Quality Assurance of Cervical Smear Slide Inspection Using a Novel Eye-Tracking Technique

By

Lee Roy Coombes

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

Doctor of Philosophy

School of Computing, Communications & Electronics
Faculty of Technology

SEPTEMBER 2004

Quality Assurance of Cervical Smear Slide Inspection Using a Novel Eye-Tracking Technique

Lee Roy Coombes

Abstract

A novel objective quality assurance system for smear slide screening is investigated in this thesis. A method of data validation was developed that compares data from an eye tracked image display, machine image colour texture analysis and expert judgements in a statistical manner to identify salient areas of cervical cytological images. These data are used to construct screener performance profiles, which have been compared to screener experience. The experimental methodology is described and how the screener performance profile is constructed. Results from a study of 10 screeners, checkers and pathologists are presented showing predicted trends of human performance. Relations to experience and strategy are also shown, though these relationships are not statistically significant. A standardised quality assurance test is developed that profiles screeners across many performance measures. Highly significant correlations were found between fixation saliency and machine colour texture (maxima density), though fixation saliency suffers from a lack of a significant statistical basis. Further fixation data is needed, however if it conforms to the existing trends then the results would support the new data validation method as a framework from which image analysis techniques applied to cytology may be objectively tested. Furthermore, this new approach to cervical cytology quality assurance would have the potential to further reduce human errors in the cervical smear inspection process by lowering levels of observer variation found in all aspects of the cervical screening process.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Thanks to:

Dr Phil Culverhouse for the supervision, support, advice and the programming help. Dr Ade Oriolowo and all the staff at the Histopathology Dept at Derriford. Dr Debbie Lee and Mr Mike Rowell for help with the image sets and far more besides. Mr Richard Winder for the encouragement and support and the EPSRC for funding this research

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was financed with the aid of a studentship form the Engineering and Physical Sciences Research Council

The word count for this thesis is 38, 086

Publications:

Coombes, L. (2002), Improving the Quality Control of Human Expert Cytological Slide Inspection, **Proceedings of the British Psychological Society**, 10 (2). 59 (Abstract Only)

Coombes, L.R. and Culverhouse, P.F. (2003) Pattern Recognition in Cervical Cytology. **Proceedings of 5th International Conference on Advances in Pattern Recognition**, Kolkata, India, 227-230.

Poster presentations:

Coombes, L. (2001) Improving the Quality Control of Human Expert Cytological Slide Inspection through the Application of Advanced Image Analysis and Pattern Recognition Methods. **SET for Britain/Young Engineers Awards**, House of Commons, LONDON. Monday 3rd December.

Coombes, L. (2002) Improving the Quality Control of Human Expert Cytological Slide Inspection. **British Psychology Society Annual Conference**, Hilton and Imperial Hotels, BLACKPOOL. 13th-16th March.

Paper Presentations:

Coombes L.R. (2003) Improving the Quality of Human Expert Cytological Slide Inspection. **Postgraduate Research Conference in Electronics, Photonics, Communications and Software (PREP)**, Exeter University, 14th -16th April.

Coombes, L.R. and Culverhouse, P.F. (2003) Pattern Recognition in Cervical Cytology. **5th International Conference on Advances in Pattern Recognition (ICAPR), Indian Statistical Institute**, Kolkata, India. 10th – 13th December.

External Contacts:

Dr. Ade Oriolowo, Consultant Histopathologist, Histopathology Department,
Derriford Hospital, PLYMOUTH. PL6 8DH

Mr Mike Rowell, South & West Regional Cervical Screening Quality Assurance
Reference Centre, Southmead Hospital, BRISTOL. BS10 5NB

Mr Richard Winder, Deputy National Coordinator, NHS Cancer Screening
Programmes, The Manor House, 260 Ecclesall Road South, SHEFFIELD.
S11 9PS

Signed ....

Date 6/6/06

Chapter 1

Introduction

1 - Introduction

1.1 Introduction

Judging the contribution that cancer research makes to reduce mortality rates is difficult. As with the majority of scientific research, there are many factors that need to be carefully considered, and their effects accounted for, before any causal relationship can be established. In recent years, the importance of public awareness has been recognized, organisation of screening programmes has improved, and diagnosis techniques that are constantly evolving have all undoubtedly saved lives. The Office of National Statistics (1998) report that incidence of the two types of cancer with the largest and most organised screening processes are in decline, with breast cancer rates falling by nine percent and cervical cancer falling by twenty-six percent in England and Wales over a period of five years. The cervical cancer rate had the most marked decrease in incidence out of all cancers, and it is hard to imagine that this is for any other reason than the improved screening of cervical smears, which leads to an earlier diagnosis at a pre-cancerous and treatable stage. This is consistent with the significant drops in related mortality rates seen throughout Europe and North America where these organised screening programs have operated in some countries now for over 30 years (Austoker and McPherson, 1992). While there are some commonly acknowledged factors that can contribute to the development of cervical cancer, such as a high number of sexual partners and smoking, primary prevention of this type of cancer is not yet possible (National Heath Service Cancer Screening Programme, n.d). This means that other methods of reducing the incidence rates have to be examined, as it remains the second most common female cancer in thee world today.

1.2 Project Outline

The current project is relatively simple to outline yet there is a great deal of depth regarding the actual research involved. The project seeks to improve the quality of cervical smear slide inspection by using a novel approach to the problems faced. There have been numerous approaches to try and automate the inspection process but the nature of these systems rules out their general use either because they're too costly (both in money and training time), or simply don't achieve an acceptable standard. Our approach to quality assurance is to look at the screener rather than the slide.

When a slide is scanned the viewer will be basing their strategy on explicit rules taught to them through their training and implicit rules which have been learned over a period of time through experience gained from colleagues, improvements in the service and physically undertaking the task of judging slides on a daily basis. This means it is very difficult to come up with a solid set of rules by which a computer may be programmed to automate the task. An alternative approach would be to inspect the features that screeners view in order to make their judgements, as these are salient to the slide classification process and, through statistical means, can be isolated. By comparing the eye fixations made by a screener against features selected by a computer image analysis it should be possible to judge whether the screener has been viewing the most salient features on each slide in order to reach a decision. Later, this thesis will discuss how this can improve the quality assurance by focussing training on those who need it the most.

The question remains as to how a computer can find features if a screener cannot totally externalise the rule system by which they work in a useful way. Perhaps the best way of addressing this issue is by utilising eye tracking technology. This allows the task to be carried out by a screener using an individual internalised rule system and can provide data on how the cytological material is viewed. When a screener examines this material while wearing the eye tracker helmet a recording of the places they view on the image, prior to the decision they make regarding its diagnosis, is produced. The data on where the eye fixates is then compared an analysis of the hue, saturation and value (HSV) components and across different resolutions of the image. The machine analysis produces a list of interesting features across the multiresolutional HSV images. Aspects of the images can then be analysed to produce a list of interesting features that are closely matched to the features examined by the screener during their diagnosis. This will eventually lead to the development of an effective computer system capable of picking out relevant features based on the implicit judgements of the cytological screeners. Furthermore, by using this to examine how the screeners view slides rather than to make a judgement on the slide, some of the ethical issues that are associated with automated classification devices are no longer relevant as no decision regarding the slides classification is made.

The issue of quality controlling the existing system of cervical cancer screening can at best be described as problematic. The current project aims to address many of these issues by proposing a methodology and providing supporting experimental evidence that overcomes many of these difficulties. This thesis will first outline the project and state specifically the aims of this work. It will then discuss the nature of the task faced by cervical smear screeners and the limitations of the existing screening methods. It will consider some of the specific evidence relating to the quality assurance offered by the screening service particularly focussing on the levels of observer variation in diagnosis and expert judgments in visual classifications. This will then be discussed with relation to the advances in technology that have attempted to automate or semi-automate parts of the screening aimed at providing a better service. This will give a good basis for understanding how this project fits with current scientific thinking in this area. An experimental rationale detailing a process of data validation will be argued before the experimental evidence is presented which supports our approach. A final discussion will evaluate this evidence before relating it to the current literature. Finally, conclusions and future work will be addressed.

1.3 Contribution to knowledge

The work presented here represents a significant contribution to existing knowledge. This thesis reports extensive exploratory work undertaken in the development of a computer system capable of predicting the areas of cytological images that are salient to the human screener who makes the cytological diagnosis. A novel methodology is used that shows a lateral approach to a problem that has traditionally attracted research, but has yielded few useful applications due to implementation issues with automatic cytological classification devices. The image analysis methods used are aimed at finding salient features, rather then abnormal ones. This also represents a departure from the traditional approach towards a new cytological image analysis method. Specifically, the work presented in this thesis contributes to existing scientific knowledge by:

- Introducing eye tracking as a viable research tool for locating salient features based on cytology screener's fixations across cytological images.

- Introducing a feature marking exercise as a method of classifying those fixations.
- Introducing a novel approach to cytological image analysis that has an emphasis on locating salient features rather than abnormal ones.
- Providing evidence that using a combination of eye tracker data and feature marking data can reliably be used for the analysis of cytological images.
- Providing evidence that the image analysis methodology is applicable to cervical cytology images.
- Outlining and providing dynamic statistical software that allows the user to analyse all the data types automatically.
- Shows the effectiveness of machine colour texture analysis in predicting saliency in cytological images.
- Presents results that support both the novel methodology and analytical process being employed.

The work presented in this thesis not only provides a methodology and analysis software that has a real world application but also, provides a strong research basis for further work. It also represents a multi-disciplinary solution to a complex image analysis problem.

1.4 Project Aims

The overall aim of this project is to develop a novel methodology for the quality assurance of human cervical smear examination by trained experts and provide evidence that supports this approach. This can be broken down into a number of specific project aims relating to training, routine performance measures and online performance measures.

- To provide a training tool for quality assurance assessment using gold standard images for use by histopathology laboratories. A detailed definition of what constitutes a gold standard in this context can be found in section 6.3.1.

  - Obtain and independently verify cervical cytology slide images to provide a gold standard image set.
  - Provide a model with supporting evidence that allows objective testing and recording of classifications using an eye tracker.
  - Use the gold standard image set to assess the classification performances of cervical cytology screeners through the use of eye tracking technology.

- To provide routine performance measurement assessment of cervical cytology screening using gold standard images.

  - Use cytologist's expert knowledge to locate abnormal features within the gold standard image set.
  - Compare abnormal features with eye tracker fixations to create an index based on saliency.
  - Use eye tracking technology to provide a number of performance measures across several dimensions and provide evidence that these measures are both objective and accurate.

- To provide online performance measurement and assessment of cervical cytology screening using images that are not gold standard.

- Extend the model, to allow use of non-gold standard image presentation, through data verification with a new objective measure of salient areas
- Develop an objective measure through machine colour texture analysis.
- Provide performance measures based on non gold standard images, indicating when salient areas have not been viewed.

Chapter 2

The Cervical Cancer Screening Programme

2 - The Cervical Cancer Screening Programme

2.1 Introduction

The media's coverage of cervical cancer screening is by its very nature sensationalist. Medical advances and mistakes are highly publicised and can be misleading. For example:

"Cervical Cancer Vaccine Within 5 Years"

Jenny Hope, Medical Correspondent.

Front Page Headline, Daily Mail, Wednesday November 20th 2002

In the case of the Daily Mail headline quoted above, the article attached describes a drug that appears to have an impact on only part of the cause of cervical cancer. The high profile nature of some mistakes can also give members of the public a biased view of the success of the cervical cancer-screening programme. However, the programme is a huge success, with around 4.5 million smears examined every year and where errors do occur they are often down to individual human errors. For instance, in 1993 over 1000 women were recalled after a nurse took smears using a tongue depressor (see BBC, 2001 for details of screening errors).

To understand the fully the research being detailed here, it is important to understand the existing system of quality assurance, and to place this into context an understanding is needed of how cervical smears are diagnosed. This chapter will describe how cervical smears are most commonly screened from the initial cell samples being taken through to the final outcome. While every laboratory in every country will screen slides in a slightly different way there are a great many similarities both in the screening and QA measures that exist.

2.2 Screening Methods

Cervical cancer screening is a method of cancer prevention that is used to detect and treat abnormalities that can be a precursor of this type of cancer. There are several different methods available to those who carry out this screening. These will depend on a number of factors such as time, cost and staff training. The current system of screening in the U.K. National Health Service (NHS) aims to inspect smear slides taken from every sexually active woman over the age of twenty. This is then repeated every 3-5 years. Before discussing some of the other methods available it is worth outlining the general process used within the United Kingdom. This breaks down into three stages. Initially cells have to be collected, then the cells have to be processed to allow examination and finally the cell inspection can take place. In this section two basic methods will be described. Firstly, the traditional Papanicolaou method of screening which will give an overview of the whole screening process before discussing the differences of Liquid Based Cytology (LBC). Other methods of screening are available, however these two approaches represent the overwhelming majority of existing clinical practices.

2.2.1 Papanicolaou Method

The Papanicolaou method of cervical screening has been used as standard since the introduction of the UK screening program. It is based on the work of Dr. George Papanicolaou (American Society for Clinical Pathology, n.d.), who is seen as the predominant reason cytology became an acceptable basis for diagnosis. Cells are taken from the cervix by a general practitioner or nurse who will also visually inspect the cervix (neck of the womb). Cells from the full circumference of the cervix are collected using a spatula and cyto-brush and these are then transferred onto a thin glass slide. This is then coated with a fixative that ensures the cells do not degrade. The slide is then passed onto a laboratory along with patient details and the identification of the person taking the smear. Once at the laboratory the next step is to stain the slide using the Papanicolaou method. An example of this type of staining can be seen in figure 2.1 that shows two slides taken at different magnification.

This is a cheap and effective way of showing contrast between the cells on the slide. Once stained the slide can then be inspected for abnormalities. The primary screening will be carried out by a cytotechnologist who will use a microscope to thoroughly inspect the slide. On the basis of this inspection, one of three classifications is chosen. Where there is a problem with the slide it will be marked as inadequate. Where a slide is negative, that is that there are no apparent abnormalities on it, it will be reported as being Within Normal Limits (WNL). Re-screening of the WNL slides will then take place by rapidly inspecting the slide taking maybe a minute compared to ten for a primary screening. Other methods of re-screen include partial random re-screening and targeted re-screening and will often depend on the patient history going even as far as a full re-screening where a patient has a cytological or clinical history of abnormality. When a slide is read as abnormal by the primary screener a grade will be suggested and the slide is passed for secondary screening to a cyto-pathologist. The slide will then be given its final classification according to the system used. There are many varied classification systems that can be seen in Table 2.1 although this does not fully reflect the number of grades that exist in the U.K. screening process. There are a total of 8 different classifications that a slide may be given in the UK. These are 'Inadequate Specimen', 'Within Normal Limits' where no abnormality is found, 'Borderline Changes', 'Mild Dyskaryosis', 'Moderate Dyskaryosis', 'Severe Dyskaryosis', 'Severe Dyskaryosis/?Invasive Cancer' and 'Glandular Neoplasia'

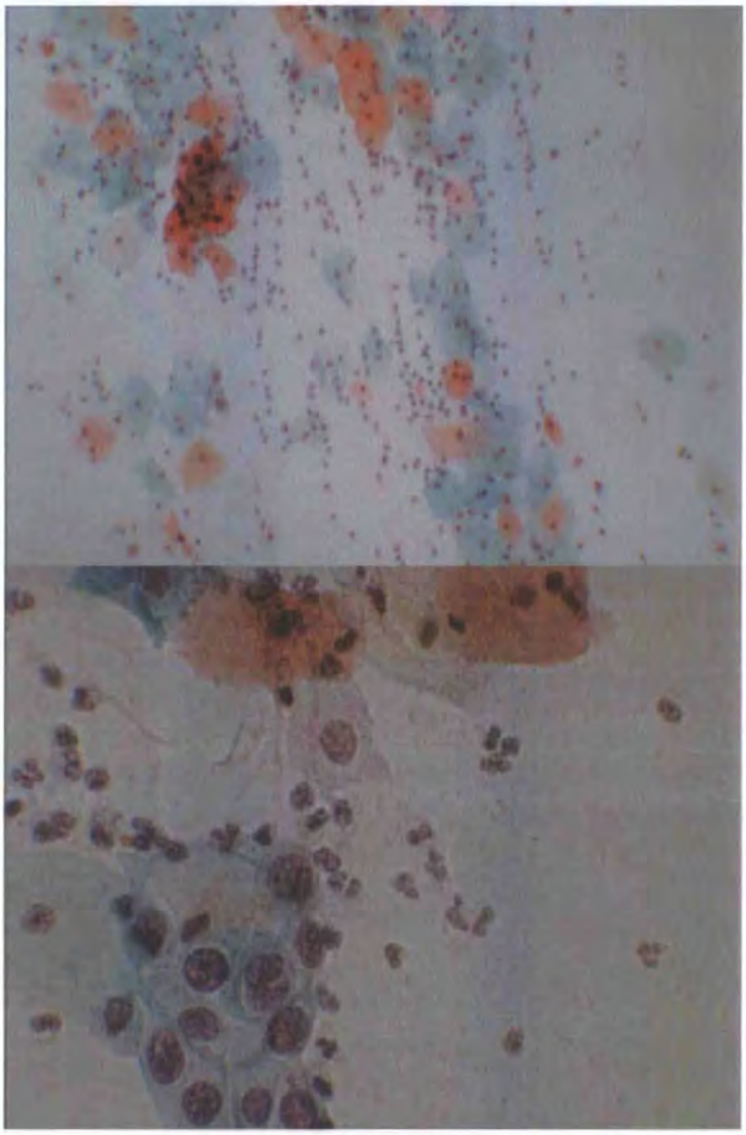


Figure 2.1 A typical Papanicolaou slide seen at x10 (top) and at x40 (bottom) magnifications. At x40 abnormal cells become far clearer.

Table 2.1 Various classification schemes for cervical cytology. Adapted from Nanda, McCrory, Myers, Bastian, Hasselblad, Hickey, & Matchar (2000)

<table>
<tr><td>Classification system</td><td colspan="9">Classification</td></tr>
<tr><td>UK</td><td rowspan="6">Normal/ Within Normal Limits</td><td colspan="2">Borderline Changes (including HPV)</td><td colspan="2">Mild Dyskaryosis</td><td>Moderate Dyskaryosis</td><td>Severe Dyskaryosis</td><td colspan="2">Severe Dyskaryosis/?Invasive Cancer</td></tr>
<tr><td rowspan="2">The Bethesda System (TBS)</td><td rowspan="2">Infection Reactive Repair</td><td rowspan="2">Ascus</td><td colspan="5">Sqaumous Intraepithelial Lesion (SIL)</td><td rowspan="5">Invasive Carcinoma</td></tr>
<tr><td colspan="2">Low Grade (LSIL) (including HPV)</td><td colspan="3">High Grade (HSIL)</td></tr>
<tr><td rowspan="2">Richart</td><td colspan="2" rowspan="2"></td><td rowspan="2">Condyloma</td><td colspan="4">Cervical Intrepithelial Neoplasia (CIN)</td></tr>
<tr><td>CIN I</td><td>CIN II</td><td colspan="2">CIN III</td></tr>
<tr><td>Reagen (World Health Organisation)</td><td colspan="2">Atypia</td><td colspan="2">Mild Dysplasia</td><td>Moderate Dysplasia</td><td>Severe Dysplasia</td><td>Carcinoma *in situ* (CIS)</td></tr>
<tr><td>Papanicolaou</td><td>I</td><td colspan="2">II</td><td colspan="3">III</td><td colspan="2">IV</td><td>V</td></tr>
</table>

2.2.2 Liquid Based Cytology Methods

While Liquid Based Cytology (LBC) is not new, it is only recently that it has begun to replace the traditional Papanicolaou method of slide preparation. In the UK, a limited trial of LBC has been carried out and currently all NHS laboratories are expecting to switch completely to using LBC methods within five years after it was shown that, while there was not a significant difference in either cost or detection rates, there was a significant drop in the number of inadequate smears (National Institute for Clinical Excellence, 2003). Overall there was an 87% reduction from 9.1% to an average of 1.6% when using LBC. While the commitment to switch to LBC is already underway, the exact method is not yet certain, as there are many companies who provide LBC technology. The different methods are currently being appraised.

In order to understand the difference between LBC and Papanicolaou preparation methods, a general outline for LBC will now be described. This will vary depending on the LBC method used, but the same principles are evident throughout. The cell sample itself is collected with a specially designed brush, which is inserted directly into the cervix. The cells are then transferred into a fixative liquid immediately, the fixative vial is sealed and then this is sent to a laboratory that prepares the final slide. To create the slide itself, the cells are spread across the surface of the slide to give a monolayer of cells, rather than the multilayers associated with the Papanicolaou method. A monolayer requires no focussing up and down the cell surface to view the different layers, as the scene is two-dimensional rather than three-dimensional. This can make the application of machine vision methods far more simplistic as the images taken from the slides are not cluttered scenes with occlusions and transparencies that are evident in traditional Papanicolaou slide images. Liquid based cytology methods will be discussed in more depth in Chapter 4 along with other technological advances in cytology.

2.3 Limitations of screening

While cervical screening programmes have undoubtedly had an impact there are still some limitations. Even though the rates of mortality and incidence of invasive cancer are declining there are still ways for women to be screened and slip through the detection process. Chamberlain (1986) looked at the reasons why women develop invasive cancers in countries that have organised screening and found that the largest group were those who had never been through the screening process. This was followed by those women who had been through the screening process and had abnormalities but had never followed up adequately. After this comes those who had long periods of time between smears, and then those with false negative slides.

A false negative occurs when a woman has been screened and the outcome is negative (clear of abnormality) when actually there are abnormal cells on the cervix. Of all the possible outcomes a false negative is the most dangerous as it is the only one that overlooks an abnormality. False negatives can be seen in relationship to the other possible outcomes in this context in Table 2.2

		Real Result	
		Positive	Negative
Diagnosis Result	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Table 2.2 shows the outcome of diagnosed and actual positive and negative results

Mistakes that are made in screening occur at one of three points during the process; smear taking, preparation and reading. Most of the errors occur where there is a problem with either taking the sample or preparing the slide. In fact, according to McCrory *et al.* (1999), this accounts for roughly two-thirds of false negative readings. The final third are where errors in detection are from actually reading the slide. To understand how these errors can occur, the example of a single abnormal cell needing to be detected can be used. In the initial stages this needs to firstly be transferred onto the spatula that is by no means guaranteed. One of the issues LBC has addressed is the number of cells that are lost during this process as the majority of those cells that are sampled are transferred into the fixative solution. However, the traditional method of sampling means that the abnormality may be left on the spatula. The next stage of the process also has room for error. The single cell would then need to be transferred onto the slide, and then this abnormality has to be found by the screener looking at many thousands of cells. In these terms it is easy to understand how detection errors can occur.

There is one final limitation that can lead to false negative results that should be mentioned. When a slide is screened and is actually negative, but an abnormality develops soon afterwards the screening itself is correct. In these circumstances the failure to detect lies with the length of time between screens rather than the process itself. The UK's 3-5 year gap between screenings is designed to ensure that even when abnormalities occur just after a negative screening, it is still in an early enough stage to be treatable.

In any of these circumstances a false negative reading could have absolutely devastating consequences. The media have not been slow to pick up on this and it's not unusual to hear an item on the news reporting mistakes that have been made. A recent case reported by Dyer (1999a, 1999b) details three women who developed adenocarcinoma resulting in hysterectomies after their smears had been treated as negative. In this case the High Court upheld a ruling that the women had been victims of medical negligence. The current increase in litigation, particularly in the United States, has lead to increased concern regarding false negatives as well as competitive laboratories trying to increase their market share. These factors have lead to a general call for ways in which the sensitivity of testing can be improved.

Costing around £130 million a year to maintain, the screening programme in the U.K. is a vast undertaking as it intends to target at least 80% of the female population. There are still some shortcomings that have been noted by the National Audit Office (1998) which relate to the programme itself rather than the specifics of the task being undertaken. There is concern about achieving the 80% target especially when the groups of women are from ethnic minorities or impoverished backgrounds and general concern about the length of time it takes to process certain slides. The report stresses that steps should be taken to minimise errors and for quality assurance to be improved so that when errors do occur they can be detected at the earliest opportunity.

There are many problems with the current system of screening that could be improved with the use of effective quality control. However the form that this may take is open to debate. Currently, quality control is undertaken regionally and involves assessment on grading a set of gold-standard slides in order to measure the abilities of each of the individuals involved. Although this performs an adequate test of capability, this kind of performance assessment is far from perfect as there are more issues that relate to the human elements of this type of task that could also be relevant to the implication of quality control. Koss, Lin, Schreiber, Elgert, and Mango (1994) believe that most errors that occur are due to psychological factors rather than training, experience or volume of work. Certainly the wealth of psychological research into expert judgement and classification tasks is highly relevant but before examining some of the evidence it is worth considering some of the worrying findings that, while hard to directly attribute to purely psychological reasons, do seem to indicate these factors are at work.

2.4 Observer Variation

The actual physical process of screening is standardised and well established; each slide being viewed by different experts to ensure that the likelihood of misclassification is minimal. The whole service naturally errs on the side of caution as much as possible. This is true of most organised screening programmes throughout the world. One of the worrying things that this method of screening highlights is the variability between screeners within the same laboratory. Even the adequacy of slides is open to different interpretations by different people. Observer variation is at the heart of many of the debates that exist in cytology screening.

Observer variation can manifest itself in a number of ways. At the beginning of the screening process, and perhaps the most basic of the decisions affected by variation, is judging whether a smear is adequate or not. In a study by Yobs *et al.* (1987) where 10,000 slides were exchanged between two departments, a total of 478 were classified as being inadequate by one or the other. However agreement on these slides only occurred in 99 instances, or approximately one in five. O'Sullivan (1998) points out that the reason for this variation is likely to be lack of guidance regarding what constitutes an adequate smear. The UK guidelines (National Health Service Cancer Screening Program, 2000) mostly present a qualitative description of what constitutes adequacy rather than specifically offering guidance as to when a smear crosses the line between adequacy and inadequacy. In the United States, the Bethesda method of classification (Solomon *et al.*, 2002) does offer some quantitative measures but again these are open to individual interpretation.

Beyond slide adequacy, variation has been well documented for a number of cytological screening tasks. The extent of variation within laboratories is illustrated in a study by Gatscha, Abadi, Babore, Chhieng, Miller, and Saigo (2001). They investigated the rescreening of slides in the U.S. as classified as Atypical Squamous Cells of Undetermined Significance (ASCUS) according to The Bethesda System for Reporting Cervical/Vaginal Cytological Disorders, shown in Table 2.1. This is one of the more diagnostically difficult categorisations to make. Two cytotechnologists and a cytopathology fellow examined slides that they were aware had been initially diagnosed as ASCUS. They found that of the 632 slides rescreened, only 200 (32%) were given the same classification by each examiner. Of these, only 91 (14%) were given the same ASCUS classification as they had previously. It was also found that the classifications were in complete disagreement for 41 (6%) of the slides. While ASCUS is one of the more difficult diagnoses to reproduce, this demonstrates how difficult it can be to achieve consistency amongst screeners. There are many examples of this type of study which all demonstrate the problems of variation amongst observers

Variance between observers is not just found within laboratories, but between them as well. For example, in a study conducted by Branca, Duca, Riti, Rossi, Leoncini, Turolla, Morosini, and the National Working Group for External

Quality Control in Cervical Screening (1996) evidence was shown of variation in smear slide classification in an experiment conducted across 15 laboratories in Italy. This experiment took part in two stages. The first of these involved two sets of smears with varying degrees of abnormality that were judged for adequacy, had a diagnosis and prognosis formulated and were judged for their degree of difficulty in classification. The second phase involved two further sets of slides, which were presented after the first sets had been discussed amongst those taking part. The variability between the laboratories is described as 'striking' by the authors, both in terms of diagnosis and prognosis. Large variability was also found in the difficulty ratings given to each slide. Lessons are constantly being learnt as to how these variances can be reduced. This is reflected in the development of standardised procedures designed to give the same reliability regardless of geographical location.

Despite the best efforts of the authorities responsible for screening, variance still remains. Perhaps this is not so surprising when one considers the subjective nature of the task at hand. A screener will look for certain features in a slide that indicate a diagnosis, but these features may be missed or overlooked. When they are located, further problems arise. Because slides are being viewed that range from normal to cancerous, with every variation in between, the boundaries between classifications are arbitrary and open to interpretation. This is not unique to this situation, as it occurs for every task involving categorisation of items that form a continuum. Even the adequacy of each slide is judged to be different between observers. Because each slide is a novel image there is no benchmark with which to compare success. Where true mistakes are made, often it is only time that reveals a problem with diagnosis.

The management of patients also suffers from problems with observer variation that are evident throughout the screening process. O'Sullivan, Ismail, Barnes, Deery Gradwell, Harvey, *et al.* (1996) showed 10 observers a set of 100 smears on two occasions. Five of these people were histopathologists and the other five were cytopathologists. It was found that the cytopathologists reported endocervical cells and wart virus infections with greater regularity than the histopathologists. Both groups also showed poor inter-observer agreement in all the parameters measured and there were many changes of management recommendation between the two rounds. The authors note that most of these reflected the changes of opinion on the degree of dyskaryosis. They also note though, that in 24 of the examined cases the initial assessment on both viewings had been identical, but a different strategy for management was recommended. They note that it would be difficult to attribute these changes of opinion to anything other than human factors. In particular they point to the evidence that the levels of intra-observer agreement were good by comparison and suggest that this is an indication that personal criteria were applied and this remained constant over time.

There have been many attempts with varying degrees of success to ensure the quality of screening. For instance, some automated systems are now being used to screen slides (Broadstock, 2001). The implementation of any automated system is fraught with ethical and moral dilemmas. This means that even before use they are often restricted to a limited role. Where they are in use they are more likely to be checking through negative slides for missed cases (false negatives) and refer these back for a human rescreen. However, the problem of human variation remains.

One way to reduce the variation is with further training of those staff involved in categorisation. Jones, Thomas and Williamson (1996) looked at whether attending training courses or discussing the criteria through which slides were diagnosed reduced this variation. Nine cytotechnologists screened 100 cervical smears and the results were recorded. Approximately six months later this process was repeated. In this six-month period, two of the cytotechnologists had attended a training course, while two others had discussed other cases in-depth with the aim of reducing the variation between their diagnostic criteria. They found that both training and discussion increased the agreement in some areas between the two pairs of participants. Unfortunately, training is expensive both in time and cost to the individual, the laboratory and the health authority. Furthermore, with such heavy workloads as are generated by inclusive screening programmes, the opportunity to discuss diagnostic criteria in real depth is also limited.

2.5 Conclusions

The effect of observer variation in the task of screening cervical cytology slides is evident in every step of the process from the initial judgment of adequacy through to the recommended management of the patient. Guidelines that have been in place both in the UK and abroad are designed to allow more uniformity both within and between laboratories but these still remain open to some interpretation. The levels of variation are certainly worrying but the programme acknowledges that these differences exist and address it by always operating on the side of caution. Any question regarding the classification of a slide will mean that it is scrutinised until either a decision has been reached, or in the event a decision cannot be made, the patient will be recalled to give another cell sample for the laboratory to examine. This means that there are genuinely few false negative slides that go undetected. The individuals who make the diagnosis and prognosis decisions on the slides are all experts at the task, and as such the way they make their decisions comes under close scrutiny whenever a problem occurs.

Chapter 3

Human Factors

3 – Human Factors

3.1 Introduction

Studies of observer variation often suggest that it is human factors that are a main factor in the observed differences. O'Sullivan *et al.* (1996) suggests that because their results were far more consistent within as opposed to between observers, that it was difficult to attribute this to anything other than human factors. But what are the human factors that can cause the observed differences? There are a number of psychological factors that may be contributing to the levels of observer variation that have been demonstrated in the previous chapter. This is because of the nature of the task being carried out. It involves an expert judgment to be made, interpreting a qualitative guideline and applying it based on observed features, and It involves maintaining a high concentration level while making these classifications.

3.2 Expert Judgement

Expert judgement in visual classification is a deceptively complex area of study. When the question "What makes an expert decision better than a novice decision?" is asked, how an expert is actually defined needs to be considered. How can the decisions taken by an individual be assessed to discover if they are of an expert standard or not? This chapter will discuss expert judgement, before considering expertise in cytology. The underlying psychological processes involved in the categorisation process are then discussed, before vigilance in decision-making tasks is discussed in relation to the screening process.

In order to consider what makes an expert decision, we must first define exactly what is meant by expert judgement. Shanteau and Stewart (1992) define expert judgement as something that "applies in situations where there are grounds for saying that some judgements are better than others" (page 95). They state that there are at least three reasons why experts are worthy of study and these illustrate the wide diversity of expertise.

The first reason is that of generalisability of research. For researchers, it is vital that work can be generalised to other populations than the one under examination. In the domain of expertise this becomes more problematic as experts may or may not be governed by the same rule system being used. For instance, an expert in cognition can still see visual illusions regardless of their level of understanding of the psychological processes that underlie the effect. In contrast, skilled tasks such as cytological screening require both training and experience to achieve expertise and so an expert may be using a different rule system to a novice at these types of tasks.

The second reason given by Shanteau and Stewart is to provide a basis from which expert systems can be built. The study of expertise can show the knowledge and decision rules in practice and these can then be transferred to computer systems. There are a large number of cytological screening expert systems, some of which will be critiqued in the next chapter, but it is only through the study of experts and how they demonstrate that expertise that such systems can be developed.

The third reason given is that experts are worthy of study in their own right. Expertise may be shown in wildly different skills and tasks that utilise different types of knowledge and decision making processes. An expert juggler will demonstrate their expertise through their motor skills, while an expert computer programmer will show their expertise through their problem solving skills and subsequent application. In cytology, the expert will demonstrate their expertise in their ability to classify novel cellular matter based solely on visual information. It is this wide variety of domains that makes the question posed at the beginning of this chapter such an interesting and intriguing yet complex area to research.

3.3 Expertise in Cervical Cytology Screening

The observer variation found in cervical cytology screening reveals something about the nature of expertise being used. The levels of variation that can be observed imply that the rule system each screener uses in order to assess each slide is subjective and implicit in nature. If the process of slide examination and classification were completely objective then levels of observer variation would be minimal. Furthermore, if the rule system being used by experts could be externalised then novices would easily be able to emulate the screening task. As this is not the case, the question of why some people are more expert than others remains.

