


SEGMENTATION METHODS AND FEATURE EXTRACTION
FOR CERVICAL CELL RECOGNITION

by

 Nam G. Nguyen

School of Computer Science

McGill University

March 1983

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Master of Science

Abstract

In order to achieve better accuracy and speed in the automatic analysis and recognition of cervical cells, several large research efforts are being made to improve each of four main subproblems: image acquisition and enhancement, scene segmentation, feature extraction, and classification. The goal of this thesis is to investigate possible improvements in scene segmentation and feature extraction processes.

A threshold selection segmentation technique which selects the density threshold based on the stability of the cellular area was used. Images of cervical cells scanned at 530 nm were used to segment the cells or clumps of cells from the background and images scanned at 570 nm were used to separate the nuclei from the cytoplasm. Experimental results showed that the segmentation technique worked better with two color images (530 nm and 570 nm) than with single color images of cervical cells (570 nm). To study the problem of overlapping cells, an algorithm for generating overlapping cells was devised. This algorithm which creates random overlaps from single cell data, was used to develop and evaluate procedures for detecting overlapping cells.

The feature extraction software system developed previously at the Macdonald Stewart Biomedical Image Processing Laboratory, Department of Pathology, McGill University, was expanded to include several two-dimensional histogram features and additional Fourier and Granlund shape features.

Three feature selection procedures were investigated. The selected features were compared by means of the classification error rates obtained by using the minimum Mahanalobis distance classifier to classify cervical cells into 16 subclasses.

The new two-dimensional histogram features devised in this research were found to be the best among all feature categories studied. A reduction of 57.88% in the classification error rate (from 5.2% to 2.19% for the random partitioning method) was achieved using the 13 features selected by the forward sequential search procedure (including the two-dimensional histogram features) in comparison with the 6 features (with no color information) previously used by Oliver et al[Oli78a].

Résumé

Dans le but d'améliorer l'analyse et l'identification de cellules cervicales par ordinateur, les efforts de recherche se concentrent sur quatre sous-problèmes majeurs: l'acquisition et l'enrichissement des images, la segmentation des images, l'extraction des caractères et la classification. Le but de la recherche ci-décrite a été d'évaluer différents procédés pour effectuer la segmentation de régions et l'extraction de caractéristiques afin d'obtenir de meilleurs résultats.

Une méthode de segmentation basée sur la sélection du seuil de densité déterminé par la stabilité de la surface cellulaire a été utilisée. Des images enregistrées à 530nm ont été utilisées pour segmenter les cellules ou les amas de cellules de l'arrière-plan; des images enregistrées à 530nm et 570nm ont servi à la segmentation des noyaux et du cytoplasme. Les expériences de segmentation utilisant des images de deux couleurs (530nm, 570nm) ont donné des résultats supérieurs à celles utilisant des images d'une seule couleur (530nm). Pour étudier le problème de la superposition des cellules, un algorithme a été produit pour créer au hasard des cellules superposées à partir d'images de cellules isolées. Cet algorithme a servi à développer et évaluer des méthodes pour découvrir les cellules superposées.

Le logiciel développé auparavant au Laboratoire de Traitement d'Images Macdonald Stewart, Département de Pathologie, Université McGill, pour l'extraction de caractères a été étendu pour pouvoir utiliser des histogrammes de caractères bi-dimensionnels en plus des caractères morphologiques Fourier et Granlund.

Trois méthodes de sélection de caractères ont été examinées. Les caractères choisis ont été comparé d'après les taux d'erreur de classification obtenus en utilisant le classificateur de distance minimum Mahalanobis sur un ensemble de cellules cervicales à classifier en 16 sous-classes.

Les nouveaux histogrammes bi-dimensionnels développés dans cette recherche ont donné des résultats supérieurs à tous les autres caractères choisis. Le taux d'erreur de classification a été réduit de 57.88% (de 5.2% à 2.19%) en utilisant les 13 caractères qui ont été sélectionnés par la méthode de recherche séquentielle avancée (incluant les histogrammes de caractères bi-dimensionnels) en comparaison des 6 caractères (sans information de couleurs) utilisé par Oliver et al (01178a).

Acknowledgements.

I wish to thank Dr. Ronald Poulsen, Director of the Macdonald-Stewart Biomedical Image Processing Laboratory, for providing facilities, for his guidance, advice, and especially for so much of his time devoted to this research. I also wish to thank Professor Godfried T. Toussaint for his advice and comments, and Miss Wendy Rogers for her meticulous reading of several drafts of this thesis. Finally, I would also like to thank Miss Claude Louis, consulting Cytotechnologist at the Image Processing Laboratory, for her support in the Cytology field. The 3000 cell images used in this research were scanned, manually segmented, and classified by her.

Financial support for parts of this work was provided by the Natural Sciences and Engineering Research Council of Canada (A-4516) and the Macdonald-Stewart Foundation, Montreal, Quebec, Canada. The thesis manuscripts were prepared by using the DEC PDP11 text program RUNOFF.

To my parents and my wife
for their wonderful love

Table of Contents

	page
Abstract	i
Résumé	iii
Acknowledgements	v
Dedication	vi
Chapter 1 INTRODUCTION	
1.1 Introduction to Image Processing and Pattern Recognition system	1
1.2 Introduction to the system in Biomedical Image Processing Laboratory	5
1.3 Introduction to cervical cell data base	7
1.4 Scope of the thesis	9
Chapter 2 SEGMENTATION METHODS APPLIED TO CERVICAL CELL RECOGNITION	
2.1 Introduction	11
2.2 Survey of scene segmentation techniques	13
2.3 Segmentation technique chosen for cervical cell recognition	16
2.4 Segmentation results on single cervical cells	18
2.5 The problem of overlapping cells	20
2.5.1 Introduction	20
2.5.2 Data set	22
2.5.3 Algorithm for generating cells having	

a uniformly distributed overlap degree	22
2.5.4 Overlapping-cell detection method	24
2.5.5 Overlapping-cell detection results	28
2.6 Conclusions	32

Chapter 3 FEATURE COMPUTATION

3.1 Introduction	36
3.2 The IPS feature extraction system	37
3.3 Features computed by the expanded system	
3.3.1 Color and density features	38
3.3.1.1 Derived from multi-dimensional histograms of multi-color image data	38
3.3.1.2 Derived from several one-dimensional histograms of multi-color image data	42
3.3.2 Geometric features	45
3.3.2.1 Features indicating the positions of nuclei in the cells	45
3.3.2.2 Size features	46
3.3.2.3 Shape features	46
a) Moment features	46
b) Fourier shape features	48
c) Granlund's shape features	49
3.3.3 Texture features	50
3.4 Summary	53

Chapter 4 FEATURE SELECTION

4.1 Introduction	54
4.2 Survey of feature selection approaches	56
4.2.1 Feature evaluation criteria	56
4.2.2 Feature search procedures	59
4.3 Feature selection procedures applied to cervical cell recognition problem	63
4.3.1 Feature evaluation criterion	63
4.3.2 Feature subset search procedures	63
a) Forward sequential search procedure	63
b) Parallel search procedure	64
c) Feature clustering procedure in pattern space	64
4.4 Experimental results for the three feature search procedures	64
4.4.1 Feature subsets selected by three procedures	64
4.4.2 Classification performance estimates	66
a) Forward sequential search procedure	66
b) Parallel search procedure	68
c) Feature clustering procedure in pattern space	69
4.5 Discussion and conclusions	70

Chapter 5 FEATURE EVALUATION

5.1 Introduction	72
5.2 Evaluation of each feature category	73
5.2.1 Feature search procedure used	73

5.2.2 Classification performance estimation of feature subsets selected from each feature category	73
5.3 Experimental results	84
5.3.1 Two-dimensional histogram features	84
5.3.2 One-dimensional histogram features	87
5.3.3 Geometric features	87
5.3.4 Texture features	88
5.4 Conclusions	89

Chapter 6 SUMMARY AND CONCLUSIONS

6.1 Summary	92
6.2 Conclusions and suggestions for future studies	93
6.2.1 On scene segmentation	93
6.2.2 On feature extraction and feature selection	94
REFERENCES	97

Chapter 1

INTRODUCTION

Ever since the "Pap smear" was introduced as a reliable diagnostic tool for cervical cancer, there has been a large and rapid increase in its use as a screening method for cervical cancer. As a consequence, several research groups have been working towards an automated system for analysing cervical smears. At the Macdonald-Stewart Biomedical Image Processing laboratory (BIPLAB), Poulsen et al[Pou77a,78a,81a,Tou79a], Cahn et al[Cah77a,77b], Oliver et al[Oli77a,78a,78b] have undertaken research to develop image processing and pattern recognition algorithms for cervical cell recognition. However, to build a system for practical use, better classification accuracy is still needed. A continuous effort is being made in the BIPLAB to develop better image processing and pattern recognition techniques suitable for cervical cell analysis and classification.

In the following subsections of this chapter, a typical image processing and pattern recognition system, the system used in the BIPLAB and the data base of cervical cell images are described. Also, the major contributions of this thesis to the scene segmentation and feature extraction processes are briefly mentioned.

1.1 Image Processing and Pattern Recognition systems

Four main subproblems involved in image processing and pattern recognition are image acquisition and enhancement, scene segmentation, feature extraction, and pattern classification as shown in figure 1.

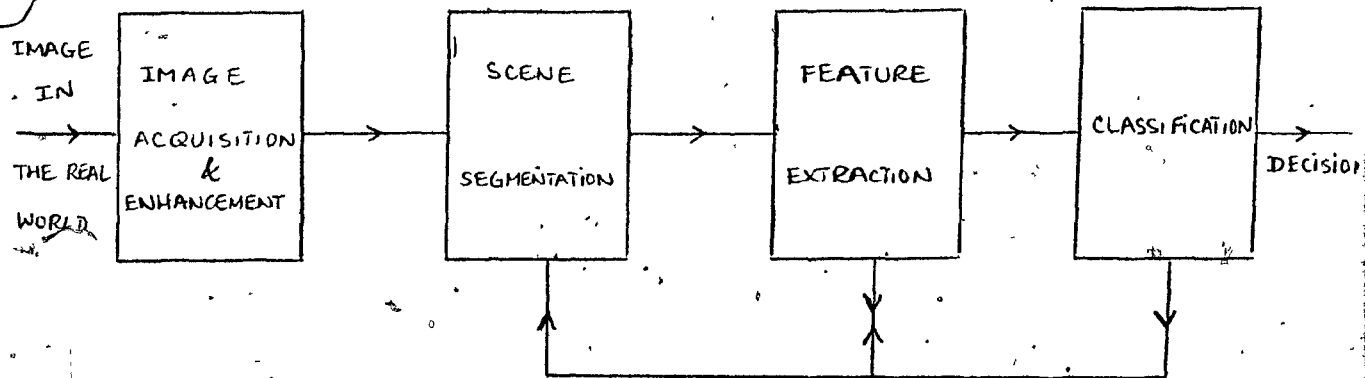


Figure 1. Block diagram of processes in a typical image processing and pattern recognition system.

1.1.1 Image Acquisition and Enhancement

- Image Acquisition

In order to use computer software to analyze an image in the real world, that image must be converted into a form acceptable to the computer. The form widely used now is an array of picture points, or pixels, with each pixel density quantized into gray level values. The level of detail that a digital image can represent depends on the resolution of the image, (i.e. the magnification and the total number of pixels), and the number of possible gray level values assigned to each pixel. Typically, for cell recognition systems, a minimum of 64 gray levels (or optical density levels) is required along with a pixel spacing of 1 micron or less.

- Image Enhancement

The image acquired by the scanning system sometimes has the following defects:

a) Shading problem: For example, an image of constant density may appear to be lighter in the middle and darker around its edges. This problem is mainly caused by the uneven illumination across the specimen and/or the variation in the sensitivity across the field of the scanner. For a sophisticated image processing and pattern recognition system, a hardware shading correction unit and/or, a software shading correction package are usually included to overcome this shading problem.

b) High fluctuations of gray values: In some systems, high fluctuations of gray values of pixels occur either because of the noise in the system or the nature of the specimens themselves. Thus, fast and effective digital smoothing software is sometimes required to remove these undesired fluctuations.

c) Low contrast images: The images acquired in some cases have very low contrast and hence are very difficult to analyze visually. To solve this problem, software techniques such as gray scale modification techniques have been developed to improve the contrast of the images.

d) Blurring: Even in the absence of external noise, the acquired images are still degraded to a certain extent according to the modulation transfer function of the system. This kind of degradation can be compensated for by applying an inverse filter to the acquired images:

Assuming that the degraded image of a point is independent of the position of that point and if $H(u,v)$, the Fourier transform of the point spread function $h(x,y)$, is known and different from zero, then one can restore the original image. This is done by first multiplying the Fourier transform of the acquired image by $1/H(u,v)$ and then applying the inverse Fourier transform to the resultant as described in [Ros76b]. In practice, only a limited correction is possible due to noise considerations.

1.1.2 Scene segmentation:

Scene segmentation involves separation of interesting regions of the scene by considering common properties of the pixels within each region. For cervical cells, it involves the separation of cells or clumps of cells from the background, of nuclei from cytoplasm, and of one cell from others. Features extracted from badly segmented regions usually deviate greatly from those extracted from well segmented regions. Consequently, poor segmentation generally adversely affects classification results. Because of the influence of the segmentation results on subsequent processes, feedback from the feature extraction and/or classification processes can be used for evaluating and selecting good segmentation algorithms. For example, one can choose from among several segmentation algorithms, the one which produces the least deviation of the important features from the corresponding ones obtained using a reference segmentation (ie human segmentation). Alternately, one can choose the one which produces the least classification error rate. Research on segmentation error measurement is being done currently at the BIPLAB, and results will be reported in the near future.

1.1.3 Feature Extraction:

Since the goal of the pattern recognition system is to discriminate among images of different classes, it is unnecessary and prohibitive, in terms of computation time and memory storage, to classify images based on differences among individual images. Instead, one must classify images based on important measurements of properties which vary less among images of the same class than among images of different classes. Moreover, to further optimize the cost, a search procedure for effective subsets of important features should be developed. Because the classification error rates obtained depend on how much information the selected features contain, feedback of classification error rate to the feature search procedure can be used to select an effective subset of all features considered.

1.1.4 Classification:

The classification process involves the decision to assign images to different classes based on the values of features selected in the feature extraction process. The best classifier should be the one which gives the least confusion among classes of images, when applying the same set of selected features and the same data set for testing.

1.2 Introduction to the system in the Biomedical Image Processing Laboratory (BIPLAB)

1.2.1 Image acquisition and enhancement

The system used for this thesis' research is the one at the

Macdonald Stewart Biomedical Image Processing Laboratory (BIPLAB), Department of Pathology, McGill University. The images were scanned using a Quantimet 720D Plumbicon scanning system to produce a digital image of 144x144 pixels of 64 optical density levels. Even though image enhancement could be performed by the system, when applied to cervical cell recognition, no correction was made for optical or electronic shading so that the quality of the scanned data would be consistent with what would be expected in a practical laboratory instrument. However, as pointed out by Poulsen et al [Pou77a], the optical and electronic shading of this system is minor compared to the variations in background density found in routine slides. Also, no image enhancements such as smoothing or deblurring were made due to the fact that the digitized images represent the original images very well.

Scene segmentation, feature extraction, and classification processes have been developed at the BIPLAB over several years. Following are brief descriptions of these processes:

1.2.2 Scene segmentation

One scene segmentation algorithm, developed by Cahn [Cah77a,77b], selects an appropriate density threshold based on the stability of the boundary of the segmented regions as the threshold is varied. Presently, techniques for generating and detecting overlapping cervical cells are being developed.

1.2.3 Feature extraction

Presently, the feature extraction system can be used to extract

several important features such as color features, density features, geometric features, and texture features. Moreover, feature search procedures for effective subsets of features and feature category evaluation are under investigation.

1.2.4 Classification

The minimum Mahanalobis distance classifier developed by Oliver et al [Oli77a,78a,78b] is used under the assumption that the density function of the multi-dimensional feature vectors has a Gaussian distribution.

The scene segmentation and the feature extraction processes will be discussed in more detail in the next sections of this thesis.

1.3 Introduction to the cell data base:

In this research, 3000 cells from routine Papanicolaou stained cervical specimens were scanned at 0.7 micron resolution, higher resolution than in previous studies (1.0 micron resolution), to improve the quality of the acquired digital images. Also, the cells were scanned at three different wavelengths (530 nm, 570 nm, 620 nm), instead of at the one wavelength (using a Zeiss VG-9 filter) used before to provide color information for scene segmentation and feature extraction studies. The cells were divided into 16 classes: 11 classes of normal cells and 5 classes of abnormal cells (listed in Table 1 on the next page).

NORMAL CELL CLASSES

1. SSQ	Superficial Squamous	351
2. ISQ	Intermediate Squamous	318
3. NAV	Navicular Squamous	179
4. PAR	Parabasal Squamous	263
5. ENLI	Endometrial (Stromal)	93
6. ENM*G	Endometrial (Glandular)	66
7. ENC*S	Endocervical (Secretory)	152
8. ENC*C	Endocervical (Ciliated)	147
9. HIS	Histiocytes	139
10. MET*A	Metaplastic (Acidophillic)	182
11. MET*B	Metaplastic (Basophillic)	115
Sub total (Normal)		2005

ABNORMAL CELL CLASSES

12. DYS-MLD	Mild Dysplasia	204
13. DYS-MOD	Moderate Dysplasia	181
14. DYS-SEV	Severe Dysplasia	187
15. CIS	Carcinoma in Situ	148
16. INV	Invasive Carcinoma	275
Sub total (Abnormal)		995

Table 1 : High resolution data base of cervical cells

1.4 Scope of the thesis:

Chapter two contains the major contribution of the thesis with respect to the scene segmentation process. In this chapter, an algorithm which selects the density threshold based on the stability of the area of the segmented regions is described. Cell images scanned at 530 nm were used for separating cells or clumps of cells from the background and the same cell images scanned at 570 nm were used for separating nuclei from cytoplasm. To study the problem of overlapping cells, two algorithms were developed: -- a) an algorithm, for generating a data base of overlapping cells having a uniformly distributed overlapping percentage, for use in evaluating overlapping-cell detection and segmentation algorithms, -- b) an overlapping-cell detection algorithm using Fourier shape descriptors and cell density information.

Chapters three, four, and five contain the major contributions of the thesis related to the feature extraction process:

Chapter three describes the previously developed feature extraction system and the computation of the set of 209 features (including new color features and additional shape features) of the expanded feature extraction system.

