	$F21 = \sum_{i,j} \lvert P(i,j) - P(j,i) \rvert$

$$\mu_x = \sum_i i \sum_j P(i,j), \mu_y = \sum_j j \sum_i P(i,j);$$

$$\sigma_x = \sum_i (i-\mu_x)^2 \sum_j P(i,j), \sigma_y = \sum_j (j-\mu_y)^2 \sum_i P(i,j);$$

$$P_x(i) = \sum_j P(i,j), P_y(j) = \sum_i P(i,j);$$

$$P_{x+y}(k) = \sum_{i,j \mid i+j=k} P(i,j), P_{x-y}(k) = \sum_{i,j \mid |i-j|=k} P(i,j);$$

HX and HY are the entropies of $P_x(i)$ and $P_y(j)$ respectively;

$$HXY1 = -\sum_{i,j} P(i,j) \log(P_x(i) P_y(j)).$$

Appendix F

Quadratic Surface Fitting to Discrimination Matrix

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Maple V code for quadratic surface fitting
% of lower half of the GLCM matrix
% f(i,j) is the surface to be fit
% g(i,j) is the fitted surface
% A,B,C,D,E,F are surface coefficients
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Surf_Error:=sum(sum((f(i,j)-g(i,j))^2,j=1..i), i=1..N):
g(i,j):=A*i^2+B*j^2+2*C*i*j+2*D*i+2*E*j+F:

%Differentiation
dA:=diff(Surf_Error,A):
dB:=diff(Surf_Error,B):
dC:=diff(Surf_Error,C):
dD:=diff(Surf_Error,D):
dE:=diff(Surf_Error,E):
dF:=diff(Surf_Error,F):

%Solve for the 6 coefficients
exp2:=solve({dA,dB,dC,dD,dE,dF},{A,B,C,D,E,F}):
assign(exp2):


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Matlab code for quadratic surface fitting
% of lower half of the GLCM matrix, using the six
% closed-form solutions for A,B,C,D,E,F from Maple
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [maskfit,se]=Fit_quad(mask);
%function [maskfit,se]=Fit_quad(mask);
%Fits a quadric surface estimate to an inputted mask;
%'maskfit' is the fitted surface
%'se' is the squared error

N=16; %co-occurrence matrix order
se=0; %mean square error
z=zeros(16);
f=mask;
```

```
%define constants which simplify the expressions
%for A,B,C,D,E,F from Maple
s1=0; s2=0; s3=0; s4=0; s5=0; s6=0;
for i=1:N
 for j=1:i
  s1=s1-2*f(i,j);
  s2=s2-2*f(i,j)*i;
  s3=s3-2*f(i,j)*j;
  s4=s4-2*f(i,j)*j^2;
  s5=s5-2*f(i,j)*i^2;
  s6=s6-2*f(i,j)*i*j;
 end;
end;

%Here are the closed-from expressions from Maple
A= 90*(-s1*N^2+s1*N-4*s3*N+6*s2*N-s3-6*s5-s4+6*s6)/(-N-2+2*N^2+N^3)
    /(N^2+N-6)/N;
B= -90*(3*s2+s5-6*s6+2*s1-6*s3+6*s4)/N/(-5*N^3-15*N^2+4*N+12+N^5+3*N^4);
C= -30*(8*s2*N+2*s3*N-3*s1*N+6*s2-3*s3-18*s6+9*s4-4*s1*N^3-10*N^2*s3
    +20*N^2*s2-s1*N^2+12*s6*N-18*s5*N)/N/(-5*N^3-15*N^2+4*N+12+N^5+3*N^4);
D= 30*(-20*N^2*s3+10*N^2*s2+10*s1*N^2-12*s5*N+24*s6*N-2*s2*N+10*s1*N
    -20*s3*N+18*s4-3*s5-18*s3+ 6*s1+3*s2-6*s6)/(-N-2+2*N^2+N^3)
    /(N^2+N-6)/N;
E= 6*(-6*N^4*s1+20*N^3*s2-12*s1*N^3+5*N^2*s2-28*s1*N^2-15*s5*N^2
    +50*N^2*s3+50*s3*N+15*s2*N-22*s1*N-60*s6*N+15*s5*N-30*s4
    +30*s3-12*s1)/(N+1)/(N^4+2*N^3-7*N^2-8*N+12)/N;
F= -180*(2*s2*N-4*s3*N+2*s1*N-3*s2-3*s5+s3+8*s6-3*s4)/N
    /(-5*N^3-15*N^2+4*N+12+N^5+3*N^4);