There is a wealth of information available to each observer engaged in the task of making effective decisions during a slide examination and all the relevant sources of information will be taken into consideration when a classification is decided upon. Decision-making processes generally rely on the use of heuristics which simplify the task (Tversky and Kahneman, 1974). Because of this, when a decision is made the sources of information being used may be inappropriate and the amount of information used may also be suboptimal.

Shanteau (1992) suggests that often there is a misguided assumption that experts are simply using more information when making their decisions than novices. In a review of five studies of expert judgements it was shown that in fact experts and novices use the same amount of information prior to a decision, but there is a difference in the importance of the information being used. While an expert may use the same or fewer cues than a novice, those cues are more relevant to the decision being made. The implication of this for cytology screening is that the information used by experts which pertains to the diagnosis and possible classification will be more salient than that of a novice screener. If this is indeed the case, evidence of this could be found under experimental conditions. This point in particular is highly relevant to the work contained within this thesis

3.4 Categorisation

Another consideration when looking at experts involved in a visual classification task is the way in which humans categorise. Defined by Medin and Aguilar (1999) as the process by which distinct entities are treated as equivalent, the structure of natural object categories has been the focus of a lot of research (reviewed in Rips 1990; Komatsu 1992) however there is still some debate as to the underlying processes by which humans classify.

Perhaps the most instinctive of the existing theories is that similarity is used as the principle behind our organisation of categories. The critical issue here is the extent to which similarity can provide an account for our ability to conceptually categorise the world. While at first a similarity based account seems logical, as a poodle and a terrier (both dogs) are more similar than a poodle and a horse (not a dog), similarity based models have proved to be controversial. Rosch (1975) states that objects in the world can be clustered together and that this will be by using a number of correlated attributes. The cluster of these attributes leads to a formation of a prototype concept, so in cytology a screener would be expected to have an internal concept of the idealized category member and any judgement would be made by comparing any novel stimuli to the prototypes for each category to see which it is closest to. There is some debate as to the nature of this internal representation as it is unlikely to be simply the best example but rather an abstracted concept with some going as far to suggest that the model only need include a set of attributes. However, a consequence of similarity based models such as those using prototypes and exemplars is that the world is already organised for us and that it is our categories that map onto this reality (Rosch & Mervis, 1975).

The concept of similarity as an objective organising explanatory principle is not without its criticism. Goodman (1972) suggests that although similarity is based on shared properties of the two objects being categorised, any two objects can share an unlimited number of properties. A poodle and a horse may be considered similar because they are both animals, but also because they have four legs, are mammals, have hearts, make noise, and many other similarities besides. In these terms, the concept of similarity seems infinitely flexible and this makes similarity far too unconstrained to be useful as a method of explaining categories. In order to be a useful explanatory principle, it would need to be definable within constrained parameters. For this reason, Medin and Aguilar (1999) suggest that we may see things as similar because they belong to the same category, rather than basing our categories on similarity.

An alternative to this type of theory is summarised by Rips (1989). It is suggested that the way in which typicality and similarity are determined is different from the process that is used to determine category membership. Our internal representation of categories consists of concepts comprised of properties and features that go beyond identification and classification. Further work has argued that underlying principles, which are often causal, help us to decide relevant features and discover how they might be interrelated (Komatsu, 1992). This can be considered in cytological terms by examining a case of a normal slide. A slide is normal when there are no abnormalities found, but to be typical of a normal slide it would have to fill all the stereotypical criteria that define normality in this case. Where a slide may not be at all typical it can still be normal.

In the context of cervical smear categorisation the task is problematic because the slides are chaotic, each being unique and novel to the screener. Grades are defined by guidelines but because the level of abnormality will vary from one slide to the next it is difficult to set concrete rules by which they can be judged. The cells will range from normal through to cancerous and this means the grades should be viewed as being placed along a continuum. A rudimentary problem with any judgement of this nature is where to draw the line between one grade and the next. Even the line between abnormal and normal is difficult to define. This is another example of the effect individual differences can have on a person's approach to screening. The decision is a subjective one, which will depend on each individual's interpretation of the guidelines and how these should be applied. As the guidelines are qualitative there is little help provided to establish where the line between each category is.

To demonstrate the difficulties of feature-based visual classification, Sokal (1974) used pictures of imaginary creatures known as Caminalcules, shown in Figure 3.1.

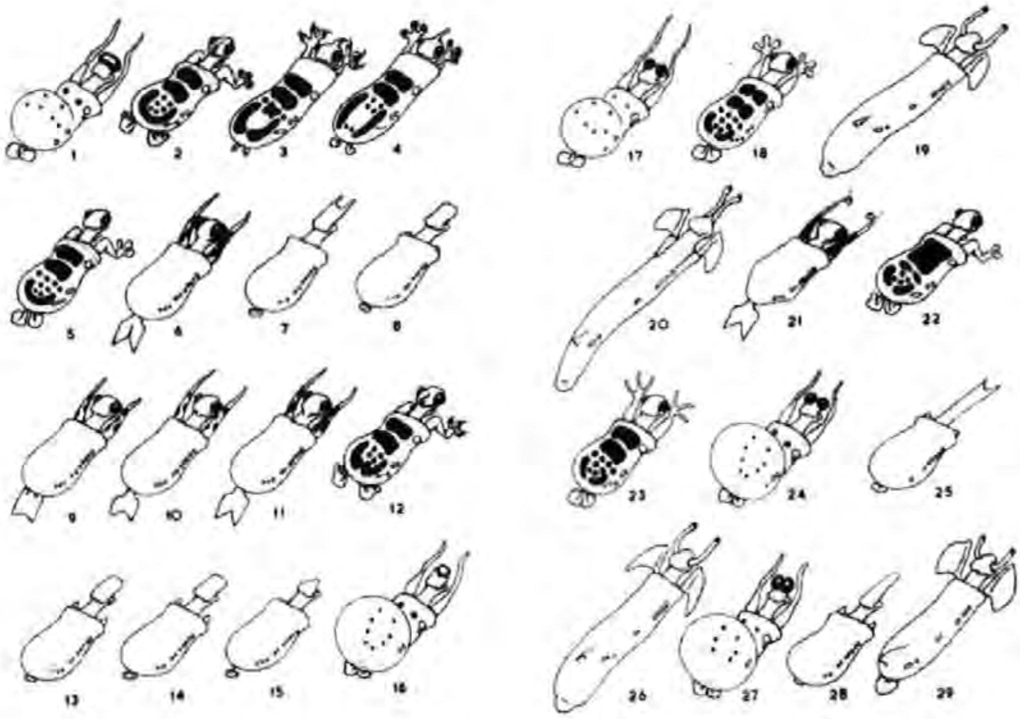


Figure 3.1 Imaginary creatures known as Caminalcules illustrate the difficulties of visual feature based classification (from Sokal, 1974)

Three taxonomists (A, B and C) were asked to group the creatures together based on their similarities. While A and C thought that 13 was more similar to 8, B believed it was a closer match to 28. Taxonomist C placed 5 and 18 together, while A grouped 22 with 5 and 18 with 23. B did not group any of these Caminalcules together. While A described 17 as most similar to 1 and C described it as most similar to 27, B described all three as equally similar. Many more differences were found, and analysis showed that there wasn't a single feature that was salient to all three participants. It is noted that each individual stressed different aspects of the creatures. While this is a simplistic approach to demonstrating feature-based classification differences, the effect is so strong that the exercise of classifying Caminalcules is still used in many universities today where it is given to undergraduate biology students to emphasise the difficulties of taxonomic work.

The issues raised by this study are easily extended to the cytology classification domain as one of the sources of observer variation. To address this, fewer categories are used in an attempt to reduce the variation. Traditionally all classification systems use mild, moderate or severe dyskaryotic categories, or low or high grade abnormalities. Doekler and Morris (2003) argue that the use of fewer categories in order to reduce observer variation may be misguided. The classification of slides is based on a subjective judgement that then may be contradicted either by a second person or by the original screener reviewing the material. This has lead to the recommendation that fewer categories should be employed, but they argue that a more logical approach to this would be to increase the number of categories. Miller (1956) showed that as the number of information channels is increased the information being retained levels out at around seven items – Miller's magical number seven plus or minus two. Doekler and Morris point out that the levels of information being transmitted do not fall if the number of categories is beyond seven. Someone attempting to process information from one hundred channels will take in the same amount of information as someone processing seven. It is this principal that they use to demonstrate the logic of changing the cytology classification system from fewer channels to a 100-point scale. In order to assess subjective judgements on a uni-dimensional continuum, they used a simple task of estimating the position of a dot placed between two lines. A total of 24 participants took part in this study, each classifying the dots between one and a hundred, depending on their position. In most cases, the mean and median values of the estimates were within one point of its true value with the maximum deviation being five. Because of the use of a 100-point scale, it allows for the calculation of confidence intervals and as the number of individuals giving estimates increases, so the confidence interval to be reduced.

While their suggested application is within pathological continua, of which cervical dyskaryosis is one, it would be difficult to put such a method into practice within the framework of the existing screening programme. Slides may only be viewed by two or three people, and only be examined thoroughly once. There would still be difficulty surrounding borderline cases between categories, and defining the exact point at which a smear classification becomes abnormal from normal, or moderate from mild, would still cause observer variation. Furthermore, there could be confusion regarding exactly what would be defined as either 1% or 100% dyskaryosis. The authors conclude that because there is no penalty for increasing the number of categories used, compared to information that may be lost by using too few, this is a logical step forward as long as confidence intervals are given with each classification. This would imply the accuracy of the classification.

What is clear from this work is that, as more observers view a slide the accuracy of its classification increases. The case for increasing the number of categories being used is compelling but even with limited categories, such as simply high and low grade, the accuracy of the diagnosis will still increase as more people view and provide a classification for each slide. Relating this work back to the work has already been considered in the expert judgement literature, the implication is that examining a number of experts and the way they reach their diagnostic conclusions can help examine the best strategy for examining a slide and reaching a diagnosis. This can also help with understanding which areas of each slide are most important when reaching a classification decision.

3.5 Vigilance in Screening

Before moving away from the topic of psychological influences in screening, there is one further issue that needs to be taken into consideration. There is no doubt that one of the influences on cytology judgement is fatigue. This is because cytological screening involves keeping a high level of attention for a sustained period of time.

Warm (1984) describes these types of tasks that involve prolonged vigilance as being related by the following dimensions:

- Prolonged and continuous for over 30 minutes
- Signals for detection are usually clearly perceivable when the observer is alerted to them, but are weak to most observers because they are not compelling changes in the observers operating environment.
- The signals to be detected occur infrequently, aperiodically and without forewarning.
- The observers response typically has no effect upon the probability of the appearance of critical signals

The immediate problem is that with a task that requires inspection, often there is a reliance on a sample rather than 100% coverage. This is true of cytological screening because the slide itself only contains a sample and each slide contains many thousands of cells. It would be simply impossible to inspect every cell. Evidence shows that it is better to carry out a limited careful inspection and generalise this to the sample than a 100% inspection that causes factors like fatigue to influence the person carrying out the task (Tsao, Drury and Morawaski 1979).

Vigilance research is difficult to generalise from the laboratory to the real world because of the large number of tasks that require vigilance. Everything from simple manufacturing through to running a nuclear plant require monitoring of some description but rarely will any two tasks from different environments have the same attributes. Linking vigilance research to cytology inspection is no easier but some general observations may be made to demonstrate the scope and impact of seemingly unrelated factors on inspection performance.

Factors relating to the facilitation or hindrance of the task take many forms. The stimulus itself may help if the display contains one cell type, is well preserved and well stained and where the density of the signals the screener is looking for is high but the display is impoverished as opposed to where there are a variety of cell types which are paler and smaller with a low signal density in a very cluttered scene. There are also factors relating to the differences in the workplace such as how long the slide is screened for and how many are examined a day which can have an influence, as does the number and type of breaks from the task that the job allows. Environmental factors are also a huge influence with variables such as temperature and noise levels helping or hindering the task. Finally there are what psychologists term individual differences. This is where factors like personality make a difference, an introverted personality being more likely to facilitate the task than an extroverted one. Likewise someone who drinks a lot of coffee, which is a stimulant, will generally perform better than someone who has consumed alcohol, a depressant. Even the time of day can play a part in an individual's alertness based on their circadian rhythms, with tiredness cited as a contributing factor in disasters such as Three Mile Island, Chernobyl, the Exxon Valdez, Challenger and the Herald of Free Enterprise (Dement, 1999)

Beyond environmental factors, it has been shown that during a monotonous vigilance task, alertness can decrease as much as 80% in one hour (Colquhoun, 1976). This phenomenon known as 'boredom fatigue', is likely to occur during the screening process so limitations are set on the length of time that a screener may repeatedly view slides. It is important to remember though that these are only generalisations. They cannot and should not be taken as truths about vigilance tasks because there will always be exceptions to generalised rules.

3.6 Conclusions

There are a number of human factors that can effect expert judgement of visual categorisation. The methods by which an expert reaches their decision have a significant bearing on their capabilities as an expert, just as the categorisation method being used and the vigilance level brought to the task. Expert judgement during the process of visual classification is still an area in need of research. Because expertise is domain specific, generalising from one area to another is problematic, although there are still similarities from which implications can be drawn. While it is evident that expert judgement and classification research has many shortcomings, it does lead to one very useful conclusion. The study of experts and classification should not be concerned with the externalised methods being used but rather the internal ones. Even if an expert can verbalise the rules by which they classify, they are likely to be flawed and not the same as the rules being used in practice. The heuristics being used are likely to be subconscious and maybe even involuntary. It may be argued that it is these involuntary heuristics that leads a competent cytological screener into bad practice and poor performances even if the individual's belief is that they are performing well.

The study of such experts then becomes a more complicated issue, as it requires examining their actions and deriving information from them. In cytology, this would mean examining the physical actions of a screener during the screening process and deriving information regarding their strategy and approach from these actions. Recent technological advances have attempted to reduce the variation that is seen in this type of task by automating some or part of the screening process or otherwise removing areas of the screening process which are seen as causes of variation. However, with so many factors able to have a bearing on a human's decision, the contribution of such systems and methods is debatable.

Chapter 4

Technological Advances

4 – Technological Advances

4.1 Introduction

The existing screening programme has been effective but there are still many problems that need to be addressed. One approach to answering some of the problems that relying on human judgment presents is to introduce fully or semi automated systems into the screening process, or otherwise introduce technology that aims to improve the quality of service. The use of automated systems is of particular interest to larger laboratories because of ongoing shortages of qualified cytotechnologists (Fetterman, Pawlick, Koo, Hartinger, Gilbert and Connell, 1999).

Automated systems for the analysis of cervical smear slides have been researched for over 40 years. Early systems such as TICAS (Wied, Bartels, Barh & Oldfield, 1968, 1970), SAMBA (Brugal, Garbay, Giroud & Adelh, 1979), and CERVIFIP (Tucker & Shippey, 1983) and many more beside have all failed to make an impact on cervical screening. While various systems exist, and have been the focus of constant research and development, only a small number have made any serious impact.

Automated systems have been the subject of extensive and continuing research in the U.K., where they are yet to be implemented despite the agreement that it could increase both productivity and quality of the existing screening program. Because of the impact using such a system would have on the nature of U.K. screening it was felt that the long-term benefits needed to be further justified (Potter, 1999). Furthermore, the benefits would need to outweigh the cost of equipment and training before being accepted into NHS laboratories.

4.2 Slide Preparation Devices/Techniques

While traditional Papanicolaou slides are the cheapest and most widely available for analysis by automated systems, other slide preparations claim to be more sensitive to the characteristic cellular changes that manual screeners and automated systems search for. A variety of systems exist and there are constant developments in the field toward replacing Papanicolaou smears. These have largely been concerned with 'monolayer' or 'thinlayer' slide preparations. A number of different preparation devices and techniques are available, but very few can claim to be supported by independent research evidence. Those systems are discussed here.

4.2.1. SurePath (developed by TriPath Imaging, Inc.)

The SurePath method, developed by TriPath Imaging Inc, requires a sample of cells to be taken using a sampling device which is retained in a transport fluid filled proprietary SurePath collection vial. This is so that the cell sample in its entirety can be forwarded to the laboratory where the vial is vortexed and centrifuged. Subsequent preparation of the sample and slide is then automated using a purpose built Prepstain machine (National Institute for Clinical Excellence, 2003)

4.2.2 Cytoscreen (developed by Seroa)

The collection method for Cytoscreen is similar to that of SurePath with a sample taken using a collection device, and placed into proprietary transportation fluid. This is then vortexed before a photometric reading is taken to estimate sample cellularity. An aliquot of the sample is then centrifuged onto a glass slide where it can be stained using the same method as staining of Papanicolaou slides.

This means that Papanicolaou cytopathology laboratories can use their existing staining procedures (National Institute for Clinical Excellence, 2003).

4.2.3 Labonard Easy Prep (developed by Labonard)

Easy Prep differs from SurePath and Cytoscreen collection methods because instead of using a transport fluid, it uses a fixative fluid once cells have been collected using a proprietary sample collection device. An aliquot of the fluid is then placed into a separation chamber. This chamber is attached to a glass slide containing absorbent paper and the cells sediment onto it in a thin layer. Again this type of slide means that cytopathology laboratories can use their existing Papanicolaou staining procedures (National Institute for Clinical Excellence, 2003).

4.2.4 ThinPrep (developed by the Cytyc Corporation)

ThinPrep is one of the bigger names within LBC and can provide semi- or fully-automated sample preparation methods. A sample of tissue is taken in the conventional way, but rather than being applied directly to the slide the collection device is rinsed in a transportation solution. This solution is processed by specialist ThinPrep equipment in such a way that a slide is created with just a single layer of cells. These slides can then be stained using standard laboratory procedures. Microscopic evaluation of the slides is also similar to conventional methods (National Institute for Clinical Excellence, 2003).

4.2.5 AutoCyte PREP (developed by AutoCyte, Inc.)

AutoCyte PREP aims to provide a representative sample of the specimen in order for it to be easier to screen. A subsample of the cells are washed from the collection device into a preservative fluid. This is then processed through a number of steps such as vortexing and sedimenting until finally a slide is produced with a 13mm disc of thin layered cells on it. (Australian Health Technology Advisory Committee, 1998)

4.3 Image Analysis Devices

Once a slide is prepared is must be inspected and classified. This is another part of the screening process where automated and semi-automated devices are being applied to provide an alternative or compliment to the existing human screening.

4.3.1 AutoCyte SCREEN (AutoCyte Inc)

Designed primarily for use with the AutoCyte PREP device, the AutoCyte SCREEN device also accepts Papanicolaou stained thinlayer preparations. Slides are robotically positioned on the stage of a microscope and then the stage movements and focussing of the slide is computer controlled. Up to 300 slide images per day can be captured at a maximum of 3000 x 2000 pixels. These high resolution images are evaluated using an assessment of the cell population histograms that involves extraction of features and a decision-tree analysis. The results are based on statistical classifiers (Kobler, 1996). This device then picks selected images from the slides which contain the most significant cellular findings and abnormalities for a manual video review by a human cytologist. This is then followed by a full manual rescreening for abnormal cases. During the analysis process, AutoCyte SCREEN also gives each case a classification that can be compared to that of the human screener (Australian Health Technology Advisory Committee, 1998).

4.3.2 AutoPap (Neopath)

The AutoPap system uses a high-speed video microscope and purpose built computer software to collect conventional Papanicolaou smear slide images that are selected for quality control rescreening by being scored and ranked. This process follows a set of complex algorithms that are designed to detect abnormal features, and classify the slides in one of four ways. If the score is below the rescreening threshold for quality control then there is no review. If the specimen is inadequate because there is, for instance, scant cellularity, then the slide is reviewed. If the score is above the threshold then it is chosen for quality control rescreening. Finally, if the system cannot review the slide because of a technical problem such as contamination, the process is reviewed. Some of these slides (10%) are then randomly picked for quality control rescreening (Fetterman *et al.*, 1999).

The AutoPap system was approved by the U.S.A.'s Food and Drug Administration (FDA) in both a primary screening capacity and for quality control rescreener and when used in both modes will designate slides as either Review, or No Further Review (NFR). NFR slides are considered to be WNL and are not manually reviewed. Review slides are ranked according to the likelihood that they are abnormal and then manually reviewed by a screener who will be aware of the ranking the slide has been given (Broadstock, 2001).

4.3.3 PAPNET (Neuromedical Systems Inc)

The PAPNET system requires those slides that have been screened and determined to be negative to be sent to a facility operated by the manufacturers, where the PAPNET instrument examines the slides for abnormalities that have been overlooked by the initial examination. Digitised video pictures are sent back to the original laboratory for further examination. High-resolution images are presented on computer screen, for review by expert personnel (Koss *et al.*, 1994).

There are two major components that make up the PAPNET system. The first of these is the screening apparatus which scans the slide automatically using a microscope with a low powered scanning objective and high powered function objectives governed by computer software. The second is the review station where the final images are stored for human analysis. Areas of interest are selected by eliminating objects from further consideration through a process of dilation and erosion based on size, shape and optical density. The process described here is known as the reversed top hat, or well algorithm. This primary classification process selects between 20,000 and 50,000 objects from a digitised video image (512 x 480 pixels) by following the assumption that all slides contain a number of abnormal cells that are isolated, or a cluster that can indicate a neoplastic event. Each object has its centre located by a shrinking procedure and the centroids are passed, along with the surrounding 24 x 24 pixel field, to a neural network for automated analysis. This neural network is trained under conditions of supervised learning, using digitised images of a wide variety of abnormal cells. Digitised images of overlapping cell clusters, neutrophiles and debris are used as negative control images. Values are assigned to each of the areas selected by the primary classifier ranging from 0.1 for the negative images and 0.9 for the abnormal cells. The system then passes those objects with the highest assigned value on for human review (Koss *et al.*, 1994).

4.5 Evaluation of Automated and Semi-Automated Devices

All of the systems and preparations discussed here have been heavily scrutinized due to the nature of the problem that they address. This is particularly true where approval from the FDA has been sought as this means the systems and preparations are being used on real people for real evaluation.

4.5.1 Evaluation of Liquid Based Cytology

There is no doubt that thinlayer preparation techniques significantly increase the quality of the slide for analysis (e.g. Lee, Ashfaq, Birdsong, Corkill, McIntosh & Inhorn, 1997) and in accordance with this finding the FDA have approved both ThinPrep and AutoCyte PREP for use in preparing cellular samples (Bishop, Cheuvront, & Sims, 2000). ThinPrep's efficiency has been compared to that of Papanicolaou slides by Tezuka, Oikawa, Shuki and Higashiiwai (1996) with very favourable results. The study involved taking a sample of tissue that was then split to create a Papanicolaou slide and a set of 10 ThinPrep slides for each patient. There was direct agreement for 95.3% for diagnosis from both preparations, with a 99.5% agreement within one diagnostic grade. The final diagnosis, in the case of the ThinPrep slides, took half the screening time using a quarter of the screening area and one tenth of the epithelial cells.

While the screening time for ThinPrep slides is shorter (Knowles, Bur, Otis *et al.*, 1992; Ferenczy, Robitaille, Franco, Arseneau, Richart, and Wright, 1996) one study that suggested this demonstrated that the cost of each slide screened was higher than for conventional slides (Bur, Knowles, Pekow, Corral, and

Donovan, 1995). In the U.K., the cancer screening programme has acknowledged this fact but claims that the extra cost of producing slides is offset by fewer inadequate slides requiring another sample to be taken (National Institute for Clinical Excellence, 2003). In their evaluation of ThinPrep, the Australian Health Technology Advisory Committee (1998) found that there were indeed fewer smears rated as unsatisfactory and that ThinPrep was superior for the detection of minor non-specific changes. They also warn that although the screening time is shorter than for conventional Papanicolaou slides, additional staff are required to prepare the slides.

AutoCyte PREP has also been shown to have a shorter screening time than conventional preparation methods, however it has been studied far less than ThinPrep (Australian Health Technology Advisory Committee, 1998). In a study by Bishop (1997), readings from over 2000 AutoCyte PREP and conventional slides were compared with each other and with the consensus diagnosis. A total of 148 squamous intraepithelial lesions (SIL) were found by either method and of these 85% were detected by AutoCyte PREP and 58.5% by the Papanicolaou method. Compared to the consensus diagnosis, AutoCyte PREP had a sensitivity of 86.7% for SILs and 99.7% specificity compared to 63.6% and 99.7% for conventional smears. When the consensus diagnosis was reviewed, 1.4% of AutoCyte PREP and 1.8% of conventional slides were upgraded to SIL.

A comparison between ThinPrep and AutoCyte PREP carried out by McGoogan and Reith (1996) investigated differences in cost, operator time, ease of use and performance for the two methods. They conclude that while consumables for AutoCyte PREP were more expensive, operator time for ThinPrep was more expensive and encountered more mechanical problems.

ThinPrep was also considered to be more tedious to use. Neither method produced slides that were deemed to be inadequate. They also suggest that real conclusions could not be drawn without exhaustive and extensive laboratory and field trials. This is partly attributed to the fact that lack of familiar markers of disease or their alteration when thinlayer methods are used may cause the learning period to be significant.

A recent evaluation of LBC methods carried out by the U.K. Cervical Cancer Screening Programme suggested that SurePath had no impact on detection rates of borderline, mild or moderate dyskaryotic smears. On severe dyskaryotic smears there was a reduction in detection, although there may be a number of reasons for this such as the effects of training and the different sampling techniques being used. While a drop in the rate of inadequate smears from 9% to 1-2% was noted due to the introduction of LBC, this was shown to be lower for SurePath than for ThinPrep. The long term effect of the reduction was not able to be assessed given the existing data. The question of cost effectiveness has also been examined, with suggested additional costs of transferring to LBC between £17,700 and £70,200 per year based on a laboratory processing 30,000 slides per year and dependent on which technique is used. In conclusion, It is stressed that at present there is not enough evidence to make an objective informed choice regarding which of the available methods should be adopted by the NHS (Moss, Gray, Legood, Henstock, 2003; Moss, Gray, Marteau, Legood, Henstock and Maissi, 2004).

4.5.2 Evaluation of Image Analysis Devices

Koss *et al*'s (1994) original evaluation of the PAPNET system examined an alpha and beta version of the machine. The beta version of the system outperformed the alpha version when presented with archived abnormal slides showing the entire range from low-grade lesions to invasive cancer. Following this, the beta version was presented with 500 further archived slides. A total of 140 of these slides were recommended for rescreening due to either the discovery of atypical cells or because the slide was considered inadequate. This review found three cases of LSIL in slides that were previously categorized as negative, and three further cases in slides previously classified as atypical. Two further cases were recommended for colposcopy without a revision of the atypical diagnosis. The system did miss three cases, one endometrial adenocarcinoma and two squamous neoplastic lesions. This lead the author to suggest that there was a place for PAPNET as an efficient quality control system for reducing false negative smears.

The efficiency of the PAPNET system has been further examined by Veneti, Papaefthimiou, Symiakaki and Ioannida-Mouzaka (1999). They selected 24 slides from patients who had developed a pre-cancerous lesion or cancer in a short time after a negative smear. These were then rescreened by PAPNET and re-evaluated by two observers. A blind manual re-evaluation by a third observer also took place. The automatic screening found one false negative smear that was also re-classified by manual screening. PAPNET took around one minute to interpret a slide whereas manual screening took around five minutes, leading to the conclusion that PAPNET was indeed fast and efficient. Further extensive testing has also shown that PAPNET is a reliable system and, when used with microscopy it improves the accuracy of cervical cytology (Denaro, Herriman and Shapira, 1997).

The benefits to both clinics and those they treat seem obvious, but some studies question the usefulness of automated review stations. Brotzman, Kretchner, Ferguson, Gottlieb and Stowe (1999) looked at the usefulness of having an automated rescreening process in place at a community hospital. Their principal findings question whether PAPNET can have a serious impact on detection rates. Of 1200 slides, 8 were identified with ASCUS by the PAPNET system. This was a similar rate to that already established at the laboratory through a manual rescreening of 10% of slides. The mean turnaround time was also a lot longer for the PAPNET review, taking 13.9 days to process compared to the average of 3.9 days for manual review. A similar study carried out by O'Leary, Tellado, Buckner, Ali, Stevens and Ollayas (1999) shows that after screening over 5000 slides, the PAPNET system picked 29% for review. Of these, only eleven cases were identified as having previously undiagnosed abnormal cells. This finding also indicates that the use of PAPNET is not likely to significantly reduce the rate of false negatives when compared to manual rescreening.

The AutoPap system has also been subjected to extensive testing. Fetterman *et al* (1999) found that detection of false negative results increased greatly, with its use. Compared to the practice of randomly rescreening a small percentage of the slides, they conclude this was a far more efficient and reliable way of selection. Overall findings indicated a greater specificity using the automated system when compared to the current practices within the laboratories. However, the performance of the machines tested varied greatly. It shows how important it is for laboratories to establish baselines and monitor performance upon the introduction of any new equipment in any role.

Colgan, Patten and Lee (1995) rescreened a set of over 3,000 WNL slides both manually and using AutoPap. Their manual rescreening found 106 abnormal slides and the review process confirmed abnormalities in 80%, and then 86% at the second review. This was then used to provide a baseline for the performance of AutoPap. Using a 10% review rate, AutoPap found 241 abnormalities of which 207 came from those that the manual review had not picked out. This represents a 4.3 to 5.0 fold improvement over the 10% random rescreening method used in the manual review.

One further problem encountered in the studies of both O'Leary *et al.* and Brotzman *et al.* was that the cost of implementing such a rescreening process further negated its usefulness in the laboratory. This has to be a concern as, should these systems be accepted on a wider scale, it is possible that not all laboratories will be able to afford them. An extensive study by Brown and Garber (1999) took a detailed look at the cost effectiveness of two automated systems, AutoPap and PAPNET, and the ThinPrep method of slide preparation. They searched MEDLINE for all relevant papers published between January 1987 and December 1997 and hand searched relevant journals for the same period of time. They also obtained unpublished articles from the manufacturers of the three technologies. The information from these studies was then pooled together provided that the papers included the number and results of all cytological slides taken, reported the FDA approved use of one of the technologies, used biopsy or review of discrepant results by a panel of at least three cytopathologists to validate all the positive findings, and included slides with validated LSIL, HSIL, or cancerous diagnoses. This amounted to nearly 200 studies. Using a hypothetical treatment programme that served a cohort of 20 to 65 year old women, they investigated the costs of each technology if each woman had joined the screening programme at the same age and the patients as a whole were representative of the general population. Their findings can be seen in Table 4.1.

Table 4.1. Selected results from Brown & Garber (1999)

		Lifetime Costs per Woman Screened		Lifetime Health Effects per Woman Screened		
Screening Strategy		No of Screenings	Health Care Costs*	% Developing Cervical Cancer	% Dying From Cervical Cancer	Additional Days Of Life
Quadrennial	Pap Smear – 10% Rescreen	12	446	0.33	0.10	23.91
	ThinPrep – 10% Rescreen	12	505	0.28	0.09	25.07
	Pap Smear – AutoPap Rescreen	12	476	0.27	0.08	25.32
	Pap Smear – Papnet Rescreen	12	508	0.26	0.08	25.47
Triennial	Pap Smear – 10% Rescreen	16	614	0.28	0.09	24.93
	ThinPrep – 10% Rescreen	16	695	0.25	0.07	25.73
	Pap Smear – AutoPap Rescreen	16	657	0.24	0.07	25.89
	Pap Smear – Papnet Rescreen	16	700	0.23	0.07	26.00
Biennial	Pap Smear – 10% Rescreen	23	939	0.24	0.08	25.72
	ThinPrep – 10% Rescreen	23	1059	0.22	0.07	26.19
	Pap Smear – AutoPap Rescreen	23	1005	0.22	0.07	29.29
	Pap Smear – Papnet Rescreen	23	1068	0.22	0.07	26.35
Annual	Pap Smear – 10% Rescreen	46	1955	0.20	0.06	26.56
	ThinPrep – 10% Rescreen	46	2194	0.19	0.06	26.80
	Pap Smear – AutoPap Rescreen	46	2089	0.19	0.06	26.86
	Pap Smear – Papnet Rescreen	46	2212	0.18	0.06	26.90

*In 1996 US Dollars

They concluded that the incremental cost effectiveness ratios of AutoPap and PAPNET assisted rescreening was comparable to conventional methods when screening occurred every three or four years, or less frequently. This finding should be accepted cautiously because of the nature of the literature that was reviewed, the authors admitting that it is often incomplete and can be contradictory in nature.

In one of the largest reviews of the current literature, Broadstock (2001) looked at both the effectiveness and cost effectiveness of automated and semi-automated cervical screening devices when compared to the traditional Papanicolaou method. Some of the problems inherent to the work of Brown and Gardner also become evident. The author reviewed over 700 articles from which only 26 met the criteria for inclusion. It was concluded that: -

- Test sensitivity and test effectiveness could not be reliably determined and provided no evidence for improved detection rates
- These estimates were the main source of uncertainty for establishing cost effectiveness
- Increases in sensitivity may lead to decreased specificity. This would add to cost by producing a higher false positive rate
- Higher quality research is needed to generate valid estimates of sensitivity and specificity
- Promotional information for new devices needs to be balanced with independent reports.
- Missed abnormalities on Papanicolaou smears will be detected at subsequent screens presuming adequate performance levels in the

laboratory, preventing 93% of cervical cancer assuming total screening coverage. Therefore the Papanicolaou smear should remain.

- Introduction of new devices cannot be recommended
- Resources should be targeted to other ways of improving the screening program
- Resources should be directed at appropriate monitoring of the program

It is interesting to note that after considering the evidence presented by Broadstock, the New Zealand Health Authority who commissioned it decided against the introduction of both automated screening and LBC technology.

4.6 Ethical Considerations

Something that underlies the use of any automated methods, or slide preparation methods, are the ethical and legal issues related to their introduction. Often omitted from papers introducing new technology, there is a fundamental problem with removing the human aspect from any part of the screening process. It is for this reason that new technology is thoroughly tested prior to introduction in any medical field. In cytology, the decision made when categorising a slide could be a life and death decision if cancerous cells are missed. Because of this more than one person views each slide in order to minimise the possibility of human error. In the case of a fully automated system engaged in primary screening that also misses a case, then the ramifications would go far beyond the legal and ethical issues that would certainly arise.