Chapter four describes the performance of three feature selection procedures. Each procedure was evaluated by determining the classification error rate obtained using a minimum Mahalanobis distance classifier together with a random partitioning test method.

Chapter five describes the performance of features of different types such as two-dimensional histogram features, one-dimensional histogram features, geometric features, and texture features. They are investigated to determine their power in discriminating different classes of cervical cells.

In chapter six, the segmentation methods, the feature search procedures, and the feature evaluation are summarized and discussed. Also, suggestions for future improvements on the scene segmentation and the feature extraction processes are proposed.

Chapter 2

SEGMENTATION METHODS APPLIED TO CERVICAL CELL RECOGNITION

2.1 Introduction:

In pattern recognition, whether the ultimate goal is to derive features or to classify objects according to their common patterns, a critical step is to segment images into meaningful regions with common properties. When applied to cervical cells, the problem involves separating single cells or clumps of cells from the background (scene segmentation), nuclei from cytoplasm (cell segmentation), and one cell from others (overlapping cell segmentation). High segmentation error tends to produce feature values that deviate greatly from those obtained using a reference segmentation (human segmentation) and generally significantly increases the classification error rate. Several segmentation algorithms have been developed for the cervical cell segmentation problem in particular. One of the segmentation techniques which proved to work satisfactorily is the one developed by Cahn et al[Cah77a,77b]. In our present research, a modified version of their technique was used. Moreover, images of cervical cells scanned at 530 nm were chosen for use in scene segmentation and images of cervical cells scanned at 570 nm were chosen for use in cell segmentation. Previously, images of cervical cells scanned using a Zeiss VS-9 filter were used for both scene and cell segmentation.

For the problem of overlapping cells, the threshold selection technique based on the stability of cellular area was used to separate clumps of

overlapping cells or touching cells from the background. Fourier descriptors of cellular boundaries, together with density information, were used for detecting and segmenting overlapping cells.

In the following subsections, a brief survey of scene segmentation techniques, a short description of the segmentation technique used for the cervical cell segmentation problem, a method for generating overlapping objects having a uniformly distributed overlapping percentage, and a method for detecting overlapping objects are described and evaluated.

2.2. Survey of scene segmentation techniques:

2.2.1 Overview

Scene segmentation involves locating regions with pixels having the same properties. Several scene segmentation techniques, applied to different fields in pattern recognition, have been developed. They generally fall into four main categories: threshold selection techniques, edge detection techniques, region-growing techniques, and relaxation techniques.

a) Threshold selection techniques.

An image can be segmented using its gray level histogram (global), or its local properties. For the global approaches, the histogram of an image is first computed, then smoothed if necessary to obtain distinct peaks and valleys. Finally, the thresholds are set at the valleys found. Reasonable results were reported using this approach for segmenting white and red blood cells [Gre79a, Ben79a]. Recently, the histogram-based thresholding technique was generalized to select thresholds based on multi-dimensional histograms (histograms of images scanned at two or three different wavelengths): Aggarwal et al [Agg77a] used two-dimensional histograms (528 nm, 569 nm) to locate the cytoplasm by a ceiling-lowering clustering technique. A 88.1% success rate in isolating cytoplasm was obtained using a test set of 233 cervical cells. Aus et al [Aus77a] used the trivariate histogram (420 nm, 580 nm, 620 nm) to segment bone marrow cell images. In the local thresholding approaches, the thresholds are selected based on local properties of the segmented regions. Cahn et al [Cah77b] have developed a technique to select

the threshold based on the geometric stability of the perimeter of segmented regions as the threshold is varied. Cahn et al [Cah77b] and Lin et al [Lin81a] have applied this technique to cervical cell segmentation and obtained excellent results (Cahn et al reported 6.6% poor segmentation on 1500 cervical cells and Lin et al reported less than 0.5% really poor segmentation on 1153 cervical cells).

b) Edge detection techniques

Assuming that there exists an abrupt change in gray value or texture at the edge of an object, the boundary of the object is found by applying gradient operators, or by applying high pass spatial filtering. A survey of edge detection techniques is described by David [Dav75a].

c) Region growing techniques

Riseman et al [Ris77a] discussed three different techniques depending on different characteristics of the scenes: (1) Merging: Small pieces which have common properties are grouped together to form final regions, (2) Splitting: Large pieces of an image are broken into smaller areas until a high confidence that they are homogeneous under the features of interest is obtained, (3) Spatial and feature analysis: Histograms of various feature pairs are employed to find clusters of feature activity. These clusters are used to label local areas of the scene, followed by a spatial analysis of these labels to guide the formation of the desired regions.

d) Relaxation techniques

In the relaxation techniques, each pixel is initially assigned a set of membership indices relating that pixel to a region in the image. These indices are then iteratively updated for each pixel by examining the membership indices of neighboring pixels.

2.2.2. Advantages and disadvantages of the scene segmentation techniques mentioned above:

a) Threshold selection techniques

- Histogram-based threshold selection: This method is inexpensive and simple but it often cannot handle complicated scenes where distinct modes usually do not stand out in the histograms.

- Local property-based threshold selection: This method is inexpensive and efficient. Moreover, it can be used to take into account any important local property such as area, perimeter, gradient, etc or a combination of them as the basis for selecting thresholds. However, the technique is still dependent on a parameter (stability percentage for area, perimeter, gradient) which is heuristically chosen.

b) Edge detection techniques

- The edge detection techniques are inexpensive and simple but they have two main disadvantages: (1) The segmented region sometimes does not have a connected boundary, and (2) the threshold value of the change in gray value has to be chosen.

c) Region growing techniques:

For both the merging and splitting techniques, connected boundary analysis is not required to identify acceptable regions. However, the techniques have difficulties when the textural variation within regions is nearly the same as the variation between different regions. In other words, these techniques are very sensitive to the threshold set to split or merge the image.

With the spatial and feature analysis method, good segmentation can result if feature pairs are chosen such that clusters of interest stand out (not always the case for complicated scenes). On the other hand, an analysis of spatial areas which have been labeled is required to find the desired connected regions.

d) Relaxation techniques

The relaxation techniques usually improve segmentation accuracy of regions segmented first by other segmentation techniques. However, a relaxation technique has two main disadvantages: (1) It is sensitive to the updating rule and the initialization of the class membership indices, and (2) it is expensive in terms of computation time and in terms of memory space required to store the indices of every pixel.

2.3 Segmentation technique chosen for cervical cell recognition:

For cervical cell recognition, a modified version of the local property-based threshold selection technique developed by Cahn et al[Cah77b] was applied. In this technique, an appropriate density threshold is selected based on the stability of the perimeter of the segmented regions as the

threshold is varied. Because we observed that cellular and nuclear areas are two of the most important features for cervical cell recognition, we modified Cahn's technique by basing threshold selection on the stability of area instead of perimeter. The importance of cellular and nuclear areas is verified in chapter 4, which shows that these two features were selected, from the total set of 209 informative features of all categories, by the forward sequential search procedure.

For scene segmentation (separating cells from the background), the following threshold selection procedure was applied:

- First, the size of the segmented regions must be reasonable: the area should be between 100 and 20000 pixel counts (50 and 10000 square microns respectively), and the perimeter should be between 50 and 550 pixel counts (35 and 355 microns respectively). This excludes small unexpected objects and large clumps of cells (large clumps of cells, composed of two or three overlapping cells, are discussed in the section of overlapping-cell problem).

- Secondly, a threshold is selected whenever the change in the area of the segmented region at that threshold, with respect to the area of the segmented region at previous threshold, is the smallest and less than 40% (the stability parameter of 40% seemed to give the best results for scene segmentation of cervical cells).

For cell segmentation (separating nuclei from cytoplasm), the area range is 25 to 2000 pixels (12.5 and 1000 square microns respectively), and the perimeter range is 15 to 200 pixels (10.5 and 140 microns respectively). The

size limitation is used in order not to select small dark regions inside the cytoplasm and not to select large clumps of overlapping nuclei or large and dark overlapping cytoplasm. The area change is restricted to less than 20% (this parameter seemed to give the best cell segmentation results for our cervical cell data).

2.4 Segmentation results on single cervical cells

To improve the segmentation accuracy, a 0.7 micron scanning resolution in three color images of cervical cells was used instead of a 1.0 micron scanning resolution in a single color as was used in previous studies. Using higher scanning resolution appeared to make scene and cell segmentation easier. Particularly, using two color images instead of single color images for segmentation improved scene segmentation very significantly. To verify this improvement, a segmentation error measurement based on the percentage of misclassified pixels was applied. The detailed description of this method is given by Yasnoff et al [Yas77a]. In the present research, human threshold selection was assumed to provide correctly segmented regions. Thus, whenever segmented regions obtained from a segmentation method did not match the reference segmented regions, the percentage of misclassified pixels was used as the segmentation error measurement.

The method was applied to both single-color and two-color image data: a) For single -color data, images of cervical cells scanned using a Zeiss VG-9 filter were used for both scene and cell segmentation, b) For two-color data, images of cervical cells scanned at 530 nm wavelength were used for scene segmentation and images of cervical cells scanned at 570 nm wavelength were

used for cell segmentation.

In the present research, 250 single cells out of the 3000-cell data base described in chapter 1 were used for testing the significance of color information in scene segmentation. When single-color images were used, 11 cells were rejected because the area stability of 40% for scene segmentation was not achieved, and the remaining 239 cells were segmented with the average percentage of misclassified pixels being 7.92% for scene segmentation and 12.8% for cell segmentation. When two color images were used, only 2 cells out of 250 were rejected and the remaining 248 cells were segmented with the average percentage of misclassified pixels being 3.6% for scene segmentation and 12.8 % for cell segmentation.

2.5 The problem of overlapping cells

2.5.1 Introduction

One key problem in scene segmentation is distinguishing binucleated cells and clumps of overlapping cells from single cells. Shape information has been used [Ecc77a, Jai80a, Tuc78a] for segmenting touching and overlapping cells. Textural information has also been considered for segmenting overlapping cells [Lin81a]. Recently, both shape information and density information has been used for detecting overlapping cell nuclei [Ben81a].

In the studies done [Ecc77a, Jai80a, Tuc78a, Lin81a, Ben81a, Syc78a, Cah77b], the detection rate achieved for recognizing overlapping objects such as cells is extremely dependent on the data base of overlapping objects used. On one hand, the degree of overlap and types of the cells studied would appear to have a considerable bearing on the inherent difficulty of the overlapping-cell detection problem. On the other, collecting a data base of overlapping cells representative of the wide variations in degree of overlap and the many cell types involved would be very difficult indeed. Thus, meaningful comparison of the large variety of algorithms in the literature is difficult. In an attempt to minimize these problems we devised an algorithm for generating a data base of overlapping cells having a uniformly distributed overlap degree by artificially combining images of single cells taken from a large single cell data base. In this algorithm two single cells first are randomly selected from the single cell data base. These cell images are then made to overlap to a percentage obtained from a uniform random number generator. The overlap degree is computed as the ratio of the area of overlap to the area of the

smaller of the two selected cells.

The 582 artificially generated overlapping cells, together with 1046 single cells and 157 naturally overlapping cells were used for evaluating an overlapping-cell detection algorithm.

In the overlapping-cell detection algorithm, both shape and density information are used for detecting overlaps. To obtain the shape information, the boundary of the object is used for computing Fourier descriptors by applying the procedure devised by Granlund[Gra72a]. The first n Fourier descriptors are then used to reconstruct a smoothed boundary of the object. The locations and curvature values of points of local maxima in concavity along the smoothed boundary can also be found analytically in terms of these n Fourier descriptors. The relative positions and values of these points can be used for detecting overlapping cells. When considering the density information, overlapping cells and multi-nucleated cells were detected whenever more than one nucleus was found within the cell in question.

2.5.2 Data set

A data base of 1203 cells (selected from the 3000-cell data base described in chapter 1), comprising 1046 single cells and 157 naturally overlapping cells, was used for generating 582 cells having a uniformly distributed degree of overlap. This data base of 1203 cervical cells together with the 582 artificially generated overlapping cells was used for evaluating the performance of an overlapping-cell detection algorithm.

2.5.3 Algorithm for generating cells having a uniformly distributed overlap degree:

In order to evaluate the performance of the overlapping-cell detection algorithm objectively, a data base of cells having a uniformly distributed overlap degree is very useful. From the existing data base of single cells, the following algorithm was devised to generate such an overlapping cell data base:

Step 0: Let N =the number of overlapping cells to be generated, $Re=5\%$, and

$De=1.5$ times the pixel spacing (1.5×0.7 microns) where Re is the maximum overlapping percentage error allowed for any generated overlap and De is the minimum travelling distance required for continuing the binary search.

Step 1: Select two cells randomly from the single cell data base

Step 2: Compute the coordinates of the centers $C1, C2$ and the effective radii $R1=\sqrt{A1/\pi}, R2=\sqrt{A2/\pi}$ of the first and second cell

respectively (where A_1 and A_2 are the areas of the first and second cell respectively).

Step 3: Apply the binary search technique to find the appropriate positions of the two cell images in the field of view such that the overlapping percentage matches the target percentage (called U) obtained from a uniform random number generator (maximum overlapping percentage error allowed is Re):

3.1 Let $D_{min}=0$, $D_{max}=1.5*(R_1+R_2)$, $D_{target}=0.75*(R_1+R_2)$, $Ovedegree=U$

D_{min} and D_{max} are the minimum and maximum distances allowed between two centers of two cells. D_{target} is the present distance set between two centers of two cells, $Ovedegree$ is the target overlapping percentage.

3.2 Keep the first cell fixed, move the second cell along the line joining C_1 and C_2 such that the distance between C_1 and C_2 is equal to D_{target} . Then measure the percentage of overlapping area R with respect to the area of the smaller of the two cells.

3.3.a) If $|R-Ovedegree| < Re$ or $|D_{max}-D_{min}| < De$ GOTO Step 4

(Exit if measured and desired overlapping percentage differ by less than Re or if travelling distance is less than De)

b) - If $R < Ovedegree$, Let $D_{max}=D_{target}$ and $D_{target}=(D_{target}+D_{min})/2$

(If measured overlapping percentage is less than the desired overlapping percentage, set the next position of the two cells closer).

- If $R > Ovedegree$, Let $D_{min}=D_{target}$ and $D_{target}=(D_{target}+D_{max})/2$

(Otherwise, set them farther from each other).

GOTO Step 3.2

Step 4: Decrease N, if N=0 STOP, else GOTO Step 1

(Stop when N overlapping cells have been generated).

In this research, the algorithm was used to create a data base of 582 overlapping cervical cells from the data base of 1203 cervical cells mentioned in part 2.5.2. These 582 overlapping cells together with the 1203 cervical cells were used for evaluating the overlapping-cell detection method.

2.5.4 Overlapping-cell detection method

First, the threshold selection technique based on the area stability was used to segment regions (single cells or clumps of overlapping cells) from the background, then the regions were tested for clumps of overlapping cells by means of shape and density information.

a) Detection based on the shape information:

The first criterion used for detecting overlapping cells was based on the shape of the cell boundary. In this study, the Fourier transform was applied to derive Fourier descriptors of the boundary of the cell. The basic approach was devised by Granlund[Gra72a]. Also, the mathematical formula of the curvature in terms of the Fourier descriptors can be determined analytically.

Consider a contour such as C represented in the complex plane (See figure 2). A point moving around the contour at a constant speed generates a complex

value function:

$$u(t) = x(t) + jy(t)$$

where t is the path length with a period T equal to the perimeter.

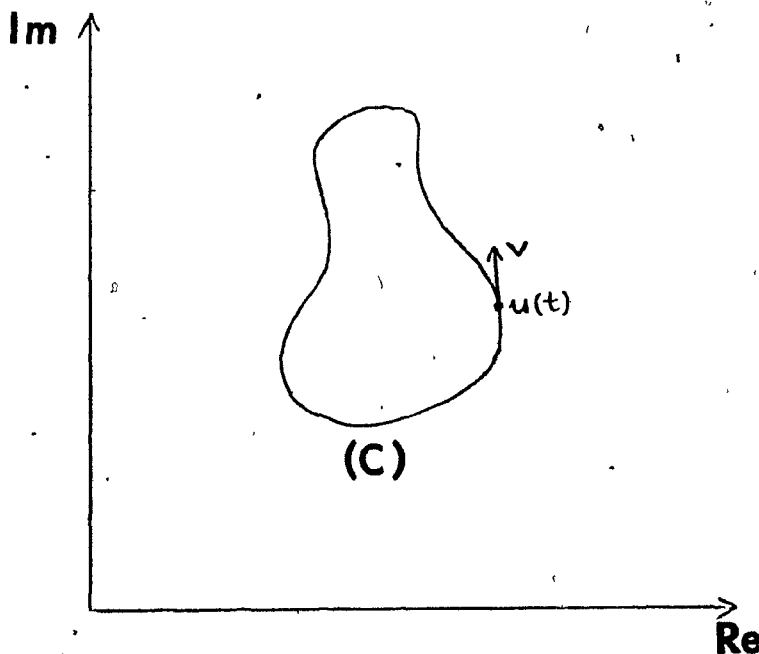


Figure 2 A representation of a shape as a contour in the complex plane. The point $u(t)$ moving around the contour at a constant velocity defines a periodic complex value function of path length t and with a period T equal to the perimeter.

Since $u(t)$ is periodic, it can be expressed as a complex Fourier series:

$$u(t) = \sum_{n=-\infty}^{+\infty} a_n \exp(j2\pi n t / T)$$

$$\text{where } a_n = 1/T \int_{-\infty}^{\infty} u(t) \exp(-j2\pi nt/T) dt$$

A truncated Fourier series

$$u_T(t) = \sum_{n=-p}^{+p} a_n \exp(j2\pi nt/T)$$

represents the smoothed boundary of the object. A more general smoothing of the boundary can be obtained by using an appropriate weighting of the Fourier descriptors.

The tangent $\mathcal{C}(t)$ and curvature $R(t)$ of a point $u(t)$ on the smoothed boundary of the object can be computed as:

$$\mathcal{C}(t) = \frac{\partial u_T(t)}{\partial t} = \sum_{n=-p}^{+p} a_n (j2\pi n/T) \exp(j2\pi nt/T)$$

$$R(t) = \frac{\partial^2 u_T(t)}{(\partial t)^2} = \sum_{n=-p}^{+p} -a_n (2\pi n/T)^2 \exp(j2\pi nt/T)$$

From $\mathcal{C}(t)$ and $R(t)$, the points of local maxima in concavity along the smoothed boundary of the objects can be located. Those are the points where the curvature magnitudes $R(t)$ are at local maxima and the phase of $R(t)$ leads that of $\mathcal{C}(t)$ as shown in Figure 3. (Note that when the phase of $R(t)$ leads that of $\mathcal{C}(t)$, the phase difference $(\mathcal{C}(t), R(t))$, where $\mathcal{C}(t)$ is the initial ray and $R(t)$ is the terminal ray, is equal to $+\pi/2$. On the other hand, when the phase of $\mathcal{C}(t)$ leads that of $R(t)$, the phase difference is equal to $-\pi/2$).