for i=1:N
 for j=1:i
  g(i,j) = A*i^2 + B*j^2 + C*i + D*j + E + F*i*j;
  g(j,i)=g(i,j);
  se=se+((mask(i,j)-z(i,j))^2)/((N*(N+1))/2);
 end;
end;
disp(sprintf('Surface coefficients are %i %i %i %i %i %i\n', A,B,C,D,E,F));
disp(sprintf('Surface error is %i\n', se));
maskfit=g;
```

Appendix G

Glossary of Medical Terms

Anaplasia A loss of differentiation of cells, and of their orientation to one another. A characteristic of tumour cells.

Anaplastic Restoring a lost or absent part. Characterised by anaplasia or reversed development.

Artifact Any artificial product. In histology or microscopy, any structure or feature that has been introduced by processing a tissue.

Benign Not malignant; nor recurrent; favourable for recovery.

Cancer A cellular tumour, the natural course of which is fatal. Cancer cells, unlike benign tumour cells, exhibit the properties of invasion and metastasis and are highly anaplastic. Cancers are divided into the two broad categories of carcinoma and sarcoma.

Carcinoma A malignant new growth made up of epithelial cells tending to infiltrate the surrounding tissues and giving rise to metastases.

Carcinoma in situ A neoplastic entity wherein the tumour cells still lie within the epithelium of origin, without invasion of the basement membrane; popularly applied to such cells in the uterine cervix.

Cervical intraepithelial neoplasia (CIN) Dysplastic changes beginning at the squamo-columnar junction in the uterine cervix which may be precursors of squamous cell carcinoma: grade 1, mild dysplasia involving the lower one-third or less of the epithelial thickness; grade 2, moderate dysplasia with one-third to two-thirds involvement; grade 3, severe dysplasia or carcinoma in situ, with two-thirds to full thickness involvement.

Chromatin The more readily stainable portion of the cell nucleus, forming a network of nuclear fibrils within the achromatin of a cell. It is a deoxyribonucleic acid (DNA) attached to a protein structure base and is the carrier of the genes in inheritance. It occurs in two interchangeable states, euchromatin and heterochromatin, and during cell division it coils and folds to form the chromosomes.

Cytology The study of cells, their origin, structure, function, and pathology.

Cytopathologist An expert in the study of cells in disease; a cellular pathologist.

Cytoplasm The protoplasm of a cell exclusive of that of the nucleus.

Dysplasia Abnormality of development; in pathology, alteration in size, shape, and organisation of adult cells.

Dysplastic Marked by dysplasia.

Endocervical Pertaining to the interior of the cervix uteri.

Epithelium The inner mucous membrane of the uterus, the thickness and structure of which vary with the phase of the menstrual cycle.

Epithelial Pertaining to or composed of epithelium.

Histology That department of anatomy which deals with the minute structure, composition, and function of the tissues.

Hyperplasia The abnormal multiplication or increase in the number of normal cells in normal arrangement in a tissue.

In Situ In the natural or normal place; confined to the site of origin without invasion of neighbouring tissues.

Intermediate Placed between; intervening; resembling, in part, each of two extremes.

Invasion 1. The attack or onset of a disease. 2. The simple harmless entrance of bacteria into the body or their deposition in the tissues, as distinguished from infection.

Invasive Having the quality of invasiveness.

Leukocyte A white blood cell.

Malignant Tending to become progressively worse and to result in death. Having the properties of anaplasia, invasion, and metastasis; said of tumours.

Metaplasia Abnormal transformation of an adult, fully differentiated tissue of one kind into a differentiated tissue of another kind; an acquired condition, in contrast to heteroplasia.

Metastases The transfer of disease from one organ or part to another not directly connected with it. The capacity to metastasise is a characteristic of all malignant tumours.

Neoplasia The formation of a neoplasm, i.e., the progressive multiplication of cells under conditions that would not elicit, or would cause cessation of, multiplication of normal cells.

Neoplasm Any new or abnormal growth; specifically a new growth of tissue in which the growth is uncontrolled and progressive. Malignant neoplasms are distinguished from benign, in that the former show a greater degree of anaplasia and have the properties of invasion and metastasis. Also called *tumour*.

Nucleus A cell nucleus: a spheroid body within a cell, consisting of a number of characteristic organelles visible with the optical microscope, a thin nuclear membrane, a nucleolus or nucleoli, irregular granules of chromatin and linin, and diffuse nucleoplasm.

Papanicolaou smear test, Pap smear test An exfoliative cytological staining procedure for the detection and diagnosis of various conditions, particularly malignant and pre-malignant conditions of the female genital tract (cancer of the vagina, cervix, and endometrium), in which cells which have been desquamated from the genital epithelium are obtained by smears, fixed and stained, and examined under the microscope for evidence of pathologic changes.