The ethical issues surrounding automation occur because there is no baseline on which to base the judgements being made, nor is there a line between classifications. Basing a system on human judgement will mean that it cannot possibly perform at a 100% success rate, because there will always be debate as to which diagnosis a slide is given. As such, automated systems designed to replace their human counterparts will quite rightly be tested extensively before being introduced. As a woman, being told that a misdiagnosis is down to human error is perhaps understandable. Being told that a misdiagnosis is due to an equipment malfunction or oversight is not.

Before widespread use of any automation is implicated, there needs to be more compelling evidence of the effectiveness of these methods. Furthermore, all of the evidence, for or against, must be handled cautiously if there is any suspicion of commercial interests and pressures by competing companies. It also means that because of the cost of using such methods, if they were to be proven beyond doubt to be effective, only those with the money may be able to afford to pay for improved healthcare. Alternatively, it may also be that the quality of care depends upon the affluence of the laboratory doing the examination and diagnosis. Companies providing automated methods of slide preparation and screening are faced with the fact that it may take several years before a system has been adequately and independently shown to be of use and often fail because this is not considered when trying to market a new product. Many of the earlier systems failed as commercial successes because of the financial pressure placed upon them for instant returns. As if to emphasise the point, the company producing PAPNET has recently gone into liquidation despite it being one of the most successful systems of its type.

4.7 Conclusions

There seems to be little doubt that the automated systems and preparations discussed here do contribute in some way to detecting false negative readings. This is particularly the case where there is a less frequent screening program in place. The use of monolayer or thin layer preparations can also improve detection rates in both manual and automated screening. So why then, are these methods not common place in all laboratories?

The ethical considerations play a large part in answering this question and often where automated systems are in place they only play a restricted role in the overall screening process. However, LBC has finally got the approval it needs to be introduced in a 5-year rollout across the UK. This is not because there is a positive advantage in using LBC for diagnosis, but because of the projected benefits of fewer inadequate smears. This should save enough time in the laboratory to make LBC superior to the Papanicolaou screening method on the grounds of laboratory productivity. It should be noted that although the NHS has made the decision to swap to LBC, the exact method is as yet undecided due to the lack of high quality assessments and comparisons of the available options (National Institute for Clinical Excellence, 2003).

What is needed is a new approach to the problem that avoids these ethical and practical issues. Given that manual random rescreening is shown by Brown and Garber (1999) to be a highly effective method already perhaps it should be this that is improved. After all, there is great variability amongst as well as within laboratories. The answer is to either modify an existing system, or develop a system that can test the abilities of a manual screener. This would mean that there are no ethical hurdles to overcome regarding the availability and cost of such a system, as it would only need to be utilised periodically. It would also reduce the variability found due to human factors. Various assumptions would naturally need to be made before a testing station is developed. Firstly, such a system would need to be able to use Papanicolaou smears, as these are the cheapest to produce and most widely used across the world. It would also need to use existing technology to analyse the pictures, such as personal computers, rather than purpose built computers. Finally, it would have to be easy to operate so that minimal or no special training is required. This last point is important, as it would allow a laboratory to test its own staff when an individual's performance is an issue and give additional training to those who are most in need of it. This can be achieved by using, as a starting point, those who are qualified to examine slides. By designing a tool for quality control purposes, an improvement on the base rate of each laboratory might be seen. This would also allow for a truer picture of the capabilities of existing automated screeners. An understanding of how slides are seen by human screeners is invaluable to guide software development and implementation of any system.

In conclusion, the future of automated systems is safe while so many issues are yet to be resolved. This is because of the promise of reductions in false negative rates, and the financial rewards it would bring, were substantial evidence backing one system or another to be produced. Until this happens, or an alternative way is found to produce desired results, money and time will still be invested.

Chapter 5

Experimental Rationale

5 - Experimental Rationale

5.1 Introduction

The work discussed so far illustrates perfectly the difficulty of providing an automated system for cervical cytological screening. All areas of the screening process are subject to observer errors and disagreement between experts. The variation is largely due to human factors, and so basing an expert system on the skills being demonstrated will inevitably also lead to variation. Ethical concerns mean it becomes very difficult to bypass human interaction when classifying these slides. In this chapter, the foundation will be presented for an alternative approach to improving the available quality of service based on a method of data verification.

As variation is to some extent an inevitable part of any human classification process, minimising its effect in cytology classification has been a goal for a very long time, and given that even merely discussing the criteria by which classifications are made can reduce it (Jones, Thomas And Williamson, 1986), there is a logical path to follow. Because minimal feedback can and does reduce variation amongst observers, an automated method of providing feedback would be of great value. The aim of this work therefore is to provide feedback to a screener of their assessment performance. Because of the difficulties in providing automatic quality assurance for such a complex and subjective task, a novel approach to the problem has been developed.

A data-driven approach has been developed, using the performance of others as a statistical baseline from which individual performance can be compared. This baseline is created using eye-tracking technology to discover areas that an expert views prior to making a classification decision of a smear image. This avoids many of the issues that can arise from using automated analysis to replace the human experts in the screening process. All of the fixations on the image are labelled for both content and importance to the diagnostic decision. An analysis of the images colour texture can be correlated against these fixation labels to test the predictive power of the colour texture measure at predicting salient areas of novel images. Finally, the colour texture analysis can be used to ensure a screener is considering the most salient information when making a slide diagnosis. An overview of this process can be seen in Figure 5.1.

This research aims to achieve a number of specific objectives. The first of these is to provide a model that will allow completely objective testing of machine colour texture analysis, and then to provide evidence supporting that model. In order to demonstrate the effectiveness of the model, it will be tested using a simple colour texture analysis. This will also show the appropriateness of the colour texture measure being used. The usefulness of saliency over abnormality as an assessment measure will be examined. Finally, the model does not remove the human element from the screening and classification process.

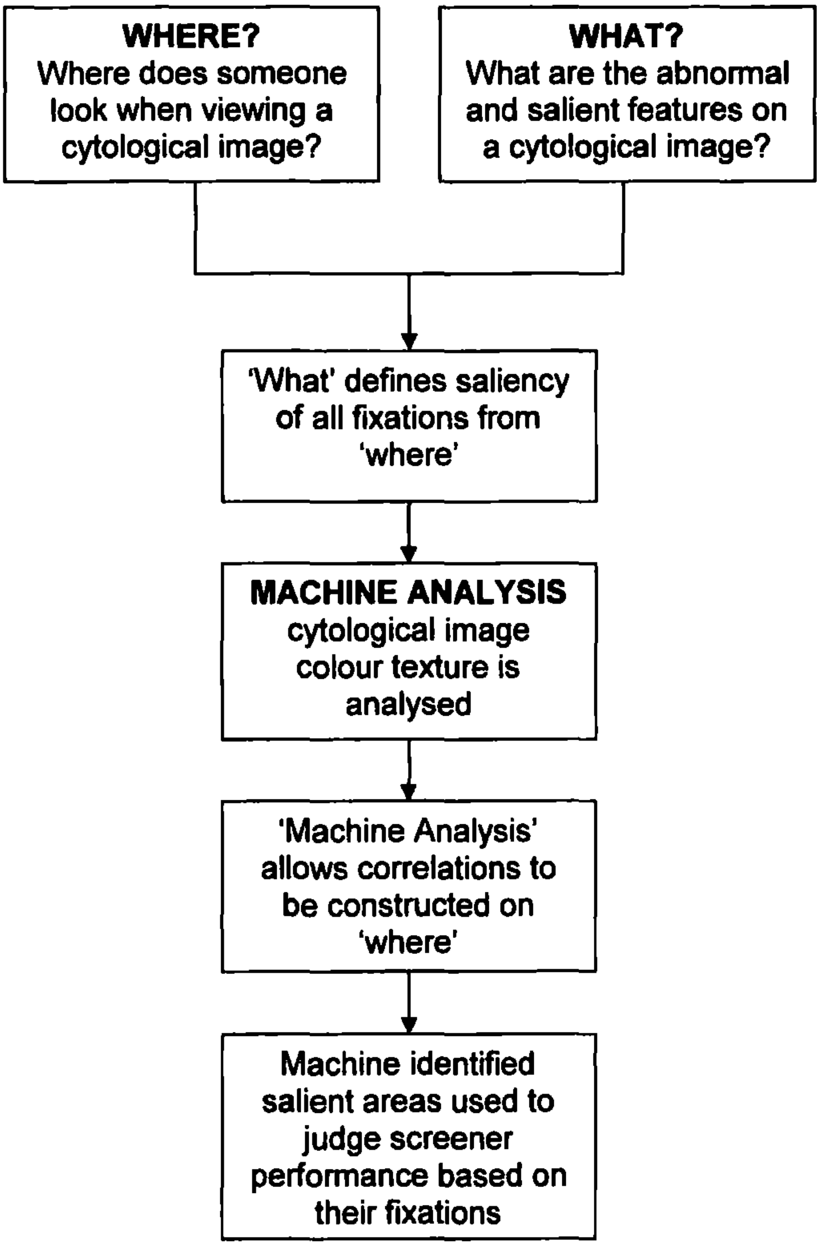

WHERE?
Where does someone look when viewing a cytological image?
WHAT?
What are the abnormal and salient features on a cytological image?
'What' defines saliency of all fixations from 'where'
MACHINE ANALYSIS
cytological image colour texture is analysed
'Machine Analysis' allows correlations to be constructed on 'where'
Machine identified salient areas used to judge screener performance based on their fixations

Figure 5.1. Data comparisons of human 'What' and 'Where' decisions with machine colour texture analysis

On the basis of these analyses, there are two further general aims. The experimental set-up being employed may allow a standardised performance test for screeners based on the performances of others, if the methodology shows evidence of being able to provide screener profiling from the data being recorded.

Providing accurate profiling will form a strong basis from which machine colour texture analysis can be assessed. Furthermore, this will allow us to examine whether, on the basis of machine analysis of images, it is possible to predict salient areas on an image in order to advise a screener if they had or had not viewed these areas prior to making the classification decision. Achieving this would allow feedback to be given to a screener had they not adequately covered a slide as part of a quality assurance process. A long-term aim of this work therefore is to provide instant real-time feedback to a screener of their performance

5.2 Eye Tracking

Eye trackers are a very useful research tool and a vital component of the work being presented here. As we have seen, expert judgements can vary from individual to individual and an expert may not be able to tell us why it is they make the decision that they do. This is due to the implicit nature of the rules they are using to make that decision. In terms of visual classification, an eye tracker allows us to directly examine where someone is looking prior to his or her classification decision. Before discussing the eye tracker, there are a number of issues relating to their use that will be discussed.

Research into eye movement and its effects on perception has shown that a number of important actions occur. Areas that have been examined are as diverse as language processing, face perception, scene perception, object recognition, dyslexia and reading music. This work has led to many variables being noted as significant indicators of ocular behaviour and these include, but are not limited to, saccades, fixations, pupil dilation and scan paths (Rayner, 1998).

There are in fact several types of eye movements of which saccades are the most relevant to the work here. When an individual is presented with a static scene to search they will make continual eye movements around the scene, fixating on various features of interest. The rapid movements between these fixations are saccades, although they are not the only type of known eye movement. Pursuit eye movements happen when the viewer is following a moving target across their visual field and, as with saccadic movement, can be affected by prior knowledge and expectations (Krauzlis and Adler, 2001). It has been shown that when pursuit eye movement is not quick enough to catch the target object, saccades are often used in order to keep up with it (White, 1976) and recent evidence suggests that saccadic and pursuit behaviour may well be different outcomes of the same sensory-motor function (Krauzlis, 2004). Of the other two types of main eye movements, vergence eye movements occur when the eyes move inwards together in order to fixate on a near object such as the end of the nose, and vestibular eye movements occur when the eyes move in response to head and body movements in order to remain fixated on an object. However, it is saccadic eye movement that remains the most important for standard information processing tasks (Rayner, 1998).

When visual attention is directed towards a specific area of the visual field, and lasts for at least 200 milliseconds, this is a fixation. This is a gaze that is spatially stable and represents the points at which information processing during a search of a static scene is most likely to occur. (i.e. Granka, Joachims, & Gay, 2004). Rayner (1998) presents evidence that we do not process any information during saccades between features and that this is because the eyes are moving so quickly, if we were able to perceive anything it would only be a blur. In fact, the saccadic suppression of visual information is so effective we do not even perceive a blur as the visual information is reduced. Pupil dilation and scan paths are also important areas of study. Measuring pupil dilation can tell us something about the viewer's interest and arousal, or fatigue. Examining the scan path can indicate the order of importance

Saccades are essential to our understanding of the world. This is due to the fact that we frequently need to take in more information than one fixation can provide. The highly receptive fovea contained within the eye needs to be focussed on whatever feature we are looking at to maximise the amount of information that can be provided by it. The visual field splits into three regions of which the fovea is the most receptive as it has the highest acuity. This foveal area or cone covers the central 2° of vision and is aligned to an area at the back of the eye that is densely packed with the receptive rods that help us to see. While the back of the eye is covered in receptive rods, there is a high concentration in the fovea. Visual acuity is not so good in the surrounding parafovea which covers the area up to 5° on each side of a fixation, and is poorer still in the periphery. The periphery is the area beyond the parafovea. To calculate the visual angle of the object being viewed a simple equation is used which is shown in Figure 5.2.

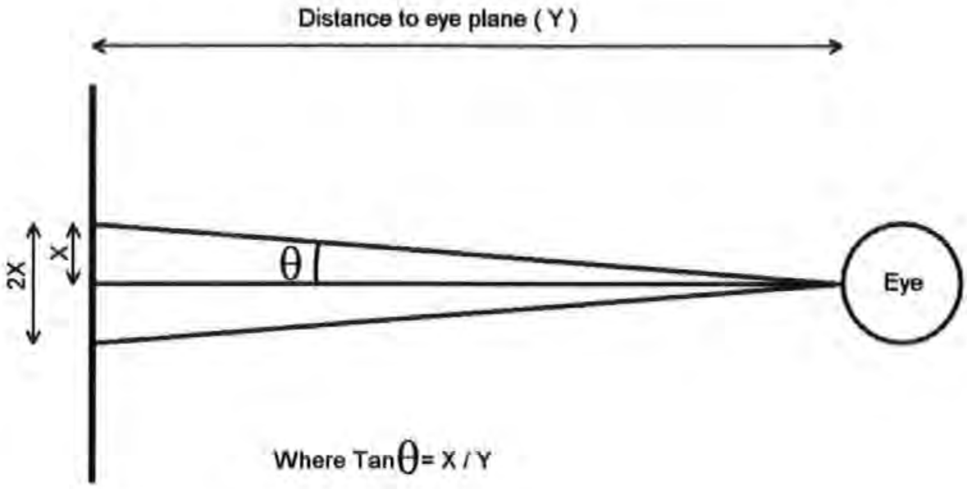


Figure 5.2 The calculation for visual angle

Visual acuity also largely depends upon the nature of the stimuli being viewed. The exact nature of something presented in the parafovea and periphery can also affect our ability to process the information it provides and whether we need to make a saccade and fixate upon it for recognition. Pollatsek, Rayner and Collins (1984) show that if an object or large letter is presented outside of the foveal area, it can often be identified without a saccade. In fact, Sanders (1993) suggests that the field of view can be divided into three areas when a person is presented with stimuli that needs identification. These are:

- where stimuli are identifiable without any action
- where stimuli are identifiable, but only after an eye movement is made
- where stimuli are identifiable, but only after a head movement is made

More recently, work investigating how we control our saccadic behaviour and choice of destination for each fixation has shown that proximity of the target is a very significant determinant of whether a saccade reaches it (Findlay, 1997; Findlay, Brown and Gilchrist, 2001). The evidence shows that saccadic destination is generally calculated from the existing fixation without the previous fixation having a carry-over effect.

In the context of cervical smear examination, this would suggest that an individual who has learnt rules on how to perform the visual search for abnormalities may well deviate from this as experience increases. If the saccadic behaviour from fixation to fixation is calculated afresh, then it can be suggested that those with more experience may well perform in a significantly different way to novice screeners. The only way to examine such behaviours is by using an eye tracker. By analysing of the end result of a visual search, the implicit rules and methods being used can be recorded rather than the explicit rules that the individual will believe they employ. In reference to the earlier Figure 5.1, this information will then fulfil the 'Where' part of the diagram. This will tell us where, when presented with cytological slide material, an individual needs to look before making the decision as to what the classification might be. The eye tracker is able to provide this information, and can indicate where the most salient areas that would need to be considered during the classification process. Although this tells us 'Where' on the image is important, it does not tell us 'What' the screener is looking at. The most salient area of a slide may be debris caught in the slide when it was created that draws the eye to it, rather than the abnormal areas that can help with the classification process. In order to find out 'What', a different type of information is required.

5.3 Feature Marking

While the eye tracker can provide information on 'where' the viewed areas of each cytology image are, it does not provide any information on why they were viewed or what they show. There are many eye-catching features contained throughout the images that do not bear any relevance to the classification but will have been examined. In effect, each individual eye tracker fixation has no direction associated with it. To make some sense of the fixations, we first need to discover what each of the fixations shows.

In order to make sense of the eye tracking data, a feature marking exercise has been devised which can provide the 'what?' part of Figure 5.1 and subsequently provide the information needed to complete this part of the validation. By asking participants to view cytological images and make a decision as to their classification, the fixation information being recorded relates to the implicit knowledge that each individual possesses. In contrast, the feature marking exercise records explicit knowledge. It is the comparison of this implicit and explicit information that defines each of the fixations that are recorded

Using a feature marking exercise to classify each of the fixations made during the eye tracker trial will not provide objective classification. However, because the classifications will result from pooling several people's data together, they should approach objectivity in the same way that a population mean approaches the true mean as the population is increased in numbers. Successful classification of each fixation is vital as this then allows testing of image colour texture analysis procedures.

5.4 Machine Colour Texture Analysis

When the eyetracker has provided information about 'Where?' and the feature-marking task has provided information on 'What?', a statistical understanding of the images being viewed can be created. This unfortunately would only apply to the images that have been processed. In order to extend this understanding to novel images we require a method of automatically assessing images and indicating salient areas. For this purpose, a machine analysis of the images is required.

5.4.1 Hue, Saturation and Value

Any image will posses a number of properties that can be exploited when trying to understand the image's content. A person's perception is co-dependant on both their memory and attention. Perception will also be directed by the properties of the visual array and things such as lighting, texture and pattern are all factors when trying to understand it. When colour is included there are a further three dimensions that need to be considered. These are hue, saturation and value.

The sensation of colour depends upon a function of the retina or optic nerve, in consequence of which rays of light produce different effects according to the length of their waves or undulations, waves of a certain length producing the sensation of red, shorter waves green, and those still shorter blue, etc. White, or ordinary, light consists of waves of various lengths so blended as to produce no effect of colour, and the colour of objects depends upon their power to absorb or reflect a greater or less proportion of the rays that fall upon them. In these terms, hue is the property of colour by which it can be perceived as ranging from red through yellow, green, and blue, as determined by the dominant wavelength of the light. Saturation can be considered to be the vividness of the hue, while value (also known as intensity or brightness) relates to the lightness/darkness of the colour. These three components of colour are illustrated in Figure 5.3. The development and research into these three dimensions has been carried out since the turn of the 20th century and is particularly relevant here because it is generally accepted that these dimensions are the most useful in terms of computer vision. The Munsell system was the first to describe the three dimensions (hue saturation and chroma) that correspond with the dimensions being employed here (Brainard, 2001).

Colour perception is important for this study because of the process the slides go through when being prepared. The staining carried out on slides is designed to highlight differences between the different types of cellular material contained on them. In the same way a screener uses this colour to aid their diagnosis, the extra colour information can also be used when designing a system to analyse images taken from the slides. The HSV dimensions relate to the way that colour perception is understood and so seems appropriate to use. Splitting an image into these three dimensions allows us to expand on the amount of information available from the initial image. While splitting an image into its HSV components can provide more information about the contents of that image, locating the features of interest within it can be further aided by using multi scale analysis.

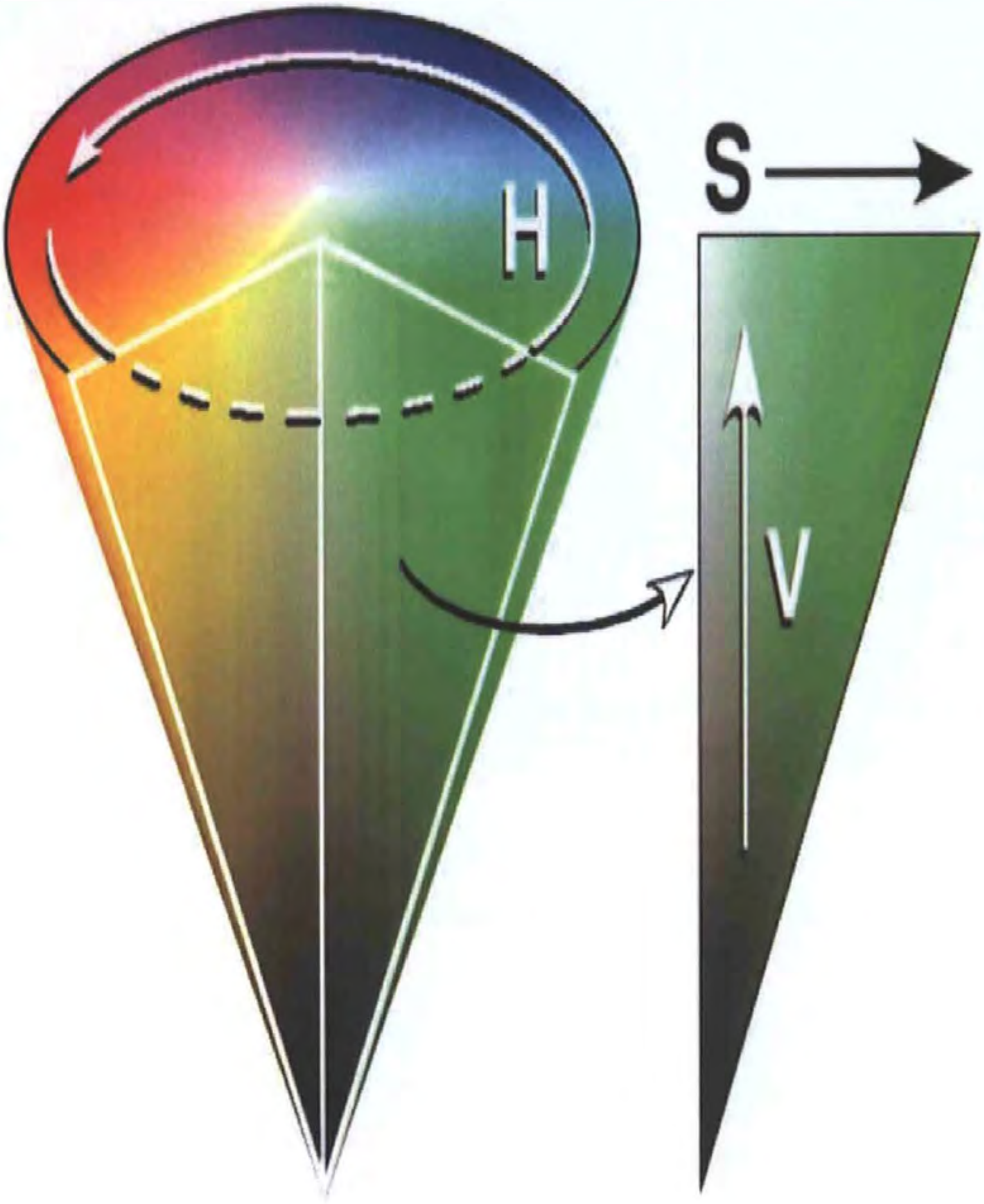


Figure 5.3 **shows the three dimensions of colour vision according to the Munsell System.**

5.4.2 Multi-Scale Image Analysis

In an image there may be a great number of features that occur at a variety of scales. While some are sharp and close together others will be more gradual and well separated. This presents a problem for computer vision when only a single fixed operator is used to view an image and capture all of the intensity changes to indicate the position of edges. The solution that developed from this is to use a number of different scales to analyse an image. For coarse-scale edged representation (low resolution) this would show only a limited number of features that would be relatively isolated. At a fine-scale (high resolution) the representation of edges that is produced is far denser. There are far more features detected at this resolution and these can be positioned very close to each other but they are different from those detected at the coarser scale. By using multiple scales to analyse an image a number of operators can be used simultaneously and can each be tuned to the different resolutions.

The concept of multi-scale analysis is not without a biological basis as there is a large body of evidence to support the idea that the visual system uses multiple channels. Wilson (1991) presents both psychophysical and physiological evidence that supports the hypothesis that the image which the photoreceptors respond with is filtered by visual mechanisms that are sensitive to patterns at different scales. Characteristics of the response are shown to be bandpass in the spatial frequency domain and reflect the variations in stimulus. Pattanaik, Fairchild, Ferwada and Greenberg (1998) list a number of appearance phenomena such as the visual systems adaptive gain control that can be explained as a result of multiscale visual processes.

It is easy to understand how multiscale descriptions taken from an image can be useful in finding salient features at a range of different resolutions. However, these descriptions can also help by directing the machine feature extraction process, and this is done in a way that is not dissimilar to how the eye uses extrafoveal information at low resolution to direct eye-movements across an image. (Rayner, 1998)

5.4.3 Colour Texture Analysis of Images

There are many methods available by which various types of features can be identified within an image and a list produced of x-y coordinates relating to these features (maxima). At present, there is a large body of work available on colour analysis, and an equally large body on texture analysis. However, it is only recently that computers have been powerful enough to handle both as a combined measure to analyse images. This is reflected in the literature by the fact that there is relatively little work on colour texture analysis prior to 1998 with an increasing volume each year since. As such a number of articles have been reviewed to confirm that the HSV/multiscale approach to image analysis being described is both acceptable and viable.

Drimbarean & Whelan (2001) tested the hypotheses that colour information can increase performance of texture analysis techniques based on overall classification performance. They show that using a colour texture measure can improve classification and that inclusion of colour does not mean significantly complicating the feature-extracting algorithm. This shows that using a combined measure for our own work is likely to improve overall performance. This is a particularly important finding because of the nature of the images we are analysing. From a human perspective, it would certainly be possible to classify cytology images if they were reproduced in grey scale but by adding colour the information available is enhanced and so we would expect a higher accuracy of classification.

Further evidence is presented by Palm, Keysers, Lehmann & Spitzer (2000). This paper uses Gabor filters to process images of different types but what makes this study different is that these were complex images. It is usual to use a standardised set that allows comparison with other research's performance. These images were also processed using a hue/saturation method that the authors state provided the best classification performance out of several options examined. This method of splitting the image into its component dimensions prior to the image analysis is very similar to our own approach although Gabor filters are not used.

These papers provide the evidence that supports our current approach to the machine analysis. Both support combining colour information with texture information and using hue and saturation dimensions to process complex images. In addition, Li and Lennie (2001) demonstrate the importance of colour texture in the human visual system by examining variations in colour and brightness in distinguishing textured surfaces. They show how, at low contrasts, observers were better able to differentiate between regions that differed in colour rather than luminance. They continue by showing how coloured noise affects the ability to distinguish certain types of textures far more than non-coloured noise does. They conclude that this equips the visual system to exploit colour even in the face of huge changes in brightness, as these coloured cues are relatively robust. Even with complex textures colour adds to the observer's ability to distinguish the world they see, albeit with a diminished effect. Because the human visual system acts as the model for computer vision, this work is important for us as it provides the biological basis for the exploitation of colour combined with texture.

5.5 Conclusions

The interaction of the data types described above should provide a basis for achieving the project aims. While each of these types of data provide useful information, by combining them implications can be made about all three data types. This method of verifying and validating the data will be described in more depth in the next chapter.

There are three mains aims the experimental work is designed to achieve. Initially, it is designed to provide a training tool for quality assurance assessment using gold standard images for use by histopathology laboratories. Then it aims to provide routine performance measurement assessment of cervical cytology screening using gold standard images. Finally it aims to provide online performance measurement and assessment of cervical cytology screening using images that are not gold standard. One final consideration that relates to all three project aims is that, at all levels of application, the model being developed reduces the ethical concerns regarding use or proposed use that can be associated with removing human judgment from automated cervical screening devices.

Chapter 6

Experimental Method

And Results

6.1 Introduction

In this chapter the experimental method and results are presented. The experiments were constrained by the availability of participants so were designed to provide as much information as possible. A number of issues raised by Potter's (1999) experience were addressed in order to maintain participant comfort.

6.2 Participants

A total of 10 participants took part in this study. In order to participate they had to be, at the time the study was carried out, actively involved in the screening and/or diagnosis of cervical cytological slides. Because the study uses an eye tracker, another pre-requisite was that they should have good short-range vision without the need for either thick-lens glasses or hard contact lenses. The participants had varying job roles and levels of experience within a histopathology laboratory

6.3 Materials

An ASL 4000 series eye tracker shown in Figure 6.1 was used along with Eyenal (eye-movement analysis) and Eyepos (eye-movement recording) software provided with the unit. This allows eye movement of 40 degrees or more vertically and 30 degrees or more horizontally depending on the optical placement and eyelids. The precision of this unit is better than half a degree and highly accurate with spatial errors between true eye position and computed measurement at less than one degree. The manufacturer notes that errors may increase but will still remain at less than two degrees in the periphery of the visual field. When using a bite-bar device designed to keep the head still, errors are estimated at half a degree of visual angle. It has a sampling and output rate of 60Hz.

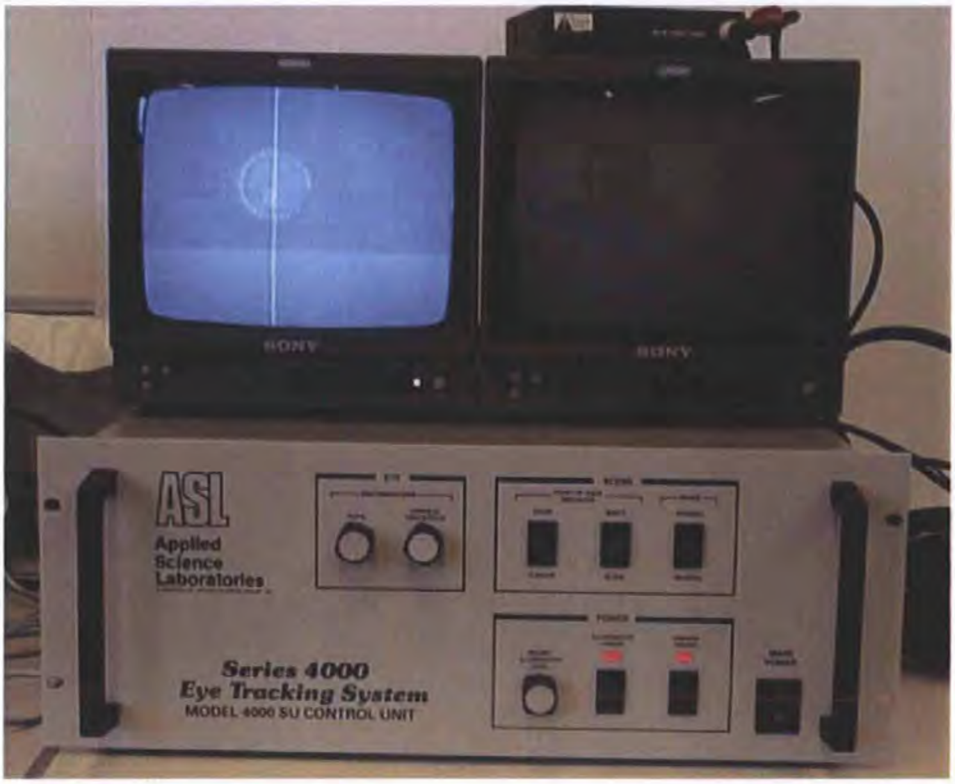


Figure 6.1 The ASL e4000 eye tracker unit. The left screen shows two crosshairs – the first locates the centre of the retina and the second shows the location of the corneal reflection.

Standardised briefing/debriefing and consent forms were used and these can be seen in (Appendices C, D and E respectively). A total of three personal computers were employed during the experiments, the first of which had a large flat-screen monitor for image display. The display monitor measured 40.8 cm (16.1inches) horizontally, 30.6cm (12.1 inches) vertically, and 51cm (20.1 inches) diagonally with a pixel pitch of 0.255mm. The display itself is an active matrix TFT LCD screen that had antiglare treatment to reduce reflections. Its maximum resolution is 1600 x 1200 pixels at 60Hz. In order to successfully run the experimental software, this computer was required to have Microsoft Windows 2000 Professional, DirectX 8.1 and Giveio.sys installed. A second computer was required to run a Microsoft Windows operating system, while the third was required to be running MS-dos. One further peripheral component was used and this was a real-time (Xdat) controller which sent a signal from the image display controls to the eye tracker indicating when the image had changed.

Purpose written software prepared by the author was used for image presentation, feature marking and statistical analysis (Appendices H, I and J respectively). The analytical software required Microsoft Excel running on a Windows based computer. The experiment also required the use of two tripods and a crossbar with a bite-bar attached shown in Figure 6.2. To prepare the bite bar for use under sterile conditions some further items were needed. These were sterilising fluid, type 1 thermoplastic impression material (green dental gum), a bowl of hot water and latex gloves. Finally, a set of images containing 25 Papanicolaou images and 25 ThinPrep images were used for the presentation with the necessary calibration and decision screen images.