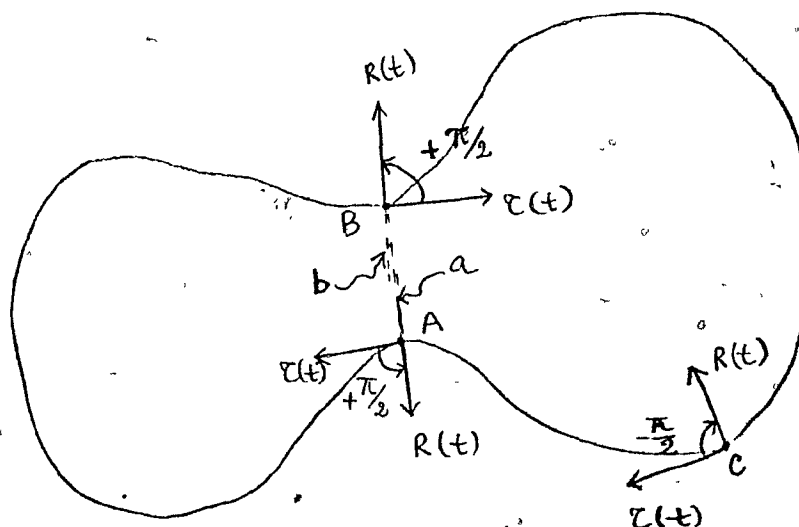


Figure 3 Examples of the phase relationship of convex and concavity points on a contour: a) at convex point C, $U(t)$ leads $R(t)$, b) at concavity points A and B, $R(t)$ leads $U(t)$.

All pairs of the maximal concavity points were tested to see if any of them was a pair formed by overlapping cells in the same manner as described by Eccles et al[Ecc77a]. In this description, to be considered as a pair formed by overlapping cells, the two maximal concavity points A,B should satisfy the following two conditions:

$$(1) (\text{min}^{\circ} \text{ path length between A and B}) / (\text{Euclidean distance } d_{AB}) > 3$$

$$(2) a + b < 60$$

where a and b are acute angles formed by $(R_A(t), \vec{AB})$ and $(R_B(t), \vec{AB})$ ($R_A(t)$

and $R_B(t)$ are curvature vectors of the points A and B respectively).

b) Detection based on the density information

In our study, density information was also used for detecting overlapping cells. If we assume that each cervical cell consists of only one nucleus, the overlapping cells can be detected whenever more than one nucleus is detected within the cytoplasm. For cell segmentation, the threshold selection technique based on a stability of the segmented area was used to segment dark regions. To test for the number of existing nuclei, the three darkest regions within the cytoplasm are segmented. If the average optical density values of the two darkest regions are both greater than twice the average optical density of the cell excluding the three darkest regions, then the two darkest regions are both considered as nuclei.

2.5.5 Overlapping-cell detection results

a) Results obtained using artificially generated overlapping cells

- Results based on cytoplasmic shape

When we considered cytoplasmic shape alone, the overlapping-cell detection rate decreased exponentially as the degree of overlap increased (See Figure 4). Overall, an overlapping-cell detection rate of 33.55% was obtained. However, when the degree of overlap was restricted to less than 20%, a 68% detection rate was obtained. Moreover, if the degree of overlap was restricted to less than 10% (touching or slightly overlapping cells), a

90.2% detection rate was obtained.

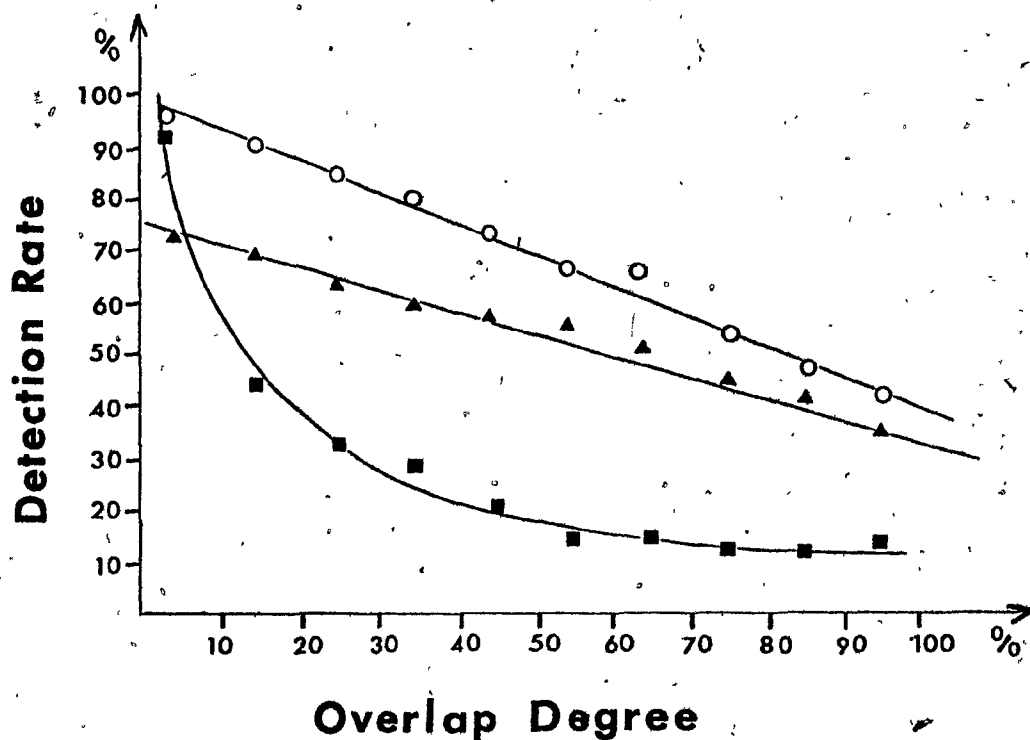


Figure 4 Figure showing the overlapping-cell detection rate as a function of the degree of overlap

a) based on the cytoplasmic shape with respect to the degree of overlap (■)

b) based on the number of nuclei detected with respect to the degree of overlap (▲)

c) based on the cytoplasmic and nuclear shape and the number of detected nuclei (using density information) (○).

-- Results based on nuclear shape

Nuclear shape was used to detect the cells having overlapping nuclei. As the degree of cell overlap increased to over 60%, there was a greater chance of nucleus overlap. Of the 582 artificially generated overlapping cells, 166 have overlapping nuclei and 58 of these 166 cells were detected as having overlapping nuclei using the Fourier shape descriptors of the nuclei (34.94%). The remaining cells exhibited such a high degree of nuclear overlapping that they were very difficult to detect even visually.

- Results based on the density information

When the overlap detection was based on the number of detected nuclei alone, the overlapping-cell detection rate decreased slightly as the degree of overlap increased. Overall, a 65% detection rate was obtained using the density criteria alone. However, if the degree of overlap was restricted to less than 20%, a 72.22% detection rate was obtained (See Figure 4).

- Results based on both shape and density information

When both the shape and density information were used as the criteria for detecting overlapping cells, the detection rate decreased slightly as the overlap degree increased (See Figure 4). Overall, a detection rate of 78.52% was obtained. However, when the degree of overlap was restricted to less than 20%, a 92.59% detection rate was obtained.

b) Results obtained using naturally occurring single and overlapping cells

When both shape and density information were used as criteria for

detecting overlapping cells in the data base of 1203 cervical cells, 146 cells out of 157 naturally overlapping cells were detected as overlapping cells (92.99%), while 146 cells out of 1046 single cells were falsely detected as overlapping cells (14.03%).

Since most of the 157 naturally overlapping cells have minimal overlap degree (less than .20%), the 92.99% detection rate is consistent with the result obtained from the artificially generated overlapping cells (92.59%) when the overlap degree was restricted to less than 20%. This close similarity assures the validity of using the overlapping-cell generation algorithm to create appropriate test situations for overlapping-cell detection studies.

2.6 Conclusions:

The following conclusions are drawn from the experimental results obtained for scene and cell segmentation, and for the overlapping cell detection problem:

(1) Using two color images of cervical cells (530 nm, 570 nm) instead of single color images produced significant improvements in the scene segmentation results. Thus, the multi color image data should be used for future research.

(2) The segmentation method which selects the density threshold based on the stability of areas of segmented regions produced very good results. Thus, it is worthwhile to investigate a more generalized version of the method which considers the stability of the combination of a variety of important features such as area, perimeter, gradient across the boundary etc.

(3) The segmentation error measurement method using the percentage of misclassified pixels is inexpensive, easy to compute, and quite effective because areas of segmented regions are very important features for cervical cell recognition. However, since several other features (describing geometry, or density, or color, or texture information) also play important roles in cervical cell recognition, it is necessary to consider a scene segmentation error measurement method which can take into account several important features at the same time. One such method which evaluates the segmentation error in the pattern space is being developed at the BIPLAB.

(4) Because the overlapping-cell generation algorithm can produce a

cervical cell data base having a uniformly distributed degree of overlap and comprising different cell types, this technique can be used for objective evaluation of alternate overlapping-cell detection algorithms.

— (5) The high detection rates of overlapping cells, with reasonable misclassification of single cells, validated the overlapping-cell detection method. However, some modifications should be made in the future for more effective handling of the following difficult cases:

a) High degree of cell overlap: As shown in the experimental results, it is very difficult to detect cells having a high degree of overlap and it is extremely difficult to detect overlapping cells when the nuclei highly overlap.

b) A small cell overlapping a much larger cell: When the large cell is more than 10 times as large as the small cell, it is very difficult to detect the overlapping cells based on their shape information.

c) Two overlapping cells with very high density contrast: When the average cytoplasmic density value of one cell is about the same or even higher than the average nuclear density value of the other cell, poor segmentation of both the nuclei and cytoplasm results. This creates a difficulty in detecting overlapping cells based on their shape and/or density information.

d) Oddly-shaped cells: Some single cell types such as moderate dysplasia, severe dysplasia, and particularly invasive squamous carcinoma cells have odd shapes which are sometimes misclassified as the shapes of overlapping cells. Some of these oddly-shaped cells may have a nucleus and a

region of the cytoplasm, nearly as dense as the nucleus, which sometimes is falsely detected as a second nucleus.

e) 'Bone' or 'bell' shaped cells: Some single cells have a 'bone' or 'bell' shape which sometimes causes them to be falsely detected as overlapping cells.

f) Pale cells: Some single cells have pale nuclear and cytoplasmic optical density with respect to the background. This often leads to poorly segmented boundaries which, if irregular, are detected as overlapping cellular or nuclear boundaries.

g) Single cells with folded cytoplasm: The optical density of folded cytoplasm is almost as dark as the nucleus and sometimes appears as another nucleus within the cell.

(6) The analytic derivation of the tangent and curvature of the boundary points, in terms of Fourier descriptors, gives us a very effective means of determining the maximal concavity points on the smoothed boundary. Using all the information of the maximal concavity points, instead of following the ad hoc procedure described by Eccles et al[Ecc77a], one can statistically analyze the relative positions and curvature values of the maximal concavity points (as in [Ben81a]) in order to derive a decision rule for finding overlapping cells.

(7) The Fourier shape descriptors can also be used to detect folded cytoplasm of cervical cells by detecting pairs of maximal convex points joined

- by a region having average density value of approximately twice that of the cytoplasm of the cell under investigation.

Chapter 3

FEATURE COMPUTATION

3.1 Introduction

A continuous effort is being made to determine the most effective features for use in cervical cell classification. Previously, at the BIPLAB, Poulsen et al [Pou77a] and Cahn et al [Cah77a] developed a feature extraction system based on several important measurements of size, shape, optical density, and texture.

Color features have been shown to play an important role in cancer cell detection. Bacus[Bac76a], Aggarwal et al[Agg77a], Bengtsson et al[Ben79a], and Holmquist et al[Hol76a] have used it for segmentation purposes. In our laboratory, Louis [Lou77a] has also indicated that color information is very important for cell segmentation. Kulkarni[Kul79a], and Taylor et al[Tay78a] have used color features for cell classification. This motivated our research on color features and led us to extend the IPS image processing software[Pou78a,Tou79a] to include the resulting new color features derived from multi-dimensional histograms of multi-color image data. Additional shape features based on the Fourier and Granlund descriptors of the boundary of cells have also been added.

In the following subsections, the feature extraction system previously developed is briefly described and the computation of all features of the expanded feature extraction system is described.

3.2 The IPS feature extraction system:

The purpose of feature extraction research is to derive new features which vary less among images of the same class than among images of different classes. Once important features have been derived, feature selection and feature evaluation in terms of cost and performance should then be investigated.

The feature extraction system which we previously used[Pou77a,Cah77a] included the following categories of features:

1) Geometric:

a) Separation: distance between the center of the nucleus and the center of the cell.

b) Size: area of the cell nucleus and cytoplasm.

c) Shape: boundary pixel count, moment of area, moment of mass, moment of perimeter, bending energy.

2) Optical Density: mean, variance, mode, skewness, kurtosis, entropy, etc of the nuclear and cytoplasmic optical density histograms.

3) Texture: cytoplasmic and nuclear texture computed as described by Haralick [Har73a].

The IPS feature extraction system has been expanded in the present research to include several new two-dimensional histogram features (consisting of density and color features) and additional shape features. Features

computed by the upgraded system are described in the following section.

3.3 Features computed by the expanded system

3.3.1 Color and density features:

3.3.1.1 Color and density features derived from multidimensional histograms:

Our present research has led to the development of new two-dimensional histogram features which are based on multi-dimensional histograms of multi-color image data.

A two-dimensional histogram P is formed as an $n \times n$ matrix where n is the maximum number of gray levels (or optical density levels) allowed and each element $p(i,j)$ represents the number of pixels with gray value i in one color image (i.e. 530 nm) and gray value j at the corresponding position in a second image (i.e. 570 nm). If the gray value of every pixel is the same for both colors, there is no color information and all non-zero elements of the two-dimensional histogram P lie on the diagonal. On the other hand, if there is color information, non-zero elements exist off the diagonal. Intense color information is represented by non-zero elements lying far from the diagonal. Several density and color measures which can be computed from the two-dimensional histogram are described below:

Considering the two-dimensional histogram $P(i,j)$, $i=1\dots n, j=1\dots n$, if we let $N = \sum_{i=1}^n \sum_{j=1}^n P(i,j)$ be the total pixel count, then $R(i,j)=P(i,j)/N$, $i=1\dots n, j=1\dots n$ is the normalized two-dimensional histogram matrix.

Now, if we let $Q=R+R^T$, where R^T is the transpose matrix of R , then Q is symmetric.

Similar to the textural features derived from the co-occurrence matrix described by Haralick et al [Har73a], the following 14 color and density features can be derived from matrix $Q(i,j), i=1\dots n, j=1\dots n$.

- 1) Logarithm of the angular second moment

$$\text{LOG} \left(\sum_{i=1}^n \sum_{j=1}^n Q^2(i,j) \right)$$

- 2) Logarithm of the contrast

$$\text{LOG} \left(\sum_{k=0}^n k^2 Q_{x-y}(k) \right) \quad \text{where } Q_{x-y}(k) = \sum_{i=1}^n \sum_{j=1}^n Q(i,j) \quad |i-j|=k$$

- 3) Logarithm of the inverse difference moment

$$\text{LOG} \left(\sum_{i=1}^n \sum_{j=1}^n Q(i,j) / (1+(i-j)^2) \right)$$

- 4) Logarithm of the expected value of first diagonal distribution

$$\text{LOG} \left(\sum_{k=0}^n k Q_{x-y}(k) \right)$$

- 5) Logarithm of the variance of first diagonal distribution

$$\text{LOG} \left(\sum_{k=0}^n (k-C4)^2 Q_{x-y}(k) \right)$$

$C4$ is the expected value of first diagonal distribution

- 6) Entropy of the first diagonal distribution

$$\sum_{k=0}^n Q_{x-y}^{(k)} \cdot \text{Log}(Q_{x-y}^{(k)})$$

7) Logarithm of the expected value of second diagonal distribution

$$\text{LOG} \left(\sum_{l=0}^{2n} l \cdot Q_{x+y}^{(l)} \right) \quad \text{where } Q_{x+y}^{(l)} = \sum_{i=1}^n \sum_{j=1}^n Q(i, j) \quad i+j=l$$

8) Logarithm of the variance of second diagonal distribution

$$\text{LOG} \left(\sum_{l=0}^{2n} (l-C7)^2 Q_{x+y}^{(l)} \right)$$

C7 is the expected value of second diagonal distribution

9) Entropy of the second diagonal distribution

$$\sum_{l=0}^{2n} Q_{x+y}^{(l)} \cdot \text{Log}(Q_{x+y}^{(l)})$$

10) Logarithm of the expected value of row distribution

$$\text{LOG} \left(\sum_{m=1}^n m \cdot Q_x^{(m)} \right) \quad \text{where } Q_x^{(j)} = \sum_{i=1}^n Q(i, j)$$

11) Logarithm of the variance of row distribution

$$\text{LOG} \left(\sum_{m=1}^n (m-C10)^2 Q_x^{(m)} \right)$$

C10 is the expected value of row distribution

12) Entropy of row distribution

$$\sum_{m=1}^n Q_x^{(m)} \cdot \text{Log}(Q_x^{(m)})$$

13) Entropy of entire distribution

$$\sum_{i=1}^n \sum_{j=1}^n Q(i,j) * \log(Q(i,j))$$

14) Correlation

$(\sum_{i,j} i*j*Q(i,j) - \mu_x * \mu_y) / (\sigma_x * \sigma_y)$ where $\mu_x, \mu_y, \sigma_x, \sigma_y$ are means and variances of $Q(i,j)$.

Also, the following two important features are included:

15) Logarithm of the chi-square value

$$\log(N * (\sum_{i=1}^n \sum_{j=1}^n Q^2(i,j) / (Q_x(i) * Q_y(j)) - 1))$$

where $Q_y(j) = \sum_{i=1}^n Q(i,j)$ and $N=n*n$

16) Logarithm of the ratio of expected value of first to second diagonal distribution

$$\log(C4/C7)$$

where C4 and C7 are expected values of first and second diagonal distributions respectively.