Papanicolaou's stain A method of staining smears of various body secretions, from the respiratory, digestive, or genitourinary tract, for the examination of exfoliated cells, to detect the presence of a malignant process.

Parabasal Pertaining to or situated beside or against a base.

Pathology That branch of medicine which treats of the essential nature of disease, especially of the structural and functional changes in tissues and organs of the body which cause or are caused by disease.

Protoplasm The viscid, translucent, polyphasic colloid with water as the continuous phase that makes up the essential material of all plant and animal cells. The protoplasm surrounding the nucleus is known as the cytoplasm, and that composing the nucleus is the nucleoplasm.

Pyknosis A thickening or condensation; specifically, a condensation and reduction in size of the cell or its nucleus, usually associated with hyperchromatosis; nuclear pyknosis is a stage of necrosis.

Sarcoma A tumour made up of a substance like the embryonic connective tissue; tissue composed of closely packed cells embedded in a fibrillar or homogeneous substance. Sarcomas are often highly malignant.

Squamous Scaly, or plate-like.

Squamous cell carcinoma A malignant neoplasm derived from stratified squamous epithelium, but which may also occur in sites, such as bronchial mucosa, where glandular or columnar epithelium is normally present;

Stain Any dye, reagent, or other material used in producing colouration, such as a substance used in colouring tissues or microorganisms for microscopical study.

Stoichiometric Relating to the proportions in which chemicals combine to form compounds and the weight relations in chemical reactions. A stoichiometric stain produces a total optical density proportional to the DNA content.

Bibliography

Abmayr, W., Burger, G. & Soost, H. J. (1979), 'Progress report of the TUDAB project for automated cancer cell detection', *The Journal of Histochemistry and Cytochemistry* **27**(1), 604–612.

Adams, R. & Bischof, L. (1994), 'Seeded region growing', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(6), 641–647.

Albregtsen, F., Kanagasingam, Y., Farrants, G. & Danielsen, H. E. (1992), Texture discrimination of normal and malignant mouse liver cell nuclei, *in* P. Johansen & S. Olsen, eds, 'Theory and Applications of Image Analysis: Selected Papers from the 7th Scandinavian Conference on Image Analysis', World Scientific Publishing Company, pp. 324–335.

AMS (1991), *User's Guide to AMSFonts Version 2.1*, American Mathematical Society, Providence, RI.

Anderson, T. L. (1994), Automatic screening of conventional papanicolaou smears, *in* Wied, Bartels, Rosenthal & Schenck (1994), pp. 306–311.

Augusteijn, M. F., Clemens, L. E. & Shaw, K. A. (1995), 'Performance evaluation of texture measures for ground cover identification in satellite images by means of a neural network classifier', *IEEE Transactions on Geoscience and Remote Sensing* **33**(3), 616–626.

Australian Bureau of Statistics (1994), *Causes of Death, Australia*, Australian Government Publishing Service, Canberra.

Australian Institute of Health (1991), *Cervical Cancer Screening in Australia: Options for Chance*, Australian Government Publishing Service, Canberra.

Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton University Press, Princeton.

Bengtsson, E. & Nordin, B. (1994), Densitometry, morphometry, and textural analysis as tools in quantitative cytometry and automated cancer screening, *in* 'The Automation of Cervical Cancer Screening', Igaku-Shoin Medical Publishing Inc., New York, pp. 21–43.

Besag, J. (1974), 'Spatial interactions and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society* **36**(2), 192–225.

Besag, J. (1986), 'On the statistical analysis of dirty pictures', *Journal of the Royal Statistical Society* **48**(3), 259–302.

Bethke, A. D. (1981), Genetic Algorithms as Function Optimizers, PhD thesis, University of Michigan.

Bhattacharyya, A. (1943), 'On a measure of divergence between two statistical populations defined by their probability distributions', *Bulletin of the Calcutta Mathematics Society* **35**, 99–109.

Bibbo, M., Bartels, P. H., Dutch, H. E. & Wied, G. L. (1984), 'Computed cell image information', *Monographs in clinical cytology* **9**, 62–100.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society* **26**, 211–252.

Bradley, A. P. (1996), Machine Learning for Medical Diagnostics: Techniques for Feature Extraction, Classification, and Evaluation, PhD thesis, University of Queensland, Australia.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Wadsworth International Group, Belmont, CA., USA.

Brodatz, P. (1966), *Textures: A Photographic Album for Artists and Designers*, Dover Publications, Inc., New York.