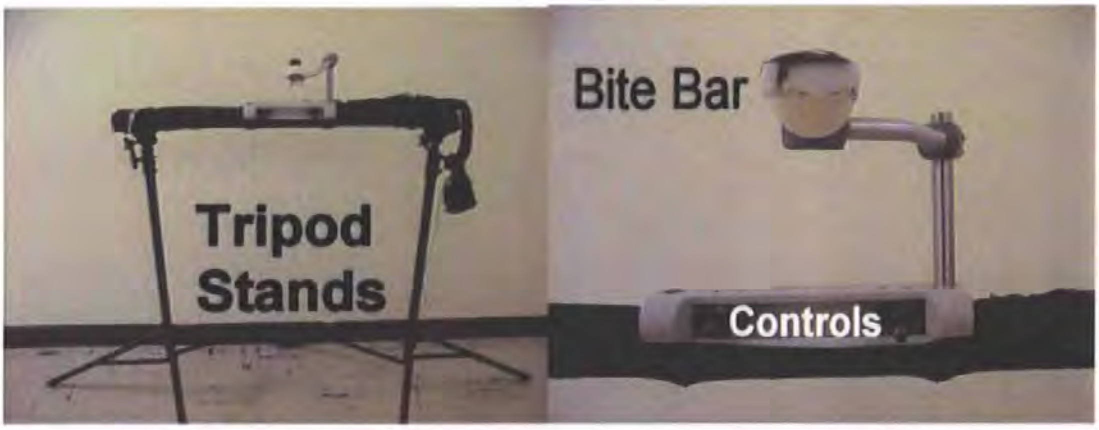


Figure 6.2 The stands (left) keep the bite bar (right) stable while the experiment is in progress. It is fully adjustable in order to be comfortable regardless of the size of the user.

6.3.1 The Image Set

The images used during the experimental work are very important. They must be representative of a number of classifications and be independently verified to ensure they accurately portray the element of each classification. The images used for the experimental work were taken from a set of 20 quality assurance slides, 10 using the Papanicolaou method and 10 using the ThinPrep method, and had been had been previously independently verified for their contents and classifications by the National Health Service and the South & West Regional Cervical Screening Quality Assurance Reference Centre. Each of the slides had areas of interest marked on them as a guide for the imaging process. All 20 slides were then imaged using a Nikon Coolpix 990 digital camera attached to a Leica DM IRB microscope. The calibration slide for the microscope at x40 magnification can be seen in Appendix A. A total of 450 high definition images were taken which were then sent back to a cytology laboratory quality assurance manager for a second verification stage.

While the slides themselves came with a predefined diagnosis using UK cytology grades, a trained cytologist did not take the images. Although areas of interest had been marked on the slides this was not a guarantee that the images would contain the cells that had been indicated for imaging. This left the possibility that, even though a slide contained abnormal cells, they could be missed during the imaging process and only normal cells would be contained within a picture taken from an abnormal slide. It was for this reason that so many images were taken when only 50 were required for the trial. All 450 images were returned to the South & West Regional Cervical Screening Quality Assurance Reference Centre where each was examined to see if they were representative of each of the slides grades. In all a total of 150 images were returned as acceptable for use within the experimental work. A total of 50 of these images were selected to represent as many classifications as were available. This set of 50 images was then sent to a Senior NHS Histopathologist for further independent verification. Once they had been viewed and the classifications confirmed for a second time they were deemed to be acceptable for use. Because of this verification process the images constitute a Gold Standard for classification within the experiment. The entire image set including those images used for trial purposes can be seen in Appendices F (Papanicolaou image set) and G (ThinPrep image set).

6.4 Design and Procedure

The study consists of three separate procedures of which the first two involved the participants. These were an eye tracking task and a feature-marking task. Separately to these a machine analysis of the images was also carried out. However, prior to the experimental work taking place, ethical clearance had to be granted.

6.4.1 Ethical Approval

The experiments described here required skilled histopathology personnel currently employed to screen cervical cytology slides for abnormalities. Because of this a hospital providing a cervical cytology screening service as part of the UK's cancer screening programme were approached to gauge their interest in participating in this work. Once they had agreed in principle to take part, ethical clearance needed to be sought and granted before any experimentation was carried out. This project was subsequently registered with the relevant hospital trust and an application for ethical clearance for the experimental work was made. A copy of the ethics forms, along with a research protocol and copies of all the relevant documents such as the briefing and debriefing were sent to the ethics committee for consideration. Ethical clearance was subsequently granted (Appendix B) to carry out the eye tracking and feature marking tasks within the histopathology department itself. Requests to carry out this experimental work in other locations were accepted by two hospitals but not within a timeframe that would allow inclusion in this thesis.

6.4.2 Eye Tracking Task Procedure

The eye-tracking task involves a forced choice image presentation designed to emulate the process of screening classification as closely as possible. During this part of the experiment, the participant wears the eye-tracking helmet so information is recorded as to where on each image the participant has viewed prior to the decision on that image's classification.

Before the experiment could begin the participants were given a briefing (Appendix C) to ensure they knew what the experiment involved and what they would be expected to do. This also gave them a chance to ask any questions before the experiment commenced and they were reminded of their right to withdraw from the study, or withdraw their data from the study at any time. This briefing covered both the eye tracking and feature marking tasks. Once the participants were fully briefed they signed a standard experimental consent form (Appendix E) before commencing any further. They were then given a short demonstration of the experiment so that they could see it running and provided they fully understood the nature of the task the experiment could begin.

The experiment began by warming the dental gum in hot water and moulding it onto a sterilised bite bar attachment. The participant was then asked to gently bite the soft dental gum in order to leave an impression of their teeth on it. After a few seconds, the dental gum solidifies and the bite bar attachment is fastened into place on a crossbar held in place by two tripods. The participant then sits on a chair between the tripods and the height of the crossbar is adjusted until they are comfortable with the positioning. This part of the procedure is carried out wearing latex gloves for hygiene purposes. The bite bar was used to maximise eye-tracker accuracy and repeatability. A picture of the complete eye tracker set-up in use can be seen in Figure 6.3. The bite bar can clearly be seen attached to a crossbar and held in place by two height adjustable tripods. The purpose of this bite bar is to reduce to an acceptable level any possible head movements that could affect the recording. The impression of teeth that is taken means that the bite bar can be held between the teeth without any pressure being applied directly and so is the most comfortable way of reducing head movement. The screen is placed at exactly 67cm from the bite bar as this means that the monitor takes up the majority of the participant's field of view. The handles on the crossbar include a button allowing the participant to control the speed of the image presentation. This can be seen clearly in Figure 6.3 with a diagram of the complete lay out of the system shown in Figure 6.4.

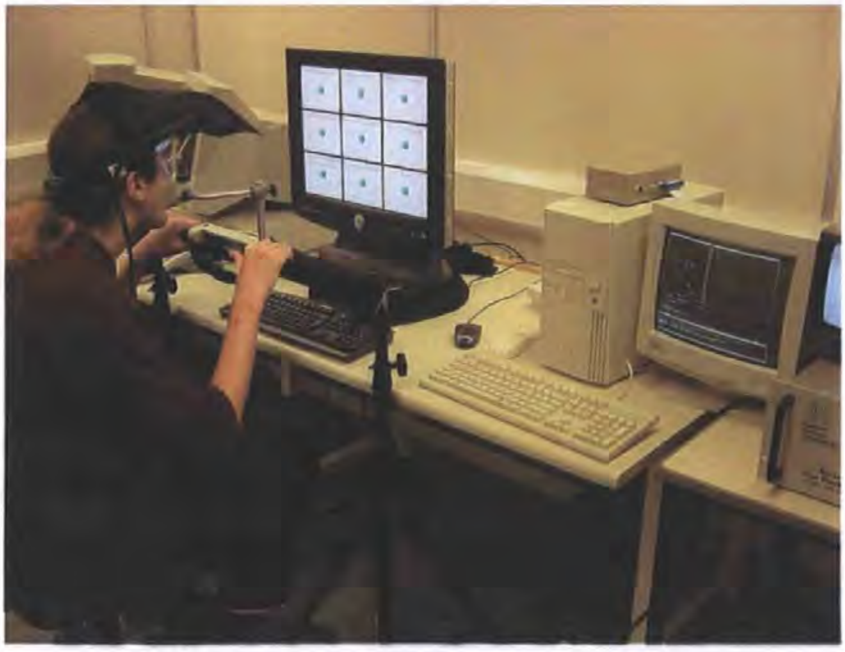


Figure 6.3 The eye-tracking equipment in use

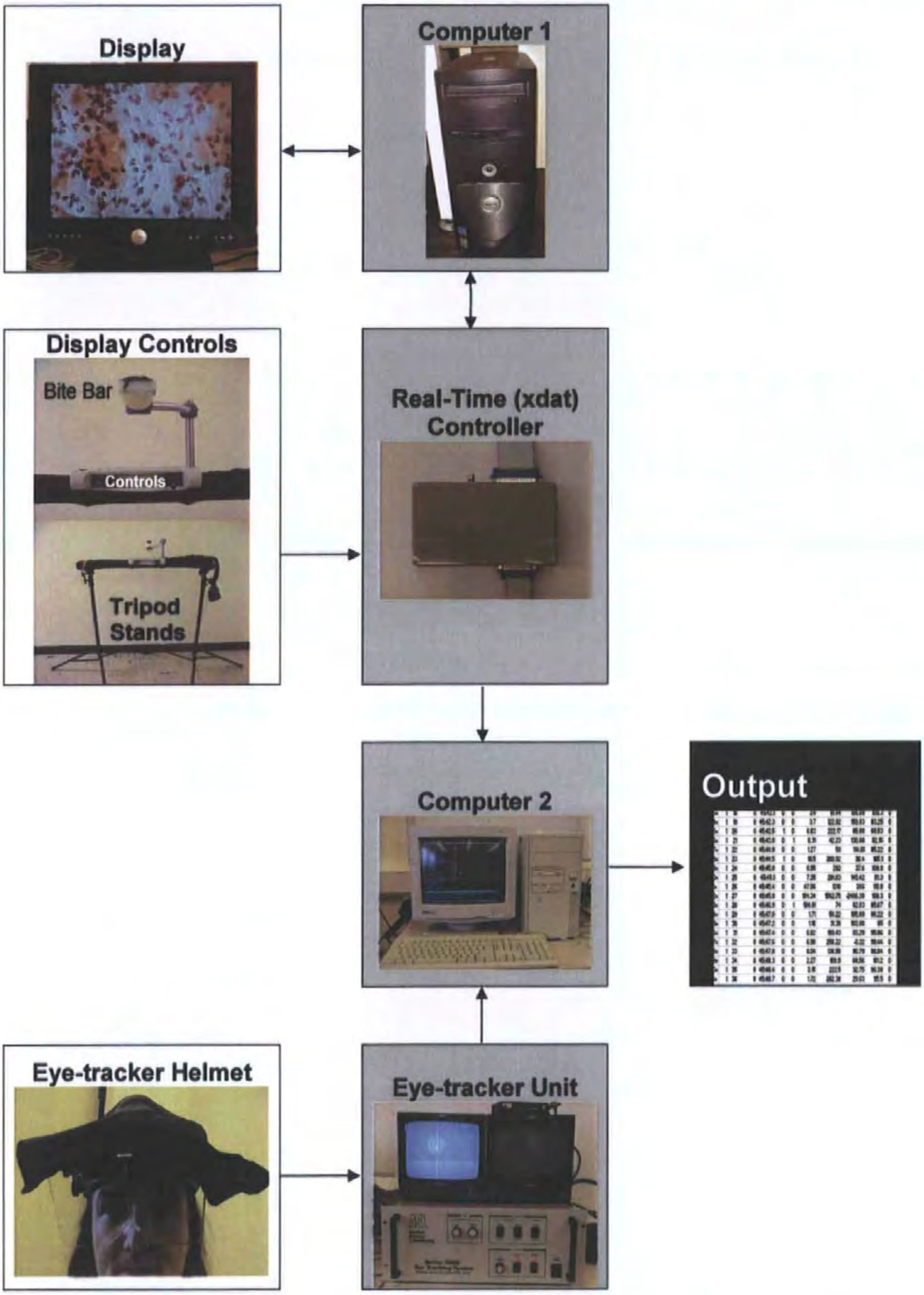


Figure 6.4 An overview of the experimental eye tracking equipment layout

When a participant had indicated that they were comfortable with the set-up, the eye tracking helmet was positioned on their head and adjusted until the retina was in the centre of the eye tracker's view and the retinal and corneal reflections were both registering on the eye tracking software. When this was achieved, the initial calibration process could begin.

Figure 6.5a shows the first of the calibration screens used during the eye tracker trial. This is a nine-point calibration screen that allows the mapping of boundaries of the area being viewed. The participant is asked to simply look at each of the points in turn as the appropriate number is read out. As each point is fixated upon, the co-ordinates are locked within the eye tracker recording software, and once all of these points have been locked, the experiment can begin in earnest.

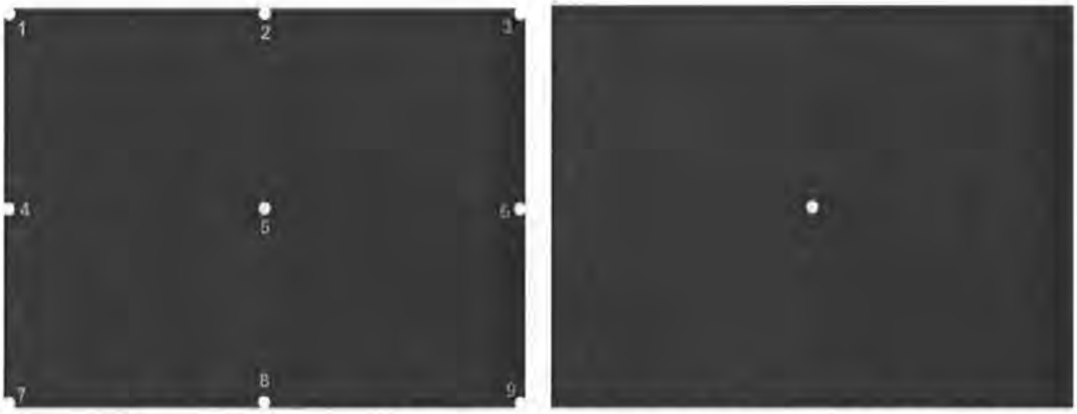


Figure 6.5 The calibration screens used during the trial, (a) 9-point, (b) single point.

When the initial calibration has been taken, the participant is informed that they can now take control of the image presentation. When the button on the crossbar is pressed they are presented with the next image in the presentation sequence. The software to present the images was written specifically for this experiment as the images were full colour and presented at a resolution of

1600x1200 pixels. It was very important that each image appeared on the screen instantaneously rather than appear gradually, and so the capabilities of Microsoft's DirectX were utilised from within a program written specifically for this task. This meant that while one image was displayed, the next was being loaded in the background ready to appear instantly when the button was next pressed.

The presentation itself consisted of a cycle of three images. The first was a single point calibration screen shown in Figure 6.5b that allowed the location of the centre point in the screen to be identified. Participants were instructed to fix their gaze on this point for a few seconds before moving onto the next screen. This was to account for any drift away from the original calibration screen as the presentation proceeded. The second screen in the cycle was the cytological image that the participant examined for as long as they needed to before making a classification. When they were ready to classify the image, they then moved on to the third image in the cycle, shown next to a sample cytological image in Figure 6.6, called the decision grid. Again participants were instructed to pause and fixate upon the correct classification for the image they had just viewed before moving on.

All possible classifications for the UK screening program are included on this slide. In order from left to right these are (top row) Inadequate Specimen, Negative (Within Normal Limits), Borderline Changes, (middle) Mild Dyskaryosis, Moderate Dyskaryosis, Severe Dyskaryosis, (bottom row) Severe Dyskaryosis/?Invasive Cancer, ?Glandular Neoplasia and a further 'other' category has been added to act as a catch-all should the participant decide that none of the existing categories are representative of the image's contents.

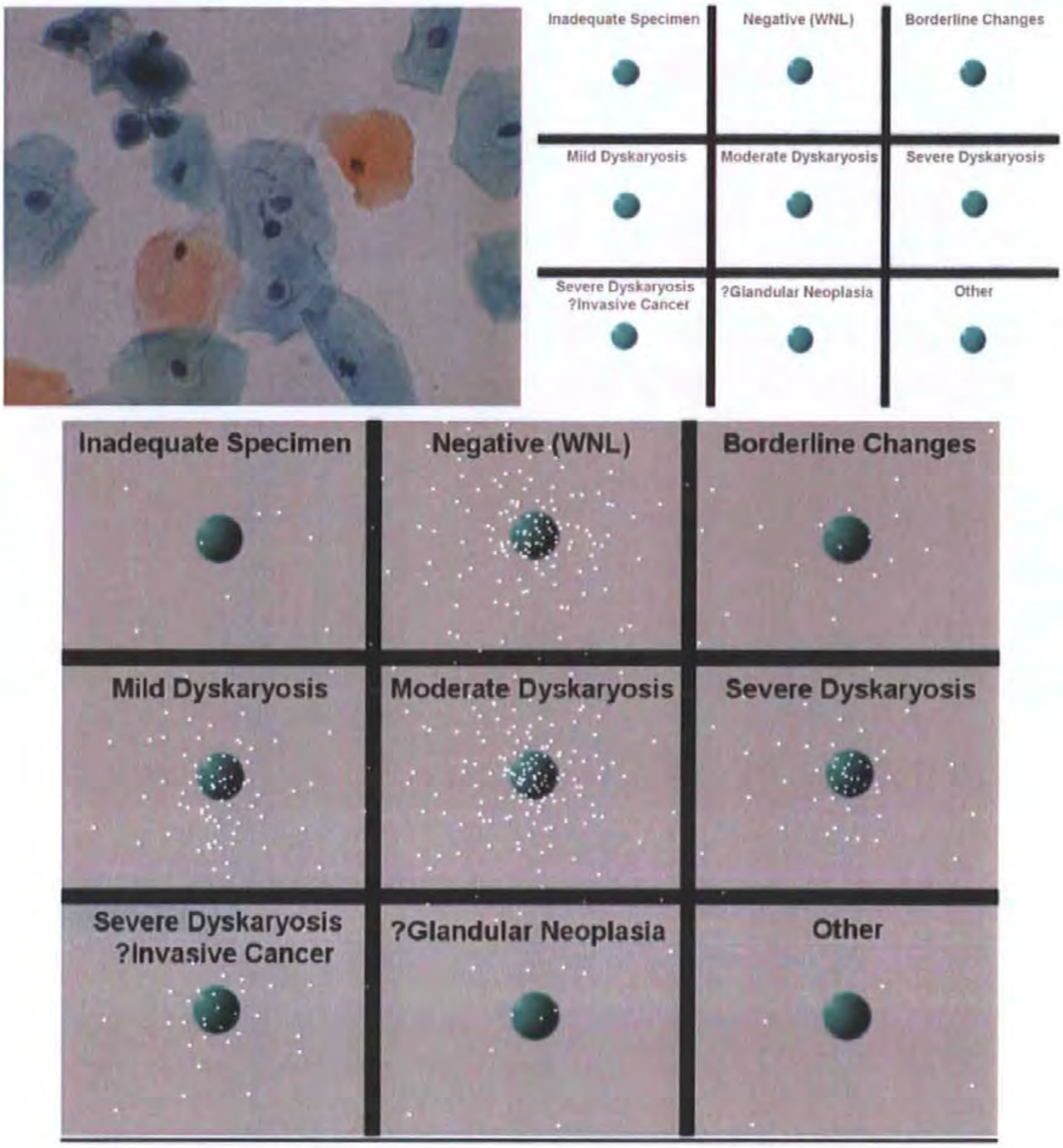


Figure 6.6 A sample image (top left), the decision grid used for recording the participants classification (top right) and an example of a decision grid overlaid with eye tracker fixations (bottom)

A total of 50 of these cycles were presented, containing 25 images taken from Papanicolaou slides and 25 taken from ThinPrep slides. These were alternated throughout to avoid any performance biases that might have existed had they been presented in larger blocks. The complete procedure took approximately 20 minutes, including the initial calibration and bite bar creation process although this varied from individual to individual as no time limit was set for responses. This allowed each person the time they needed to consider each slide properly before making a decision on it and ensured that the eye tracker recorded the areas of each image that was fixated on in order to make a completely informed decision regarding the classification.

When the participant had recorded classifications for all 50 images, the final screen notified them that the trial had ended. The eye tracker recording was stopped and the helmet removed. The bite bar was removed from the crossbar ready for the next bite bar to be put in place. The participant was then given a few minutes before continuing with the feature-marking task.

6.4.3 Feature Marking Task Procedure

Once participants had completed the eye-tracking task and had a few minutes to relax at its conclusion, they then had the briefing repeated to them to refresh the instructions for this second task. They were again reminded of their right to withdraw from the study and asked if they had any further questions. Before commencing, a short demonstration on how to operate the feature marking software was given and when the participant was happy that they knew how to complete the task, the trial began. A computer separate from the two computers used for the eye tracking experiments ran the feature marking software to make sure that the transition between the two tasks could take place quickly, easily and without disrupting the eye tracker software.

The feature-marking task again used software written specifically for the purpose that allowed the user to browse through the same image set used in the eye-tracking task and manually mark any abnormalities on them. The coordinates of each of these marks are then recorded into a text document for later analysis. Each participant was requested to mark the centre of any abnormal areas that were seen on the image. Again, there were no time limits or limit on the number of abnormalities that could be indicated as it was felt this would hinder the process. An example screenshot of this program can be seen in Figure 6.7, which shows the software has a list of the images on the left. When an image is selected from this list, it appears in the main window and can then be marked by the participant with a computer mouse. The mouse pointer, when clicked on the image, leaves a white dot behind and the coordinate information was recorded into a text file for later analysis. When a participant had marked all of the abnormalities present in the image, they moved onto the next one until all 50 from the original image set had been marked in this way.

When the task had been completed, each participant was given a full debriefing (Appendix D), which included details of the study and contact address in order to stay informed about the progress of the study. In its entirety the feature-marking task took about 10 minutes to complete although this did vary depending on the individual completing the task because no limit was set. Combined with the eye-tracking task the whole trial took approximately 30 minutes to complete.

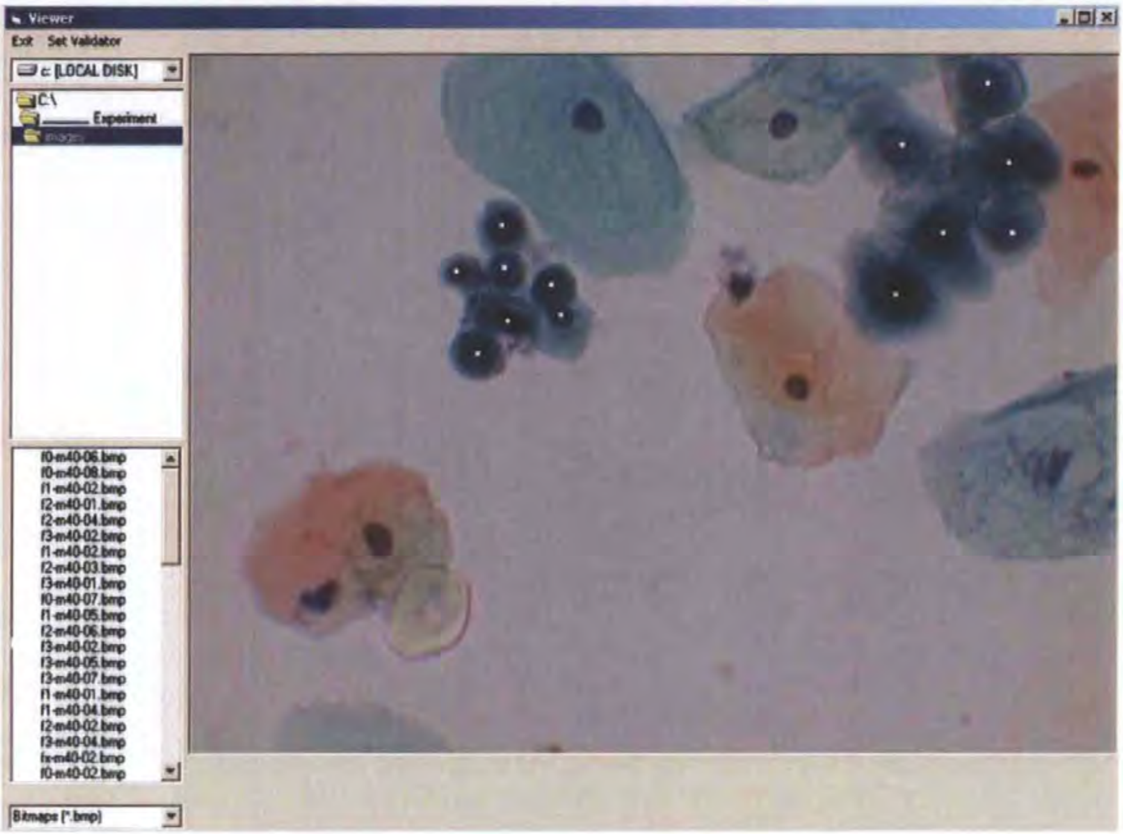


Figure 6.7 The feature marking software records on each image the location of abnormalities as marked by the participants (abnormalities are indicated by white dots).

6.4.4 Machine Colour Texture Analysis Procedure

The machine analysis aspect of the experimental work did not involve human participation but will be described here, as it is an integral part of the research procedure. The type of machine analysis employed has a vital bearing on the results that are obtained and therefore the feasibility of the system that is proposed. In order to provide a system that is capable of assessing the saliency of features in novel slide images, certainty is needed that the machine analysis is locating the most interesting features within the image data. There are many available methods for this, with more recently a larger emphasis on combining colour and texture measures. Both colour and texture are very important factors in this research. When a cervical smear is taken the cells are stained to make the task of differentiating between the different cellular matters easier. While cellular texture alone may be useful, there is no doubt that the information available to the viewer is enhanced by the use of colour.

To produce a list of texture features the image is decomposed into a hue and a saturation/value combined components. When a slide is viewed it is illuminated by the microscope's back-light and this is adjustable depending on personal preference. This means that hue is relatively stable while both saturation and value can vary considerably dependent upon the amount of illumination used. Hue texture would therefore be expected to be the superior measure.

An Atrous wavelet transformation is used to identify the location of energy maxima that relate to features at various resolutions within the images used in this study (Bijaoui, Starck and Murtagh, 1994). This method is employed in the recognition of marine microplankton from images of seawater, where successful categorisation of morphologically similar species has been demonstrated (Toth, L. and Culverhouse, 1999; Culverhouse, Williams, Reguera, Ellis and Parisini, 1996). The use of four and thirteen element vector image analysis is described in Wang and Culverhouse (2004) and applied to texture-based plankton recognition. It has been shown that this methodology may also be appropriate for cervical smear image analysis (Potter, 1999). Each of the maxima is checked against the co-ordinates of the eye-tracker fixations to look for proximity to features that are classified as abnormal, normal, or if no proximity is found then it is classed as containing unimportant features as they have not been viewed by screeners while assessing the slides.

6.4.5 Data Verification Procedure

The experimental procedure is designed in such a way that each of the three data sources recorded both verify and make inference about the other two. This verification of the data allows objectivity when results are produced. A diagrammatic overview of the model shows how the data interacts can be seen in Figure 6.8.

With reference to Figure 6.8, Initial recordings are made of each screener's eye fixations **(1)** and the abnormal features indicated by the screeners **(2)**. On its own the eye tracker data only tells us where on the image someone has fixated. In order to make sense of this data, the feature marking data is used to label each of the fixation points **(3)** depending on whether it is in the proximity of an abnormal feature, a salient feature, or no features of interest. This information is then compiled into a saliency index **(4)** where all of the fixation co-ordinates are ranked. The highest ranking is given to the co-ordinates of the most abnormal features, followed by salient features, and the lowest ranking given to fixations that do not relate to a feature of interest. Within each of these groups, the order is dictated by the number of fixations located within a five degree visual angle of these features. The highest ranking overall will be given to an area that has been marked by all participants as abnormal, and has also been viewed and fixated upon the highest number of times. The lowest ranking overall will be given to co-ordinates that have been viewed by a solitary person but are not in the vicinity of any manually marked features. During the process of compiling the first saliency index, descriptive statistics and sensitivity levels are produced for each participant.

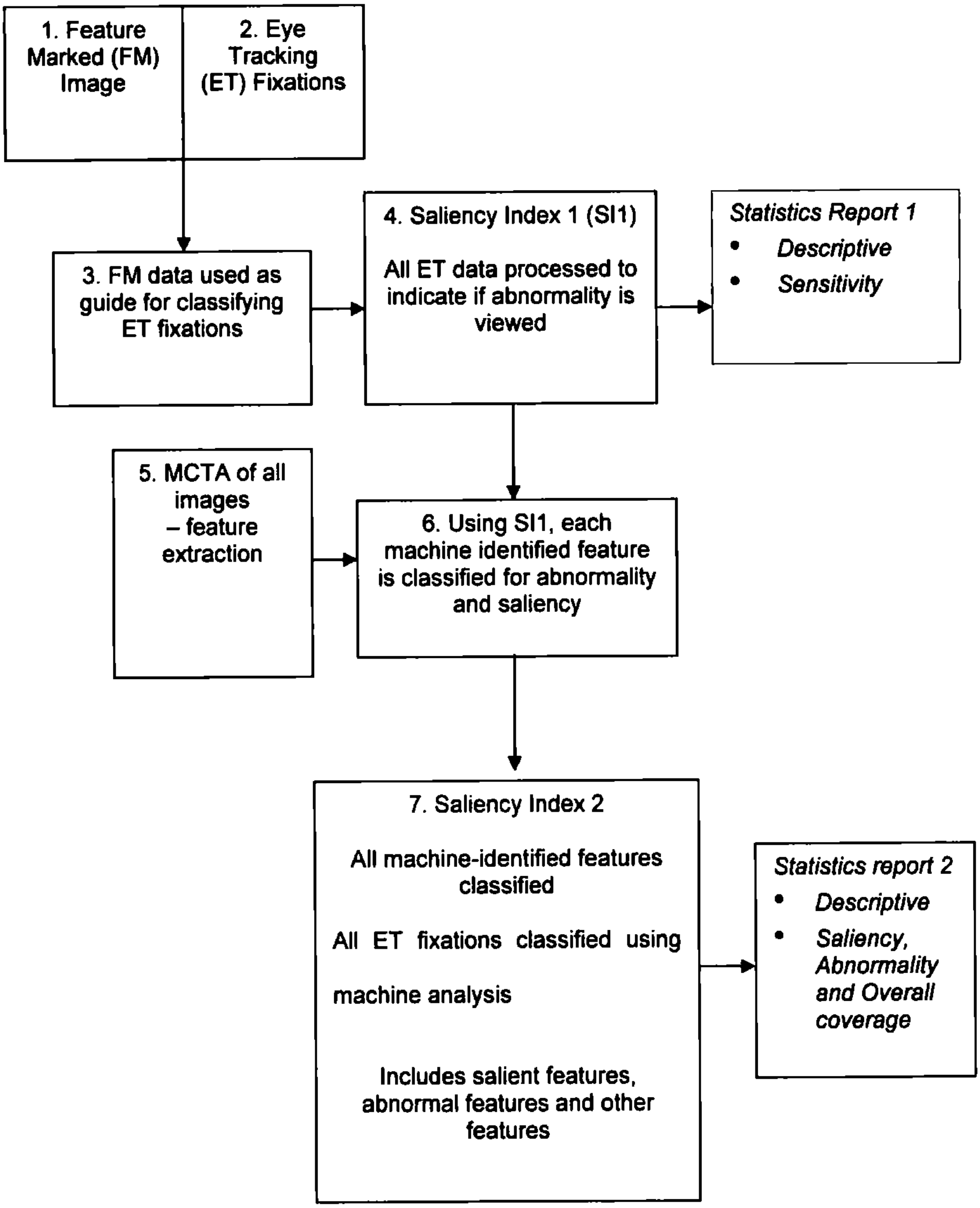

1. Feature Marked (FM) Image
2. Eye Tracking (ET) Fixations
3. FM data used as guide for classifying ET fixations
4. Saliency Index 1 (SI1)
All ET data processed to indicate if abnormality is viewed
Statistics Report 1
• Descriptive
• Sensitivity
5. MCTA of all images – feature extraction
6. Using SI1, each machine identified feature is classified for abnormality and saliency
7. Saliency Index 2
All machine-identified features classified
All ET fixations classified using machine analysis
Includes salient features, abnormal features and other features
Statistics report 2
• Descriptive
• Saliency, Abnormality and Overall coverage

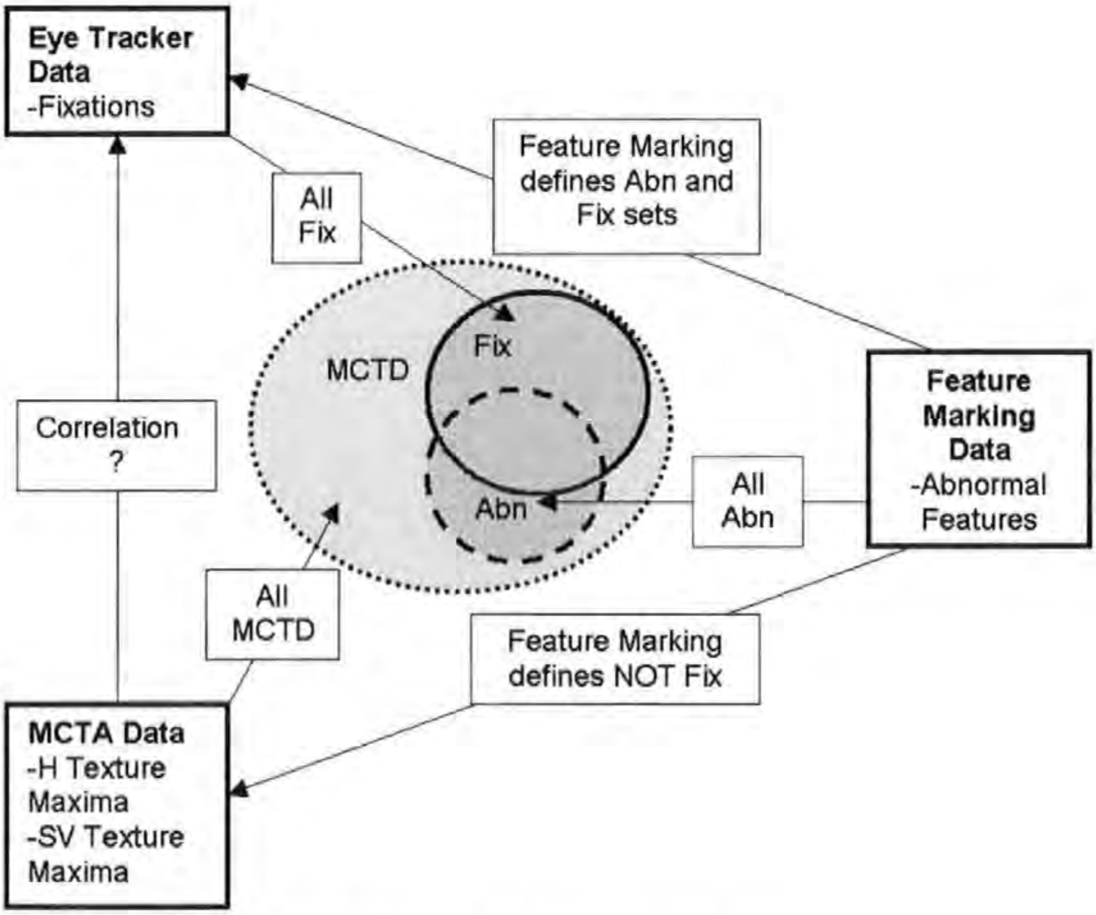
Figure 6.8. An overview of how the different data types are treated during the analysis

Separately to this process, a machine colour texture analysis (MCTA) is carried out on the images **(5)** and this produces a list of Atrous maxima texture feature co-ordinates based on image data across different spatial resolutions. This can then be cross-referenced with the saliency index **(6)**, where each of the machine identified features are ranked based upon the levels of abnormality and saliency shown in the first index. This produces a second index **(7)** which contains information on the saliency of each of the machine identified features, and saliency of each of the eye fixations. The highest ranking is given to a machine identified feature that is located in the same region as a high ranking co-ordinate from the first index. The lowest ranking will be given to a machine identified feature that is not in the area of any of the co-ordinates from the first index. During the process of producing this second index, further descriptive statistics are produced, including saliency, abnormality and overall image coverage for each of the participants.