The features 10, 11, 12 (Logarithm of the expected value, logarithm of the variance, and entropy of row distribution respectively) are density features. The rest contain either color information alone or both color and density information. For three color images of cervical cells, composed of three color images of the cytoplasm and three color images of the nuclei, there are six possible sets of 16 two-dimensional histogram features:

a) 16 features (F1 to F16) extracted from two-dimensional histograms of cell cytoplasm scanned at 530 nm and 570 nm wavelengths

b) 16 features (F17 to F32) extracted from two-dimensional histograms of cell cytoplasm scanned at 570 nm and 620 nm wavelengths.

c) 16 features (F33 to F48) extracted from two-dimensional histograms of cell cytoplasm scanned at 530 nm and 620 nm wavelengths.

d) 16 features (F49 to F64) extracted from two-dimensional histograms of cell nuclei scanned at 530 nm and 570 nm wavelengths.

e) 16 features (F65 to F80) extracted from two-dimensional histograms of cell nuclei scanned at 570 nm and 620 nm wavelengths.

f) 16 features (F81 to F96) extracted from two-dimensional histograms of cell nuclei scanned at 530 nm and 620 nm wavelengths.

3.3.1.2 Color and density features derived from one-dimensional histograms:

When considering one-dimensional histograms obtained from cell images scanned at one wavelength (using a Zeiss VG-9 filter), only density features can be extracted. But when considering all three one-dimensional histograms, color as well as density features can be extracted.

The following 9 features can be computed from each one-dimensional histogram $H(I)$, $I=1$ to n density level:

1) Mode value = most frequently occurring density level among all bin values of the histogram (value of I at which $H(I)$ is maximum).

2) Modal frequency = number of times the mode occurs divided by the area

3) Mean density =
$$\left[\sum_{I=1}^n I \cdot H(I) \right] / \text{area}$$

4) Variance of density =
$$\left[\sum_{I=1}^n (I - m)^2 \cdot H(I) \right] / \text{area}$$

5) Skewness of density =
$$\left[\sum_{I=1}^n (I - m)^3 \cdot H(I) \right] / \text{area}$$

6) Kurtosis of density =
$$\left[\sum_{I=1}^n (I - m)^4 \cdot H(I) \right] / \text{area}$$

7) Entropy of density =
$$- \sum_{I=1}^n [H(I) / \text{area}] \cdot \text{LOG}[H(I) / \text{area}]$$

8) Range of density = density at rightmost non-zero bin - density at leftmost non-zero bin

9) Median value of density = $0.5 \cdot [\text{sum of density at rightmost and leftmost non-zero bin}]$

From three one-dimensional histograms, the first set of 9 features can be extracted from any one of the three histograms. Two other sets of features can be extracted from two pairs of one-dimensional histogram features and from all three one-dimensional histogram features. The features extracted from the

C pairs are the combinations of the two corresponding one-dimensional histogram features:

- 1) Difference/sum of two mode values
- 2) Difference/sum of two modal frequencies
- 3) Difference of two optical density means.
- 4) Logarithm of sum of two variances
- 5) Logarithm of absolute value of skewness
- 6) Logarithm of kurtosis value.
- 7) Entropy.
- 8) Difference/sum of two medians.

The feature extracted from three one-dimensional histograms is:

- 1) Logarithm of sum of three optical density variances.

In this research, the following 8 sets of one-dimensional histogram features are computed:

- a) Eight features (F97 to F104) from the pair of two cytoplasmic one-dimensional histograms (530nm, 570nm).

b) Eight features(F105 to F112) from the pair of two cytoplasmic one-dimensional histograms (530nm,620nm).

c) One cytoplasmic feature (F113) from three cytoplasmic one-dimensional histograms (530nm,570nm,620nm).

d) Eight features (F114 to F121) from the pair of two nuclear one-dimensional histograms (530nm,570nm).

e) Eight features (F122 to F129) from the pair of two nuclear one-dimensional histograms (570nm,620nm).

f) One nuclear feature (F130) from three nuclear one-dimensional histograms (530nm,570nm,620nm).

g) Nine cytoplasmic features (F131 to F139) from a single cytoplasmic one-dimensional histogram (530nm).

h) Nine nuclear features (F140 to F148) from a single nuclear one-dimensional histogram (570nm).

3.3.2 Geometric features:

Several important geometric features are computed from the segmented regions:

3.3.2.1 Features indicating the position of nuclei in the cells:

$$F149 = \text{LOG}(\text{eccentricity}) = \text{LOG}\left[\sqrt{(X_c - X_n)^2 + (Y_c - Y_n)^2}\right]$$

where (X_c, Y_c) and (X_n, Y_n) are coordinates of the centers of a cell and of a nucleus

$$F150 = \text{LOG}[\sqrt{(X_{\text{cure}} - X_{\text{curn}})^2 + (Y_{\text{cure}} - Y_{\text{curn}})^2}]$$

where $(X_{\text{cure}}, Y_{\text{cure}})$ and $(X_{\text{curn}}, Y_{\text{curn}})$ are coordinates of the centers of the frames enclosing a cell and a nucleus respectively.

3.3.2.2 Size features:

$$F151 = \text{LOG}(\text{cytoplasmic area})$$

$$F152 = \text{LOG}(\text{nuclear area/cytoplasmic area})$$

$$F153 = \text{LOG}(\text{cell perimeter})$$

3.3.2.3 Shape features:

The requirements for shape features are the invariance of magnification, translation, and rotation. In addition to moment features developed before by Cahn et al [Cah77a], Fourier and Granlund shape descriptors were developed for use in cervical cell recognition.

a) Moment features:

The following moment features were defined by Cahn [Cah77a]. Given a function $f(x, y)$, the two-dimensional moments are:

$$M_{ij} = \sum_x \sum_y x^i y^j f(x, y) \quad i, j = 0, 1, 2, \dots$$

Where the summations are over all pixels in the cell image in the horizontal (x) and vertical (y) directions.

If we let $\bar{x} = M_{10}/M_{00}$ and $\bar{y} = M_{01}/M_{00}$

Then $M_{ij} = \sum_x \sum_y (x-\bar{x})^i (y-\bar{y})^j f(x,y)$ $i, j = 0, 1, 2, \dots$ are independent of translation. Moreover, invariance of magnification can be obtained if we let:

$$N_{ij} = M_{ij}/M_{00}^l, \text{ where } l=(i+j+2)/2, i+j=2, 3, 4, \dots$$

Furthermore, if we let:

$$L_1 = N_{20} + N_{02} \quad L_2 = N_{20} - N_{02} \quad L_3 = N_{30} - 3N_{12}$$

$$L_4 = 3N_{21} - N_{03} \quad L_5 = N_{30} + N_{12} \quad L_6 = N_{21} + N_{03}$$

Then the following seven moment features are invariant to translation, magnification, and rotation:

$$MF1 = L_1$$

$$MF2 = L_2^2 + 4N_{11}^2$$

$$MF3 = L_3^2 + L_4^2$$

$$MF4 = L_5^2 + L_6^2$$

$$MF5 = L_3 L_5 (L_5^2 - 3L_6^2) + L_4 L_6 (3L_5^2 - L_6^2)$$

$$MF6 = L_2 (L_5^2 - L_6^2) + 4N_{11} L_5 L_6$$

$$MF7 = -L_4 L_5 (L_5^2 - 3L_6^2) + L_4 L_6 (3L_5^2 - L_6^2)$$

In this research, three kinds of moments are considered, namely, moment of perimeter, moment of cell area, and moment of nuclear area:

* Moment of perimeter:

If in the above moment formula, the function $f(x,y)$ is set to include only points on the perimeter $p(x,y)$, then the seven moments of perimeter (F154 to F160) can be extracted.

* Moment of cell area:

If in the above moment formula, $f(x,y)$ is set to 1 when the pixel is inside the cell and zero otherwise, the seven moments of cell area (F161 to F167) are obtained.

* Moment of nuclear area:

If in the above moment formula, $f(x,y)$ is set to 1 when the pixel is inside the nucleus and zero otherwise, the seven moments of nuclear area (F168 to F174) are obtained.

b) Fourier shape features:

The additional shape descriptors evaluated in this study are based on the Fourier transform of the boundary of an object. The basic approach was devised by Granlund[Gra72a] and has been used by Holmquist et al[Hol78a] and

Chen et al[Che80a] for cervical cell classification. From the above Fourier descriptors, a_n , $n=-p$ to $+p$ (described previously in section 2.5.4.a), translational invariance is obtained by setting a_0 equal to zero and size invariance is obtained by setting $a'_n = a_n / (|a_1| + |a_{-1}|)$. Finally, the following rotationally invariant shape features can be derived:

1) Logarithm of normalized minor axis of ellipse reconstructed from the first descriptors

$$F175 = \text{LOG}(|a'_1| - |a'_{-1}|)$$

2) Logarithm of sum of power spectrum

$$F176 = \text{LOG}(\sum_{n=-p}^p (a'_n)^2 + (a'_{-n})^2)$$

3) Logarithm of sum of Magnitude

$$F177 = \text{LOG}(\sum_{n=-p}^p |a'_n| + |a'_{-n}|)$$

4) Shape factor

$$F178 = P^2/4\pi A = P^2/4\pi \sum_{n=1}^p n^3 (|a'_n|^3 - |a'_{-n}|^3)$$

5) Concavity

$$F179 = \sum_{n=1}^p n^3 * (a'^2_n - a'^2_{-n})$$

c) Granlund's shape features:

In [Gra72a] Granlund suggested that the descriptors $b_n = (a_{1+n} * a_{1-n}) / a_1^2$, $n=1,2,3,\dots$ can be used as the shape descriptors

because they are invariant to translation, magnification, and rotation. In this research, real and imaginary values of the first six Granlund's descriptors were used:

$$\begin{array}{ll}
 F180 = \text{Re}(b1) & F181 = \text{Im}(b1) \\
 F182 = \text{Re}(b2) & F183 = \text{Im}(b2) \\
 F184 = \text{Re}(b3) & F185 = \text{Im}(b3) \\
 F186 = \text{Re}(b4) & F187 = \text{Im}(b4) \\
 F188 = \text{Re}(b5) & F189 = \text{Im}(b5) \\
 F190 = \text{Re}(b6) & F191 = \text{Im}(b6)
 \end{array}$$

3.3.3 Texture features:

The technique to extract texture features from co-occurrence matrices was proposed by Haralick et al [Har73a]. In this method, first, the co-occurrence matrices $P_{d\theta}$ are computed. The element value $p_{d\theta}(i, j)$ represents the number of times two pixels at a distance d , angle θ , whose gray levels are i and j were found. For each co-occurrence matrix (corresponding to specific values of d and θ), the following 13 statistical measurements can be obtained:

$$f(1) = \sum_{i=1}^n \sum_{j=1}^n p^2(i, j) \quad - \text{"Angular second moment"} -$$

$$f(2) = \sum_{k=0}^n k^2 p_{x-y}(k) \quad - \text{"Contrast"} -$$

$$\text{where } p_{x-y}(k) = \sum_{i=1}^n \sum_{j=1}^n p(i, j) \\ |i-j| = k$$

$$f(3) = \sum_{i=1}^n \sum_{j=1}^n p(i, j) / (1 + (i-j)^2) \quad - \text{"Inverse difference moment"} -$$

$$f(4) = \sum_{k=0}^n k p_{x-y}(k)$$

- "Expected value of 1st diagonal distribution" -

$$f(5) = \sum_{k=0}^n (k-f(4))^2 p_{x-y}(k)$$

- "Variance of 1st diagonal distribution" -

$$f(6) = \sum_{k=0}^n p_{x-y}(k) \log(p_{x-y}(k))$$

- "Entropy of 1st diagonal distribution" -

$$f(7) = \sum_{l=2}^{2n} l p_{x+y}(l)$$

- "Expected value of 2nd diagonal distribution" -

$$\text{where } p_{x+y}(l) = \sum_{\substack{i=1 \\ i+j=l}}^n \sum_{j=1}^n p(i,j)$$

$$f(8) = \sum_{l=2}^{2n} (l-f(7))^2 p_{x+y}(l)$$

- "Variance of 2nd diagonal distribution" -

$$f(9) = \sum_{l=2}^{2n} p_{x+y}(l) \log(p_{x+y}(l))$$

- "Entropy of 2nd diagonal distribution" -

$$f(10) = \sum_{m=1}^n m p_x(m)$$

- "Expected value of row distribution" -

$$\text{where } p_x(i) = \sum_{j=1}^n p(i,j)$$

$$f(11) = \sum_{m=1}^n (m-f(10))^2 p_x(m)$$

- "Variance of row distribution" -

$$f(12) = \sum_{m=1}^n p_x(m) \log(p_x(m))$$

- "Entropy of row distribution" -

$$f(13) = \sum_{i=1}^n \sum_{j=1}^n p(i,j) \log(p(i,j)) \quad \text{"Entropy of entire distribution"}$$

Cahn [Cah77a] investigated and determined that the following texture features seemed to work best:

$$t1 = \sum_{d=1}^3 \sum_{\theta=0,45,90,135} f_{d\theta}(7)/12$$

$$t2 = \sum_{d=1}^3 \sum_{\theta=0,45,90,135} f_{d\theta}(8)/12$$

$$t3 = \sum_{d=1}^3 \sum_{\theta=0,45,90,135} f_{d\theta}(10)/12$$

$$t4 = \sum_{d=1}^3 \sum_{\theta=0,45,90,135} f_{d\theta}(11)/12$$

$$t5 = \sum_{d=1}^3 \sum_{\theta=0,45,90,135} f_{d\theta}(12)/12$$

$$t6 = \sum_{d=1}^3 \sum_{\theta=0,45,90,135} f_{d\theta}(13)/12$$

$$t7 = \sum_{\theta=0,45,90,135} f_{1\theta}(1)/4$$

$$t8 = \sum_{\theta=0,45,90,135} f_{3\theta}(5)/4$$

$$t9 = \sum_{\theta=0,45,90,135} f_{1\theta}(9)/4$$

The 9 texture features extracted from cytoplasm (F192 to F200) and the 9 texture features extracted from nuclei (F201 to F209) were used.

3.4 Summary:

The feature extraction system previously developed [Cah77a] was upgraded to include several two-dimensional histogram features and several Fourier and Granlund shape features to provide a more informative set of features representative of all cell categories. These new features include color, density, geometric, and texture features. The 209 features computed from the new feature extraction system are composed of:

Feature category	Number of features
Color and density features:	
- From two-dimensional histograms	96
- From one-dimensional histograms	52
Geometric features:	
- Eccentricity	2
- Size	3
- Shape:	
* Moment features	21
* Fourier descriptors	5
* Granlund descriptors	12
Texture features	18
	<hr/>
Total	209

Chapter 4

FEATURE SELECTION

4.1 Introduction

Using the upgraded feature extraction system described in chapter 3, a total of 209 meaningful features were computed from a data base of 3000 manually-segmented cervical cells. If all of the 209 features were used for classifying cervical cells, the cost, in computer time and memory storage, would be prohibitive. Also, a large number of training samples would be needed^a to train the minimum Mahanalobis distance classifier. As pointed out by Cahn [Cah77a], the ratio $2d/(f+3)$ (where d is the number of samples and f is the number of features) should be at least 3 and preferably greater than 10 to provide adequate training of the classifier. This implies that for the 209 features, we would need at least from 318 to 1000 cervical cells per class, or from about 5000 to 16000 cervical cells, for the 16 classes used. Thus, it is necessary to select an effective subset of features from the total set. To make such a selection, one needs both a criterion to evaluate the subset of features and a procedure to search for effective subsets using that criterion. In the present research, the criterion used was the probability of misclassification, and a number of feature search procedures were applied. In order to compute the probability of misclassification, the minimum Mahanalobis distance classifier developed by Oliver [Oli77a,78a,78b] and the random partitioning method were used. Because an exhaustive search of $\binom{n}{k}$ combinations (to choose a size- k optimal subset from n features) would be far too time-consuming, three heuristic feature search procedures were

investigated. The first procedure is the forward sequential search procedure, or without-replacement search procedure. In this method, the best individual feature is chosen on the first round, then the best pair including the best individual feature is chosen for the second round, etc. The second procedure is the parallel search procedure which selects k individually best discriminating features. The third procedure is the feature clustering procedure in pattern space. In this method, first the minimal spanning tree joining N points (representing N features) in the M -dimensional pattern space is formed, then k subtrees (clusters) of features are obtained by breaking the first $k-1$ longest path lengths. Finally, k features are selected where each feature is the best individual one of its corresponding subtree and where its cancer cell detection rate is higher than a preset minimum threshold. (This minimum threshold is used to protect the procedure from selecting independent but bad features).

In the following subsections, a brief survey of feature selection methods, the methods applied to the cervical cell recognition problem and their performances are given.

4.2 Survey of feature selection approaches:

4.2.1 Feature evaluation criteria:

In order to select a subset of the total set of features, one needs an effective means (criterion) of evaluating the power of any given subset of features. The feature evaluation criteria which have been applied previously include: the probability of misclassification [Ste76a, Lin80a], the Mahanalobis distance [Dud73a, You74a], the Bhatacharyya distance [Dud73a], the divergence [Cha73a], the entropy function [You74a], and the Karhunen-Loeve expansion [Muc71a, Kan74a].

4.2.1a Probability of misclassification

In this criterion, the probability of misclassifying a sample to a class different from its own is used to evaluate feature subsets (the lower the probability of misclassification the better the feature subset). In order to obtain the probability of misclassification, a certain kind of classifier is required. This criterion has two advantages: -- 1) It evaluates the subsets of features in terms of the probability of misclassification which other criteria are only indirectly related to and -- 2) It takes into account the particular classifier structure for which the features are intended whereas other criteria such as distance measures would provide the same feature subsets for different classifiers.

4.2.1b Mahanalobis distance:

If we assume that the conditional density function $p(x/i)$ of the

feature vector x of dimension d given a class i ($i=1, \dots, M$) has a d -variate Gaussian distribution, then the Mahalanobis distance, $(x-m)^t \Sigma_i^{-1} (x-m)$, (where m and Σ_i are the mean and covariance matrix for class i , t indicates transpose, -1 indicates inverse) can be used to evaluate the feature subsets (the larger the Mahalanobis distance, the better the feature subset). If $p(x/i)$ actually has a multivariate Gaussian distribution then this criterion is a monotonic function of the probability of misclassification criterion. On the other hand, the Mahalanobis distance criterion can provide poor features for a classifier using distance measures other than the Mahalanobis distance.