Brugal, G., Garbay, C., Giroud, F. & Adelh, D. (1979), 'A double scanning microphotometer for image analysis hardware, software and biomedical applications', *The Journal of Histochemistry and Cytochemistry* **27**, 144–152.

Chanen, W. (1990), 'The CIN saga—the biological and clinical significance of cervical intraepithelail neoplasia', *Australian and New Zealand Journal of Obstetrics and Gynaecology* pp. 18–23.

Chen, C.-C. (1988), Markov Random Fields in Image Analysis, PhD thesis, Michigan State University.

Chen, C.-C. & Dubes, R. C. (1990), Discrete MRF parameters as features for texture classification, *in* 'IEEE International Conference on Systems, Man and Cybernetics', IEEE.

Chen, Y. Q., Nixon, M. S. & Thomas, D. W. (1995), 'Statistical geometric features for texture classification', *Pattern Recognition* **28**(4), 537–552.

Chomet, J. & Chomet, J. (1989), *Smear Tests—Cervical Cancer: Its Prevention and Treatment*, Harper Collins, London.

Christen, R., Xiao, J., Minimo, C., Gibbons, G., Fitzpatrick, B., Galera-Dividson, H., Bartels, P. & Bibbo, M. (1993), 'Chromatin texture features in hematoxylin and eosin-stained prostate tissue', *Analytical and Quantitative Cytology and Histology* **15**(6), 383–388.

Conners, R. W. & Harlow, C. A. (1978), 'Equal probability quantizing and texture analysis of radiographic images', *Computer Graphics and Image Processing* **8**, 447–463.

Conners, R. W. & Harlow, C. A. (1980), 'A theoretical comparison of texture algorithms', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-2**(3), 204–222.

Conners, R. W., Trivedi, M. M. & Harlow, C. A. (1984), 'Segmentation of a high-resolution urban scene using texture operators', *Computer Vision, Graphics, and Image Processing* **25**, 273–310.

Cross, G. R. & Jain, A. K. (1983), 'Markov random field texture models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-5**(1), 25–39.

Danielsen, H. E., Farrants, G. & Ruth, A. (1989), 'Characterization of chromatin structure by image analysis—A method for the assessment of changes in chromatin organization', *Scanning Microscopy Supplement* **3**, 297–302.

Davis, L. S., Johns, S. A. & Aggarwal, J. K. (1979), 'Texture analysis using generalized co-occurrence matrices', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**(3), 251–259.

Department of Health, Housing, Local Government and Community Services (1991), *Screening to Prevent Cancer of the Cervix*, Australian Government Publishing Service, Canberra.

Department of Health, Housing, Local Government and Community Services (1993), *Making the Pap Smear Better: Report of the Steering Group on Quality Assurance in Screening for the Prevention of Cervical Cancer*, Australian Government Publishing Service, Canberra.

DIC (1995), *Proceedings of DICTA-95, the Third Biennial Conference on Digital Image Computing: Techniques and Applications*, Australian Pattern Recognition Society, Brisbane, Australia.

Dubes, R. C. & Jain, A. K. (1989), 'Random field models in image analysis', *Journal of Applied Statistics* **16**(2), 131–164.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathemathics, Philadelphia.

Elfadel, I. M. & Picard, R. W. (1995), 'Gibbs random fields, cooccurrences, and texture modeling', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(1), 24–37.

Foley, D. H. (1972), 'Considerations of sample and feature size', *IEEE Transactions on Information Theory* **5**, 618–626.

Friedman, J. H. (1994), An overview of predictive learning and function approximation, *in* V. Cherkassky, J. H. Friedman & H. Wechsler, eds, 'From Statistics to Neural Networks: Theory and Pattern Recognition Applications', Vol. 136 of *NATO ASI Series: Computer and Systems Sciences*, Springer–Verlag, Berlin, pp. 1–61.

Galloway, M. M. (1975), 'Texture analysis using gray level run lengths', *Computer Graphics and Image Processing* **4**, 172–179.

Garcia, G. L. (1986), Feasibility Of Contextual Analysis in an Automated Cervical Prescreening System, PhD thesis, Worcester Polytechnic Institute.

Garner, D., Ferguson, G. & Palcic, B. (1994), The cyto-savant system, *in* Grohs & Husain (1994), pp. 294–304.

Gay, J. D., Donaldson, L. D. & Goellner, J. P. (1985), 'False-negative results in cervical cytologic studies', *Acta Cytologica* **29**, 1043–1046.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc., U.S.A.

Gonzalez, R. C. & Woods, R. E. (1993), *Digital Image Processing*, Addison-Wesley Publishing, U.S.A.

Gool, L. V., Dewaele, P. & Oosterlinck, A. (1985), 'Texture analysis anno 1983', *Computer Vision, Graphics, and Image Processing* **29**, 336–357.

Gotlieb, C. C. & Kreyszig, H. E. (1990), 'Texture descriptors based on co-occurrence matrices', *Computer Vision, Graphics, and Image Processing* **51**, 70–86.

Grefenstette, J. J. (1992), Optimization of control parameters for genetic algorithms, *in* B. P. Buckles & F. E. Petry, eds, 'Genetic Algorithms', IEEE Computer Society Press, Los Alamitos, CA, U.S.A., pp. 5–11.