This process allows the eye tracker data and the MCTA data to be verified by the feature marked data. The feature marking is used to label each of the eye tracking fixations and to produce descriptive statistics that ensure that there are no unusual or unexpected trends or results. Both of these data types are then used to verify the MCTA and ensure again that there are no unexpected or unusual trends within the data that would otherwise indicate that something other than saliencies and abnormalities were being ranked into the second index. Once verified, this second index produces an objective test of the MCTA.

Shown in Figure 6.9 is a diagram that demonstrates exactly how all the data types overlap. Specifically, the following statements can be made of the data:

- Eye Tracker Data points are all contained within *Fix* (Fixations)
- Feature Marking Data points are all contained within *Abn* (Abnormal Features)
- Eye tracker data points that are not contained within *Abn* and *Fix* must be *Fix* but not *Abn*
- Furthermore, *Fix* are all contained within *MCTD* (Machine Colour Texture Data), and *Abn* are also all contained within *MCTD* so therefore *Abn* not contained within *Fix* are *Abn* that have not been indicated by the Eye Tracking Data.



Where MCTD = Machine Colour Texture Data space

Fix = Fixation space

Abn = Abnormal Feature space

Figure 6.9 The interaction of data during the verification process.

Each of the data sets (Fix, Abn and MCTA) are defined by X, Y coordinates produced during the experimental work from one of the three data sources and cover the whole of the image field. For eye tracking, the data consists of the X, Y coordinates recorded during the image presentation. For the feature marking data, it consists of X, Y coordinates manually marked by the participants across each of the images and MCTA consists of multi-element texture vectors from a set of machine-generated coordinates over the field of view.

6.5 Results

In order to make sense of the results and examine the relationships between the conditions and variables, the participants are split into two groups based on their experience. Expertise levels varied between a few months through to over 20 years of screening experience and because we would expect those with more experience to outperform those with less experience, splitting the group allows for exploring both reliability and validity of the experimental method. The exact experience level of each of the participants is not reported to preserve the anonymity of those taking part.

6.5.1 Performance Results

The sensitivity levels for each participant are derived using a calculation that is designed to emulate as closely as possible the existing method of calculating sensitivity for screeners. The original method of calculating this statistic can be seen in the UK screening guidelines (National Heath Service Cancer Screening Programme, 2000) and this is reproduced below in Table 6.1. While sensitivity and moderate+ sensitivity are calculated in a similar way, where sensitivity* = (A+B / A+B+D+E) x 100 and where moderate*+ sensitivity = (A /

A+D) x100, there is an important difference in how the final performance percentage is calculated. A slide is reviewed and a final report produced and this would also be taken into consideration when giving a final sensitivity score to a person being assessed. In this case a final clinical report on classification is not available when making the calculation so instead of using the final report, the image Gold Standard classification is used. All of the analyses presented in this thesis use the modified sensitivity* and moderate*+ sensitivity method of measuring performance unless otherwise stated.

Table 6.1 Revised sensitivity* key for performance calculations adapted

		Prior Image Classification		
		Abnormal		Normal
		Moderate+	Borderline /Mild	Negative /Inadequate
Participants Classification	Abnormal	A	B	C
	Normal	D	E	F

The result of splitting the two groups up based on their experience can be seen in Table 6.2. This shows that in every condition the more experienced participants outperformed the least experienced. Standard deviations for all of the conditions are also larger for the least experienced groups indicating a larger distribution of scores. This is further reflected in a larger standard error of measurement.

The mean differences are shown graphically in Figure 6.10 where the consistent increase in performance levels seen in the higher experience groups across all conditions are evident. Furthermore we can see that performance was better for all participants when viewing Papanicolaou slide images than for ThinPrep slide images.

Table 6.2 Sensitivity results presented according to slide contents and experience

Condition	Exp Level	N	Mean	SD	SEM
All	High	5	86	4.85	2.17
	Low	5	75.6	10.38	4.64
All Mod+	High	5	82.2	10.11	4.52
	Low	5	74.4	17.27	7.72
Pap	High	5	95.8	3.83	1.71
	Low	5	86.8	14.02	6.27
Pap Mod+	High	5	96.4	8.05	3.6
	Low	5	92	10.95	4.9
TP	High	5	75.2	11.62	5.2
	Low	5	68	18.07	8.08
TP Mod+	High	5	75.2	13.86	6.2
	Low	5	64.2	21.2	9.48

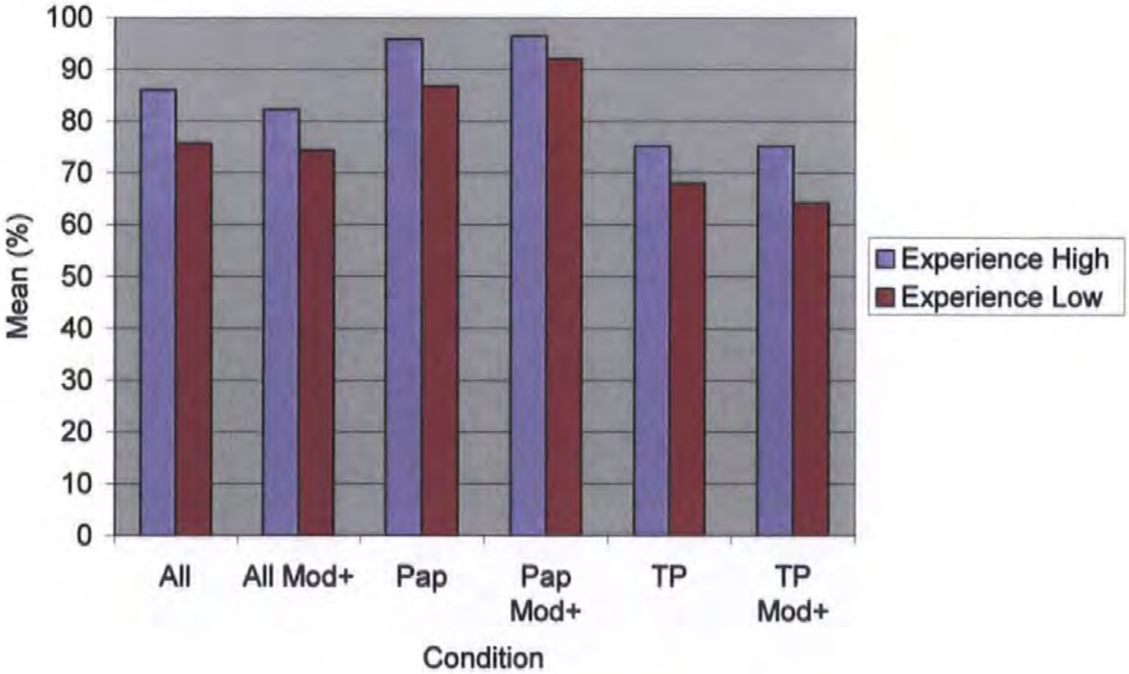


Figure 6.10 Mean differences across conditions based on experience

The data met the assumptions of an independent samples t-test and the results for this comparison between the group means can be seen in Table 6.3. This shows that there were no significant differences between the two groups for any of the conditions. The wide confidence intervals indicate that more data should be collected before any strong conclusions can be drawn.

Table 6.3 independent samples t-test results

Condition	t	df	Sig. (two tailed)	Mean Difference	Std Error Difference	95% Confidence Interval	
						Lower	Upper
All	2.029	8	.077	10.4	5.12	-1.41	22.21
All Mod+	.872	8	.409	7.8	8.95	-12.84	28.44
Pap	1.384	8	.204	9	6.5	-5.99	23.99
Pap Mod+	.724	8	.490	4.4	6.08	-9.6	18.41
TP	1.144	8	.286	11	9.6	-11.16	33.16
TP Mod+	.636	8	.543	7.2	11.33	-18.92	33.3

While no significant differences are evident between the two groups for any of the conditions we can be encouraged by the fact that every single trend in the data is in the direction that would be expected. Those with more experience recorded higher levels of sensitivity throughout and both groups performed better with Papanicolaou slide images than with ThinPrep. While ThinPrep slide images are easier to resolve visually as they do not contain clutter or occluded objects to identify, the participants had previously had very little experience of their analysis. For this reason significance testing between Papanicolaou and ThinPrep has not been carried out as this could only indicate that participants performed better on the slide images they have experience of classifying.

The sensitivity and moderate+ sensitivity scores are very positive as they verify the basic methodology for recording screeners. Given that all of the participants would be expected to score 90-95% sensitivity on Papanicolaou slides during normal quality assurance testing, the fact they achieved sensitivity over

90% overall, rising to 96% for moderate+ supports the model. In every condition it was found that the more experienced participants outscore those with least experienced, and they perform better on Papanicolaou images than on ThinPrep images. It should be noted that the participants had no specific training and very little exposure to ThinPrep slides at the time of the study, so a much lower score on these images was expected. Using this evidence it has been shown that the data recorded from the eye-tracker does indeed reflect the capabilities of the individual, and that the analysis method and calibration technique being employed do not have an adverse effect on the data. This is a key part of the research, as it suggests that, using the data from feature marked images, the saliency of each of the eye-tracked fixations can be robustly predicted.

6.5.2 Image Coverage

Image coverage refers to the fixations made by each individual while viewing each image. When an individual views the images, details of their fixations are recorded and the order in which these occurred. This list records the individual's eye scan path around the image. On each of those fixation points a black circular area is overlaid and this is repeated until all of the fixation points are accounted for. The ratio of black to white on the image is then calculated giving a percentage of total image coverage. This procedure is performed for varying visual angles, as there is some debate as to which is most appropriate. At a visual angle of 2° a participant has viewed those areas covered. At 5°, the angle that is used for most of the analysis, it is still certain that these areas will have been viewed. At 10° and 20° this becomes open to debate, and generally relies on the nature of what is being viewed. A single large object in the field of view can lend itself to being assessed at these angles and as many of the slides contain cells or clusters of cells in isolation with background filling the rest of the image, it was decided to include measurements for all possibilities. In all cases the angle refers to the total diameter of the area being viewed. For instance, at 5°, the area that is 2.5° visual angle around the central fixation point is considered.

An original image is shown in Figure 6.11 (a) and its corresponding fixation-created coverage image for 2° (b) and 10° (c) foveal areas. These examples demonstrate how it is possible to examine a screening strategy using image coverage. The cluster of points shown in the bottom left of (b) does not expand in (c) proportionally to larger coverage compared to the sparsely distributed points throughout the rest of the image. This enhances the information recorded for each screener and facilitates inferences about individual screening strategy.

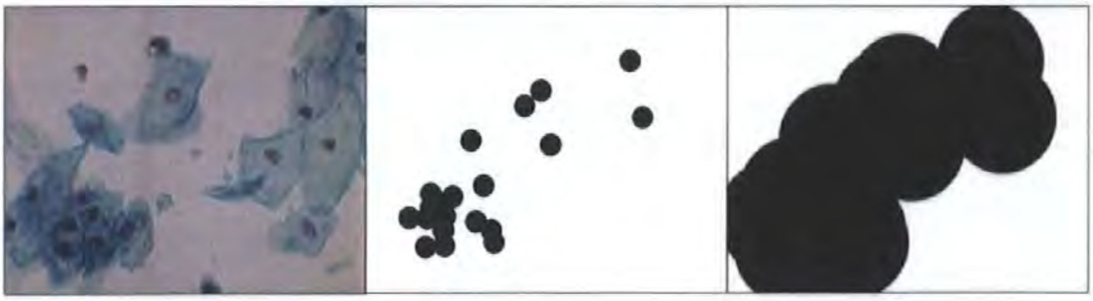


Figure 6.11 (a) An image (b) an example of its corresponding image coverage for one of the participants at two degrees of visual angle and (c) an example at ten degree visual angle coverage.

The performances for the two groups across a number of conditions can be seen in Table 6.4. This shows that generally, those with higher levels of experience cover less of the image than those with less experience. The data met the assumptions of an independent samples t-test and the results for this comparison between the group means can be seen in Table 6.5. The results show that there were no significant differences between the two groups for any of the visual angles investigated. Wide confidence intervals again indicate that more data should be collected before strong conclusions can be drawn.

Table 6.4 Image coverage percentages for each different condition across four different visual angles

	2 degrees		5 degrees		10 degrees		20 degrees	
	Experience Level							
Condition	High	Low	High	Low	High	Low	High	Low
All	4.67	5.595	21.015	23.48	48.855	52.065	83.25	85.52
All Mod+	4.69	5.65	21.1	23.61	49.185	52.035	83.505	84.905
Pap	4.69	5.81	21.45	24.48	50.14	53.53	83.53	84.93
Pap Mod+	4.765	5.895	21.73	24.505	50.695	53.59	84.52	85.18
TP	4.6	5.29	20.43	22.14	47.42	49.56	82.17	83.66
TP Mod+	4.8	5.385	21.215	22.55	48.745	50.2	82.86	84.395

Table 6.5 Independent samples t-test results for groups based on levels of experience

Visual Angle	t	df	Sig. (two tailed)	Mean Difference	Std Error Difference	95% Confidence Interval	
						Lower	Upper
2°	-.830	5.47	.441	-9.4	1.13	-3.78	1.9
5°	-.601	5.67	.571	-2.48	4.13	-12.72	7.76
10°	-.407	8	.695	-3.04	7.47	-20.27	14.18
20°	.303	8	.770	-1.98	6.5	-17.04	13.08

While Table 6.5 compares experience levels across different all visual angles, a question remains about the validity of combining the data in this way. In order to investigate further the possible relationships that exist within the data set, each visual angle has been tested under a number of conditions based on the type of image being viewed and subsequent classification. It is important to separate ThinPrep from Papanicolaou images, correctly classified and incorrectly classified, and sensitivity based on standard classifications and moderate+ classifications. The results from significance testing between experience levels for all possible combinations can be seen in Tables 6.6, 6.7, 6.8 and 6.9. This shows that there were no significant differences between any of the conditions.

Table 6.6 Independent Samples Test - 2 degrees

Condition	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
All Correct	.740	5.322	.491	.0074	.01003	-.01791	.03275
All Incorrect	.899	5.350	.408	.0111	.01231	-.01997	.04209
All M+ Correct	.707	4.950	.512	.0077	.01092	-.02044	.03588
All M+ Incorrect	.947	8	.371	.0114	.01203	-.01635	.03915
All Papanicolaou	.979	5.144	.371	.0112	.01142	-.01793	.04029
All ThinPrep	.598	5.831	.572	.0068	.01143	-.02133	.03501
Papanicolaou Correct	.074	4.621	.944	.0011	.01439	-.03685	.03897
Papanicolaou Incorrect	1.113	4.745	.319	.0142	.01276	-.01914	.04754
Papanicolaou M+ Correct	.687	5.206	.521	.0083	.01211	-.02244	.03908
Papanicolaou M+ Incorrect	1.241	5.275	.267	.0143	.01151	-.01485	.04341
ThinPrep Correct	.718	5.502	.502	.0071	.00989	-.01764	.03184
ThinPrep Incorrect	.518	8	.619	.0060	.01155	-.02066	.03262
ThinPrep M+ Correct	.500	5.197	.637	.0051	.01016	-.02074	.03090
ThinPrep M+ Incorrect	.519	8	.618	.0065	.01259	-.02250	.03558

Table 6.7 Independent Samples Test - 5 degrees

Condition	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
All Correct	.622	5.658	.558	.0226	.03635	-.06765	.11289
All Incorrect	.624	5.371	.558	.0267	.04282	-.08112	.13452
All M+ Correct	.492	5.176	.643	.0197	.04001	-.08211	.12151
All M+ Incorrect	.720	8	.492	.0305	.04243	-.06730	.12838
All Papanicolaou	.723	5.374	.500	.0303	.04194	-.07526	.13594
All ThinPrep	.422	5.848	.688	.0171	.04059	-.08282	.11710
Papanicolaou Correct	-.136	8	.895	-.0081	.05949	-.14525	.12909
Papanicolaou Incorrect	.760	4.800	.483	.0341	.04483	-.08264	.15076
Papanicolaou M+ Correct	.404	5.537	.702	.0188	.04663	-.09762	.13526
Papanicolaou M+ Incorrect	.922	5.603	.395	.0367	.03981	-.06241	.13581
ThinPrep Correct	.589	5.810	.578	.0205	.03474	-.06523	.10615
ThinPrep Incorrect	.374	8	.718	.0154	.04128	-.07977	.11061
ThinPrep M+ Correct	.288	5.624	.784	.0104	.03595	-.07906	.09978
ThinPrep M+ Incorrect	.365	8	.725	.0164	.04487	-.08711	.11983

Table 6.8 Independent Samples Test - 10 degrees

Condition	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
All Correct	.622	5.658	.558	.0226	.03635	-.06765	.11289
All Incorrect	.624	5.371	.558	.0267	.04282	-.08112	.13452
All M+ Correct	.492	5.176	.643	.0197	.04001	-.08211	.12151
All M+ Incorrect	.720	8	.492	.0305	.04243	-.06730	.12838
All Papanicolaou	.723	5.374	.500	.0303	.04194	-.07526	.13594
All ThinPrep	.422	5.848	.688	.0171	.04059	-.08282	.11710
Papanicolaou Correct	-.136	8	.895	-.0081	.05949	-.14525	.12909
Papanicolaou Incorrect	.760	4.800	.483	.0341	.04483	-.08264	.15076
Papanicolaou M+ Correct	.404	5.537	.702	.0188	.04663	-.09762	.13526
Papanicolaou M+ Incorrect	.922	5.603	.395	.0367	.03981	-.06241	.13581
ThinPrep Correct	.589	5.810	.578	.0205	.03474	-.06523	.10615
ThinPrep Incorrect	.374	8	.718	.0154	.04128	-.07977	.11061
ThinPrep M+ Correct	.288	5.624	.784	.0104	.03595	-.07906	.09978
ThinPrep M+ Incorrect	.365	8	.725	.0164	.04487	-.08711	.11983

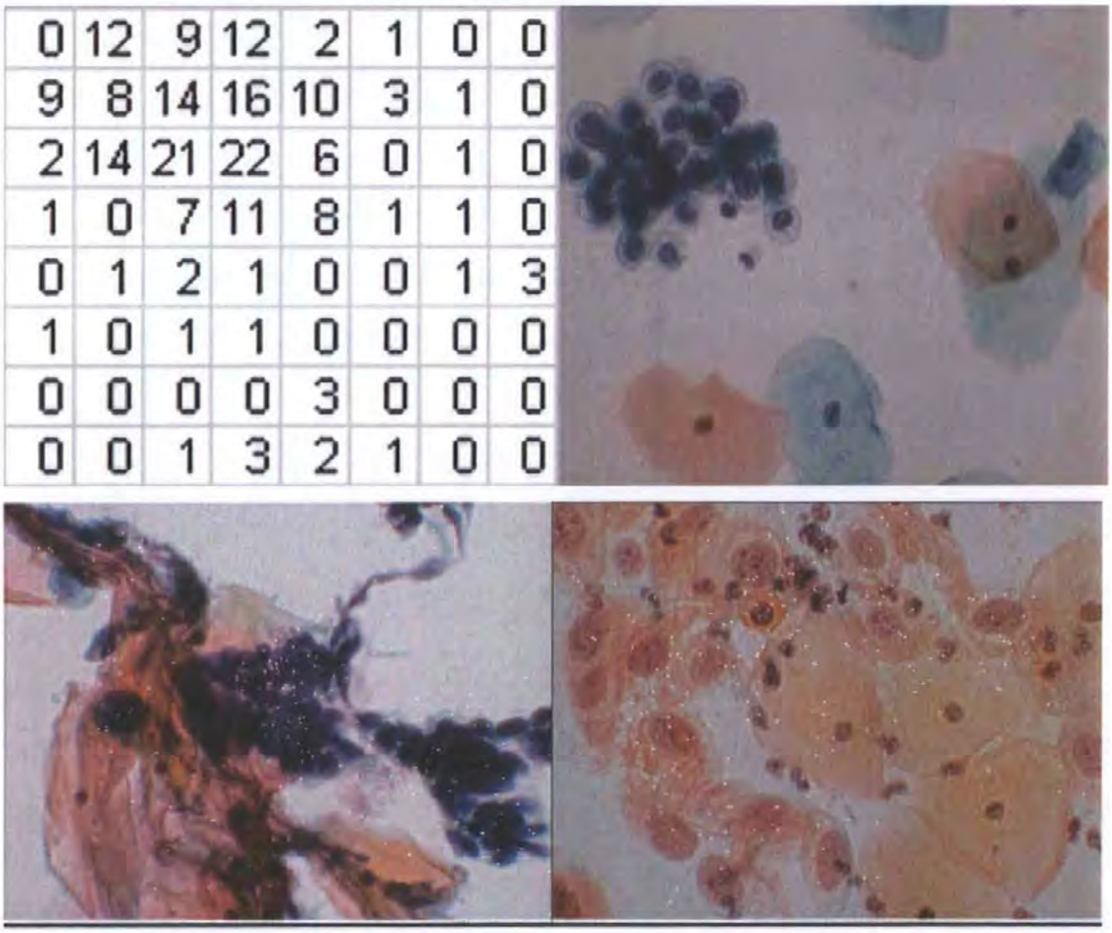
Table 6.9Independent Samples Test - 20 degrees

Condition	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
All Correct	.777	8	.459	.0397	.05104	-.07802	.15738
All Incorrect	.100	8	.923	.0057	.05688	-.12547	.13687
All M+ Correct	.206	8	.842	.0128	.06222	-.13067	.15627
All M+ Incorrect	.279	8	.787	.0152	.05449	-.11043	.14087
All Papanicolaou	.216	8	.835	.0140	.06509	-.13605	.16413
All ThinPrep	.238	8	.818	.0149	.06280	-.12991	.15975
Papanicolaou Correct	-.770	8	.464	-.1407	.18287	-.56241	.28097
Papanicolaou Incorrect	.089	8	.931	.0052	.05864	-.12998	.14046
Papanicolaou M+ Correct	-.275	8	.790	-.0225	.08178	-.21110	.16610
Papanicolaou M+ Incorrect	.813	8	.440	.0357	.04391	-.06557	.13693
ThinPrep Correct	1.116	8	.297	.0526	.04716	-.05612	.16136
ThinPrep Incorrect	.019	8	.985	.0011	.05612	-.12836	.13048
ThinPrep M+ Correct	.780	8	.458	.0390	.05003	-.07632	.15440
ThinPrep M+ Incorrect	-.126	8	.903	-.0083	.06569	-.15977	.14321

Although there were no significant differences between the groups it is interesting to note that in every condition, those with more experience viewed a smaller area of each images. This could reflect an evolving strategy toward screening that changes as more experience is gained. As reported in Chapter 3, experts do not consider more information than novices, but do select more salient information on which to base an expert judgement. This would be consistent with this finding as it would suggest that those with more experience are covering less area of each image in order to make a better decision and subsequently record the higher sensitivities shown earlier.

6.5.3 Saliency Coverage

The Saliency Coverage function allows us to examine how many of the most important areas of the image have been viewed. This is achieved by firstly deducing what constitutes a salient area. From the eye tracker data analysis it is known what areas of each slide have been viewed most frequently regardless of whether they are abnormal or not. In order to rate an individuals' performance on saliency, an 8x8 grid is created and overlaid on the fixation map for each image and then the number of fixations in each square is counted using the second saliency index (see Figure 6.8 item 7). An individual's fixation file is then opened and each fixation compared to this map. The number of areas considered to be salient are those within one standard deviation of the mean number of fixations on each image. This ensured that those images that were quicker and easier to resolve visually, such as those with a large amount of background and an isolated abnormality that attracted fewer fixations, were more stringently tested. A leave-one-out methodology to compare each screener to the map was also investigated as a way of examining the usefulness of the measure. However because of the large number of fixations required in order to make what is considered to be a salient area, the results using each 9-person map were the same as using the 10 person map. Raw data used for this analysis can be found in Appendix K. An example of the 8x8 grid can be seen in Figure 6.12 with the image that it relates to. The higher numbers reflect the cluster of cells in the top left hand corner of the image and it is these areas that are used for the saliency coverage statistics. The method used for this process is not static, and so each time the saliency coverage statistics are calculated, the grid will be updated to reflect any future data added to the original saliency index.



0	12	9	12	2	1	0	0
9	8	14	16	10	3	1	0
2	14	21	22	6	0	1	0
1	0	7	11	8	1	1	0
0	1	2	1	0	0	1	3
1	0	1	1	0	0	0	0
0	0	0	0	3	0	0	0
0	0	1	3	2	1	0	0

Figure 6.12 (top left) The saliency coverage grid with fixation numbers (top right) the image for which the grid was calculated and (bottom) example images with eye tracking fixations indicated by white dots

Individual saliency coverage performances can be seen in Table 6.10. Table 6.11 shows the average coverage percentages for each of the conditions and for both the more and less experienced groups. This shows that the largest difference is observed for the ThinPrep Moderate+ condition and this can be seen graphically in Figure 6.12. The consistent differences between those with high and low levels of experience show that those with more experience cover less of the salient areas before making a decision on the slide's classification. It could suggest that during the visual search, those with more experience locate salient areas on which to base their decision while considering less of the total salient areas than their counterparts. This may also explain why such a difference is seen for ThinPrep moderate+. Given the fact that ThinPrep images are designed to be clearer and therefore easier to interpret, in the moderate+ condition this could account for the exaggerated effect.

Table 6.10 Individual performances for Saliency Coverage

Participant	Salient Areas viewed? (%)	Average Fixations	Average Salient Fixations Per Slide	Saliency Coverage (%)
1	93.3	11.375	4	32.95
2	100	26.04	16.55	65.99
3	100	16.79	7.33	44.13
4	70	7.78	1.8	20.55
5	100	12.92	5.37	42.60
6	94	11.12	4.36	37.75
7	100	23.3	12.18	53.24
8	100	26.7	16.27	61.69
9	100	13.2	6.16	48.29
10	100	20.6	11.64	59.12

Table 6.11 Saliency coverage percentages for each condition

	Experience Level			
Condition	High	Low	Total	Difference
All Slides	45.6	47.65	46.67	2.05
All Mod+	45.6	48.19	46.89	2.59
Pap Slides	45.76	47.73	46.75	1.97
Pap Mod+	46.51	48.18	47.34	1.67
TP Slides	44.81	47.64	46.23	2.83
TP Mod+	43.24	48.75	45.99	5.51

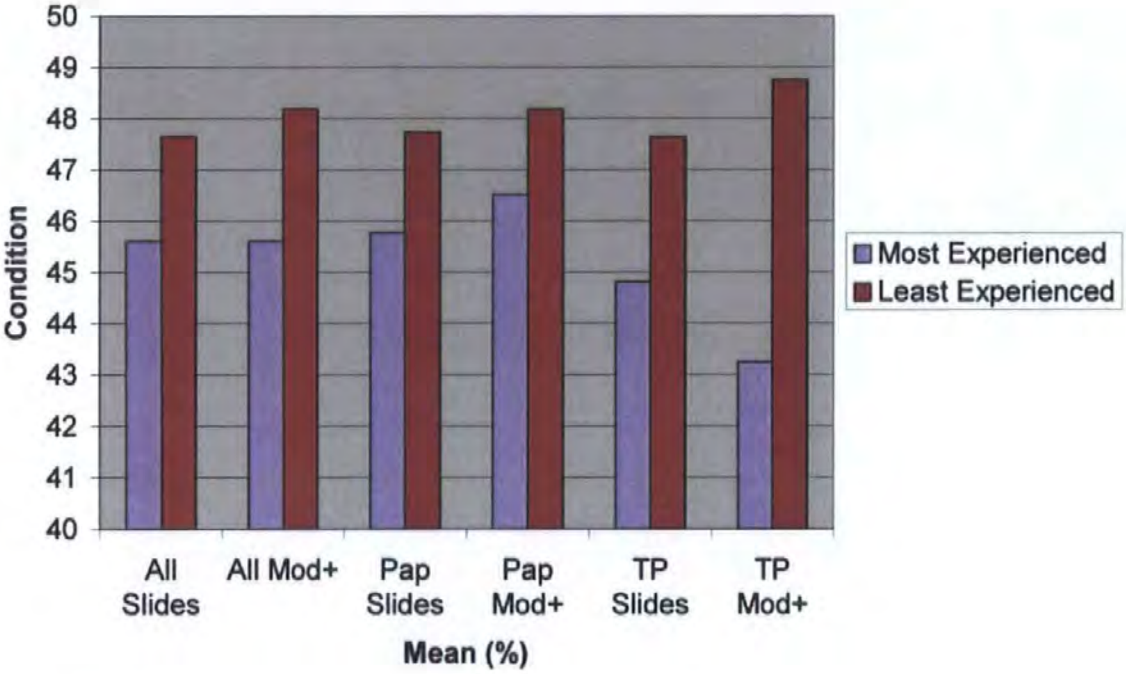


Figure 6.13 Saliency Coverage differences for each image type

The other statistic that is calculated for each participant is whether they have viewed any of the salient areas indicated on each of the slides. This can also be seen in Table 6.10 and shows that 7 of the participants viewed the salient areas on every slide they were shown. The results of Independent Samples t-testing on this and saliency coverage can be seen in Table 6.12. This shows that there are no significant differences between levels of experience for both saliency coverage and the number of salient areas viewed. A wide range for the reported confidence intervals once again demonstrates the need for more data to be recorded before any strong conclusions can be drawn from this result. In order to investigate any potential differences between experience levels further, similar conditions to those applied to the image coverage data were investigated. The descriptive statistics for these conditions are shown in Table 6.13 and presented graphically in Figure 6.14. The results of Independent samples t-tests are presented in Table 6.14.