4.2.1c Bhattacharyya distance:

Since $b = \int \sqrt{p(x/c1) * p(x/c2) * \dots * p(x/cm)} dx$, the Bhattacharyya coefficient, is a measure of how much overlap there is among class conditional density functions, the Bhattacharyya distance $-\log(b)$ can be used as a measure of goodness of the feature subsets (the larger the Bhattacharyya distance the better the feature subset). The Bhattacharyya distance criterion has two disadvantages: -- 1) It is not a direct measure of the overall probability of misclassification and -- 2) Some density estimation techniques are required to estimate the class conditional density.

4.2.1d Divergence:

For two classes, the divergence is defined as $J(x) = J1(x) + J2(x)$ where $J1(x) = \int p(x/c1) \log(p(x/c1)/p(x/c2)) dx$ and $J2(x) = \int p(x/c2) \log(p(x/c2)/p(x/c1)) dx$. It can be used for discriminating pairs of class probability densities in the feature space and thus can be used

to evaluate feature subsets (the higher the divergence value the better the feature subset). Like the Bhattacharyya distance, the divergence is not a direct measure of the overall probability of misclassification and some density estimation techniques are required to estimate the class conditional density.

4.2.1e Entropy function:

The entropy is defined as $-\int p(x/C_i) \log(p(x/C_i)) dx$ for class C_i . It is a measure of the spread of the samples of each class in the pattern space (entropy is equal to zero if all members of a class have the same vector representation). Therefore, it can be used to evaluate the effectiveness of the features in representing each class. The entropy has the same two disadvantages as the Bhattacharyya distance and the divergence: it is not a direct measure of the overall probability of misclassification and some density estimation techniques are required to estimate the class conditional density.

4.2.1f Karhunen-Loeve expansion (principal component analysis):

In this method, a principal component analysis is performed to form new features corresponding to the eigenvectors of the sample covariance matrix from the initial features. Like the entropy function, the Karhunen-Loeve expansion can be used to select the most effective features in representing each class. This criterion is the most reliable feature extractor among linear transformations. However, the reduction in the number of dimensions is sometimes offset by the increase in computation necessary to

map vectors into the selected subspace. It also requires a very large amount of training data.

In summary, as pointed out by Stearns [Ste76a], any criterion other than the probability of misclassification imposes a division between feature extraction and classification because it evaluates the subsets of features without reference to a classifier.

4.2.2 Feature search procedures

Several investigations [Tou71a,71b,Cov74a] showed that the exhaustive search over all $\binom{n}{k}$ subsets of size k was necessary to find the optimal k features from a set of n features. Because the examination of all possible subsets is impractical, several heuristic feature search procedures have been developed for use. These include the sequential search, the parallel search, dynamic programming, and the feature clustering (minimal spanning tree) in pattern space.

4.2.2a Sequential search procedure (i,j) [Ste76a]:

In this procedure, the best i features are added one-by-one according to a feature evaluation criterion, then the worst j features are removed one-by-one according to a possibly different feature evaluation criterion. The process is then repeated until the desired subset is obtained. The procedure is called bottom-up if $i > j$ and top-down if $i < j$. In the special case where $i=1, j=0$, it is equivalent to the forward sequential procedure [Muc71a] and where $i=0, j=1$, it is equivalent to the backward sequential

procedure. If both i and j are different from zero then the selected features have no without-replacement property (i.e. for the forward sequential procedure, the subset of the k selected features has to contain the $k-f$ features previously selected, and for the backward sequential procedure, any features which have been removed are not reconsidered).

The sequential search procedure considers the correlations among selected features, therefore, the selected features can be very effective. In terms of computer time, for any specified subset size k , when k is much smaller than n , the bottom-up procedures ($i > j$) require much less computer time than the top-down procedures. Conversely, when k is close to n , the top-down procedures need much less computer time. In practice, k/n is almost always less than a half, making bottom-up search preferable to top-down. When computer time is very limited, the forward sequential search procedure (which requires the least computer time) can be used.

4.2.2b Parallel selection:

In this procedure, all n features are evaluated individually and then the k features which have the highest individual discriminating power are selected. An example of this procedure can be found in [Muc71a]. The procedure is computationally simple but it has a critical disadvantage in not taking interactions between features into account. Therefore the selected feature subsets may contain redundant members.

4.2.2c Dynamic programming [Cha73a]:

The dynamic programming procedure overcomes the drawback of the

exhaustive search procedure in terms of computer time and, at the same time, is not subject to the without-replacement property of the forward or backward sequential procedures:

The dynamic programming method has the property that whatever the initial state and initial decision are, the resulting decisions must constitute an optimal policy with regard to the state resulting from the first decision.

An application of the foregoing principle of optimality together with an appropriately formulated recursive functional equation ensure that the final subset chosen may not necessarily include all of the best single features selected in the previous stages.

The procedure requires more computer time than the forward sequential procedure. Chang [Cha73a] has applied a dynamic programming procedure which required nearly twice the amount of computation as the forward sequential procedure.

4.2.2d Feature clustering in pattern space [Cah77a]:

In this procedure, each feature is represented by a point in the pattern space and k clusters of more or less redundant features are found by first forming the minimal spanning tree of the points and then breaking the first $k-1$ longest path lengths. One representative feature can be chosen from each cluster to form a subset of k features. This procedure is also very time-saving and it takes feature correlation into account. However, a criterion for selecting a representative feature from each cluster has to be

derived and as shown recently in [Rob82a], the clustering in pattern space does not necessarily yield an effective subset of features.

In summary, as a general rule, if time permits, the sequential search procedure or dynamic programming is preferable. If computation time is very limited, the parallel search procedure and the feature clustering procedure should be applied.

4.3 Feature selection procedures applied to cervical cell recognition:

4.3.1 Feature evaluation criterion

The probability of misclassification criterion was chosen to evaluate subsets of features. To compute the probability of misclassification, the minimum Mahalanobis distance classifier developed by Oliver[Oli77a,78a,78b] was used with the "random partitioning" method. In this method, 80% of the data set (2400 cervical cells) was randomly selected for training and the remaining independent samples (600 cells) were used for testing. The detailed description of this method is given by Toussaint [Tou74a]. Because of computer time constraints, the 16-class classification performance estimates were obtained by averaging the results for only 5 different partitionings of data.

4.3.2 Feature subset search procedures:

Three feature subset search procedures were investigated in this research: The forward sequential search, the parallel search, and the feature clustering in pattern space.

4.3.2a Forward sequential search

The description of this procedure can be found in subsection 4.2.2a . Instead of trying to select features directly from the total set of 209 features, we initially divided these features into 7 groups of about 30 features which belong to the same category. The forward sequential search procedure was applied to each group to select features. The selected features

from the 7 groups were then merged and were divided a second time into groups of about 30 features regardless of their feature categories. The search procedure was then applied to the new groups. The process of merging, dividing, and selecting features was continued until finally only k features ($k < 20$) were selected.

4.3.2b Parallel search

The description of this procedure can be found in subsection 4.2.2b. K was chosen to be less than or equal to 20 in our present research.

4.3.2c Feature clustering in pattern space

The description of this procedure can be found in subsection 4.2.2d. K was chosen to be less than or equal to 20 in our present research, and the best discriminating feature from each cluster was selected provided its 16-class classification detection rate was higher than a preset minimum threshold of 6%. Also, for ease of computation, the 209 computed features described in chapter 3 were initially broken into three subsets of around 70 features.

4.4 Experimental results for the three feature search procedures

4.4.1 Feature subsets selected by the three procedures

Table 2 shows the results obtained when the above three procedures were used to select features from the total set of 209 features.

Order of selection	Individually best feature selection	Forward sequential selection	Feature clustering in pattern space
1st	F152	F152	F152
2nd	F15	F151	F15
3rd	F47	F55	F91
4th	F149	F39	F178
5th	F31	F36	F62
6th	F151	F166	F149
7th	F154	F67	F132
8th	F153	F7	F175
9th	F91	F50	F81
10th	F115	F87	F150
11th	F84	F75	F141
12th	F50	F136	F46
13th	F132	F88	F63
14th	F107		F195
15th	F128		F115
16th	F79		F123
17th	F52		F50
18th	F59		F172
19th	F66		F79
20th	F49		F204

TABLE 2. Features selected by the three search procedures

F_n is as described in chapter 3.

For the forward sequential search procedure, the 16-class classification error rate reached a minimum and thus terminated the search process when the number of features was 13. This was not so for the other two procedures; the 16-class classification error rates decreased and saturated as the number of features approached 20.

In order to compare more accurately the performances of the three feature selection methods, the same minimum Mahanalobis distance classifier and the random partitioning method were applied as k , the total number of selected features, varied from 1 to 13 for forward sequential search procedure and from 1 to 20 for the other two procedures. The classification performance estimates were obtained by averaging the results over 30 different partitionings of the data instead of the 5 partitionings used when searching for feature subsets. The error rates for misclassifying cells of one class into the other fifteen classes (16-class classification error rate) and the error rates for misclassifying normal to abnormal cells and vice versa (2-class classification error rate) were both investigated.

4.4.2 Classification performance estimates:

4.4.2a Forward sequential search procedure:

This method gave significantly better results on the classification error rates, compared to the other two procedures (the parallel search and the feature clustering in pattern space). As the number of features increased and reached 13, the classification error rates decreased gradually and reached 24.94% for the 16 classes and 2.2% for the two classes as shown in Figures 5

and 6. This procedure, however, is more time-consuming than the other procedures.

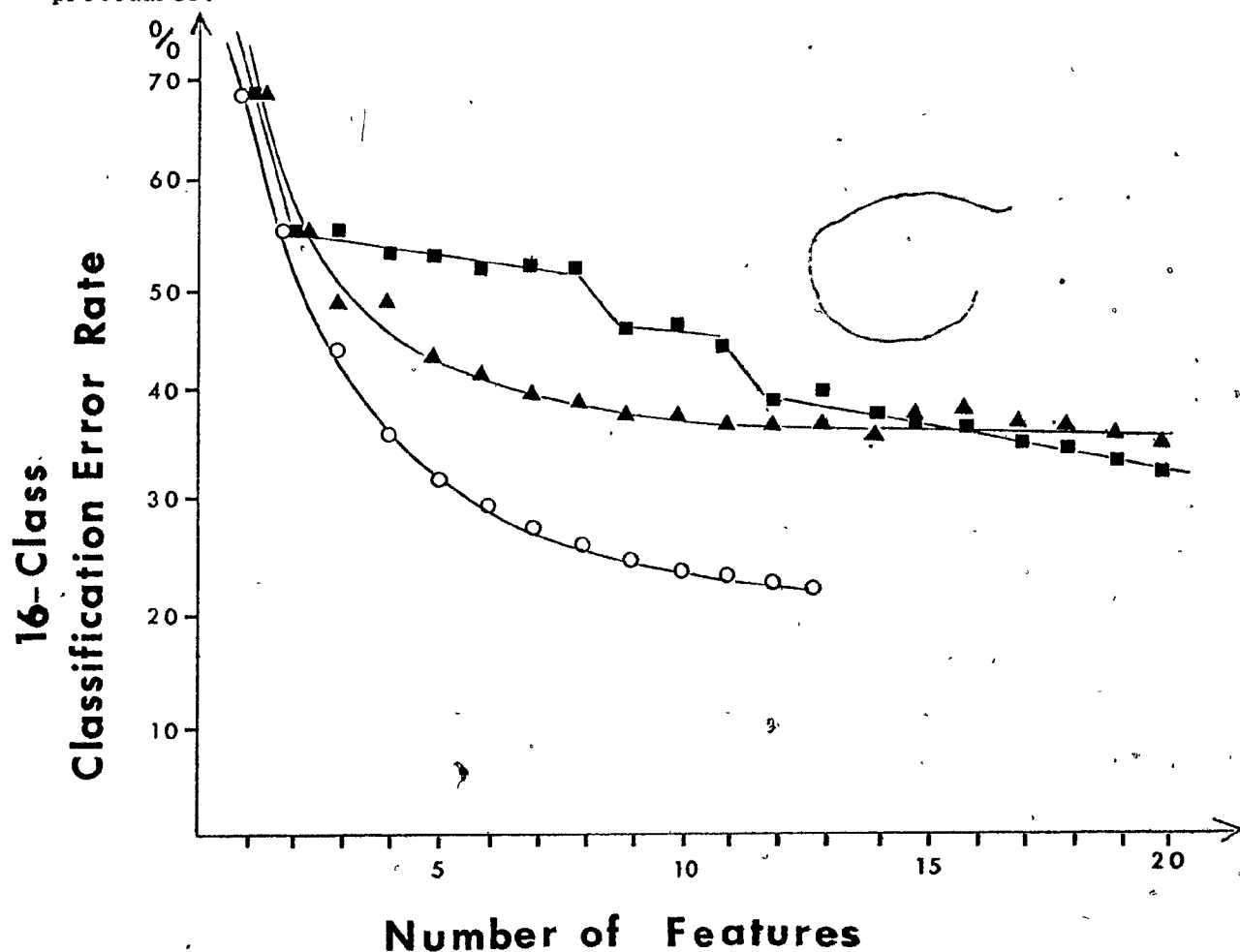


Figure 5. 16-class classification error rate for : a) the forward sequential search procedure (O), b) The parallel search procedure (■), and c) The feature clustering procedure in pattern space (▲).

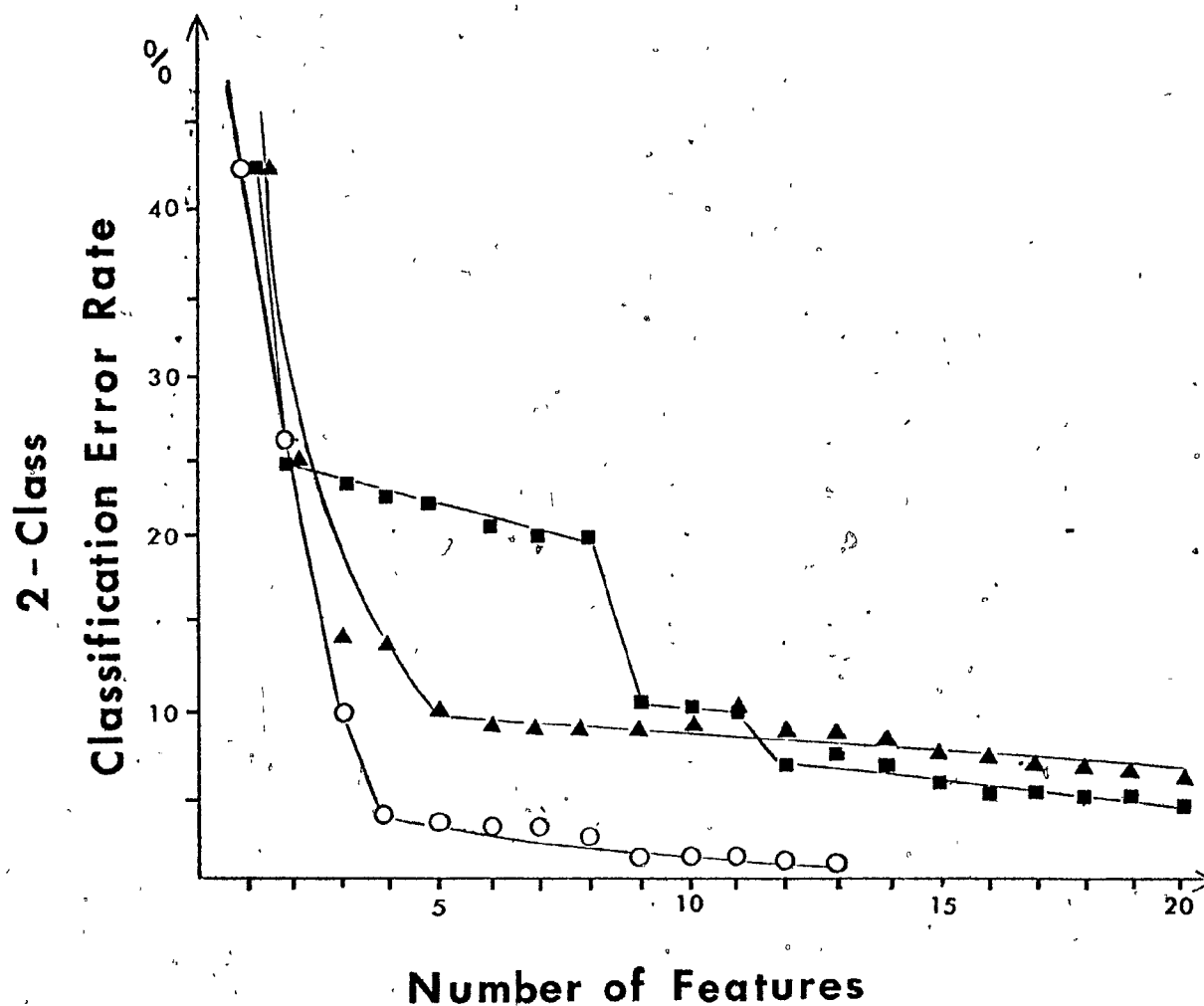


Figure 6. 2-class classification error rates for : a) The forward sequential search procedure (O), b) The parallel search procedure (■), and c) The feature clustering procedure in pattern space (▲).

4.4.2b Parallel search procedure:

This procedure gave worse results than the forward sequential search procedure. As k varied from 1 to 20, the classification error rates based on 16 subclasses and based on 2 classes (normal and abnormal) decreased and

saturated at about 35% and 4.9% respectively (at $k=20$) as shown in Figures 5 and 6.

4.4.2c Feature clustering procedure in pattern space:

This procedure gave better results than the parallel search procedure as k varied from 1 to 7 and slightly worse results when k was greater than 7. This is probably due to the fact that the procedure did not choose non-redundant features (as shown recently by Roberts et al[Rob82a]) or because of the initial breaking of the 209 features into three subsets of features for ease of computation. As k varied from 1 to 20, the classification error rates decreased and saturated at 37.56% and 7.4% for 16 subclasses and 2 classes respectively as shown in Figures 5 and 6.

4.5 Discussion and Conclusions

The forward sequential search procedure selected much more effective features than the other two procedures. When the forward sequential search procedure was applied, the classification error rates decreased significantly for the first few selected features ($k < 5$) and then decreased gradually and monotonically to a minimum value as the number of features increased. In contrast, for the parallel search and the feature clustering, that behaviour was not observed. In the parallel search procedure, significant drops of classification error rates were obtained when k was increased from 1 to 2, 8 to 9, and 11 to 12. The occurrences of significant drops were probably due to the fact that the 2nd, 9th, and 12th features added were particularly effective and non-redundant. In the feature clustering procedure, for $k < 6$, the classification error rates decreased significantly in the same manner as the forward sequential search procedure, but when k was greater than 6 the error rates decreased only slightly. This is probably because of the ineffectiveness of the procedure as pointed out recently by Roberts et al [Rob82a] or because of the initial manual breaking of the 209 computed features into three subsets of features.