Grohs, H. K. & Husain, O. A. N., eds (1994), *Automated Cervical Cancer Screening*, Igaku-Shoin, New York.

Gruner, O. C. (1916), 'Study of the changes met with the leukocytes in certain cases of malignant disease', *British Journal of Surgery* **3**, 506–522.

Hall, E. L., Kruger, R. P. & Turner, F. A. (1974), 'An optical-digital system for automatic processing of chest X-rays', *Optical Engineering* **13**, 250–257.

Hallouche, F. (1993), Image Processing and Statistical Pattern Recognition in the Computer-Aided Analysis of Cell Malignancy, PhD thesis, Newcastle University.

Hand, D. J. (1981), *Discrimination and Classification*, John Wiley and Sons, USA.

Hand, D. J. (1982), 'Branch and bound in statistical data analysis', *The Statistician* **30**(1), 1–13.

Hanselaar, A., MacAulay, C., Palcic, B., Garner, D. & LeRiche, J. (1992), 'Discrimination between progressive and regressive cervical intraepithelial neoplasia (CIN) by DNA-cytometry', *Analytical Cellular Pathology* **4**, 165.

Haralick, R. M. (1979), 'Statistical and structural approaches to texture', *Proceedings of the IEEE* **67**(5), 786–804.

Haralick, R. M., Shanmugam, K. & Dinstein, I. (1973), 'Textural features for image classification', *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**, 610–621.

Harms, H., Gunzer, U., Baumann, I. & Serbouti, S. (1993), 'Malignancy-associated changes in monocytes and lymphocytes in acute leukemias measured by high-resolution image processing', *Analytical and Quantitative Cytology and Histology* **15**(6), 371–379.

Hauta-Kasari, M., Parkkinen, J., Jaaskelainen, T. & Lenz, R. (1996), Generalized co-occurrence matrix for multispectral texture analysis, *in* 'Proceedings 13th IAPR International Conference on Pattern Recognition', IEEE Computer Society Press, Los Alamitos, CA, Vienna, Austria, pp. B785–789.

Hjort, N. (1986), Notes on the theory of statistical symbol recognition, Report 778, Norwegian Computing Centre, Oslo.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Michigan.

Hong, T. H. & Rosenfeld, A. (1984), 'Compact region extraction using weighted pixel linking in a pyramid', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, 222–229.

Jackway, P. T. & Walker, R. F. (1995), Scaled gradient watersheds and cell feature extraction, *in* DIC (1995), pp. 68–73.

James, M. (1985), *Classification Algorithms*, Collins, London.

Julesz, B. (1962), 'Visual pattern discrimination', *IRE Transactions on Information Theory* **IT-8**(2), 84–92.

Julesz, B. (1981), 'Textons, the elements of texture perception, and their interactions', *Nature* **290**, 91–97.

Julesz, B. & Bergen, J. R. (1983), 'Textons, the fundamental elements in preattentive vision and perception of textures', *The Bell System Technical Journal* **62**(6), 1619–1645.

Kashyap, R. L., Chellappa, R. & Khotanzad, A. (1982), 'Texture classification using features derived from random field models', *Pattern Recognition Letters* **1**, 43–50.

Kittler, J. (1978), Feature set search algorithms, *in* C. H. Chen, ed., 'Pattern Recognition and Signal Processing', Sijthoff and Noordhoff, The Netherlands.

Kittler, J. (1986), Feature selection and extraction, *in* T. Y. Young & K.-S. Fu, eds, 'Handbook of Pattern Recognition and Image Processing', Acedamic Press, San Diego, pp. 60–83.

Knesel Jr, E. A., Geyer, J. W., Gahm, T., Nguyen, T., Fischer, J. & Dorrer, R. (1994), The Roche cytology systems: Cyto-Rich and AutoCyte, *in* Grohs & Husain (1994), pp. 294–304.

Knuth, D. E. (1986), *The TEXbook*, Adison-Wesley Publishing Co., Reading. Ma.

Komitowski, D. & Janson, C. (1990), 'Quantitative features of chromatin structure in the prognosis of breast cancer', *Cancer* **65**(12), 2725–2730.

Komitowski, D. & Zinser, G. (1985), 'Quantitative description of chromatin structure during neoplasia by the method of image processing', *Analytical and Quantitative Cytology and Histology* **7**(3), 178–182.

Kopp, R. E., Lisa, J., Mendelsohn, L., Pernick, B., Stone, H. & Wohlers, R. (1976), 'Coherent optical processing of cervical cytologic samples', *The Journal of Histochemistry and Cytochemistry* **24**, 122–137.

Kruger, R. P., Thompson, W. B. & Turner, F. A. (1974), 'Computer diagnosis of pneumoconiosis', *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-4**, 40–49.

Kurman, R. J. & Solomon, D. (1994), *The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses*, Springer-Verlag, New York.

Lachenbruch, P. A. (1975), *Discriminant Analysis*, Hafner Press, New York.

Lamport, L. (1994), *LATEX: A Document Preparation System*, 2nd edn, Adison-Wesley Publishing Co., Reading. Ma.

Latecki, L. (1993), 'Topological connectedness and 8-connectedness in digital pictures', *CVGIP: Image Understanding* **57**(2), 261–262.

Laws, K. I. (1979), Texture energy measures, *in* 'Proceedings of the Image Understanding Workshop', pp. 47–51.