Table 6.12 Independent samples t-test results

Condition	t	df	Sig. (two tailed)	Mean Difference	Std Error Difference	95% Confidence Interval	
						Lower	Upper
Saliencies Viewed	2.029	8	.077	10.4	5.12	-1.41	22.21
Saliency Coverage	.872	8	.409	7.8	8.95	-12.84	28.44

Table 6.13 Saliency Coverage percentages across conditions

	Most Experienced	Least Experienced
All Correct	46.92%	47.39%
All Incorrect	44.47%	47.91%
All M+ Correct	47.12%	48.98%
All M+ Incorrect	44.08%	47.39%
All Papanicolaou	45.76%	47.73%
All ThinPrep	44.81%	47.64%
Papanicolaou Correct	47.66%	43.32%
Papanicolaou Incorrect	45.41%	47.61%
Papanicolaou M+ Correct	46.48%	48.98%
Papanicolaou M+ Incorrect	46.53%	47.38%
ThinPrep Correct	46.67%	48.06%
ThinPrep Incorrect	42.96%	50.28%
ThinPrep M+ Correct	45.32%	49.61%
ThinPrep M+ Incorrect	41.16%	47.89%

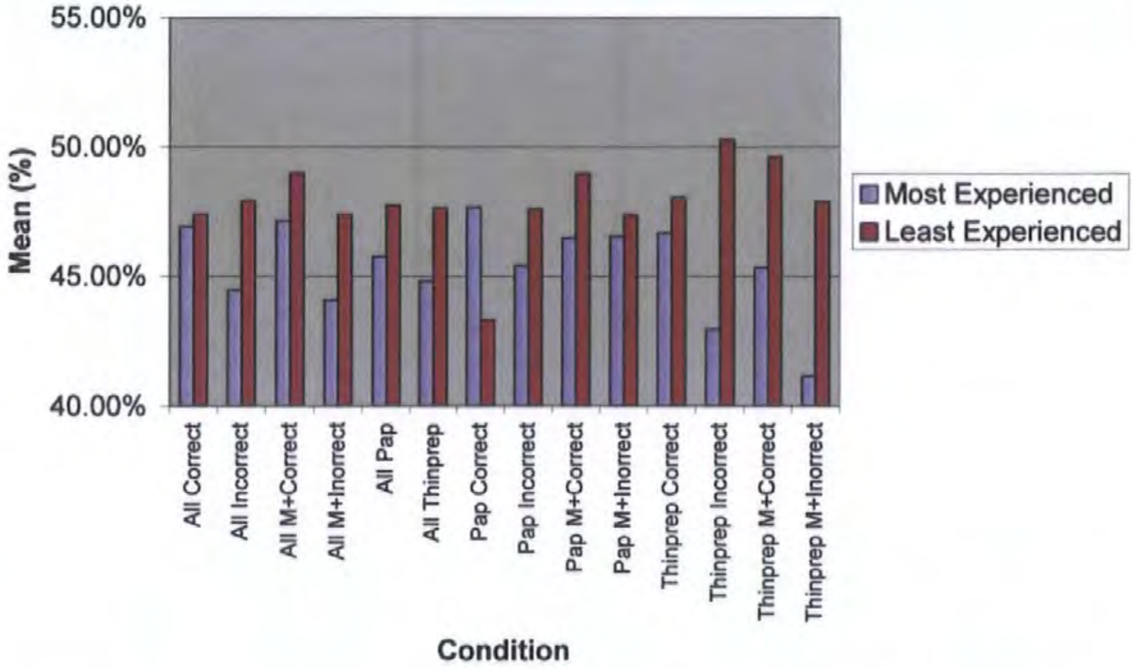


Figure 6.14 **Saliency Coverage percentages across conditions.**

Table 6.14 **Independent Samples Test**

Condition	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
All Correct	.046	8	.964	.0047	.10232	-.23121	.24069
All Incorrect	.392	8	.706	.0344	.08771	-.16790	.23662
All M+ Correct	.200	8	.846	.0185	.09250	-.19476	.23184
All M+ Incorrect	.351	8	.735	.0330	.09413	-.18402	.25011
All Papanicolaou	.227	8	.826	.0197	.08690	-.18067	.22011
All ThinPrep	.276	8	.790	.0283	.10279	-.20870	.26537
Papanicolaou Correct	-.349	8	.736	-.0434	.12440	-.33023	.24349
Papanicolaou Incorrect	.248	8	.810	.0220	.08871	-.18257	.22654
Papanicolaou M+ Correct	.311	8	.764	.0249	.08011	-.15981	.20968
Papanicolaou M+ Incorrect	.077	8	.940	.0084	.10942	-.24387	.26075
ThinPrep Correct	.125	8	.903	.0139	.11095	-.24194	.26975
ThinPrep Incorrect	.746	8	.477	.0732	.09812	-.15305	.29949
ThinPrep M+ Correct	.432	5.36	.682	.0429	.09924	-.20715	.29292
ThinPrep M+ Incorrect	.772	8	.463	.0673	.08725	-.13387	.26852

While the saliency coverage statistics shown in Table 6.14 are not significant, the results appear positive. The fact that the majority of screeners viewed salient areas according to the experimental criteria supports the experimental design being employed. Furthermore, Figure 6.14 presents two noteworthy differences relating to the coverage displayed on correctly classified Papanicolaou images compared to ThinPrep images. The largest differences seen between groups relate to ThinPrep conditions, and suggest that more experienced screeners view less of the salient areas before making a classification decision. In fact this trend is seen for all of the conditions but the difference is less marked than with ThinPrep images. However, the trend is reversed for correctly classified Papanicolaou images indicating that those with less experience viewed less of the slide. While these differences are not significant, and not consistent enough to draw any conclusions, it is possible that this is evidence of an emerging strategic difference depending on the image contents. In order to examine this theory further, more data would need to be collected to either increase the likelihood that these differences become significant or show that these differences are merely random.

6.5.4 Abnormality Coverage

The method used to objectively assess whether the images abnormal areas have been viewed or not is similar in nature to the method used for saliency coverage. In Figure 6.12 we can see how an 8x8 grid is constructed as a way to designate salient areas. A similar grid is constructed to allow the designation of abnormal areas based this time on the abnormal features marked during the feature marking exercise. A participant's fixation file is then compared to the grid generated from the feature marking for each image. The number of areas considered to be abnormal again depends on the standard deviation taken from the number of fixations made on each screen. Again a leave-one-out protocol was investigated but, as with the saliency analysis, it was found that each 9-person map/grid generated the same results as the 10 person map. Raw data used for this analysis can be found in Appendix K. A breakdown of individual performance of abnormality coverage can be seen is in Table 6.15. This shows that levels of abnormality coverage varied far more than they did for saliency coverage. In particular, the simple yes/no question of whether someone had viewed a slides abnormalities produced variance suggests that perhaps the methodology used for saliency coverage is not appropriate for abnormality. This is surprising, as the abnormality coverage measure's use of the feature marking data should provide a stronger basis than the saliency coverage measures use of eye tracking data. Despite this, the variance seen suggests that perhaps a more implicit process is in effect and that it is this implicit categorisation process that the saliency coverage measure utilises.

Table 6.15 Individual Performances for Abnormality Coverage

Participant	Abnormal Areas Viewed? (%)	Average Fixations	Average Abnormal Fixations Per Slide	Abnormality Coverage (%)
1	53.6	11.61	1.72	13.08
2	100	27.97	12.7	46.21
3	75	19.125	2.64	13.37
4	55.2	6.93	1.03	15.21
5	93.1	13	3.53	27.3
6	89.7	12	2.8	24.87
7	100	25.83	8.63	34.21
8	100	26.76	11.6	41.28
9	60.7	13.14	2.03	12.78
10	100	21	8.7	40.92

Further investigation of this abnormality coverage measure reveals a trend that those with more experience of screening cover fewer abnormal areas before making a decision than those with less experience. This is shown in Table 6.16, and represented graphically in Figure 6.15. These differences are again evident when various conditions are imposed on the data as shown in Table 6.17 and Figure 6.16. Independent samples t-tests, shown in Table 6.18, show that all but one of these differences are not significant. The one significant difference that exists is in the ThinPrep Moderate+ Incorrect classifications conditions.

Table 6.16 **Abnormality Coverage average Percentages**

		Mild Dysk	Mod Dysk	Sev Dysk	Sev ?Inv	?GlanNeo
All Screeners	All	27.15	29.90	29.03	29.68	26.75
	Papanicolaou	27.53	33.59	n/a	29.68	n/a
	ThinPrep	25.26	28.51	29.03	n/a	26.75
Most Experienced	Papanicolaou	21.78	31.84	n/a	26.93	n/a
	ThinPrep	21.83	26.01	26.33	n/a	19.18
Least Experienced	Papanicolaou	33.29	35.97	n/a	32.40	n/a
	ThinPrep	28.01	30.49	31.32	n/a	34.20

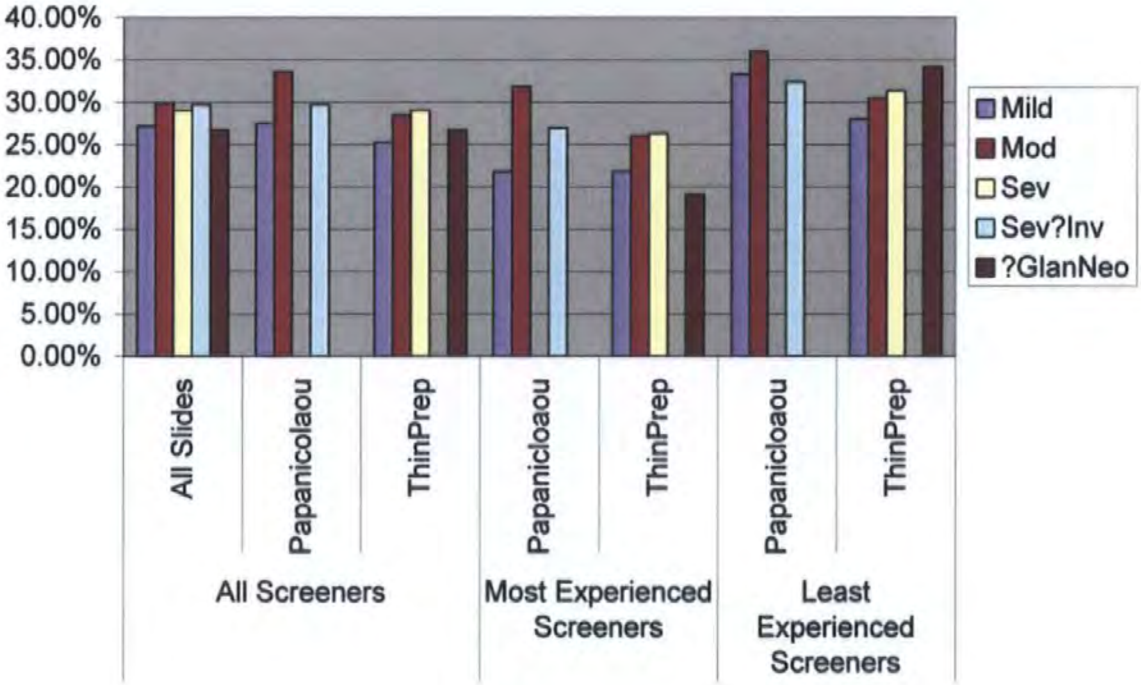


Figure 6.15 **Abnormality Coverage average Percentages**

Table 6.17 Abnormality Coverage percentages across conditions

	Most Experienced	Least Experienced
All Correct	24.13%	28.96%
All Incorrect	20.67%	31.97%
All M+Correct	21.26%	31.22%
All M+Incorrect	22.16%	32.11%
All Papanicolaou	20.91%	32.19%
All ThinPrep	22.68%	30.64%
Papanicolaou Correct	23.73%	30.05%
Papanicolaou Incorrect	20.24%	31.29%
Papanicolaou M+Correct	20.79%	33.16%
Papanicolaou M+Incorrect	22.97%	29.54%
ThinPrep Correct	24.40%	27.00%
ThinPrep Incorrect	19.71%	32.35%
ThinPrep M+Correct	23.40%	27.18%
ThinPrep M+Incorrect	14.26%	33.56%

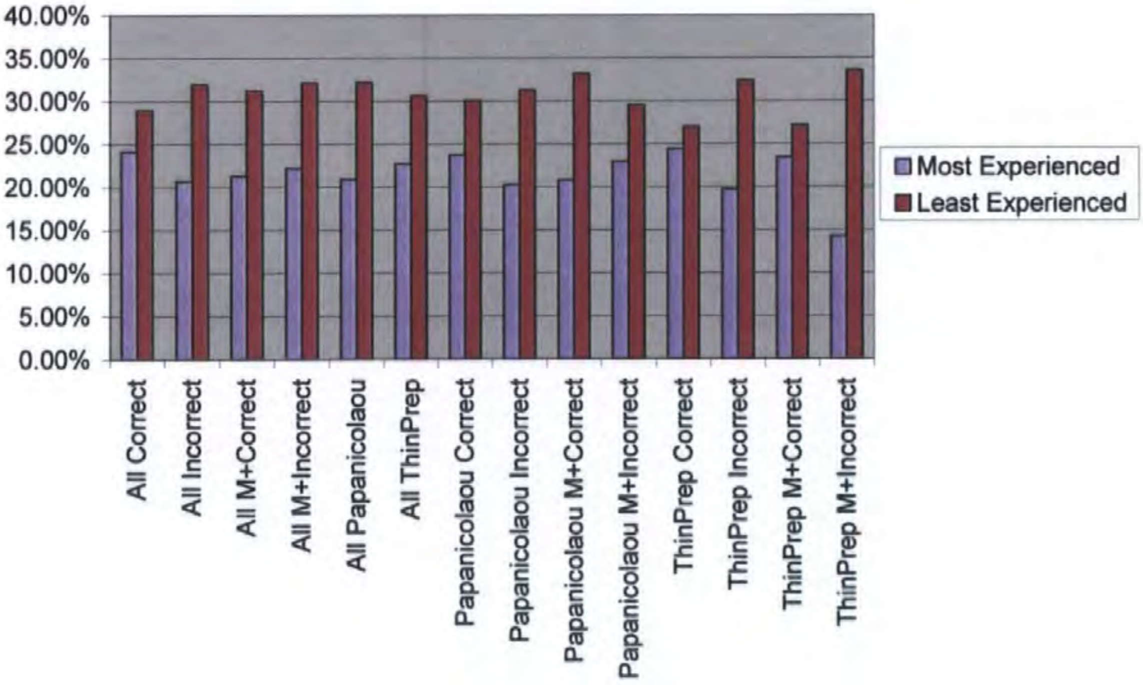


Figure 6.16 Abnormality Coverage average Percentages

Table 6.17 Abnormality Coverage percentages across conditions

	Most Experienced	Least Experienced
All Correct	24.13%	28.96%
All Incorrect	20.67%	31.97%
All M+Correct	21.26%	31.22%
All M+Incorrect	22.16%	32.11%
All Papanicolaou	20.91%	32.19%
All ThinPrep	22.68%	30.64%
Papanicolaou Correct	23.73%	30.05%
Papanicolaou Incorrect	20.24%	31.29%
Papanicolaou M+Correct	20.79%	33.16%
Papanicolaou M+Incorrect	22.97%	29.54%
ThinPrep Correct	24.40%	27.00%
ThinPrep Incorrect	19.71%	32.35%
ThinPrep M+Correct	23.40%	27.18%
ThinPrep M+Incorrect	14.26%	33.56%

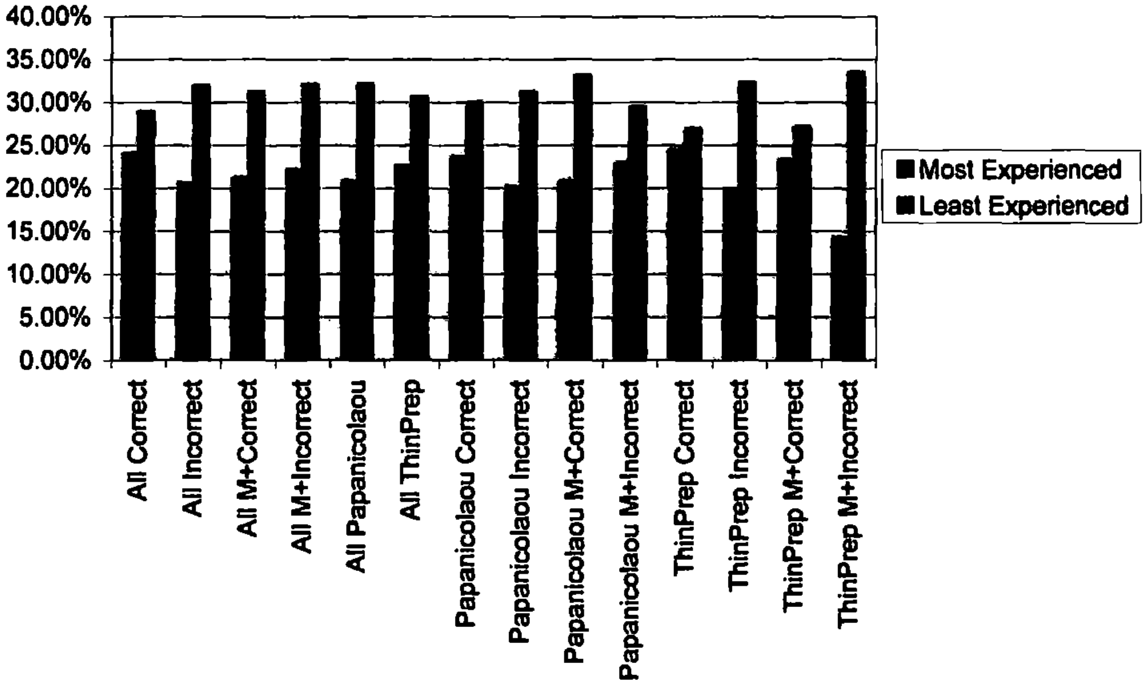


Figure 6.16 Abnormality Coverage average Percentages

Table 6.18 Abnormality Coverage Independent T-tests

Condition	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
All Correct	.433	8	.676	.0484	.11156	-.20891	.30562
All Incorrect	1.389	8	.202	.1130	.08132	-.07455	.30048
All M+ Correct	1.233	8	.253	.0996	.08078	-.08671	.28583
All M+ Incorrect	1.082	8	.311	.0994	.09187	-.11242	.31127
All Papanicolaou	1.167	8	.277	.1129	.09668	-.11010	.33581
All ThinPrep	1.075	8	.314	.0796	.07404	-.09117	.25032
Papanicolaou Correct	.479	8	.645	.0632	.13194	-.24106	.36744
Papanicolaou Incorrect	1.219	8	.257	.1105	.09065	-.09851	.31959
Papanicolaou M+ Correct	1.352	8	.213	.1237	.09151	-.08731	.33472
Papanicolaou M+ Incorrect	.569	8	.585	.0657	.11553	-.20071	.33213
ThinPrep Correct	.244	8	.813	.0260	.10624	-.21904	.27095
ThinPrep Incorrect	1.389	8	.202	.1265	.09106	-.08354	.33645
ThinPrep M+ Correct	.458	8	.659	.0378	.08251	-.15245	.22809
ThinPrep M+ Incorrect	2.445	8	.040	.1930	.07893	.01100	.37503

The reason why this difference is significant could be because, as suggested above, the abnormality measure does not correctly identify abnormal areas with which to judge performance. However, a more likely explanation is that for moderate+, the limited number of images available may have played a part by exaggerating the effect for ThinPrep images. In particular, while the image set represents most classifications, some were not available for this study as shown in Table 6.16. It may be that because of the poorer sensitivity performance levels seen for ThinPrep images, the incorrect classification groups are inflated and consequently the significant difference shown is because this represents a larger amount of participant recording data again supporting the notion that more participant data is needed before any strong conclusions can be drawn.

6.5.5 Machine Colour Texture Analysis Results

The machine analysis methodology was assessed by looking for a correlation between the eye-tracker fixations and the number of local maxima clustered around each of the features being viewed. The Atrous wavelet transformation locates features of interest at different resolutions so we might expect increased density of maxima surrounding salient areas. A Spearman's Rho correlation was performed between eye-tracker data, ranked using the feature marking data as a guide, and the maxima density detected in the hue and combined saturation/value texture components surrounding the fixated points of each image within a five degree visual angle. The results of this process for both a four and thirteen element vector can be seen in Tables 6.19, 6.20 and 6.21. Significant correlations at p<.05 are highlighted in bold type, and those that are significant at p<.01 are highlighted in bold italics.

There are a number of interesting results from this analysis. The first of these is that the four element vector shows only one significant condition. This shows that there was a positive significant correlation between maxima detected on the four element vector and correctly classified Papanicolaou image fixations. This could suggest that the correct/incorrect status of each image and the prior training on Papanicolaou slides that each participant had as part of the pre-requisite for participation in this experiment are important factors. It is also shown that this correlation exists for the hue of the images, but not the saturation/value component. As hue is a particularly important aspect of the image classification process, it would be expected to show a stronger relationship than the saturation and value. However, this result in isolation can also be easily explained as a type II error given the number of correlations that have been performed. Correctly classified images and Papanicolaou images show no other significances for four element vectors which reinforces the idea that this could be an anomalous result.

Table 6.19 Spearman's Rho Correlations

Eye-Tracked Condition		4 elements		13 elements	
		Hue	SatVal	Hue	SatVal
All Slides	Correlation Coefficient	0.007	0.002	***0.051***	-0.013
	Significance (2-tailed)	0.526	0.884	***0.000***	0.245
	N	8103	8103	***8103***	8103
Correct Classifications	Correlation Coefficient	0.032	-0.007	***0.057***	-0.031
	Significance (2-tailed)	0.068	0.697	***0.001***	0.074
	N	3286	3286	***3286***	3286
Incorrect Classifications	Correlation Coefficient	-0.008	0.008	***0.044***	-0.005
	Significance (2-tailed)	0.555	0.574	***0.002***	0.718
	N	4817	4817	***4817***	4817
Moderate+ Correct Classifications	Correlation Coefficient	0.025	-0.005	***0.058***	-0.016
	Significance (2-tailed)	0.087	0.728	***0.000***	0.291
	N	4537	4537	***4537***	4537
Moderate+ Incorrect Classifications	Correlation Coefficient	-0.015	0.011	**0.042**	-0.011
	Significance (2-tailed)	0.382	0.512	**0.013**	0.508
	N	3566	3566	**3566**	3566
Most Experienced	Correlation Coefficient	0.012	-0.009	***0.051***	-0.009
	Significance (2-tailed)	0.482	0.601	***0.003***	0.592
	N	3356	3356	***3356***	3356
Least Experienced	Correlation Coefficient	0.003	0.010	***0.051***	-0.016
	Significance (2-tailed)	0.815	0.496	***0.000***	0.284
	N	4747	4747	***4747***	4747
Most Experienced Correct Classifications	Correlation Coefficient	0.040	-0.019	***0.077***	-0.032
	Significance (2-tailed)	0.175	0.513	***0.009***	0.280
	N	1157	1157	***1157***	1157
Least Experienced Correct Classifications	Correlation Coefficient	0.028	0.002	**0.047**	-0.029
	Significance (2-tailed)	0.191	0.928	**0.031**	0.185
	N	2129	2129	**2129**	2129
Most Experienced Incorrect Classifications	Correlation Coefficient	-0.002	-0.004	0.036	0.000
	Significance (2-tailed)	0.908	0.865	0.095	0.985
	N	2199	2199	2199	2199
Least Experienced Incorrect Classifications	Correlation Coefficient	-0.014	0.018	***0.051***	-0.009
	Significance (2-tailed)	0.490	0.357	***0.009***	0.636
	N	2618	2618	***2618***	2618
Most Experienced Moderate+ Correct Classifications	Correlation Coefficient	0.036	-0.037	***0.063***	-0.024
	Significance (2-tailed)	0.133	0.125	***0.009***	0.324
	N	1741	1741	***1741***	1741
Least Experienced Moderate+ Correct Classifications	Correlation Coefficient	0.018	0.019	***0.057***	-0.009
	Significance (2-tailed)	0.329	0.322	***0.003***	0.635
	N	2796	2796	***2796***	2796
Most Experienced Moderate+ Incorrect Classifications	Correlation Coefficient	-0.015	0.022	0.039	0.006
	Significance (2-tailed)	0.550	0.383	0.118	0.822
	N	1615	1615	1615	1615
Least Experienced Moderate+ Incorrect Classifications	Correlation Coefficient	-0.015	0.002	**0.045**	-0.025
	Significance (2-tailed)	0.520	0.916	**0.048**	0.275
	N	1951	1951	**1951**	1951

Table 6.20 Spearman's rho correlations (Papanicolaou)

Eye-Tracked Condition		4 elements		13 elements	
		Hue	SatVal	Hue	SatVal
All Papanicolaou	Correlation Coefficient	-0.001	0.012	***0.054***	0.008
	Significance (2-tailed)	0.952	0.427	***0.001***	0.597
	N	4175	4175	***4175***	4175
Papanicolaou Correct Classifications	Correlation Coefficient	**0.052**	0.021	**0.067**	-0.020
	Significance (2-tailed)	**0.047**	0.428	**0.011**	0.456
	N	**1448**	1448	**1448**	1448
Papanicolaou Incorrect Classifications	Correlation Coefficient	-0.026	0.011	**0.045**	0.018
	Significance (2-tailed)	0.168	0.573	**0.020**	0.359
	N	2727	2727	**2727**	2727
Papanicolaou Moderate+ Correct Classifications	Correlation Coefficient	0.029	0.014	***0.058***	0.011
	Significance (2-tailed)	0.150	0.478	***0.005***	0.589
	N	2405	2405	***2405***	2405
Papanicolaou Moderate+ Incorrect Classifications	Correlation Coefficient	-0.039	0.012	**0.047**	0.003
	Significance (2-tailed)	0.102	0.625	**0.046**	0.903
	N	1770	1770	**1770**	1770
Papanicolaou Most Experienced	Correlation Coefficient	0.004	-0.011	***0.077***	0.021
	Significance (2-tailed)	0.858	0.661	***0.002***	0.384
	N	1677	1677	***1677***	1677
Papanicolaou Least Experienced	Correlation Coefficient	-0.004	0.030	0.036	-0.001
	Significance (2-tailed)	0.846	0.135	0.072	0.950
	N	2498	2498	2498	2498
Papanicolaou Most Experienced Correct Classifications	Correlation Coefficient	0.085	-0.006	***0.140***	0.054
	Significance (2-tailed)	0.085	0.904	***0.005***	0.280
	N	407	407	***407***	407
Papanicolaou Least Experienced Correct Classifications	Correlation Coefficient	0.040	0.035	0.036	-0.048
	Significance (2-tailed)	0.194	0.265	0.248	0.123
	N	1041	1041	1041	1041
Papanicolaou Most Experienced Incorrect Classifications	Correlation Coefficient	-0.021	-0.012	0.055	0.011
	Significance (2-tailed)	0.444	0.671	0.050	0.683
	N	1270	1270	1270	1270
Papanicolaou Least Experienced Incorrect Classifications	Correlation Coefficient	-0.031	0.032	0.035	0.023
	Significance (2-tailed)	0.240	0.229	0.179	0.372
	N	1457	1457	1457	1457
Papanicolaou Most Experienced Moderate+ Correct Classifications	Correlation Coefficient	0.040	-0.028	**0.083**	0.043
	Significance (2-tailed)	0.241	0.410	**0.014**	0.204
	N	870	870	**870**	870
Papanicolaou Least Experienced Moderate+ Correct Classifications	Correlation Coefficient	0.024	0.043	0.044	-0.010
	Significance (2-tailed)	0.350	0.095	0.086	0.695
	N	1535	1535	1535	1535
Papanicolaou Most Experienced Moderate+ Incorrect Classifications	Correlation Coefficient	-0.033	0.007	**0.071**	-0.001
	Significance (2-tailed)	0.343	0.845	**0.045**	0.976
	N	807	807	**807**	807
Papanicolaou Least Experienced Moderate+ Incorrect Classifications	Correlation Coefficient	-0.043	0.014	0.027	0.009
	Significance (2-tailed)	0.179	0.660	0.407	0.774
	N	963	963	963	963

Table 6.21 Spearman's rho correlations (ThinPrep)

Eye-Tracked Condition		4 elements		13 elements	
		Hue	SatVal	Hue	SatVal
All ThinPrep	Correlation Coefficient	0.016	-0.010	***0.047***	-0.036
	Significance (2-tailed)	0.329	0.537	***0.003***	0.024
	N	3928	3928	***3928***	3928
ThinPrep Correct Classifications	Correlation Coefficient	0.016	-0.028	**0.051**	-0.040
	Significance (2-tailed)	0.494	0.227	**0.030**	0.084
	N	1838	1838	**1838**	1838
ThinPrep Incorrect Classifications	Correlation Coefficient	0.015	0.005	**0.044**	-0.035
	Significance (2-tailed)	0.493	0.826	**0.046**	0.106
	N	2090	2090	**2090**	2090
ThinPrep Moderate+ Correct Classifications	Correlation Coefficient	0.021	-0.028	***0.058***	**-0.049**
	Significance (2-tailed)	0.340	0.201	***0.007***	**0.025**
	N	2132	2132	***2132***	**2132**
ThinPrep Moderate+ Incorrect Classifications	Correlation Coefficient	0.010	0.011	0.036	-0.025
	Significance (2-tailed)	0.677	0.653	0.125	0.295
	N	1796	1796	1796	1796
ThinPrep Most Experienced	Correlation Coefficient	0.021	-0.007	0.018	-0.043
	Significance (2-tailed)	0.388	0.775	0.458	0.080
	N	1679	1679	1679	1679
ThinPrep Least Experienced	Correlation Coefficient	0.011	-0.012	***0.070***	-0.031
	Significance (2-tailed)	0.588	0.568	***0.001***	0.141
	N	2249	2249	***2249***	2249
ThinPrep Most Experienced Correct Classifications	Correlation Coefficient	0.015	-0.026	0.044	**-0.078**
	Significance (2-tailed)	0.674	0.470	0.227	**0.032**
	N	750	750	750	**750**
ThinPrep Least Experienced Correct Classifications	Correlation Coefficient	0.017	-0.030	0.057	-0.008
	Significance (2-tailed)	0.587	0.322	0.060	0.788
	N	1088	1088	1088	1088
ThinPrep Most Experienced Incorrect Classifications	Correlation Coefficient	0.026	0.009	-0.003	-0.016
	Significance (2-tailed)	0.428	0.777	0.926	0.622
	N	929	929	929	929
ThinPrep Least Experienced Incorrect Classifications	Correlation Coefficient	0.007	0.001	***0.078***	-0.050
	Significance (2-tailed)	0.804	0.966	***0.008***	0.091
	N	1161	1161	***1161***	1161
ThinPrep Most Experienced Moderate+ Correct Classifications	Correlation Coefficient	0.032	-0.045	0.044	**-0.097**
	Significance (2-tailed)	0.341	0.181	0.196	**0.004**
	N	871	871	871	**871**
ThinPrep Least Experienced Moderate+ Correct Classifications	Correlation Coefficient	0.011	-0.013	**0.070**	-0.005
	Significance (2-tailed)	0.692	0.652	**0.012**	0.852
	N	1261	1261	**1261**	1261
ThinPrep Most Experienced Moderate+ Incorrect Classifications	Correlation Coefficient	0.008	0.039	-0.014	0.021
	Significance (2-tailed)	0.828	0.266	0.687	0.543
	N	808	808	808	808
ThinPrep Least Experienced Moderate+ Incorrect Classifications	Correlation Coefficient	0.012	-0.010	**0.069**	-0.055
	Significance (2-tailed)	0.711	0.765	**0.029**	0.084
	N	988	988	**988**	988

The difference between hue and saturation/value becomes more evident for the thirteen element vector. It can be seen in Table 6.19 that all of the hue conditions that were tested were significantly correlated with the eye-tracker fixations with the exception of the most experienced group's incorrect classification across all images, and moderate+ images. Again, this might suggest a difference between correct and incorrect classifications or between the different levels of experience if it were not for the fact that many of the incorrect conditions are shown to also be significant with the overall incorrect classifications condition shown to be highly significant. Those with less experience are also shown to be highly significantly correlated across a number of different conditions. It is interesting to note that significant relationships for ThinPrep conditions are seen for less experienced screeners but not for their more experienced counterparts across similar conditions. This could be explained in terms of experience and prior knowledge. Those who have more experience of screening Papanicolaou slides could be using their knowledge from this to make judgements based on that experience. This may be inhibit their ability to objectively view and classify a ThinPrep image, while those with less expertise may not suffer from a similar inhibition. The only non-significant relationship for the less experienced participants viewing ThinPrep images was with the thirteen element vector's hue component and correctly classified images, although this does become significant when only moderate+ images are considered.

It is difficult to draw a conclusion from this result as to whether the correct or incorrect classification given to an image by each participant has an effect on the data though it does suggest that further investigation is required. However, it is the consistent difference seen between the hue and saturation/value conditions for the thirteen element vector that is of most interest. The fact that this can be seen in all but two of the conditions in Table 6.19 would indicate that the thirteen element vector is locating the salient features in the image set. Further examination of how these conditions break down into their component groups can be seen in Tables 6.20 and 6.21. The significant correlations with hue can be seen across the ThinPrep and Papanicolaou conditions when all images, correctly classified images, incorrectly classified images and moderate+ correctly classified images are considered. The significant correlations with Papanicolaou images continue for the most experienced group's correctly classified images, moderate+ correctly classified images and moderate+ incorrectly classified images. For the ThinPrep images, significant correlations are shown for the less experienced participants across all images, for incorrectly classified images, for moderate+ correctly classified images and for moderate+ incorrectly classified images. For the ThinPrep conditions we can also see negative correlations between the saturation value component and moderate+ correct classifications, most experienced group's correctly classified images and the same group's moderate+ images. Further examination of the results for ThinPrep conditions and the saturation/value components shows that all but one of the relationships, that with the most experienced group's moderate+ incorrect image fixations, was negative. It may be that while the hue element of the images shows higher numbers of maxima around interesting features, the saturation value shows decreased levels of maxima. While the significances here are not consistent enough to draw any conclusions, this is certainly an area that would benefit from future investigation.

The strongest conclusion to be drawn from these correlations is that for the thirteen element vector, the features that are found in the hue element of the image have a relationship with the salient features that are viewed by the participants. In some cases these correlations are very significant. While there are correlations found in both Papanicolaou and ThinPrep conditions, the combined analysis suggests that the relationship becomes stronger as more data is available for the analysis. This finding is consistent with the earlier conclusions from the abnormal, saliency and image coverage that increasing the amount of data will give a clearer picture and allow stronger conclusions to be drawn.

As we have seen, the relationship between items that are correct and incorrect is unclear from the analysis. The classification of images only took place in order to add realism into the experimental process. Without asking a participant to classify an image, they would effectively have no purpose on viewing the image presentation. There are two distinct benefits from asking for classifications. It means that each image is viewed in a meaningful manner and provides a means of testing the validity and reliability of the experimental procedure. Although the differences between the conditions are not systematic enough to support separating the correct and incorrectly classified image fixations, there is certainly an argument for extending this investigation to see whether there are any differences when correct and incorrect fixations are directly compared to each other.

In order to perform this analysis, Fisher's Z transformation needs to be applied. This converts Spearman's rho into a Z score allowing comparisons between conditions using independent samples t-tests. Each of the correlations for fixation data from correctly and incorrectly classified can then be compared for both the four and thirteen element vectors, and for both hue and saturation/value. The raw data for this analysis can be seen in Appendix K. The results from this analysis can be seen in Tables 6.22 and 6.23. Significant correlations at p<.05 are highlighted in bold type, and those that are significant at p<.01 are highlighted in bold italics.

In Table 6.22 we can see that there is a significant difference between correctly and incorrectly classified Papanicolaou images and for moderate+ Papanicolaou images. In both cases this is for the four element vectors hue component. There is a further significant difference between correct and incorrectly classified images for the more experienced participants on moderate+ ThinPrep images. In this case it is for the thirteen element vectors saturation value component. These results do not support separating the fixation data based on their correct or incorrect status. Because such a large amount of statistical work is involved in producing these comparisons, without more systematic significant differences it cannot be concluded that those differences that are seen are any more than spurious results. Furthermore there is little consistency with the significant correlations shown in Tables 6.19, 6.20, and 6.21 suggesting that it is appropriate to treat fixation data as one regardless of whether the image was correctly or incorrectly classified.

There is one final comparison that can be made in order to validate the experimental process. Conditions involving Papanicolaou images need to be compared to their corresponding ThinPrep. Much of the work described here uses fixations from images of both slide preparations. In order to demonstrate that this does not have an effect on the conclusions based on the data analysis, a comparison is made using the same Fisher's Z transformation described above. Similar conditions can then be compared using independent samples t-tests.