The forward sequential procedure is more time-consuming than the parallel search and feature clustering procedures and it still contains the without-replacement property. If more computer time is allowed, the without-replacement property can be avoided by applying the sequential search procedure (i, j) with i, j different from zero or by applying the $(1, 1)$ procedure to the k features selected by the forward sequential search

procedure as in [Lin80a]. The IPS image processing software is being modified so that the manual division of features into categories can be avoided. When this research was conducted, the IPS system could only support feature selection with feature sets of up to 40 features.

Chapter 5

FEATURE EVALUATION

5.1 Introduction

The significance of each feature category can probably be determined by the proportion of the features of that category in the final subsets obtained from feature search procedures. Another method to evaluate features of the same category (i.e. two-dimensional histogram features, one-dimensional histogram features, geometric features, and texture features) is to evaluate the selected subset of features from each category. In this research, the forward sequential search procedure was chosen because of the reasons described in chapter 4. In addition, to verify the significance of the newly-developed two-dimensional histogram features, the performance of the 13 features selected from the whole set of 209 features and the performance of the set of 6 features previously used by Oliver[Oli78b] were included for comparison.

5.2 Evaluation of each feature category:

In order to evaluate each feature category, first, a search procedure was applied to all features from each category to select an effective subset representative of each feature category, and then its classification performance was estimated and compared to the classification performances of other subsets representing different feature categories.

5.2.1 Feature search procedure used for each feature category

The forward sequential search procedure with the probability of misclassification criterion was applied to each feature category (method described in chapter 4).

5.2.2 Classification performance estimation of feature subsets selected from each feature category

The same minimum Mahalanobis distance classifier used in selecting features, and two classification performance measurement methods, "resubstitution" and "random partitioning", were applied. In the resubstitution method, the same set of cells is used for training and testing. This method is regarded as optimistic. In the random partitioning method, similar to the feature selection procedure, 80% of the data set (2400 cells) was randomly selected for training and the remaining independent samples (600 cells) were used for testing. However, the classification performance estimates were obtained by averaging the results for 30 different partitionings of the data (instead of only 5 partitionings used in selecting features). This method is regarded as somewhat pessimistic. The results are

shown in the form of classification performance curves and/or confusion matrices. The classification performance curves are produced by altering the a priori probability ratio settings for normal to abnormal cell types from 10:1 to 1:30 and recording the corresponding points in the false positive, false ~~negative~~ classification performance plot. The confusion matrix corresponds to the point on the classification curve where this ratio is 1:1.

For the purpose of comparison, the performance of the features selected from one category was compared to that of the features selected from another, and to the performance of the 13 features selected from the whole set (which includes all feature categories). Also, to verify the significance of the two-dimensional histogram features in particular, the performance of the 6 features previously used by Oliver [Oli78b] was investigated. These 6 features were obtained from the same cells but scanned in a single color and at 1.0 micron resolution.

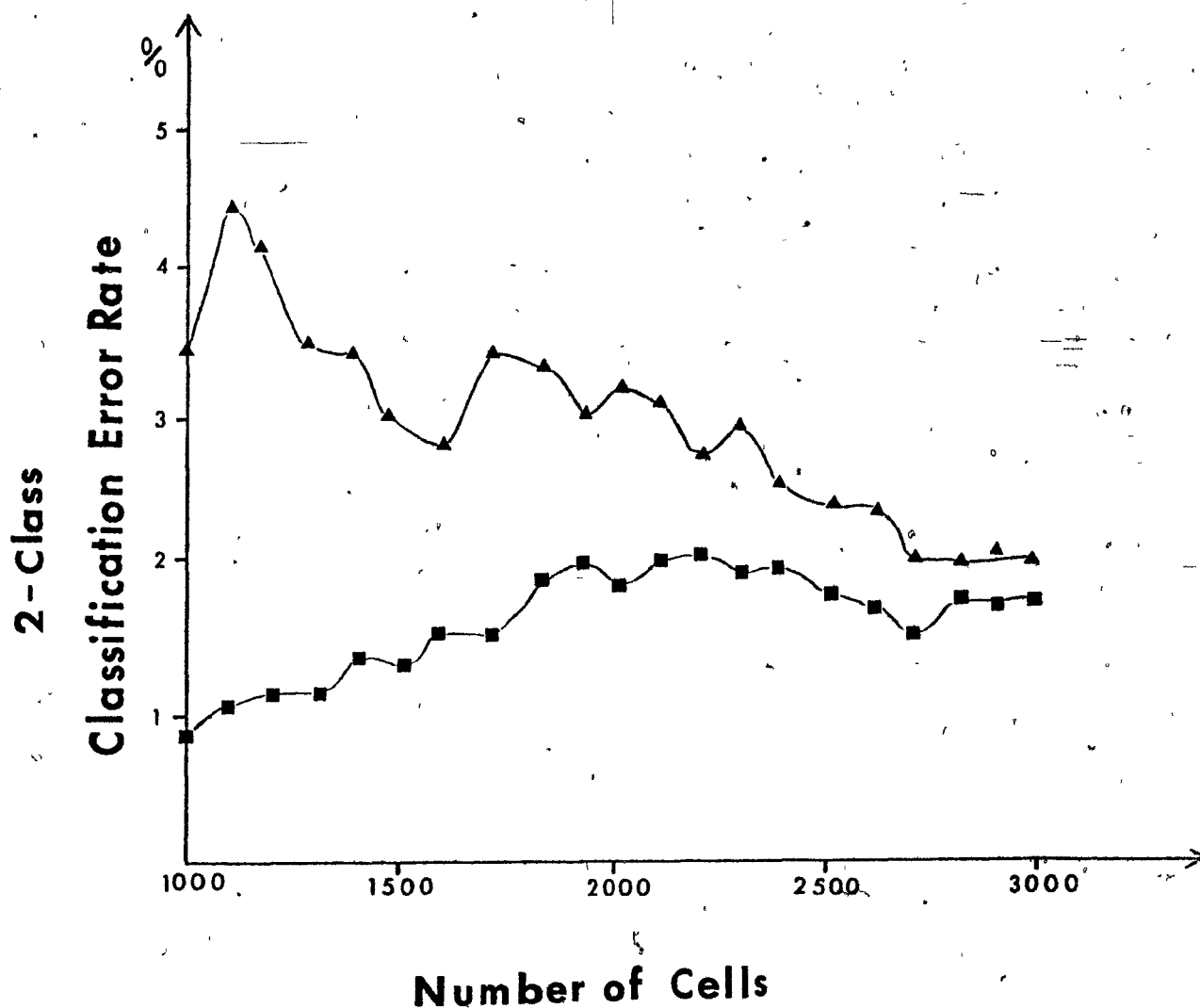


Figure 7

Classification error rates obtained when using the 13 features selected by the forward sequential procedure and by applying --a) the resubstitution method (■) and --b) the random partitioning method (▲) as the functions of number of cells. Note that constant values are reached for both curves when the number of cells used approaches 3000.

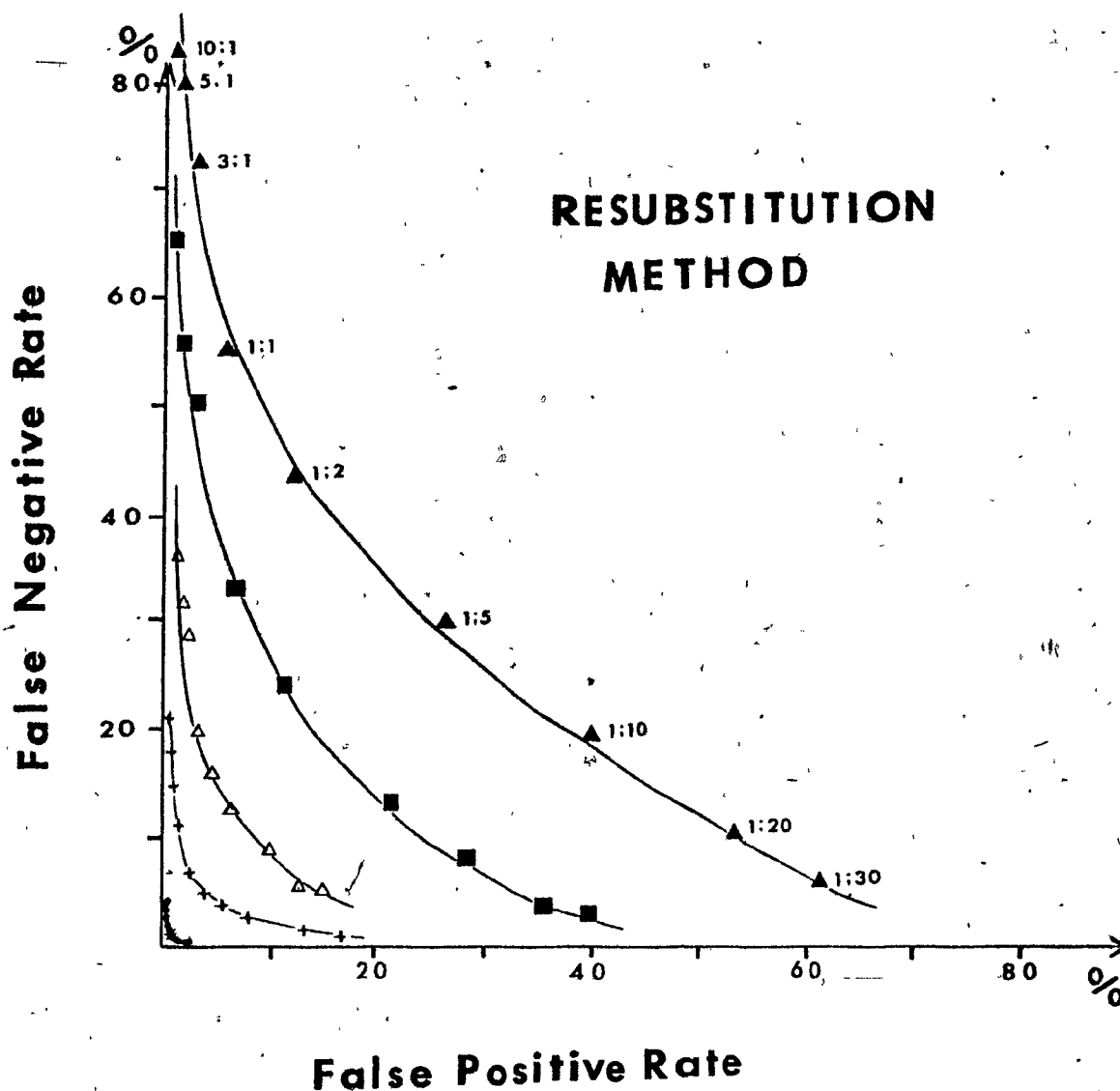


Figure 8

Classification performance curves obtained by applying the resubstitution method when using --a) 13 features selected from the total data set of 209 features (●), --b) 18 two-dimensional histogram features (●), --c) the 6 features previously chosen by Oliver (+), --d) 18 one-dimensional histogram features (Δ), --e) 10 geometric features (■), --f) 6 texture features (▲). Note that the performance curve of the 18 two-dimensional features is nearly on top of that of the best 13 features.

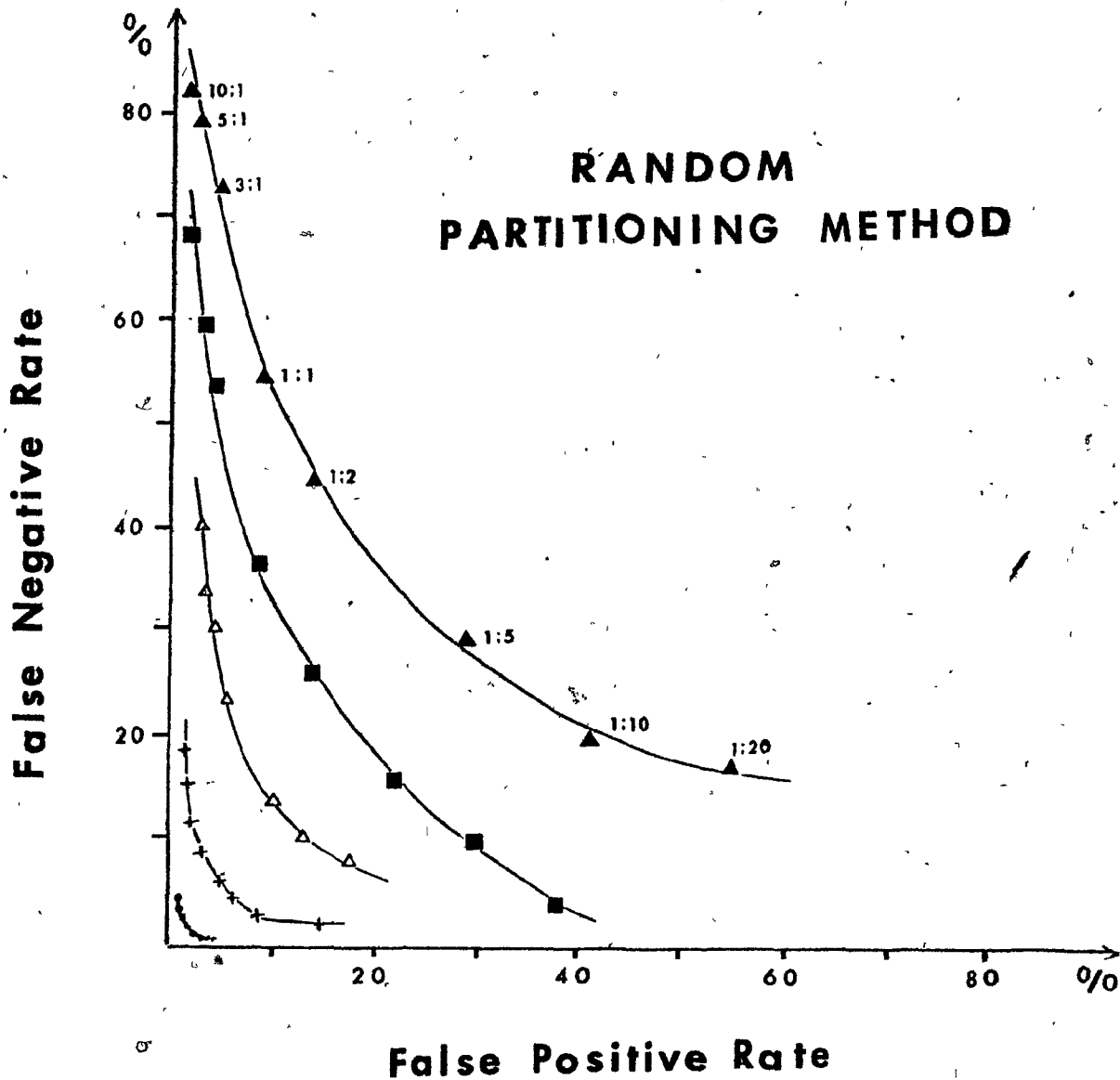


Figure 9

Classification performance curves obtained by applying the random partitioning method (20% holdout over 30 partitions) when using --a) 13 features selected from the total data set of 209 features (•), --b) 18 two-dimensional histogram features (•), --c) the 6 features previously chosen by Oliver (+), --d) 18 one-dimensional histogram features (Δ), --e) 10 geometric features (■), --f) 6 texture features (▲). Note that the performance curve of the 18 two-dimensional histogram features is nearly on top of that of the best 13 features.

PERFORMANCE OF THE BEST SET OF 13 FEATURES

	SSQ	ISQ	NAV	PAR	ENM	ENMG	ENCS	ENCC	HIS	MET	METB	MLD	MOD	SEV	CIS	INV
SSQ	87	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISQ	8	85	4	3	0	0	0	0	0	0	0	1	0	0	0	0
NAV	3	4	85	3	0	0	0	0	0	2	2	0	0	0	0	1
PAR	0	1	3	85	0	0	0	0	2	6	2	1	0	0	0	0
ENM	0	0	0	2	73	7	1	1	11	0	0	0	0	2	4	0
ENMG	0	0	0	1	12	77	0	2	3	3	0	2	0	1	0	0
ENCS	0	0	0	0	1	0	77	10	9	1	1	0	1	1	0	0
ENCC	0	0	0	2	0	0	10	82	3	3	0	0	0	0	0	0
HIS	0	0	0	4	8	3	4	4	69	3	0	0	1	3	1	0
MET	0	0	2	4	0	0	1	2	2	77	10	1	0	0	0	0
METB	1	1	5	1	0	0	0	1	1	13	72	2	1	0	0	0
MLD	0	0	0	0	0	0	0	0	1	3	0	74	17	0	0	3
MOD	0	0	0	0	0	0	0	1	0	0	0	16	55	13	1	13
SEV	0	0	0	0	0	0	0	0	2	1	0	1	15	56	12	13
CIS	0	0	0	0	2	0	1	0	1	0	0	0	2	12	72	9
INV	0	0	0	0	0	0	0	0	0	0	0	2	7	6	10	75

FALSE POSITIVE RATE 1.74%

FALSE NEGATIVE RATE 2.64%

Table 3

Normalized confusion matrix showing classification results obtained by applying the random partitioning method (20% holdout over 30 partitions) and by using the 13 features, selected from the 209 computed features, to classify 3000 cells of 16 cell types scanned at 0.7 micron resolution. Each row indicates the classifier's decisions concerning cells of a particular type (given at left), expressed as a percent of the number of cells of that type. The matrix is partitioned to emphasize the two-class (normal-abnormal) decision, with the false positive errors in the upper right corner and the false negative errors in the lower left.