Lee, Y. H. (1985), 'Algorithms for mathematical morphological operations with flat top structuring elements', *SPIE Applications of Digital Image Processing* **8**, 33–45.

Lendaris, G. O. & Stanley, G. L. (1970), 'Diffraction pattern sampling for automatic pattern recognition', *Proceedings of the IEEE* **58**, 198–216.

Liu, S. S. & Jernigam, M. E. (1990), 'Texture analysis and discrimination in additive noise', *Computer Vision, Graphics, and Image Processing* **49**, 52–67.

Makins, M., ed. (1992), *Collins English Dictionary — Australian Edition*, Harper Collins, Sydney.

Mango, L. J. & Herriman, J. M. (1994), The PAPNET cytological screening system, *in* Wied et al. (1994), pp. 320–334.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

Meyer, F. (1980*a*), Feature extraction by mathematical morphologogy in the field of quantitative cytology, *in* J. Sklansky & J. Bisconte, eds, 'Biomedical Images and Computers', Vol. 17 of *Lecture Notes in Medical Informatics*, Springer-Verlag, Berlin, pp. 56–65.

Meyer, F. (1980*b*), 'Quantitative analysis of the chromatin of lymphocytes: An essay on comparative structuralism', *Blood Cells* **6**, 159–172.

Meyer, F. (1981), 'Cytology automation with mathematical morphology', *Biol. Cell* **41**, 1–5.

Meyer, F. (1982), The perceptual graph: A new algorithm, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing', IEEE, Paris, France, pp. 1932–1935.

Meyer, F. & Beucher, S. (1990), 'Morphological segmentation', *Journal of Visual Communication and Image Representation* **1**(1), 21–46.

Meyer, F. & van Driel, A. (1980), 'Automatic screening of papancolaou stained cervical smears with the T.A.S.', *Microscopica Acta* **Supplement 4**, 82–95.

Neiburgs, H. E. (1968), 'Recent progress in interpretation of malignancy associated changes (MAC)', *Acta Cytologica* **12**, 445–453.

Nguyen, N. G., Poulsen, R. S. & Louis, C. (1983), 'Some new color features and their application to cervical cell classification', *Pattern Recognition* **16**, 401–411.

Noguchi, Y. (1985), Studies on Automated Cancer Cell Detection System Using Multispectral Cell Images, PhD thesis, Tokyo.

Ohanian, P. P. & Dubes, R. C. (1992), 'Performance evaluation for four classes of textural features', *Pattern Recognition* **25**(8), 819–833.

Ojala, T., Pietikainen, M. & Harwood, D. (1996), 'Quantitative description of chromatin structure during neoplasia by the method of image processing', *Pattern Recognition* **29**(1), 51–59.

Paget, R. & Longstaff, D. (1996), A nonparametric multiscale Markov random field model for synthesising natural textures, *in* 'Proceedings of ISSPA-96, Fourth International Symposium on Signal Processing and its Applications', Gold Coast, Australia.

Paget, R., Longstaff, D. & Lovell, B. (1997), Texture classification using nonparametric Markov random fields, *in* 'Proceedings of DSP'97, the 13th International Conference on Digital Signal Processing', Santorini, Greece, pp. 67–70.

Palcic, B. & MacAulay, C. (1994), Malignancy associated changes—can they be employed clinically?, *in* Wied et al. (1994), pp. 157–165.

Papanicolaou, G. N. (1945), 'Diagnosis of uterine cancer by the vaginal smear', *New York Street Journal of Medicine* **45**, 1336–1338.

Papanicolaou, G. N. & Traut, H. F. (1943), 'Diagnosis of uterine cancer by vaginal smear', *New York Commonwealth Fund* .

Payne, P. W., Lam, S., LeRiche, J. C., MacAulay, C., Ikeda, N. & Palcic, B. (1994), 'Image cytometry: Prediction of progression and regression of dysplasia in bronchial biopsies', *Analytical Cellular Pathology* **6**(3), 205.

Philipp, S. & Smadja, M. (1994), 'Approximation of granular textures by quadric surfaces', *Pattern Recognition* **27**(8), 1051–1063.

Picard, R., Graczyk, C., Mann, S., Wachman, J., Picard, L. & Campbell, L. (1995), 'Vistex vision texture database', http://www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html.

Pitts, D. E., Premkumar, S. B., Houston, A. G., Babaian, R. J. & Troncoso, P. (1993), Texture analysis of digitized prostate pathologic cross-section, *in* 'Proceedings of the SPIE—Medical Imaging 1993', Vol. 1898, SPIE, Newport Beach, California, pp. 465–470.

Ploem, J. S., Goyarts-Veldstra, L., van Driel-Kulker, A. M. J., Zaaner, J. J. & Meyer, I. (1983), 'Progress report of qualitative and analytical cytology using LEYTAS in clinical applications', *Analytical and Quantitative Cytology* **5**, 215.

Ploem, J. S., Werwoerd, N., Bonnet, J. & Koper, G. (1979), 'An automated microscope for quantitative cytology combining television image analysis and stage scanning microphotometry', *The Journal of Histochemistry and Cytochemistry* **27**(1), 136–143.

Poulsen, R. S. (1973), Automated Pre-screening of Cervical Cytology, PhD thesis, McGill University, Montreal, Canada.

Pressman, N. J. (1976), 'Markovian analysis of cervical cell images', *The Journal of Histochemistry and Cytochemistry* **24**, 689–695.

Pressman, N. J. (1986), Optical Texture Analysis for Automatic Cytology and Histology: A Markovian Approach, PhD thesis, University of Pennsylvania.