The results from a series of comparisons between maxima surrounding ThinPrep and Papanicolaou image fixations, for four and thirteen element vectors, and for hue and saturation/value components, are shown in Table 6.23. This shows significant differences between Papanicolaou and ThinPrep fixations when all fixations, fixations when images are correctly classified, moderate+ fixations when images are correctly classified and moderate+ fixations when images are correctly classified by more experienced screeners are considered.

Table 6.22 Independent samples t-testing comparing eye-tracked data using local maxima density

Slide Type	Independent samples t-test		4 element t-values		13 element t-values	
	Condition 1	Condition 2	Hue	SatVal	Hue	SatVal
Combined	Correct Classifications	Incorrect Classifications	1.78	-0.66	0.58	-1.15
	Moderate+ Correct Classifications	Moderate+ Incorrect Classifications	1.79	-0.72	0.73	-0.20
	Most Experienced	Least Experienced	0.39	-0.84	-0.01	0.28
	Most Experienced Correct Classifications	Most Experienced Incorrect Classifications	1.17	-0.43	1.13	-0.86
	Least Experienced Correct Classifications	Least Experienced Incorrect Classifications	1.43	-0.55	-0.13	-0.67
	Most Experienced Moderate+ Correct Classifications	Most Experienced Moderate+ Incorrect Classifications	1.47	-1.69	0.69	-0.85
	Least Experienced Moderate+ Correct Classifications	Least Experienced Moderate+ Incorrect Classifications	1.12	0.55	0.41	0.53
Papanicolaou	Correct Classifications	Incorrect Classifications	**2.42**	0.31	0.68	-1.14
	Moderate+ Correct Classifications	Moderate+ Incorrect Classifications	**2.18**	0.09	0.33	0.26
	Most Experienced	Least Experienced	0.26	-1.29	1.31	0.71
	Most Experienced Correct Classifications	Most Experienced Incorrect Classifications	1.87	0.10	1.50	0.74
	Least Experienced Correct Classifications	Least Experienced Incorrect Classifications	1.75	0.08	0.01	-1.76
	Most Experienced Moderate+ Correct Classifications	Most Experienced Moderate+ Incorrect Classifications	1.50	-0.71	0.25	0.90
	Least Experienced Moderate+ Correct Classifications	Least Experienced Moderate+ Incorrect Classifications	1.63	0.69	0.41	-0.47
ThinPrep	Correct Classifications	Incorrect Classifications	0.03	-1.03	0.22	-0.15
	Moderate+ Correct Classifications	Moderate+ Incorrect Classifications	0.34	-1.20	0.69	-0.74
	Most Experienced	Least Experienced	0.30	0.16	-1.62	-0.36
	Most Experienced Correct Classifications	Most Experienced Incorrect Classifications	-0.22	-0.73	0.96	-1.27
	Least Experienced Correct Classifications	Least Experienced Incorrect Classifications	0.22	-0.74	-0.50	0.98
	Most Experienced Moderate+ Correct Classifications	Most Experienced Moderate+ Incorrect Classifications	0.50	-1.73	1.19	**-2.42**
	Least Experienced Moderate+ Correct Classifications	Least Experienced Moderate+ Incorrect Classifications	-0.02	-0.08	0.02	1.17

significant @ 0.5 Probability > 1.96
significant @ 0.1 Probability > 2.58

Table 6.23 Independent samples t-testing comparing Papanicolaou to ThinPrep eye tracked data using local maxima density

Slide Type	Independent samples t-test		4 element t-values		13 element t-values	
	Condition 1	Condition 2	Hue	SatVal	Hue	SatVal
ThinPrep vs Papanicolaou	All Papanicolaou	All ThinPrep	-0.74	1.00	0.28	**1.99**
	Papanicolaou Correct Classifications	ThinPrep Correct Classifications	1.03	1.39	0.46	0.59
	Papanicolaou Incorrect Classifications	ThinPrep Incorrect Classifications	-1.42	0.21	0.03	1.82
	Papanicolaou Moderate+ Correct Classifications	ThinPrep Moderate+ Correct Classifications	0.29	1.42	-0.02	**2.00**
	Papanicolaou Moderate+ Incorrect Classifications	ThinPrep Moderate+ Incorrect Classifications	-1.46	0.03	0.34	0.82
	Papanicolaou Most Experienced	ThinPrep Most Experienced	-0.48	-0.11	1.71	1.85
	Papanicolaou Least Experienced	ThinPrep Least Experienced	-0.53	1.44	-1.18	1.03
	Papanicolaou Most Experienced Correct Classifications	ThinPrep Most Experienced Correct Classifications	1.14	0.33	1.56	**2.14**
	Papanicolaou Most Experienced Incorrect Classifications	ThinPrep Most Experienced Incorrect Classifications	-1.10	-0.49	1.34	0.64
	Papanicolaou Least Experienced Correct Classifications	ThinPrep Least Experienced Correct Classifications	0.55	1.49	-0.49	-0.92
	Papanicolaou Least Experienced Incorrect Classifications	ThinPrep Least Experienced Incorrect Classifications	-0.97	0.77	-1.09	1.86
	Papanicolaou Most Experienced Moderate+ Correct Classifications	ThinPrep Most Experienced Moderate+ Correct Classifications	0.16	0.36	0.82	**2.92**
	Papanicolaou Most Experienced Moderate+ Incorrect Classifications	ThinPrep Most Experienced Moderate+ Incorrect Classifications	-0.82	-0.65	1.70	-0.45
	Papanicolaou Least Experienced Moderate+ Correct Classifications	ThinPrep Least Experienced Moderate+ Correct Classifications	0.33	1.45	-0.70	-0.13
	Papanicolaou Least Experienced Moderate+ Incorrect Classifications	ThinPrep Least Experienced Moderate+ Incorrect Classifications	-1.22	0.52	-0.94	1.42

significant @ 0.5 Probability > 1.96
significant @ 0.1 Probability > 2.58

These significant differences exist between the conditions correlated with the thirteen element vectors saturation/value component and, in the case of the most experienced moderate+ correctly classified image fixations, this difference is highly significant (p<.01). This reinforces the earlier finding that this particular machine colour texture analysis is finding something other than salient features in the ThinPrep images. The fact that these correlations are not seen consistently throughout again suggests that more fixation information needs to be recorded before a strong conclusion can be drawn. As such, any further results based on ThinPrep image analysis of the thirteen element vectors saturation/value component should be handled cautiously. While it is not in doubt that more fixation data would give a clearer picture of the relationships that may exist between conditions, the fact that there were so few significant differences seen throughout Tables 6.22 and 6.23 is reassuring. The evidence shows that there are no differences between the majority of the conditions tested. Because of the significant correlations that can be seen with the machine colour texture analysis, the lack of significant differences between the conditions suggests that the maxima from the thirteen element hue aspect of the images is locating salient features regardless of condition or image type.

The predictive abilities of each of the conditions can be tested using a canonical discriminant function analysis. This form of multivariate analysis allows the ranked saliency data to be compared to the machine analysis in order to find the combination of variables that maximises separation between groups. In this case the groups are based on the saliency index with maxima surrounded by highly salient and abnormal features in the first group, maxima surrounded by salient and normal features in the second group, and a final group of the remaining maxima.

In order to assess the usefulness of the Atrous machine colour texture analysis the predictive capabilities need to be examined. Although the evidence shows that under some circumstances there is a significant correlation between the maxima and features of interest, in particular where the thirteen element hue vector is considered, salient areas would need to be predicted on the basis of this machine analysis. In order to assess the capabilities under each of the different conditions, canonical discriminant function analyses were performed. Based on this analysis, group membership can be predicted if there is enough discrimination between the groups. In effect, this would allow a colour texture to be analysed and rated for saliency. A summary of this analysis can be seen in Table 6.24, with detailed analysis for each condition shown in Figures 6.17 to 6.29, and Tables 6.25 to 6.38. The raw data for this analysis including the conditions not reported here can be seen in appendix K.

Groups are for the discriminant function analysis are defined as follows:

Group 1 – Salient and Abnormal (Red)

Group 2 – Salient and Normal (Green)

Group 3 – Not Salient (Blue)

Table 6.24 Canonical Discriminant Function group membership prediction summary

Image Component	Image type	Condition	Correctly Classified (%)
Hue	Papanicolaou	All	36.8
		Most Experienced	35.1
		Least Experienced	40.8
		Correct	12.1
		Incorrect	45.4
		Moderate+ Correct	15.1
		Moderate+ Incorrect	49.7
	ThinPrep	All	32.6
		Most Experienced	31.9
		Least Experienced	31.5
		Correct	28.4
		Incorrect	37.2
		Moderate+ Correct	38.5
		Moderate+ Incorrect	26.9

Table 6.25 Canonical Discriminant Function Analysis – Hue x Papanicolaou x All Classifications

		ALL_GRPS	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	330	168	303	801
		2.00	4738	4444	4874	14056
		3.00	66257	62703	76270	205230
	%	1.00	41.2	21.0	37.8	100.0
		2.00	33.7	31.6	34.7	100.0
		3.00	32.3	30.6	37.2	100.0

a 36.8% of original grouped cases correctly classified.

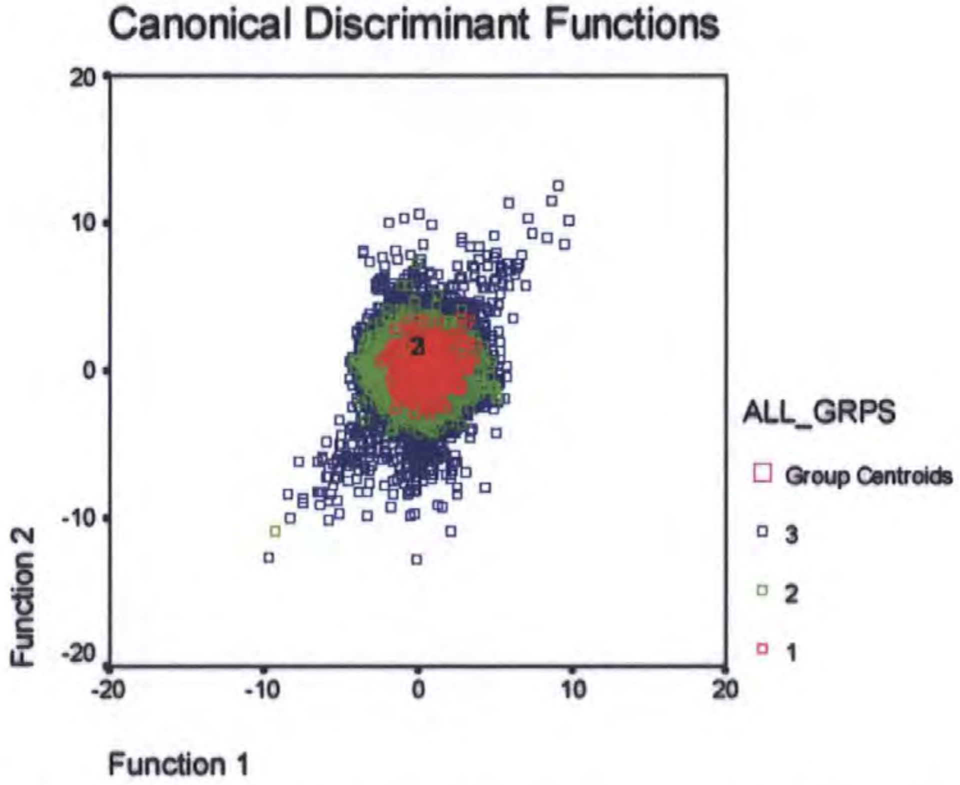


Figure 6.17 Canonical Discriminant Function Analysis – Hue x Papanicolaou x All Classifications

Table 6.26 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Most Experienced Screeners

		EXP_GR PS	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	179	87	154	420
		2.00	1892	2098	1919	5909
		3.00	65849	72856	75053	213758
	%	1.00	42.6	20.7	36.7	100.0
		2.00	32.0	35.5	32.5	100.0
		3.00	30.8	34.1	35.1	100.0

a 35.1% of original grouped cases correctly classified.

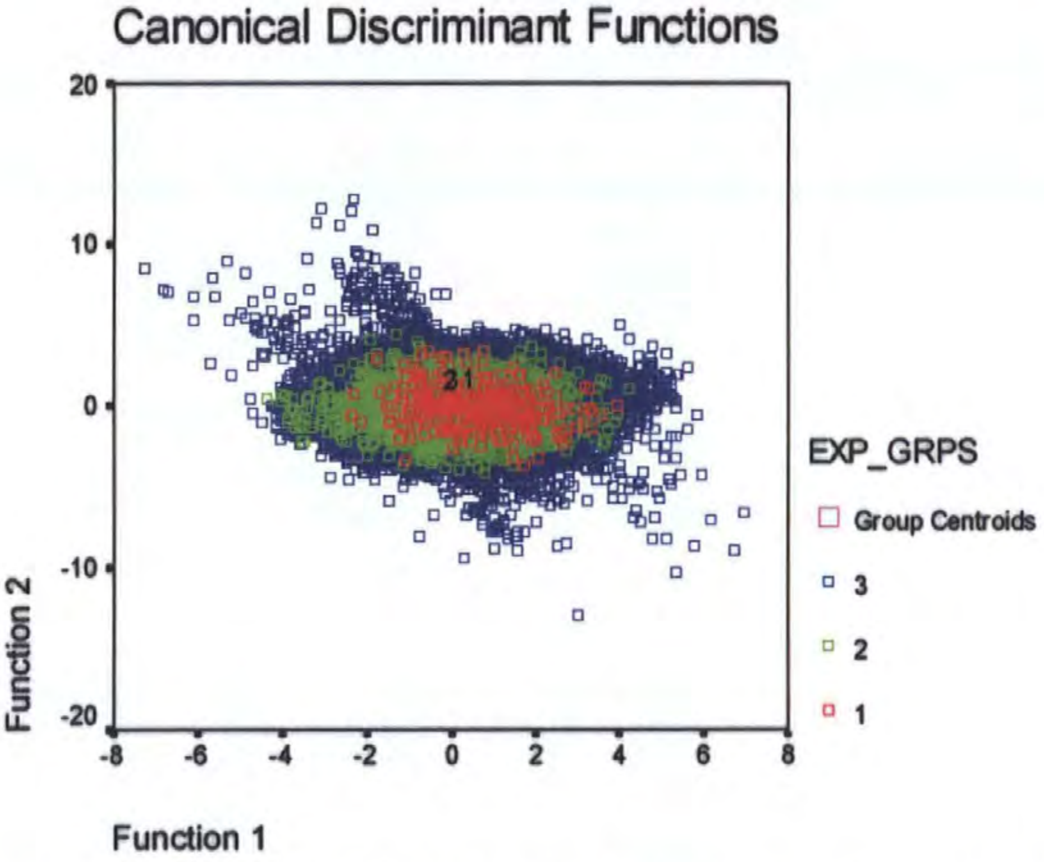


Figure 6.18 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Most Experienced Screeners

Table 6.27 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Least Experienced Screeners

		NOV_GRPS	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	172	89	165	426
		2.00	2919	2392	3455	8766
		3.00	67596	56002	87297	210895
	%	1.00	40.4	20.9	38.7	100.0
		2.00	33.3	27.3	39.4	100.0
		3.00	32.1	26.6	41.4	100.0

a 40.8% of original grouped cases correctly classified.

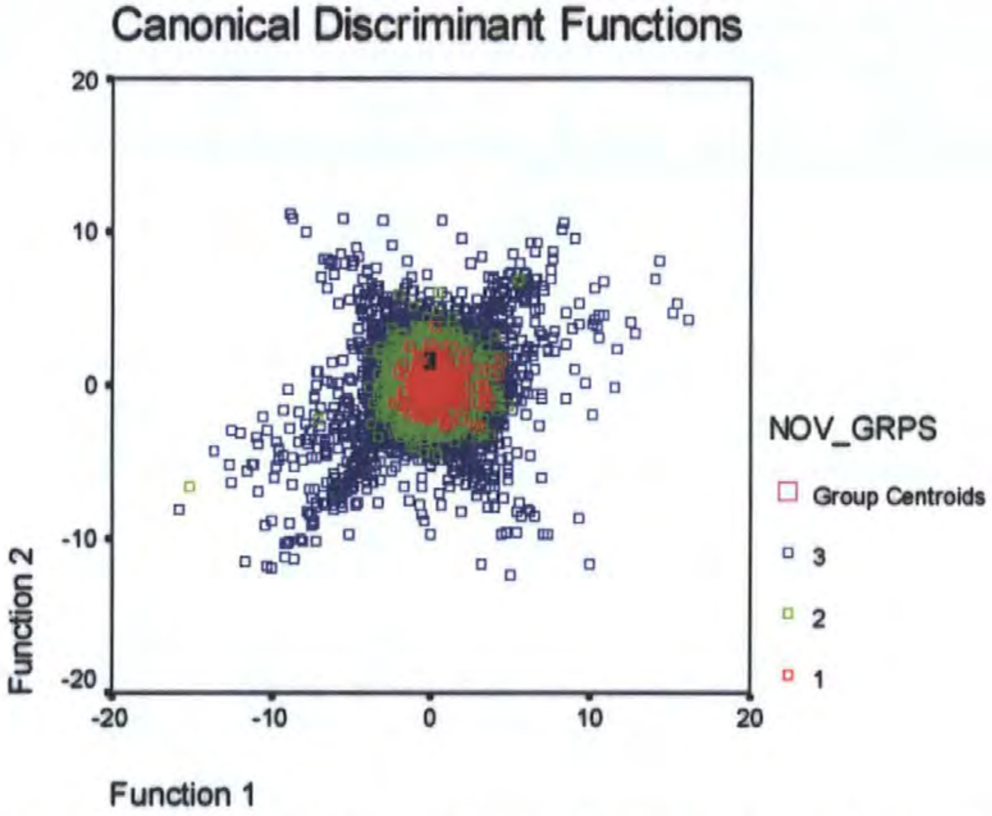


Figure 6.19 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Least Experienced Screeners

Table 6.28 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Correct Classification

		COR_GRPS	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	95	134	19	248
		2.00	1218	2999	456	4673
		3.00	71300	120265	23601	215166
	%	1.00	38.3	54.0	7.7	100.0
		2.00	26.1	64.2	9.8	100.0
		3.00	33.1	55.9	11.0	100.0

a 12.1% of original grouped cases correctly classified.

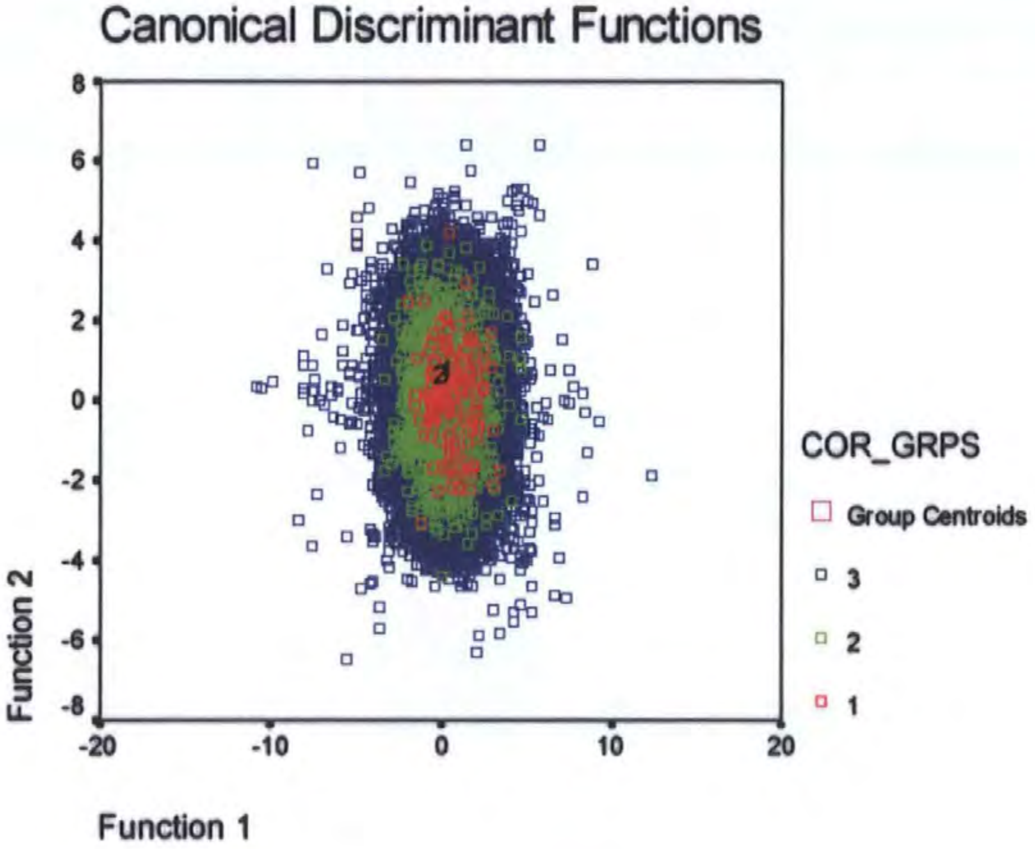


Figure 6.20 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Correct Classifications

Table 6.29 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Incorrect Classifications

		INC_GRP S	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	239	117	230	586
		2.00	3399	2684	3630	9713
		3.00	64717	48154	96917	209788
	%	1.00	40.8	20.0	39.2	100.0
		2.00	35.0	27.6	37.4	100.0
		3.00	30.8	23.0	46.2	100.0

a 45.4% of original grouped cases correctly classified.

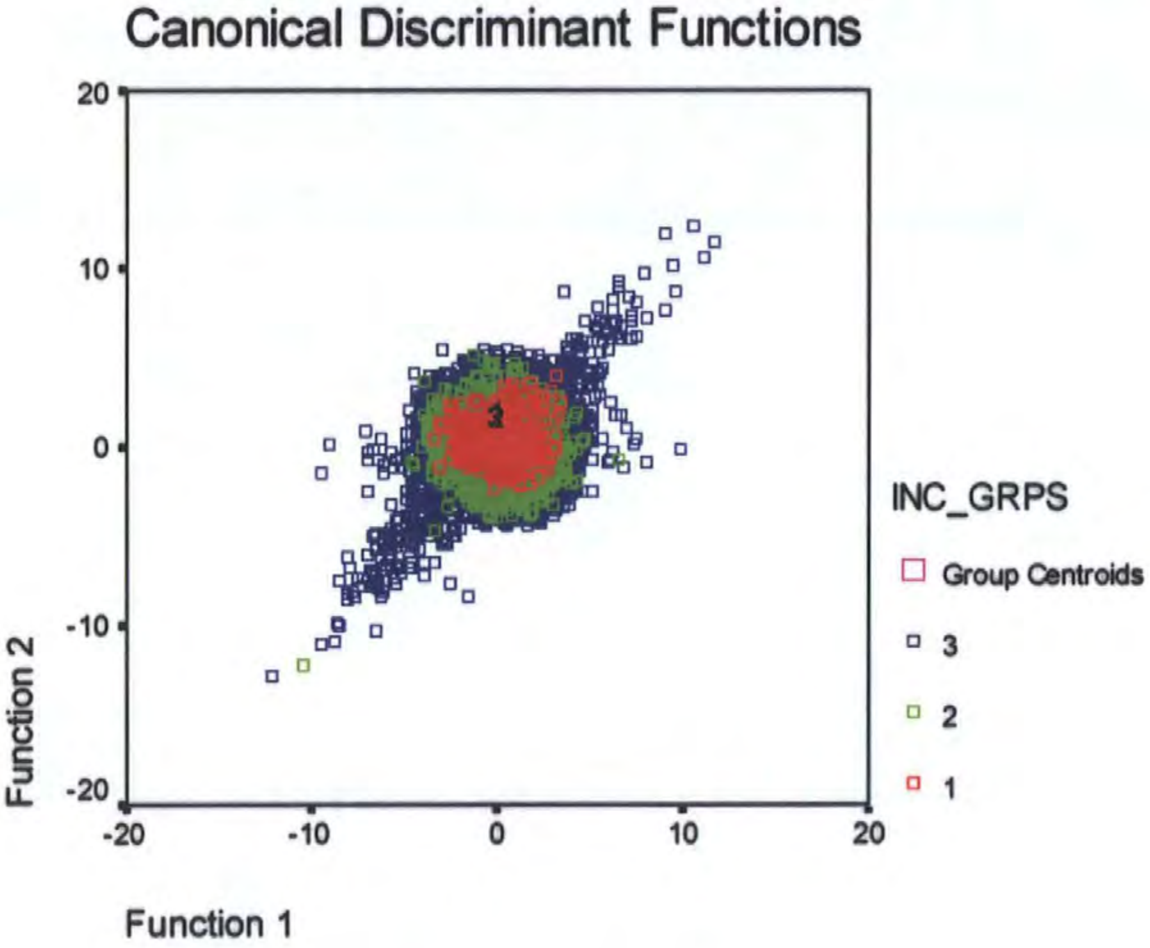


Figure 6.21 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Incorrect Classifications

Table 6.30 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Moderate+ Correct Classifications

		MCOR_GRP	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	197	176	77	450
		2.00	2586	4437	1111	8134
		3.00	71077	111898	28528	211503
	%	1.00	43.8	39.1	17.1	100.0
		2.00	31.8	54.5	13.7	100.0
		3.00	33.6	52.9	13.5	100.0

a 15.1% of original grouped cases correctly classified.

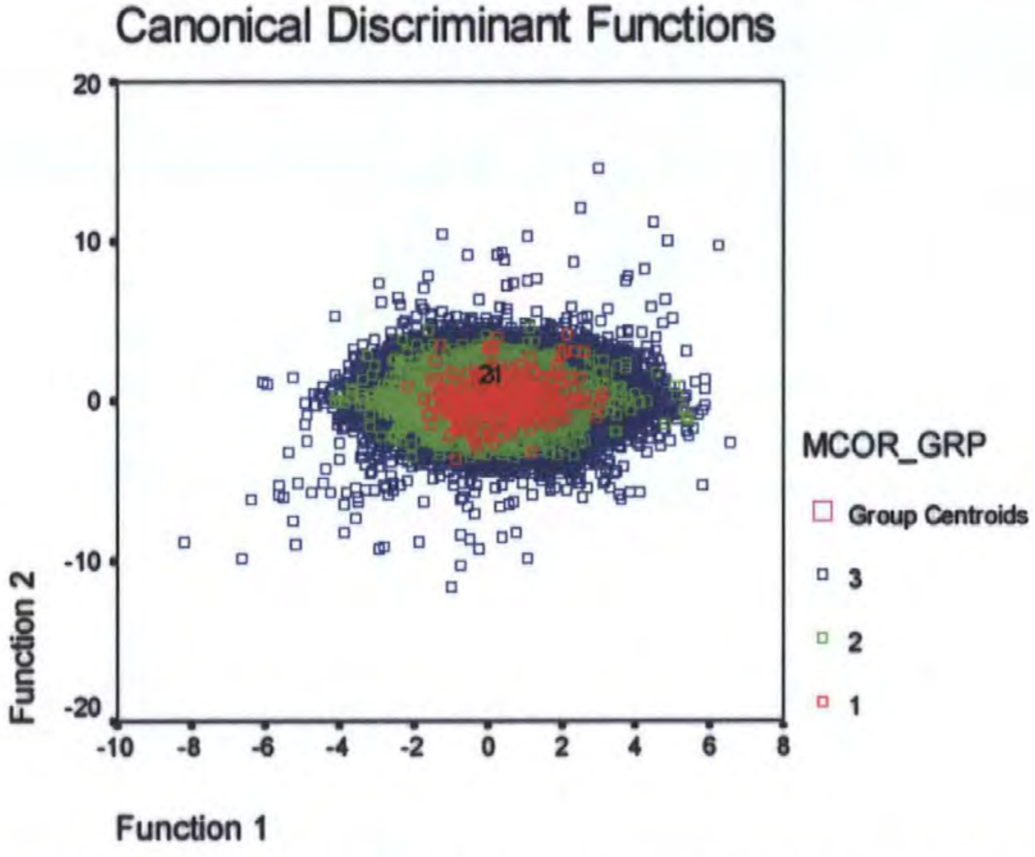


Figure 6.22 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Moderate+ Correct Classifications

Table 6.31 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Moderate+ Incorrect Classifications

		MINC_GRP	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	150	68	188	406
		2.00	2181	1473	2783	6437
		3.00	63574	41982	107688	213244
	%	1.00	36.9	16.7	46.3	100.0
		2.00	33.9	22.9	43.2	100.0
		3.00	29.8	19.7	50.5	100.0

a 49.7% of original grouped cases correctly classified.

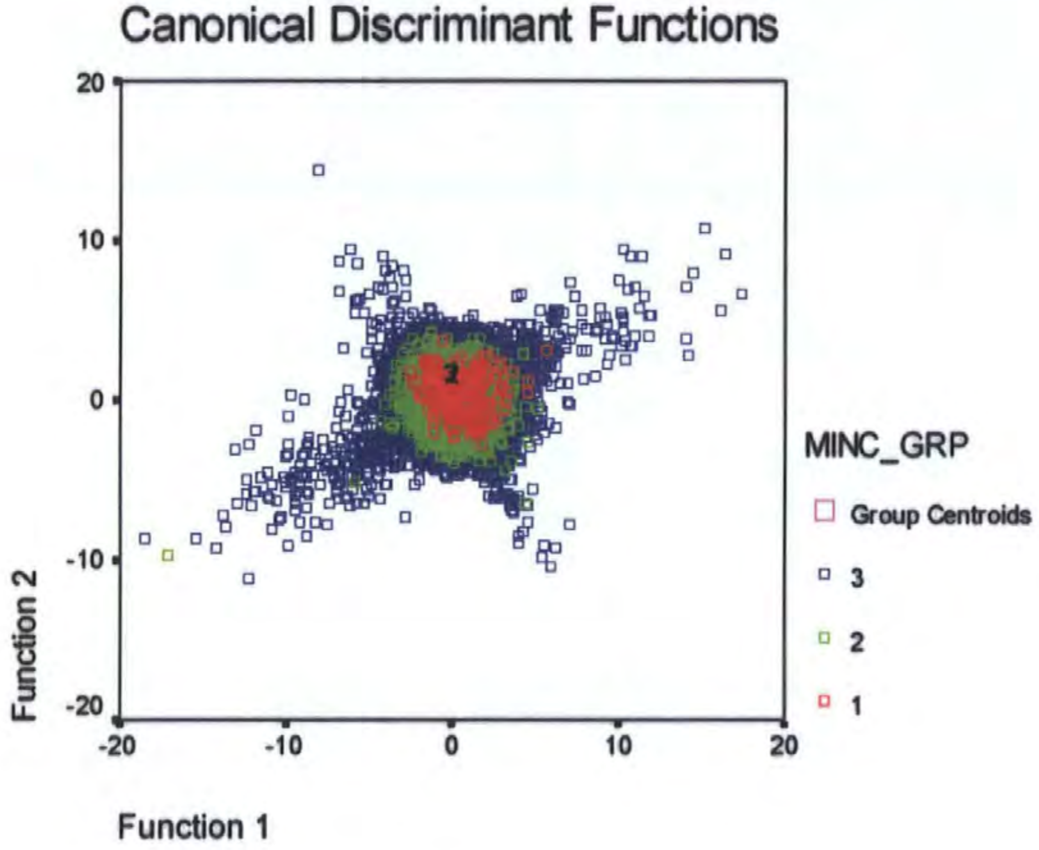


Figure 6.23 Canonical Discriminant Function Analysis – Hue x Papanicolaou x Moderate+ Incorrect Classifications

Table 6.32 Canonical Discriminant Function Analysis – Hue x ThinPrep x All Classifications

		ALL_GRPS	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	210	253	172	635
		2.00	2924	4484	3190	10598
		3.00	47775	65357	53221	166353
	%	1.00	33.1	39.8	27.1	100.0
		2.00	27.6	42.3	30.1	100.0
		3.00	28.7	39.3	32.0	100.0

a 32.6% of original grouped cases correctly classified.

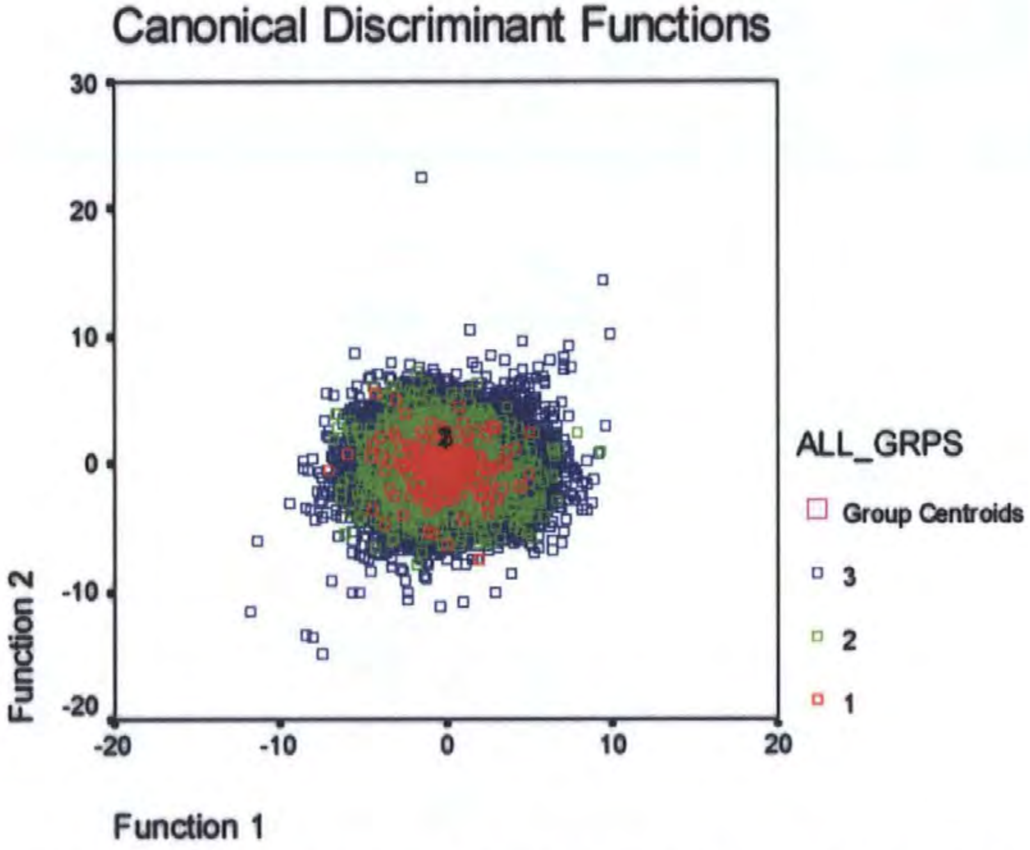


Figure 6.24 Canonical Discriminant Function Analysis – Hue x ThinPrep x All Classifications

Table 6.33 Canonical Discriminant Function Analysis – Hue x ThinPrep x Most Experienced Screeners

		EXP_GRPS	Predicted Group Membership			
			1.00	2.00	3.00	Total
Original	Count	1.00	115	86	75	276
		2.00	1621	1782	1426	4829
		3.00	57782	59935	54764	172481
	%	1.00	41.7	31.2	27.2	100.0
		2.00	33.6	36.9	29.5	100.0
		3.00	33.5	34.7	31.8	100.0

a 31.9% of original grouped cases correctly classified.