PERFORMANCE OF THE 18 TWO-DIMENSIONAL HISTOGRAM FEATURES

	SSQ	ISQ	NAV	PAR	ENM	ENMG	ENCS	ENCC	HIS	MET	METB	MLD	MOD	SEV	CIS	INV
SSQ	86	12	1	0	0	0	0	0	0	0	0	1	0	0	0	0
ISQ	6	84	3	6	0	0	0	0	0	0	0	1	0	0	0	0
NAV	2	6	83	2	0	0	0	0	0	2	3	1	0	0	0	1
PAR	0	1	2	85	0	0	0	0	2	5	2	2	0	0	0	0
ENM	0	0	0	0	68	9	1	1	14	1	0	0	0	2	5	0
ENMG	0	0	0	0	10	80	0	2	5	2	0	1	0	1	0	0
ENCS	0	0	0	0	0	0	79	7	9	1	0	1	1	0	0	1
ENCC	0	0	0	3	0	0	8	82	3	2	0	0	0	1	0	0
HIS	0	0	0	3	8	2	4	7	70	3	0	0	0	3	0	0
MET	0	0	0	7	0	0	0	2	2	77	8	1	0	0	0	0
METB	0	1	4	5	0	0	0	0	0	14	70	3	1	1	0	1
MLD	0	2	0	2	0	0	0	0	1	2	0	71	16	1	0	5
MOD	0	0	0	0	0	0	0	0	1	0	1	17	50	17	1	11
SEV	0	0	0	0	0	0	0	2	1	0	0	1	16	53	18	10
CIS	0	0	0	0	1	0	2	0	2	0	0	0	1	13	66	13
INV	0	0	0	0	0	0	0	0	0	0	0	4	5	6	9	76

FALSE POSITIVE RATE 2.27% FALSE NEGATIVE RATE 3.58%

Table 4

Normalized confusion matrix showing classification results obtained by applying the random partitioning method (20% holdout over 30 partitions) and by using the 18 two-dimensional histogram features selected by the forward sequential search procedure to classify 3000 cells of 16 cell types scanned at 0.7 micron resolution. Each row indicates the classifier's decisions concerning cells of a particular type (given at left), expressed as a percent of the number of cells of that type. The matrix is partitioned to emphasize the two-class (normal-abnormal) decision, with the false positive errors in the upper right corner and the false negative errors in the lower left.

PERFORMANCE OF THE 18 ONE-DIMENSIONAL HISTOGRAM FEATURES

	SSQ	ISQ	NAV	PAR	ENM	ENMG	ENCS	ENCC	HIS	MET	METB	MLD	MOD	SEV	CIS	INV
SSQ	80	11	1	0	0	0	0	0	0	0	0	3	1	1	0	1
ISQ	7	67	3	12	1	0	1	1	1	2	0	3	0	0	1	0
NAV	3	5	60	5	0	0	4	1	2	8	4	1	1	3	1	0
PAR	0	8	5	61	2	0	3	2	2	11	4	1	0	0	1	0
ENM	0	0	3	3	48	9	3	6	15	1	2	0	2	2	5	1
ENMG	0	1	0	0	7	51	3	5	18	3	5	1	2	4	1	0
ENCS	1	3	2	2	5	3	62	5	10	1	0	4	1	1	1	0
ENCC	0	2	1	3	5	2	6	65	6	4	2	2	1	0	1	0
HIS	0	4	0	4	16	4	7	7	53	0	0	3	0	0	1	0
MET	0	2	6	8	4	0	1	2	3	58	14	0	0	1	1	0
METB	1	0	11	6	2	3	1	1	1	15	49	0	1	1	6	1
MLD	8	8	4	3	2	0	4	3	8	1	2	29	16	3	4	3
MOD	3	1	2	2	3	0	5	0	3	2	3	14	23	15	15	7
SEV	1	1	2	1	2	1	2	0	3	2	2	6	12	22	31	12
CIS	0	1	2	1	8	1	3	1	4	1	4	2	7	14	38	14
INV	1	0	0	0	0	0	0	0	0	1	1	5	9	8	17	57

FALSE POSITIVE RATE 5.57%

FALSE NEGATIVE RATE 22.03%

Table 5

Normalized confusion matrix showing classification results obtained by applying the random partitioning method (20% holdout over 30 partitions) and by using the 18 one-dimensional histogram features selected by the forward sequential search procedure to classify 3000 cells of 16 cell types scanned at 0.7 micron resolution. Each row indicates the classifier's decisions concerning cells of a particular type (given at left), expressed as a percent of the number of cells of that type. The matrix is partitioned to emphasize the two-class (normal-abnormal) decision, with the false positive errors in the upper right corner and the false negative errors in the lower left.

PERFORMANCE OF THE 10 GEOMETRIC FEATURES

	SSQ	ISQ	NAV	PAR	ENM	ENMG	ENCS	ENCC	HIS	MET	METB	MLD	MOD	SEV	CIS	INV
SSQ	88	8	3	0	0	0	0	0	0	0	0	1	0	0	0	0
ISQ	11	77	6	1	0	0	0	0	0	0	1	4	0	0	0	0
NAV	5	3	73	15	0	0	0	0	0	0	2	2	0	0	0	0
PAR	0	1	14	58	0	0	3	0	2	8	8	1	2	2	1	0
ENM	0	0	0	0	56	26	2	1	4	0	0	0	0	3	6	1
ENMG	0	0	0	0	38	43	0	0	7	0	0	0	0	0	13	0
ENCS	0	0	0	5	0	0	40	39	1	2	2	0	5	3	0	3
ENCC	0	0	0	0	1	0	9	79	3	0	0	0	1	4	0	3
HIS	0	0	0	2	7	4	10	15	36	6	1	0	2	8	6	1
MET	0	0	2	17	0	0	6	1	4	30	14	6	7	10	2	0
METB	0	0	22	12	0	0	5	3	0	15	28	9	4	1	0	1
MLD	0	6	7	7	0	0	1	0	1	9	9	43	14	0	0	3
MOD	0	0	1	6	0	1	7	4	4	10	7	10	31	12	1	5
SEV	0	0	0	8	3	1	10	7	5	5	2	0	5	32	16	6
CIS	0	0	0	1	10	16	1	1	3	0	0	0	0	6	61	2
INV	0	0	0	3	5	3	7	5	3	3	2	4	11	12	25	18

FALSE POSITIVE RATE 8.40% FALSE NEGATIVE RATE 36.24%

Table 6

Normalized confusion matrix showing classification results obtained by applying the random partitioning method (20% holdout over 30 partitions) and by using the 10 geometric features selected by the forward sequential search procedure to classify 3000 cells of 16 cell types scanned at 0.7 micron resolution. Each row indicates the classifier's decisions concerning cells of a particular type (given at left), expressed as a percent of the number of cells of that type. The matrix is partitioned to emphasize the two-class (normal-abnormal) decision, with the false positive errors in the upper right corner and the false negative errors in the lower left.

PERFORMANCE OF THE 6 TEXTURE FEATURES

	SSQ	ISQ	NAV	PAR	ENM	ENMG	ENCS	ENCC	HIS	MET	METB	MLD	MOD	SEV	CIS	INV
SSQ	24	33	4	8	0	0	2	2	1	0	2	2	4	7	1	8
ISQ	2	72	1	15	0	0	4	1	1	0	1	1	1	0	0	0
NAV	7	15	5	17	2	1	6	3	1	7	22	3	2	4	2	4
PAR	0	20	0	53	1	1	16	1	2	0	2	2	0	1	0	0
ENM	0	7	0	2	37	25	5	1	7	4	3	1	0	4	3	2
ENMG	0	0	0	0	19	54	1	9	10	1	3	0	0	3	1	0
ENCS	1	4	0	31	2	5	37	1	12	1	5	0	0	1	0	0
ENCC	0	2	2	7	1	20	29	20	7	1	1	1	1	3	2	2
HIS	0	18	0	27	6	8	25	1	9	0	4	0	0	1	0	1
MET	1	8	1	27	2	8	26	8	2	6	5	0	0	3	1	2
METB	3	5	3	16	0	1	26	4	1	8	23	1	1	3	2	3
MLD	10	16	3	13	4	1	12	7	1	4	12	3	3	5	2	6
MOD	11	7	1	15	0	3	15	3	1	2	7	0	7	13	1	14
SEV	4	0	1	5	3	7	12	1	4	2	10	0	2	24	3	21
CIS	1	0	0	0	16	15	5	1	8	0	1	0	1	22	7	22
INV	10	1	2	0	4	4	3	3	1	0	3	1	4	18	3	42

FALSE POSITIVE RATE 8.54% FALSE NEGATIVE RATE 53.48%

Table 7

Normalized confusion matrix showing classification results obtained by applying the random partitioning method (20% holdout over 30 partitions) and by using the 6 texture features selected by the forward sequential search procedure to classify 3000 cells of 16 cell types scanned at 0.7 micron resolution. Each row indicates the classifier's decisions concerning cells of a particular type (given at left), expressed as a percent of the number of cells of that type. The matrix is partitioned to emphasize the two-class (normal-abnormal) decision, with the false positive errors in the upper right corner and the false negative errors in the lower left.

PERFORMANCE OF THE 6 FEATURES PREVIOUSLY USED BY OLIVER [01178a]

	SSQ	ISQ	NAV	PAR	ENM	ENIG	ENCS	ENCC	HIS	MET	METB	MLD	MOD	SEV	CIS	INV
SSQ	84	12	3	0	0	0	0	0	0	0	0	1	0	0	0	0
ISQ	12	82	2	2	0	0	0	0	0	0	1	0	0	0	0	0
NAV	7	6	79	2	0	0	0	0	0	2	4	0	0	0	0	0
PAR	0	1	3	82	0	0	4	0	3	4	2	0	0	0	1	0
ENM	0	0	0	0	64	19	0	1	11	0	0	0	0	0	5	1
ENIG	0	0	0	0	22	62	0	3	5	0	0	0	0	1	8	0
ENCS	0	0	0	2	1	0	52	21	17	3	1	0	1	1	1	0
ENCC	0	0	0	0	0	0	15	68	9	1	2	0	0	3	1	1
HIS	0	0	0	0	11	5	4	8	67	3	1	0	0	0	2	0
MET	0	0	4	11	0	0	1	3	2	49	21	3	2	2	0	1
METB	0	0	14	7	0	0	0	1	0	15	53	5	2	0	1	0
MLD	0	1	1	1	0	0	0	0	0	2	6	71	14	1	0	3
MOD	0	0	1	0	0	0	1	0	0	2	2	13	50	19	1	10
SEV	0	0	0	0	0	1	1	0	2	1	1	0	10	55	15	13
CIS	0	0	0	0	2	10	1	1	0	0	0	0	0	12	68	6
INV	0	0	0	0	0	2	0	0	0	1	0	2	9	16	19	51

FALSE POSITIVE RATE 2.86%

FALSE NEGATIVE RATE 7.61%

Table 8

Normalized confusion matrix showing classification results obtained by applying the random partitioning method (20% holdout over 30 partitions) and by using the 6 features, previously used by Oliver [01178b], to classify 3000 cells of 16 cell types scanned at 0.7 micron resolution. Each row indicates the classifier's decisions concerning cells of a particular type (given at left), expressed as a percent of the number of cells of that type. The matrix is partitioned to emphasize the two-class (normal-abnormal) decision, with the false positive errors in the upper right corner and the false negative errors in the lower left.

5.3 Experimental results

As mentioned in chapter 1, 3000 cells of 16 classes were used for testing. This relatively large number was required to obtain meaningful classification error estimates. To verify this requirement, the best 13 features selected by the forward sequential selection, the random partitioning method, and the resubstitution method were applied to obtain two classification error rate curves as the number of cells varied from 1000 to 3000 (See Figure 7). It is apparent that, as the size of the cell data set approaches 3000, the random partitioning test results and the resubstitution test results become relatively stable: both curves approach constant error rates (about 2.2% for the random partitioning method and 1.7% for the resubstitution method).

From the classification performance curves in Figures 8 and 9 (for the resubstitution method and the random partitioning method respectively) and from confusion matrices resulting from applying the random partitioning method (20% holdout over 30 partitions) to different selected subsets of features, the following results were observed:

5.3.1 Two-dimensional histogram features

The newly developed two-dimensional histogram features described in chapter 3 which contain both density and color informations proved to be very useful features for cervical cell classification. By applying the forward sequential search procedure to the two-dimensional histogram features, the following 18 features were selected:

F7	F9	F13	F15	F28	F36	F38	F39	F40
F50	F55	F60	F67	F75	F79	F81	F87	F88

Where F_n is the feature numbered n and can be found in the definition and listing of the 209 features in chapter 3 (Feature computation).

The performance of the two-dimensional histogram features is slightly worse than the best set of 13 features selected from the 209 features: With the a priori probability ratio setting of 1:1 -- a) The classification error rate of 1.98% was obtained for the 18 two-dimensional histogram features versus 1.69% for the 13 features when the resubstitution method was applied. -- b) a classification error rate of 2.92% versus 2.19% was obtained when the random partitioning method was applied. -- c) As can be seen in Tables 3 and 4, the individual-subclass error rates obtained using the 13 features selected from the 209 features were either the same or only slightly better than the error rates obtained by using the 18 selected two-dimensional histogram features. Moreover, for endometrial (glandular) cells and histiocytes, the error rates obtained by the two-dimensional histogram features were even better than the ones obtained by the 13 features (2% versus 3% for glandular endometrial cells and 3% versus 5% for histiocytes).

The performance of the 18 two-dimensional histogram features is much better than the performance of the best set of features selected from each of the other categories: -- a) The features from the one-dimensional histogram category (classification error rates of 11.49% and 13.8% for the resubstitution and the random partitioning methods respectively), the features from the geometric category (19.67% and 22.32% respectively), the features

from the texture category (30.82% and 31.01% respectively). -- b) As can be seen in Tables 4, 5, 6, and 7, the individual-class error rates obtained using the two-dimensional histogram features were significantly lower than those obtained using one-dimensional histogram, geometric, or texture-features. The only exceptions were the texture feature error rates obtained for endocervical cells (secretory) and histiocytes (3% versus 1% for secretory endocervical cells and 3% versus 2% for histiocytes).

The performance of the 18 two-dimensional histogram features is also better than that of the 6 features previously used by Oliver [01178b], which include geometric, density, and texture features and yielded classification error rates of 4.69% and 5.2% for the resubstitution and the random partitioning methods respectively. As can be seen in Tables 4 and 8, the error rates obtained by using the two-dimensional histogram features were much better than the 6 features (The error rates for the two-dimensional histogram features were slightly worse than for the 6 features only for 6 cell subclasses: intermediate squamous, navicular squamous, parabasal squamous, endometrial (stromal), endometrial (glandular), and histiocytes. For the other 10 subclasses, the error rates for the two-dimensional histogram were much lower than those for the 6 features).

The fact that the 18 two-dimensional histogram features performed the best among all feature categories is consistent with the large proportion of the two-dimensional histogram features (9 features) contained in the set of 13 features selected from the total set of 209 features from all categories.

5.3.2 One-dimensional histogram features

The one-dimensional histogram features described in chapter 3 include density features (when only one of the one-dimensional histograms is considered) and color features (when the other two of the one-dimensional histograms for the two other color images are also considered). When applying the forward sequential feature selection method to the one-dimensional histogram features, the following 18 features were selected:

F100 F102 F103 F104 F110 F112 F115 F117 F119
F120 F122 F128 F129 F130 F133 F136 F142 F143

where F_n is as given in chapter 3.

Even though these features include both density and color information, they are much less powerful than the 18 two-dimensional histogram features: Classification error rates of 11.49% versus 1.98%, and 13.8% versus 2.92%, were obtained for the resubstitution and the random partitioning methods respectively. However, the one-dimensional histogram features gave much better results than geometric features (19.67% for the resubstitution method and 22.32% for the random partitioning method) and texture features (30.82% for the resubstitution method and 31.01% for the random partitioning method).

5.3.3 Geometric features:

The geometric features include size and shape features. When applying the forward sequential feature selection method to the geometric features, the following 10 were selected:

F151	F152	F162	F164	F169
F171	F175	F176	F177	F178

Where F_n is as described in chapter 3.

The performance of the 10 geometric features (classification error rates of 19.67% and 22.32% for the resubstitution and the random partitioning methods respectively) is much worse than the performance of the 18 two-dimensional histogram features (1.98% and 2.92% respectively) or the performance of the 18 one-dimensional histogram features (11.49% and 13.80% respectively). However, it is better than the performance obtained for texture features (30.82% and 31.01% respectively).

5.3.4 Texture features:

When applying the forward sequential feature selection method to the texture features, the following 6 features were selected:

F192	F195	F198	F204	F207	F208
------	------	------	------	------	------

Where F_n is as described in chapter 3.

The performance of the 6 texture features turned out to be the worst of all categories (classification error rates of 30.82% and 31.01% for the resubstitution and the random partition methods respectively). This is consistent with the fact that no texture features were contained in the set of 13 features selected from the total set of 209 features from all categories.

5.4 Conclusions:

When each individual feature category was considered, the two-dimensional histogram features significantly outperformed any other feature categories (one-dimensional histogram, geometric, and texture features). Also, the performance of the two-dimensional histogram features was only slightly worse than the performance of the 13 features selected from all feature categories (2.92% versus 2.19%) and significantly better than the 6 features previously used by Oliver [Oli78b] (2.92% versus 5.2%). This indicates that the density and color features in general and the two-dimensional histogram features in particular are of prime importance in the cervical cell recognition problem. The results are also consistent with the fact that a large proportion of the two-dimensional histogram features was selected when applying the forward sequential search procedure to the total set 209 features (out of 13 features selected, 9 were two-dimensional histogram features).

The features extracted from the one-dimensional histograms of images scanned at three different wavelengths, although they contain density and color information, performed significantly worse than the two-dimensional histogram features in terms of classification error rates. This is also consistent with the fact that out of the 10 density and color features selected to form the final set of 13 features, only 1 of the one-dimensional histogram features was selected. The remaining 9 features were two-dimensional histogram features.

The geometric features, even though recognized as very important

features, performed much worse than the two-dimensional and even the one-dimensional histogram features (22.32% in comparison with 2.92% and 13.8% for two- and one-dimensional histogram features respectively). This is consistent with the fact that only 3 out of the final set of 13 features selected by the forward sequential search procedure were geometric features.

The texture features performed the worst in comparison with other categories. This is probably why not a single texture feature was selected in the final set of 13 features. This is a fortunate sign for cervical cell recognition because texture features are usually more expensive to compute than other feature categories including two-dimensional histogram and shape features.

We noted in particular that the power of the two-dimensional histogram features was most evident when features from all three combinations of 2 color images were used. The observed necessity to include features from all three combinations is probably due to the fact that the individual two-dimensional matrices were made symmetric to facilitate feature computation.

The classification performance estimates made in this study assume that all cell types occur in equal fractions -- normal and abnormal cells alike. In actual fact, this is far from reality and in general the normal squamous cell types (superficial, intermediate, navicular, and parabasal) are present in larger numbers than the remaining cell types. Because most of the errors do not involve the normal squamous cell types, the performance values cited here are quite pessimistic. If it is assumed that the normal squamous cell types are ten times more abundant than the remaining cell types, the

classification performance curve for the 13 selected features is altered as shown in Figure 10. This curve is more realistic for comparison purposes.