Prewitt, J. M. S. (1972), Parametric and non-parametric recognition by computer: An application to leukocyte image processing, *in* M. Rubinoff, ed., 'Advances in Computers', Academic Press, New York.

Pudil, P., Novovicova, J. & Kittler, J. V. (1994), 'Floating search methods in feature selection', *Pattern Recognition Letters* **15**, 1119–1125.

Quenouille, M. (1949), 'Approximate tests of correlation in time series', *Journal of the Royal Statistical Society* **11**, 18–84.

Rosenfeld, A. (1970), 'Connectivity in digital pictures', *Journal of the Association for Computing Machinery* **17**(1), 146–160.

Rosenfeld, A. & Kak, A. (1982), *Digital Picture Processing*, Vol. 2, Academic, Orlando, Florida.

Sakia, R. M. (1992), 'The Box-Cox transformation technique: A review', *The Statistician* **41**, 169–178.

Salembier, P., Gasull, A., Marques, F. & Sayrol, E. (1992), Morphological detection based on size and contrast criteria: Application to cell detection, *in* J. Morucci, R. Plonsey, J. Coatrieux & S. Laxminarayan, eds, 'Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society', IEEE, pp. 1930–1931.

Seber, G. A. F. (1984), *Multivariate Observations*, John Wiley & Sons, New York.

Shen, H. C. & Bie, C. Y. C. (1992), 'Feature frequency matrices as texture representation', *Pattern Recognition Letters* **13**, 195–205.

Shen, H. C., Bie, C. Y. C. & Chiu, D. K. Y. (1993), 'A texture-based distance measure for classification', *Pattern Recognition* **26**(9), 1429–1437.

Siew, L. H., Hodgson, R. M. & Wood, E. J. (1988), 'Texture measures for carpet wear assessment', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(1), 92–105.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Spann, M. & Wilson, R. (1985), 'A quad-tree approach to image segmentation that combines statistical and spatial information', *Pattern Recognition* **18**, 257–269.

Stallings, W. (1975), 'The morphology of Chinese characters: A survey of models and applications', *Computing in the Humanities* **9**.

Sun, C. & Wee, W. G. (1983), 'Neighboring gray level dependence matrix for texture classification', *Computer Vision, Graphics, and Image Processing* **23**, 341–352.

Tanaka, N., Ikeda, H., Ueno, T., Mukawa, A., Watanabe, S., Okamoto, K., Hosoi, S. & Tsunekawa, S. (1987), 'Automated cytologic screening system (CYBEST model 4): An integrated image cytometry system', *Applied Optics* **26**(16), 3301–3307.

Tanaka, N., Ikeda, H., Ueno, T., Watanabe, S., Imasato, Y., Tsunekawa, S., Okamoto, Y., Kashida, R. & Mukawa, A. (1979), Experimental practical use of automated screening system (CYBEST) for the mass-screening of gynecological samples, *in* 'Compendium on Diagnostic Cytology', Tutorials of Cytology, Chicago, pp. 123–134.

Tanaka, N., Ueno, T., Ishikawa, A., Konoike, K., Shimaoka, Y., Yamauchi, K., Hosoi, S., Okamoto, Y. & Tsunekawa, S. (1980), 'New automated cytologic screening system: CYBEST model 3 its outline and field test trials', *Analytical and Quantitative Cytology* **2**, 306.

Trivedi, M. M., Harlow, C. A., Conners, R. W. & Goh, S. (1984), 'Object detection based on gray level co-occurrence', *Computer Vision, Graphics, and Image Processing* **28**, 199–219.

Tucker, J. H. (1979), 'An image analysis system for cervical cytology automation using nuclear DNA content', *The Journal of Histochemistry and Cytochemistry* **27**(1), 613–620.

Tucker, J. H. & Shippey, G. (1983), 'Basic performance tests on the CERVIFIP linear array prescreener', *Analytical and Quantitative Cytology* **5**, 129–137.

Unser, M. (1986), 'Sum and difference histograms for texture classification', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-8**(1), 118–125.

van Driel-Kulker, A. M. J. & Ploem, J. S. (1982), 'The use of Leytas in analytical and quantitative cytology', *IEEE Transactions on Biomedical Engineering* **29**(2), 92–100.

Velleman, P. F. & Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury Press, Boston, Mass.

Vincent, L. & Beucher, S. (1989), The morphological approach to segmentation: An introduction, Internal Report C-08/89/MM, School of Mines, Paris.

Walker, R. F. (1995), Improving co-occurrence matrix feature discrimination, Internal technical report, The University of Queensland, St Lucia, Brisbane, Queensland, Australia.

Walker, R. F. (1996), Statistical geometric features—refinements for cytological cell analysis, Internal technical report, The University of Queensland, St Lucia, Brisbane, Queensland, Australia.

Walker, R. F. (1997*a*), Adaptive multi-scale GLCM, CSSIP Technical Report TR1/97, Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide, South Australia.

Walker, R. F. (1997*b*), Genetic algorithm optimisation of adaptive multi-scale GLCM features, CSSIP Technical Report TR2/97, Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide, South Australia.