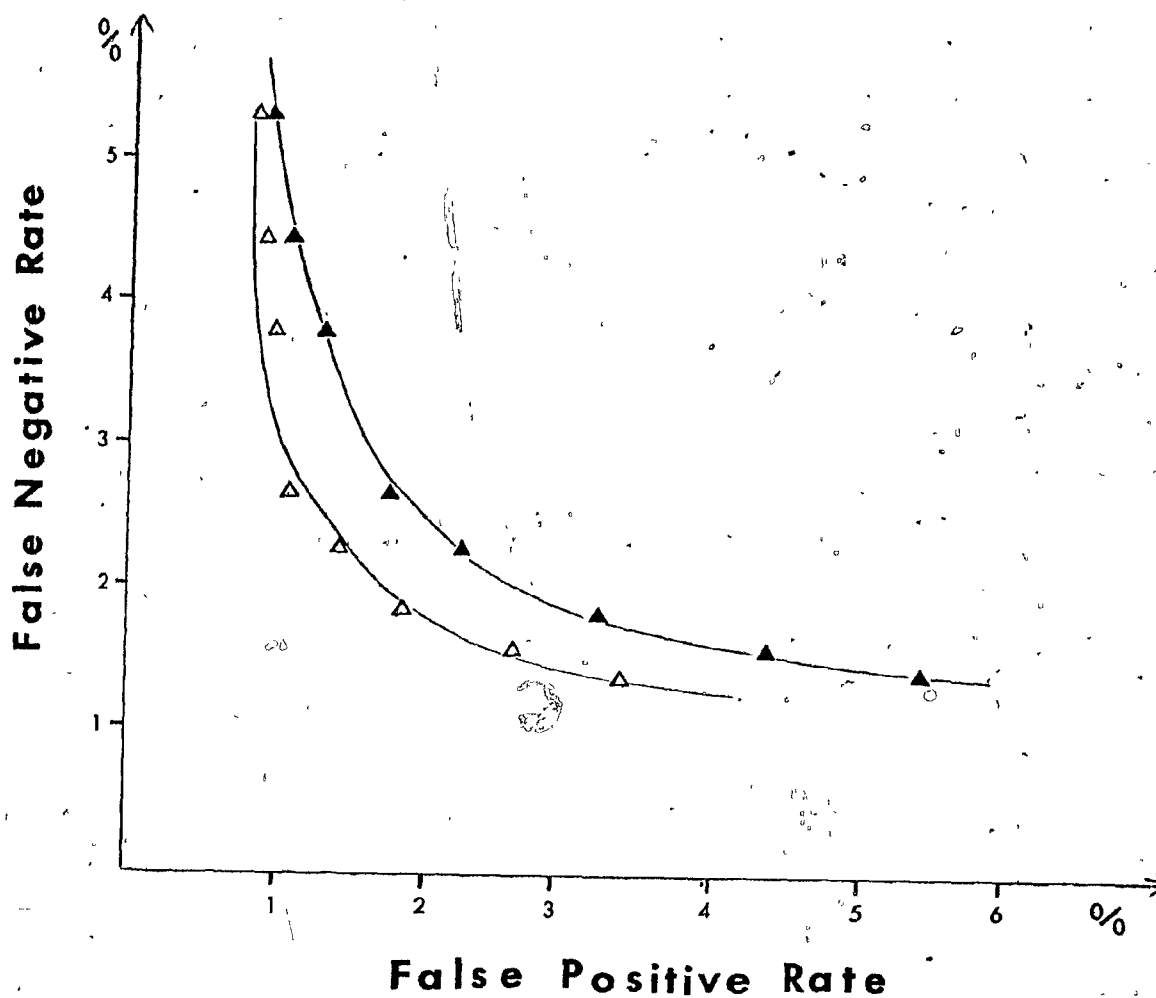


Figure 10

Classification performance curves obtained by applying the random partitioning method when all cell types occur in equal fractions (▲) and when the normal squamous cell types (superficial, intermediate, navicular, and parabasal) are ten times more abundant than the remaining cell types (△).

Chapter 6

SUMMARY AND CONCLUSIONS

6.1 Summary

A significant reduction in classification error rate was obtained in comparison with previous results[Cah77a,01i78b] due to both the improvement in the scene segmentation and the feature extraction. For scene segmentation, the threshold selection technique based on the stability of area was applied to the color images of cells (to segment cells from the background, the images of cells scanned at 530 nm wavelength were used, and to separate nuclei from cytoplasm, the images of cells scanned at 570 nm wavelength were used). In addition, for the problem of overlapping cells, algorithms for artificially generating a set of overlapping cells with a uniformly distributed overlap degree, and for overlapping-cell detection were devised and evaluated. These techniques can be applied to select single cells for further analysis.

For the feature extraction process, the system developed previously at the BIPLAB by Cahn et al[Cah78b] was expanded to include new two-dimensional histogram features and Fourier and Granlund shape features. For feature selection, the probability of misclassification criterion and three feature search procedures were applied to the 209 features computed by the new system. The best subset of features obtained so far is the set of 13 features obtained by applying the forward sequential search procedure to the total set of 209 features. Finally, the feature categories were evaluated by comparing their corresponding performances when each category was considered alone.. The

performance of the 13 features selected from the 209 features of all categories and the 6 features previously used by Oliver [Oli78b] were also included for comparison purposes.

6.2 Conclusions and suggestions for future studies

6.2.1 On scene segmentation

The use of two-color images of cells (530nm, 570nm) instead of single-color images produced significant improvements in scene segmentation. Thus, research should be continued to make full use of multi-color images to improve the scene segmentation even further.

The segmentation method which selects the density threshold based on the stability of areas of segmented regions produced very good results. Therefore, it is worthwhile to investigate a generalized version of the method which considers the stability of combinations of a variety of important features such as area, perimeter, gradient, etc.

The segmentation error measurement method based on the percentage of misclassified pixels is inexpensive, easy to compute, and fairly effective because of the fact that areas of segmented regions are very important features for cervical cell recognition. However, a generalized segmentation error measurement method should be developed to take into account several important features at the same time.

The overlapping-cell generation algorithm is very useful in producing a cervical cell data base with cells of uniformly distributed

overlap degree and different cell types. This kind of data base can be used to evaluate any overlapping-cell detection algorithm objectively.

The overlapping-cell detection algorithm using the Fourier shape descriptors, and the density information, produced very encouraging results especially for cells with relatively small overlap. Also, the analytic derivation of tangent and curvature of the boundary points in terms of Fourier descriptors can be applied to determine the maximal concavity points on the smoothed boundary. The relative positions of these maximal concavity points and their curvature values should be statistically analyzed in future research to derive a decision rule for finding overlapping cells more effectively.

6.2.2 On feature extraction and feature selection

When considering each feature category separately, the new two-dimensional histogram features significantly outperformed any other feature categories and even the 6 features previously used [Oli78b] (which were selected from all feature categories except the two-dimensional histogram feature category). Also, the two-dimensional histogram features performed only slightly worse than the 13 features selected from the 209 features composed of all feature categories. The one-dimensional histogram features performed worse than the two-dimensional histogram features. The geometric features performed even worse than the one-dimensional histogram features and the texture features performed the worst of all.

Due to the high discriminating power of features from all three

combinations of two-dimensional histograms, it would be productive to further investigate methods to compute these types of density and color features directly from the multi-dimensional histograms instead of using all combinations of two-dimensional histograms. Moreover, other density and color features such as density texture and color texture features should also be considered for future use. An example of one such density/color feature follows: First, apply a whitening transformation to the original images scanned at two different wavelengths to obtain separate "color" and "density" images. One method for doing this has been described by Bacus [Bac76a]. Then, apply the procedure described by Haralick et al [Har73a] to the "density" and "color" images to compute their corresponding co-occurrence matrices and density and color texture features respectively from the co-occurrence matrices.

The forward sequential search procedure proved to perform better than the parallel search procedure and the feature clustering procedure. However, whenever computer time restriction is not so severe, a more generalized sequential search procedure ($i, j \neq 0$) should be applied to avoid the without-replacement property of the forward sequential search procedure.

The research on new features, particularly color features described herein, has led to a considerable improvement in classification of cells from cervical smears than previously reported by our group. It would be particularly beneficial test of these new features if other groups, doing research on cervical cell recognition and other problem areas, would evaluate

them using their data and other types of classifiers.

The classification error rates obtained in this study are in reasonably close agreement with the recent result reported by Lin et al [Lin80a)83a] (0.8% false positive and 2.6% false negative on 1153 cells) using a binary tree classifier and the resubstitution testing. In our research, when 1100 or 1200 cells were used, 1.4% false positive and 0.75% false negative error rates were obtained using the resubstitution method. When 3000 cells were used, 1.15% false positive and 2.21% false negative error rates were obtained using the resubstitution method, and 1.7% false positive and 2.6% false negative error rates were obtained using the random partitioning method.

REFERENCES

- Agg77a R.K. Aggarwal, J.W Bacus "A multi-spectral approach for Scene Analysis of Cervical Cytology Smears", J. Histochem Cytochem Vol 25, No.7, pp 668-680, 1977.
- Aus77a H.H. Aus, A. Ruter, V. Termeulen, U. Gunzer, and R. Rurnberger, "Bone marrow cell scene segmentation by computer-aided color cytophotometry", J. Histochem Cytochem, Vol.25, No.7, pp 662-667 1977.
- Bac76a J.W Bacus, "A Whitening Transformation for Two-Color Blood Cell Images", Pattern Recognition, Pergamon Press, Vol 8, pp 53-60, 1976.
- Ben79a E. Bengtsson, O. Ericksson, J. Holmquist, B. Nordin, B. Stenkvis
"High resolution segmentation of Cervical Cells", J. Histochem Cytochem, Vol 27, No.1, pp 621-628, 1979.
- Ben81a E. Bengtsson, O. Ericksson, J. Holmquist, T. Järkrans, B. Nordin, and B. Stenkvis, "Segmentation of Cervical Cell: Detection of Overlapping Cell Nuclei", Computer Graphics and Image Processing 16, pp 382-394, 1981.
- Cah77a R.L Cahn "Feature Extraction and Evaluation for Cervical Cell Recognition", M.Sc Thesis, McGill University, August 1977.
- Cah77b R.L. Cahn, R.S. Poulsen, and G.T. Toussaint, "Segmentation of cervical cell images", J. Histochem Cytochem, Vol.25, No.7, pp 681-688, 1977.
- Cha73a Chieng-Yi Chang, "Dynamic Programming as Applied to Feature Subset Selection in a Pattern Recognition System", IEEE trans. Systems Man Cybernet, Vol. SMC-3, No.2, pp 166-171, 1973.

- Che80a C.J. Chen, Q.Y. Shi, "Shape Features for Cancer Cell Recognition",
In Proc of the 5th International Conference on Pattern Recognition,
Miami Beach, pp 579-581, 1980.
- Cov74a T.M. Cover, "The best two Independent Measurements are not the two
best", IEEE Trans. Systems Man Cybernet, Vol SMC-4, pp 116-217,
1974.
- Cov77a T.M. Cover, J.M.V. Campenhout, "On the Possible Orderings in the
Measurement Selection Problem", IEEE Trans. Systems Man Cybernet,
Vol SMC-7, No.9, pp 657-661, 1977.
- Dav75a L.S. Davis, "A survey of edge detection techniques", Computer
Graphics and Image Processing 4, pp 248-270, 1975.
- Dud73a R.O. Duda and P.E. Hart : Pattern classification and scene
analysis, Wiley, New York, 1973.
- Ecc77a M.J. Eccles, H.P.C. McQueen, and D. Rosen, "Analysis of the Digitized
Boundaries of Planar Objects", J. Pattern Recognition, Vol.9, pp
31-41, 1977.
- Gra72a G. Granlund, "Fourier Preprocessing for Hand Print Character
Recognition", IEEE Trans on Computers, pp 195-201, 1972.
- Gre79a J.E. Green, "Rapid analysis of hematology image data. The ADC-500
preprocessor", J. Histochem Cytochem, Vol.27, No.1, pp 174-179,
1979.
- Har73a R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features
for Image Classification", IEEE Trans. Systems Man Cybernet.
SMC-3, pp 610-621, 1973.
- Hol76a J. Holmquist, Y. Inasato, E. Bengtsson, B. Olsen, B. Stenkvist,

- "A Microspectro Photometric Study of Papanicolaou-Stained Cervical Cells as an Aid in Computerized Image Processing", J. Histochem Cytochem, Vol.24, No.12, pp 1218-1224, 1976.
- Hol78a J. Holmquist, E. Bengtsson, O. Ericksson, B. Nordin, B. Stenkvis, "Computer Analysis of Cervical Cells: Automatic Feature Extraction and Classification", J. Histochem Cytochem, Vol.26, No.11, pp 1000-1017, 1978.
- Jai80a A.K. Jain, S.P. Smith, E. Backer, "Segmentation of Muscle Cell Pictures: A Preliminary Study", IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. PAMI 1-2, No.3, pp 232-242, 1980.
- Kan74a L. Kanal, "Patterns in Pattern Recognition: 1968-1974", IEEE Trans on Information Theory, Vol. IT-20, No.6, pp 697-722, 1974.
- Kul79a A.V. Kulkarni, "Effectiveness of Feature Groups for Automated pairwise Leukocyte class discrimination", J. Histochem Cytochem, Vol.27, No.1, pp 210-216, 1979.
- Lin80a Y.K. Lin, K.S. Fu, "An Application of Pattern Recognition Techniques to Pap Smear Inspection", Ph.D Thesis, School of Electrical Engineering, Purdue University, West Lafayette, Indiana, October 1980.
- Lin81a Y.K. Lin, K.S. Fu, "Segmentation of Papanicolaou Smear Images", Analytical and Quantitative Cytology, Vol.3, No.3, pp 201-206, 1981.
- Lin83a Y.K. Lin, K.S. Fu, "Automatic Classification of Cervical Cells using a Binary Tree Classifier", J. Pattern Recognition, Vol.16, No.1, pp 69-80, 1983.
- Lou77a C. Louis, "An Investigation of Staining Techniques for Use in

Automated Cervical Cytology Screening", Cand. J. Med. Tech. 39:1977.

- Muc71a A.N. Mucciardi, E. Gose, "A comparison of Seven Techniques for choosing Subsets of Pattern Recognition Properties", IEEE Trans on Computers, Vol. C-20, No.9, pp 1023-1031, 1971.
- Oli77a L.H. Oliver, R.S. Poulsen, and G.T. Toussaint, "Estimating false positive and false negative error rates in cervical cell classification", J. Histochem Cytochem, Vol.25, No.7, pp 696-701, 1977.
- Oli78a L.H. Oliver, R.S. Poulsen, and G.T. Toussaint, "Classification of atypical cells in the automatic cytoscreening for cervical cancer", Proc. IEEE Computer Society Conference on Pattern Recognition and Image Processing, 1978, Chicago, Illinois.
- Oli78b L.H. Oliver, "Automatic Image Processing and Pattern Recognition for Biomedical Research", Ph.D Thesis, Dept. of Computer Science, McGill University, Canada, 1978.
- Pou77a R.S. Poulsen, L.H. Oliver, R.L. Cahn, C. Louis, and G. Toussaint, "High Resolution Analysis of Cervical Cells -- A Progress Report", J. Histochem Cytochem, Vol.25, No.7, pp 689-695, 1977.
- Pou78a R.S. Poulsen, L.H. Oliver, "IPS - A modular Interactive Software System for Image Processing and Pattern Recognition Research", Proc. CIPS SESSION'78 Canadian Computer Conference Edmonton, Alberta. May 23,24,25, 1978.
- Pou81a R.S. Poulsen, L.H. Oliver, G.T. Toussaint, C. Louis, "Evaluation of Single-cell Classification Schemes for Computer Classification of Cervical Cells", Analytical and Quantitative Cytology, Vol.3,

No.3, pp 207-215, 1981.

- Ris77a E.M. Riseman and M.A. Arbib, "Computational techniques in the visual segmentation of static scene", Computer Graphics and Image Processing 6, pp 221-276, 1977.
- Rob82a S.J. Roberts and R. Hanka, "An interpretation of Mahanalobis distance in the dual space", J. Pattern Recognition, Vol.15, No.4, pp 325-333, 1982.
- Ros76a A. Rosenfeld, R.A. Hummel, and S.W. Zucker, "Scene labelling by relaxation operations", IEEE trans. Syst., Man, Cybern. 6, pp 420-433, 1976.
- Ros76b A. Rosenfeld, A.C. Kak, Digital Picture Processing, Academic Press, New York - San Francisco - London, 1976.
- Ste76a S.D. Stearns, "On selecting features for pattern classifiers", In Proc. Third Int. Joint Conf. Pattern Recognition, Coronado, CA, Nov. 1976, IEEE Computer Society, pp 71-75.
- Syc78a J.J. Sychra, P.H. Bartels, M. Bibbo, J. Taylor, and G.L. Wied, "Computer Recognition of Binucleation with Overlapping in Epithelial Cells", Acta Cyto. 22, pp 22-28, 1978.
- Tay78a J. Taylor, J. Puls, J.J. Sychra, P. Bartels, M. Bibbo, G.L. Wied, "A System for Scanning Biological Cells in Three Colors", Acta Cytologica, Vol.22, No.1, pp 29-35, 1978.
- Tou71a G.T. Toussaint, "Note on optimal selection of independent binary-valued features for pattern recognition", IEEE Trans. Inform. Theory (Corresp.), Vol.IT-17, pp 618, 1971.
- Tou71b G.T. Toussaint, "Comments on 'A modified figure of merit for

feature selection in pattern recognition'", IEEE Trans. Inform.

* Theory (Corresp.), Vol.IT-17, pp 618-620, 1971.

Tou74a G.T. Toussaint, "Bibliography on Estimation of Misclassification",
IEEE Trans on Information Theory, Vol. IT-20, No.4, pp 472-479, 1974.

Tou79a G.T. Toussaint, R.S. Poulsen, "Some New Algorithms and Software
Implementation Methods for Pattern Recognition Research", Proc.
IEEE Computer Software and Applications Conference, Chicago,
pp 55-63, 1979.

Tuc78a J.H. Tucker, P. Eason, and M. Stark, "Ellipse test for the
Reduction of False-Positive Signals in Automated Cytology", Acta
Cytol. 22, pp 370-376, 1978.

Yas77a W.A. Yasnoff, J.K. Mui, and J.W. Bacus, "Error measures for scene
segmentation", J. Pattern Recognition, Vol.9, pp 217-231, 1977.

You74a T.Y. Young, T.W. Calvert : Classification, Estimation and Pattern
Recognition, American Elsevier Publishing Co., Inc., 1974.