Walker, R. F. & Jackway, P. T. (1996), Statistical geometric features—extensions for cytological cell analysis, *in* 'Proceedings of ICPR, the 13th International Conference on Pattern Recognition', IEEE Computer Society Press, Technical University, Vienna, Austria, pp. 790–794.

Walker, R. F. & Jackway, P. T. (1997), 'Adaptive multi-scale GLCM', *Pattern Recognition* . (to be submitted).

Walker, R. F., Jackway, P. T. & Longstaff, I. D. (1995), Improving co-occurrence matrix feature discrimination, *in* DIC (1995).

Walker, R. F., Jackway, P. T. & Longstaff, I. D. (1997*a*), 'Genetic algorithm optimisation of adaptive multi-scale GLCM features', *IEE Proceedings: Vision, Image and Signal Processing* . (to be submitted).

Walker, R. F., Jackway, P. T. & Longstaff, I. D. (1997*b*), 'Image texture analysis via co-occurrence methods—review and extensions', *IEE Proceedings: Vision, Image and Signal Processing* . (to be submitted).

Walker, R. F., Jackway, P. T. & Longstaff, I. D. (1997*c*), Recent developments in the use of the co-occurrence matrix for texture recognition, *in* 'Proceedings of DSP'97, the 13th International Conference on Digital Signal Processing', Santorini, Greece, pp. 63–65.

Walker, R. F., Jackway, P. T. & Lovell, B. (1995), Cervical cell classification via co-occurrence and Markov random field features, *in* DIC (1995), pp. 294–299.

Walker, R. F., Jackway, P. T., Lovell, B. & Longstaff, I. D. (1994), Classification of cervical cell nuclei using morphological segmentation and textural feature extraction, *in* 'Proceedings of the Australian and New Zealand Conference on Intelligent Information Systems', IEEE, Brisbane, Australia, pp. 297–301.

Wang, Z. & Abmayr, W. (1982), Fourier transform for chromatin structure and shape in cell image analysis, *in* 'Proceedings of the 6th International Conference on Pattern Recognition', Müchen, p. 1212.

Weiss, S. M. & Kulikowski, C. A. (1991), *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, Morgan Kaufmann, San Mateo.

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

Weszka, J. S., Rosenfeld, A., Carton, E. J., Kirby, R. L. & Mohr, J. M. (1975), A comparative study of texture measures for terrain classification, Technical Report TR-361, Computer Science Center, University of Maryland.

Wied, G. L. (1994), The inevitable and mandatory computerization of our laboratories, *in* Wied et al. (1994), pp. 1–12.

Wied, G. L., Bartels, P. H., Barh, G. F. & Oldfield, D. G. (1968), 'Taxonomic Intracellular Analytic System (TICAS) for cell identification', *Acta Cytologica* **12**, 180–204.

Wied, G. L., Bartels, P. H., Rosenthal, D. L. & Schenck, U., eds (1994), *Compendium on the Computerized Cytology and Histology Laboratory*, Tutorials of Cytology, Chicago, Illinois.

Wittekind, D., Hilgarth, M., Kretschmer, V., Seiffert, W. & Zipfel, E. (1983), 'The new and reproducible papanicolaou stain: Morphologic and spectrophotometric observations on the influence of stain composition on staining results', *Analytical and Quantitative Cytology* **4**(4), 286–294.

Wu, C. M. & Chen, Y. C. (1992), 'Statistical feature matrix for texture analysis', *Computer Vision, Graphics, and Image Processing* **54**, 407–419.

Wu, C.-M., Chen, Y. C. & Hsieh, K.-S. (1992), 'Texture features for classification of ultrasonic liver images', *IEEE Transactions on Medical Imaging* **11**(2), 141–152.

Yogesan, K. (1995), Texture Analysis as a Prognostic and Diagnostic Tool in Tumor Pathology, PhD thesis, University of Oslo.

Yogesan, K., Albregtsen, F. & Danielsen, H. E. (1994), Gray level variance matrix: A new approach to higher order statistical texture analysis, *in* 'Proceedings of the 3rd International Conference on Automation, Robotics and Computer Vision', Vol. 2, pp. 658–663.

Yogesan, K., Albregtsen, F., Reith, A. & Danielsen, H. E. (1993), Co-occurrence and run length-based texture analysis of experimental liver carcinogenesis in mice, *in* 'Proceedings of the 8th Scandinavian Conference on Image Analysis', pp. 227–234.

Yogesan, K., Jorgensen, T., Albregtsen, F., Tveter, K. J. & Danielsen, H. E. (1996), 'Entropy-based texture analysis of chromatin structure in advanced prostate cancer', *Cytometry* **24**(3), 268–276.

Young, I. T., Verbeek, P. W. & Mayall, B. H. (1986), 'Characterization of chromatin distribution in cell nuclei', *Cytometry* **7**, 467–474.

Young, L. S., Bevan, I. S. & Johnson, M. A. (1989), 'The polymerase chain reaction: a new epidemiological tool for investigating cervical human papillomavirus infection', *British Medical Journal* **298**, 14–18.

Yu, B. & Yuan, B. (1993), 'A more efficient branch and bound algorithm for feature selection', *Pattern Recognition* **26**(6), 883–889.

Zahniser, D. J. (1994), Cytyc corporation thinprep processor and CDS-1000 cytology workstation, *in* Grohs & Husain (1994), pp. 279–293.

Zahniser, D. J., Oud, P. S., Raaijmakers, M. C. T., Vooys, G. P. & van de Walle, R. T. (1979), 'BioPEPR: A system for the automatic prescreening of cervical smears', *The Journal of Histochemistry and Cytochemistry* **27**(1), 635–641.