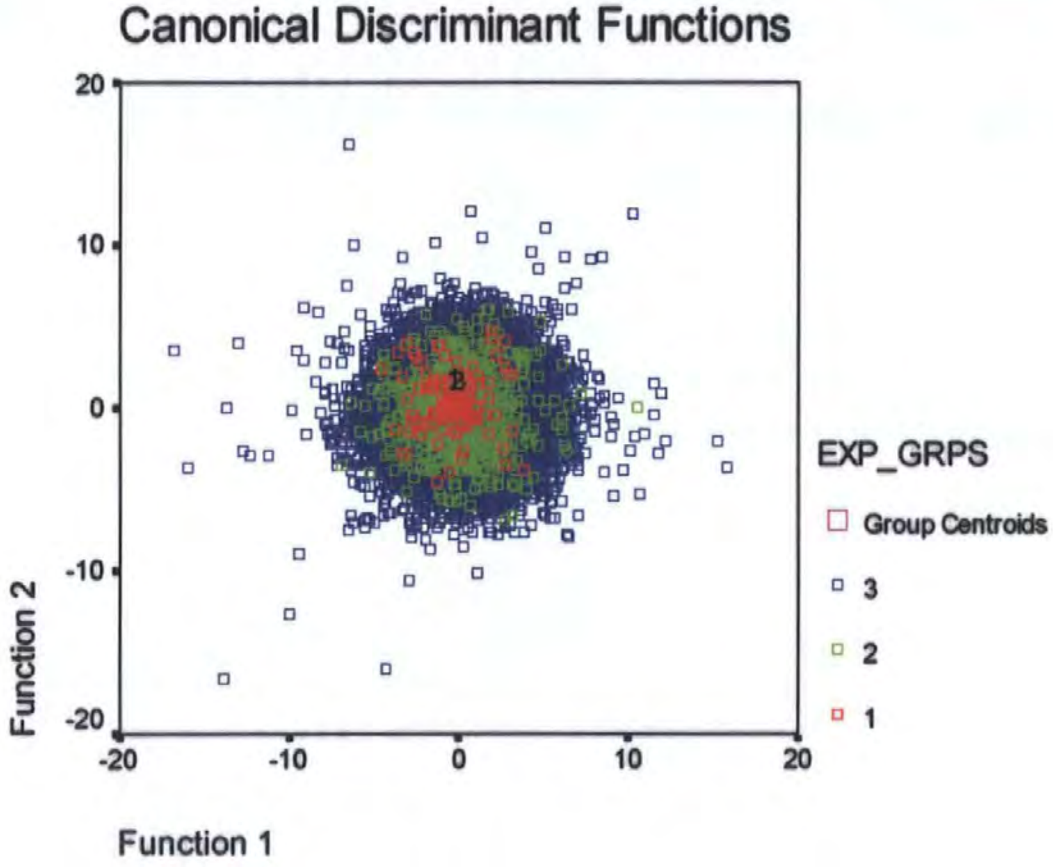


Figure 6.25 Canonical Discriminant Function Analysis – Hue x ThinPrep x Most Experienced Screeners

Table 6.34 Canonical Discriminant Function Analysis – Hue x ThinPrep x Least Experienced Screeners

		NOV_GRPS	Predicted Group Membership			
			1.00	2.00	3.00	Total
Original	Count	1.00	141	151	112	404
		2.00	1772	2655	1930	6357
		3.00	51314	66344	53167	170825
	%	1.00	34.9	37.4	27.7	100.0
		2.00	27.9	41.8	30.4	100.0
		3.00	30.0	38.8	31.1	100.0

a 31.5% of original grouped cases correctly classified.

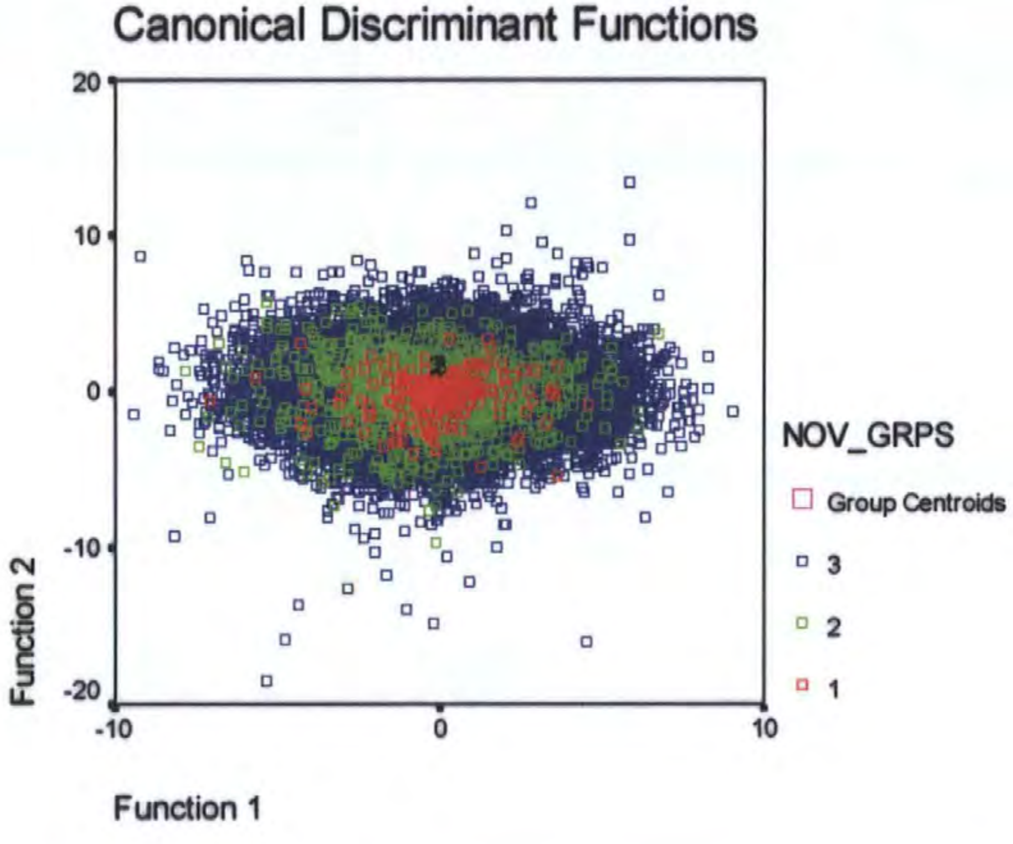


Figure 6.26 Canonical Discriminant Function Analysis – Hue x ThinPrep x Least Experienced Screeners

Table 6.35 Canonical Discriminant Function Analysis – Hue x ThinPrep x Correct Classifications

		COR_GRPS	Predicted Group Membership			
			1.00	2.00	3.00	Total
Original	Count	1.00	93	120	57	270
		2.00	1570	2197	1361	5128
		3.00	54673	69381	48134	172188
	%	1.00	34.4	44.4	21.1	100.0
		2.00	30.6	42.8	26.5	100.0
		3.00	31.8	40.3	28.0	100.0

a 28.4% of original grouped cases correctly classified.

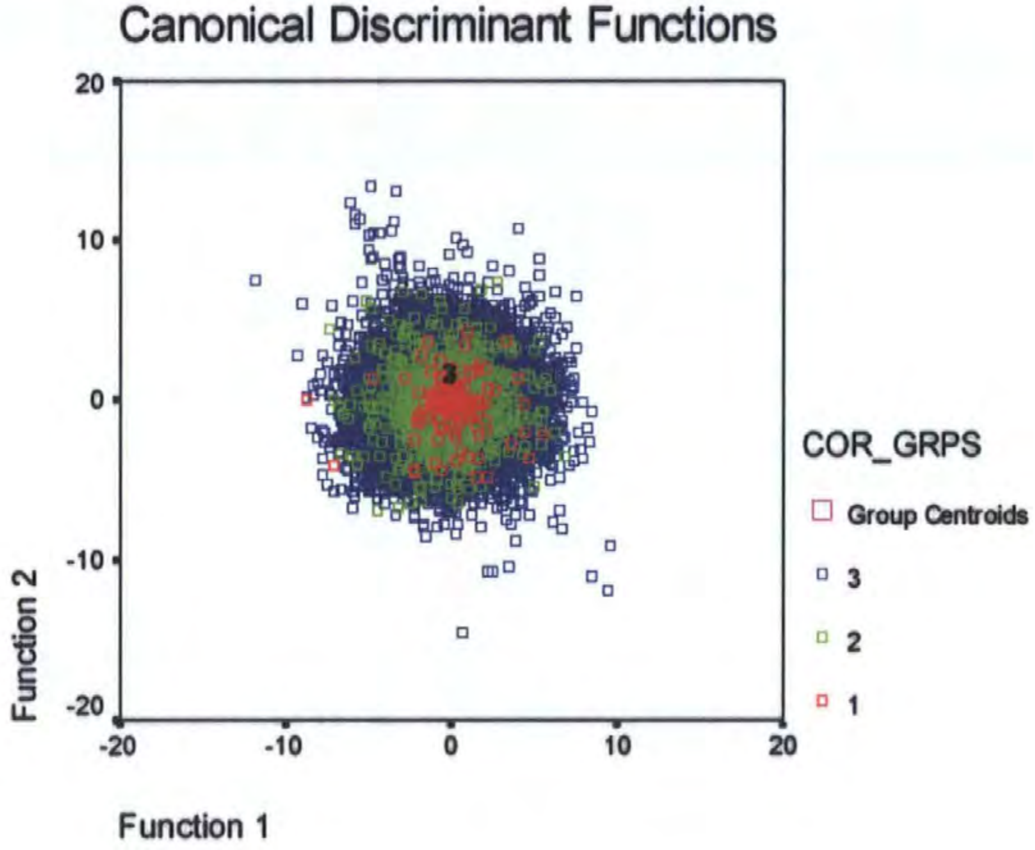


Figure 6.27 Canonical Discriminant Function Analysis – Hue x ThinPrep x Correct Classifications

Table 6.36 Canonical Discriminant Function Analysis – Hue x ThinPrep x Incorrect Classifications

		INC_GRPS	Predicted Group Membership			
			1.00	2.00	3.00	Total
Original	Count	1.00	137	129	129	395
		2.00	1658	2249	2056	5963
		3.00	47060	60572	63596	171228
	%	1.00	34.7	32.7	32.7	100.0
		2.00	27.8	37.7	34.5	100.0
		3.00	27.5	35.4	37.1	100.0

a 37.2% of original grouped cases correctly classified.

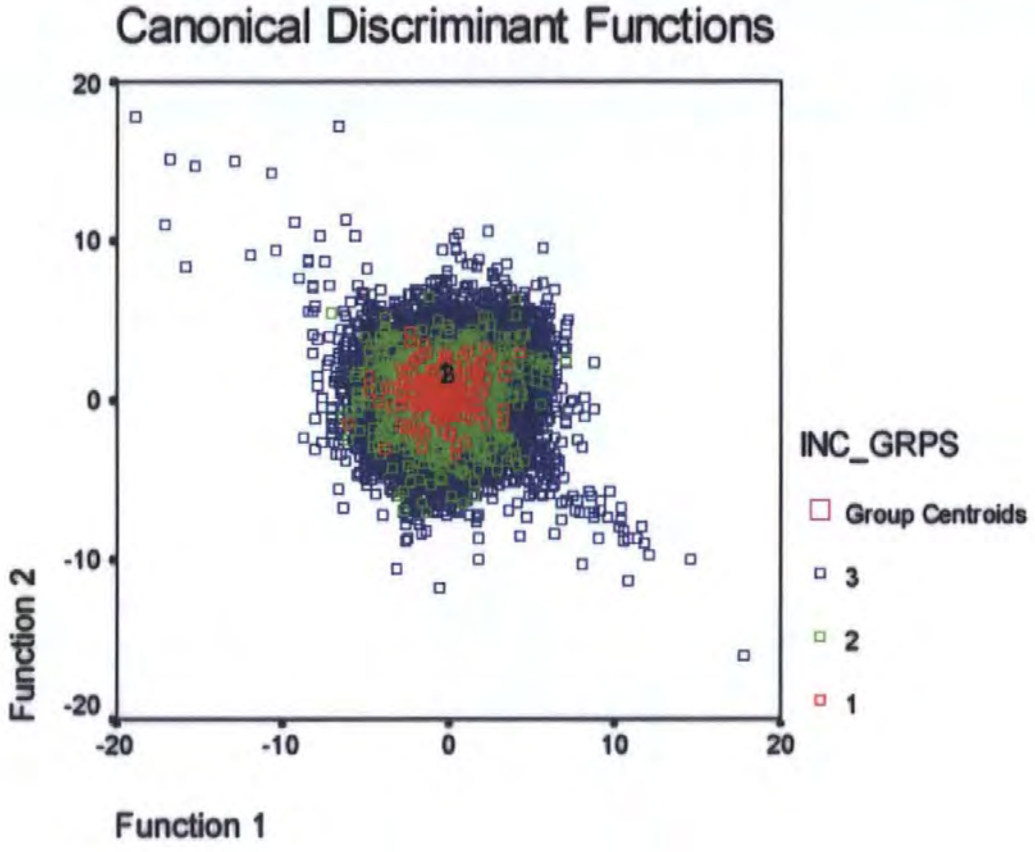


Figure 6.28 Canonical Discriminant Function Analysis – Hue x ThinPrep x Incorrect Classifications

Table 6.37 Canonical Discriminant Function Analysis – Hue x ThinPrep x Moderate+ Correct Classifications

		MCOR_GRP	Predicted Group Membership			
			1.00	2.00	3.00	Total
Original	Count	1.00	110	102	118	330
		2.00	1593	2064	2365	6022
		3.00	48057	56947	66230	171234
	%	1.00	33.3	30.9	35.8	100.0
		2.00	26.5	34.3	39.3	100.0
		3.00	28.1	33.3	38.7	100.0

a 38.5% of original grouped cases correctly classified.

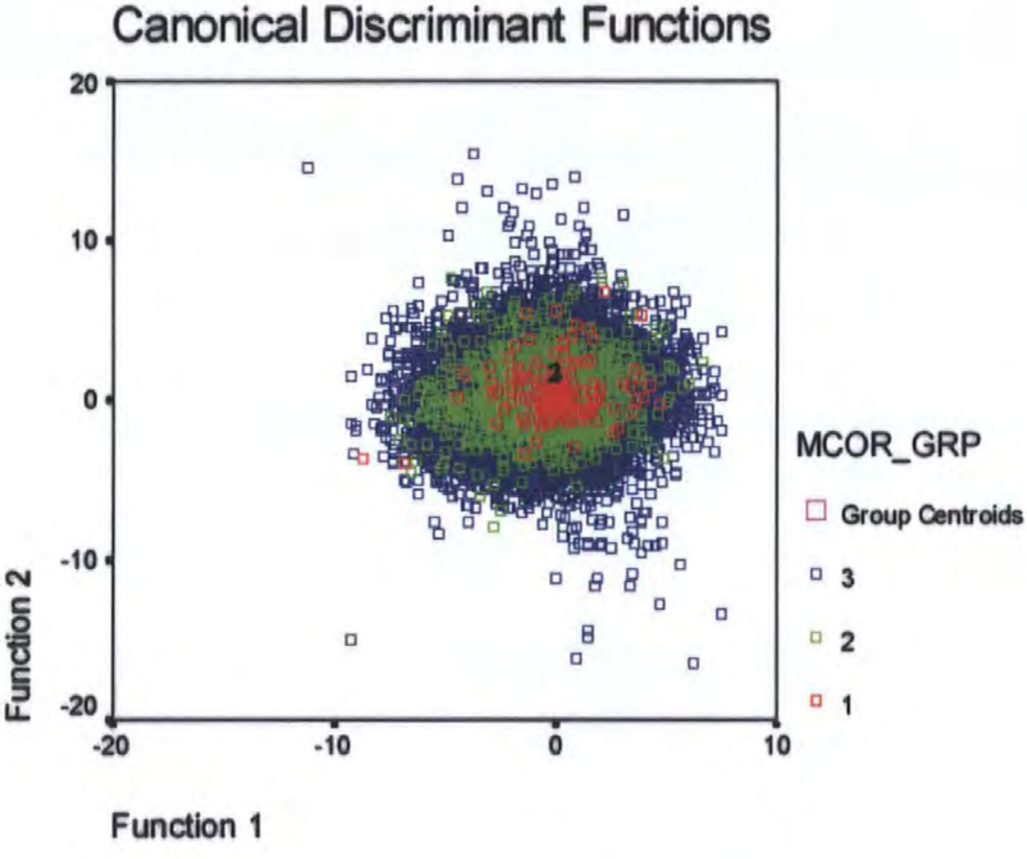


Figure 6.29 Canonical Discriminant Function Analysis – Hue x ThinPrep x Moderate+ Correct Classifications

Table 6.38 Canonical Discriminant Function Analysis – Hue x ThinPrep x Moderate+ Incorrect Classifications

		MINC_GRP	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	124	130	65	319
		2.00	1739	2217	1155	5111
		3.00	57073	69629	45454	172156
	%	1.00	38.9	40.8	20.4	100.0
		2.00	34.0	43.4	22.6	100.0
		3.00	33.2	40.4	26.4	100.0

a 26.9% of original grouped cases correctly classified.

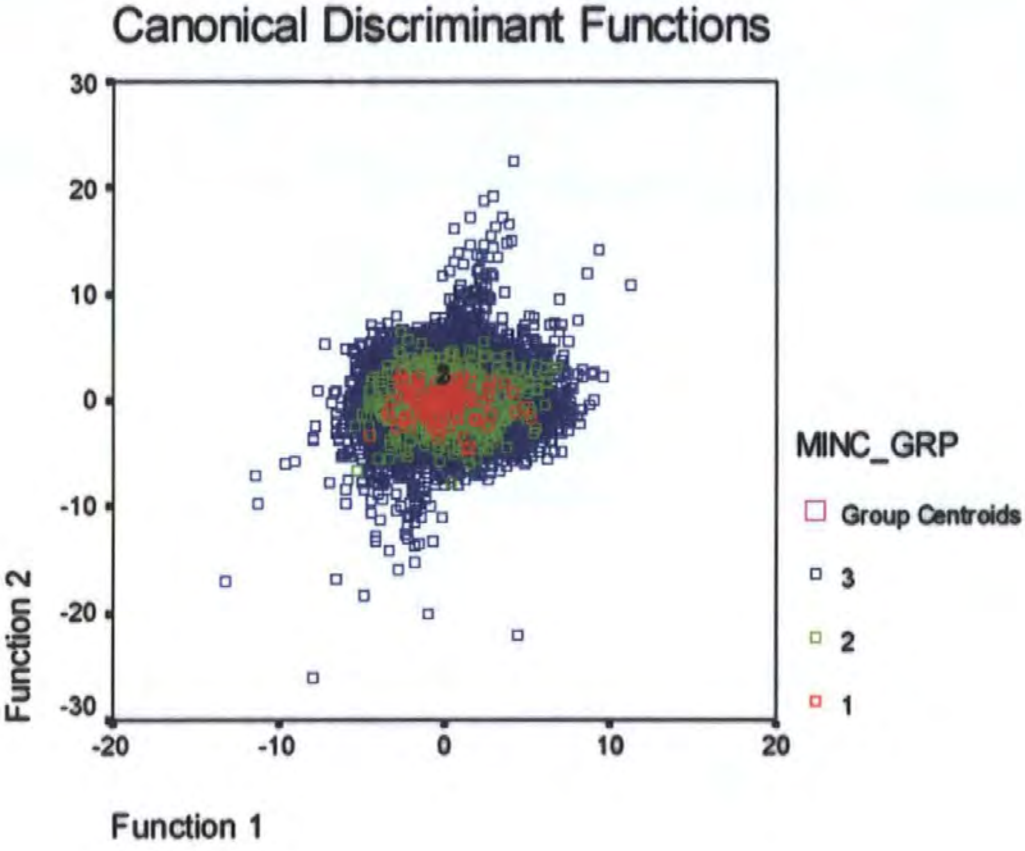


Figure 6.30 Canonical Discriminant Function Analysis – Hue x ThinPrep x Moderate+ Incorrect Classifications

The results in Figures 6.17 to 6.29 and Tables 6.25 to 6.37 show that the prediction of group membership ranges between 12.1% and 49.7%. For all of those using ThinPrep images the correctly classified groups are around the 33% mark that would be achieved by chance. This is not the case for Papanicolaou images which shows a higher predictive ability for incorrectly classified images and incorrectly classified moderate+ images when compared to the correct. While this result supports the notion of combining all of the data for monolayer ThinPrep images, it could suggest that data should be separated for Papanicolaou images. It can be argued that the misclassified images present an easier classification problem as the fixation data references the wrong colour textures. Therefore correctly classified images present a harder problem as the colour textures are more accurately and tightly defined. However, on the basis of these results, there is currently no predictive capability of the methodology described here to take a colour texture from an image and judge how salient this area of the image might be. Where predictive levels are increased they are at best only slightly higher than chance alone. This might be because the methodology does not provide a basis for predicting salient areas. However, an alternative explanation is that the evident inability to differentiate between salient areas is the result of the highly skewed group memberships. The limited number of fixations on each image results in a small proportion of the maxima being flagged as relating to salient areas of the image. Without further recording of screening fixations it would be hard to discard the methodology completely as statistically, there would have to be a very pronounced difference between the areas of the image in order for a difference to both exist and then become statistically significant. The closeness of the group centroids across all the conditions is testimony to the fact that at present, there is no differentiation. Although there is no evidence of differences between the groups across conditions, one positive conclusion is that again, the data has shown that each of the treatments has not had a significant effect on the results. While colour texture maxima density positively correlates with many of the thirteen element conditions, Atrous colour texture feature analysis does not provide a basis for predicting group membership.

6.5.6 Classification Results

Image classifications were recorded during the experimental process in order to recreate the cognitive act of making an expert judgement. While the aim of the experiments reported here was to investigate the validity of the experimental framework and test the validity and appropriateness of the machine colour texture analysis, the process also allows investigation of classification statistics. As has already been reported, these classifications have been used to ensure that there is no bias regarding correct or incorrect classifications and to examine important relationships between different conditions. Further examination of the classification statistics are also revealing in their own right.

The overall spread of classification decisions across both ThinPrep and Papanicolaou preparation methods and for all 10 participants is presented in Table 6.39. Although a category for 'other' was included on the decision grid to ensure that whatever the contents of the slide there would be an appropriate category for it to be recorded in, none of the participants used this category and so it is not reported in this table. However, a column for unrecorded decisions that occur due to limitations in the recording equipment has been included. A Borderline Changes category was also included and images that were classified as such during the experiment were processed as containing an abnormality. This mirrors the standard sensitivity calculation method currently used by the NHS.

Table 6.39 Decision classification confusion table

		Participants Classification								
		Inad	WNL	BC	Mild	Mod	Sev	S/I	GN	UR
Classification	WNL	8	***119***	13	2	58	9	0	5	16
	Mild	1	9	13	***13***	20	3	4	2	5
	Mod	1	19	7	8	***44***	3	0	2	8
	Sev	1	1	2	0	3	***7***	0	3	1
	S/I	0	1	1	7	22	8	***0***	4	7
	GN	0	8	0	0	8	4	1	***16***	3

Classification Key:	Inad	– Inadequate
	WNL	– Within Normal Limits
	BC	– Borderline Changes
	Mild	– Mild Dyskaryosis
	Mod	– Moderate Dyskaryosis
	Sev	– Severe Dyskaryosis
	S/I	– Severe Dyskaryosis/?Invasive Cancer
	GN	– ?Glandular Neoplasia
	UR	– Unrecorded

The emboldened numbers in this table show the correct number of classifications for each type of image and, with the exception of mild dyskaryotic images, the correct classification scored highest of all the possibilities. In the case of Mild, we can see that the two classifications either side, Borderline Changes and Moderate, score in similar numbers. A further breakdown of the data shown in the table can be seen in Table 6.40. This table shows each of the correct image classifications in green, while each of the acceptable classifications for moderate+ are indicated in a lighter green colour. For example, an image that was deemed to be 'Mild Dyskaryosis' would be considered correctly classified if the viewer had classified the image as such. To be considered correct under the conditions for moderate+ sensitivity both 'Borderline Changes' and 'Moderate Dyskaryosis' would also be acceptable. Likewise an image which had a classification 'Moderate/Severe Dyskaryosis' would be considered to be correctly classified if the viewer classified it as either. However, under the moderate+ sensitivity conditions, any classification between 'Borderline Changes' and '?Glandular

Neoplasia' would be accepted. While this may seem wide ranging, this methodology is in line with current National Health Service practice for rating screener sensitivity.

The spread of classifications shown in both of these tables present some interesting questions. One of the issues raised by the spread across classifications shown in is why such a high number of WNL images were classified as Moderate Dyskaryosis (58). This may be a genuine recording, indicating a human tendency to misclassify (i.e. err to false positive), and there is no doubt that for some of these recordings this will certainly be the case. What is also possible is that this is an error from the eye-tracker. The areas on the decision grid that were fixated upon to record the diagnosis for either classification are adjacent to each other and as such a drift in eye-tracker calibration may also account for some of this spread. It is also worth noting that the calibration screen, a single point presented in the middle of the screen after the decision grid, has the point on which the participant fixates in the same area as that for a Moderate classification. Although unlikely, what we might be seeing is a manifestation of someone pre-empting this screen, in which case we would record a moderate judgement for the appropriate image. However, if eye tracker errors were responsible we would have expected other errors of classification with a similar calibration bias for other categories such as Inadequate and Mild or Borderline Changes and Severe. As we can see, although there are far fewer images that fall into these categories, this is not the case. In addition it is unknown why Moderate classification should show a tendency towards Within Normal Limits. Reducing the errors potentially created by the eye tracking equipment at future experiments would certainly go some way to solving this issue or at the very least reduce the number of potential errors that could be created.

Table 6.40 Distribution of correctly classified images

image number	Inadequate Specimen	Within Normal Limits	Borderline Changes	Mild Dyskaryosis	Moderate Dyskaryosis	Severe Dyskaryosis	Severe Dyskaryosis/ ?Invasive Cancer	?Glandular Neoplasia	Other	Unrecorded
1	0	0	0	3	3	0	1	1	0	2
2	0	7	0	0	0	1	0	1	0	1
3	0	0	0	0	4	3	0	1	0	2
4	0	4	0	0	4	0	0	0	0	2
5	2	2	1	0	3	0	0	0	0	2
6	0	6	1	0	1	1	0	0	0	1
7	0	0	1	3	3	1	0	1	0	1
8	0	4	0	1	3	0	0	0	0	2
9	0	2	3	2	1	1	0	0	0	1
10	0	5	0	1	3	0	0	0	0	1
11	2	1	1	0	3	2	0	0	0	1
12	0	5	1	0	2	1	0	0	0	1
13	0	6	0	0	2	1	0	0	0	1
14	0	3	2	1	3	0	0	0	0	1
15	1	4	1	0	2	1	0	0	0	1
16	0	2	0	0	3	3	0	1	0	1
17	0	0	1	5	2	0	0	0	0	2
18	1	0	0	0	4	3	0	0	0	2
19	0	0	1	0	6	2	0	0	0	1
20	0	4	0	0	3	0	0	1	0	2
21	1	6	0	0	3	0	0	0	0	0
22	0	4	1	0	4	0	0	1	0	0
23	0	0	0	0	6	3	0	0	0	1
24	0	6	0	0	3	0	0	0	0	1
25	0	1	2	1	5	0	0	0	0	1
26	0	2	0	0	2	1	0	5	0	0
27	1	9	0	0	0	0	0	0	0	0
28	1	3	2	0	3	0	0	1	0	0
29	0	2	1	2	4	1	0	0	0	0
30	0	1	1	0	8	0	0	0	0	0
31	0	0	1	1	6	0	0	1	0	1
32	1	1	1	0	3	2	0	2	0	0
33	0	6	0	0	3	0	0	0	0	1
34	0	6	0	0	3	1	0	0	0	0
35	0	1	0	0	5	1	0	2	0	1
36	0	6	0	1	3	0	0	0	0	0
37	0	1	0	2	5	1	0	1	0	0
38	0	6	0	0	2	1	0	0	0	1
39	1	3	2	0	3	0	1	0	0	0
40	0	0	0	0	3	1	1	4	0	1
41	0	7	0	0	2	0	0	0	0	1
42	0	2	0	0	3	1	0	3	0	1
43	0	1	2	1	5	0	0	0	0	1
44	0	6	1	0	3	0	0	0	0	0
45	0	1	3	1	3	0	2	0	0	0
46	0	6	2	0	2	0	0	0	0	0
47	0	1	2	0	6	0	0	1	0	0
48	0	5	1	0	3	0	0	0	0	1
49	0	6	0	0	2	1	0	1	0	0
50	0	4	1	1	4	0	0	0	0	0
Total	11	158	36	26	162	34	5	28	0	40

Perhaps more representative is Figure 6.30, which shows the mean scores for increasingly experienced groups of individuals. These are calculated by ranking the participants in order of experience with the least experienced occupying first place. The first points on the graph are then plotted using the people ranked first, second and third. When this has been completed, the person in first place is replaced by the person in fourth and the second points are plotted. This continues until the last points are those representing mean performance scores for the most experienced participants, those ranked as eighth, ninth and tenth. This measure is therefore a rolling average.

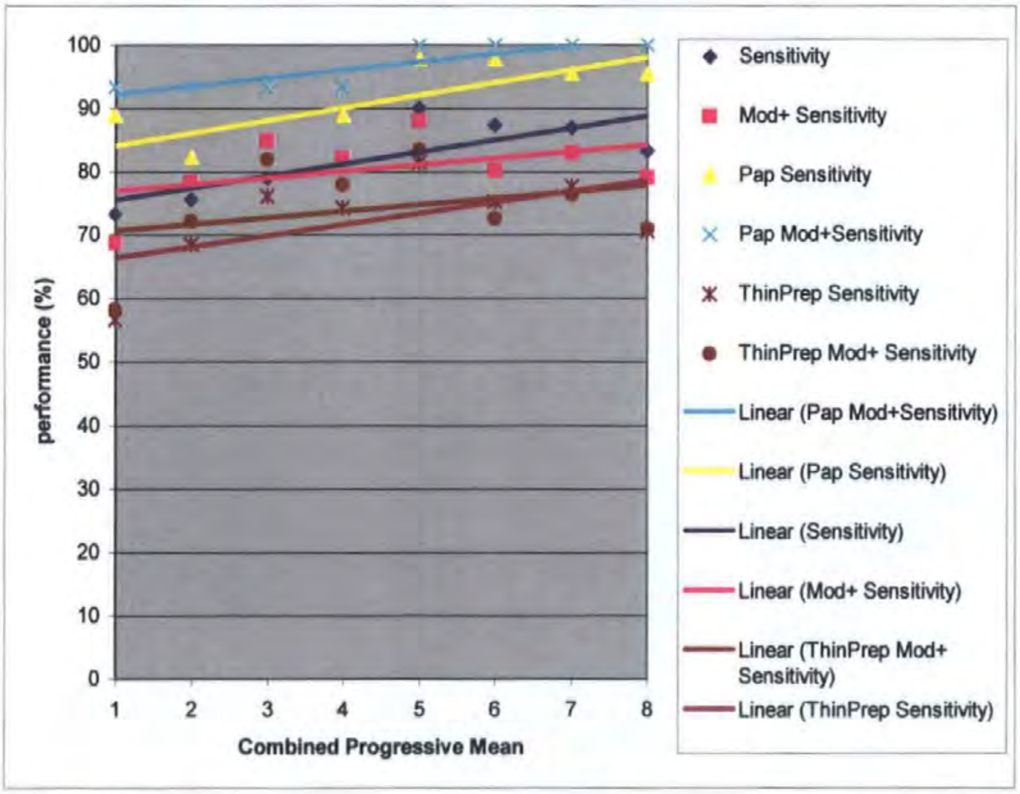


Figure 6.31 Shows the average sensitivity for progressively increasing experience levels

This method of examining the mean performances allows exploration of the effects of subject training and experience. A best-fit line across these rolling averages shows that there is a steady improvement across all types of sensitivity as screeners become more experienced. This is hardly surprising, but does help to support the method of recording fixation data and subsequent analysis presented in this thesis as the expected improving performance gained with experience is evident. While sensitivities recorded are below the standard that is required by the NHS, accounting for noise in the data created by small calibration errors a slight drop in sensitivity performance can be accounted for.

6.5.7 Pupil Diameter Results

The data recorded by the eye tracker also provided information regarding the pupil diameter of each participant throughout the experiment. A linear regression for this data was carried out using time as a sequential independent variable. It showed that for participant one there were no significant differences in pupil diameter throughout the trial (F(1133)=1.06, p>.05) but for the others this was not the case. Participant ten showed a significant decrease in pupil diameter (F(1754)=4.34, p>.05), and all but one of those remaining showed a very significant downward trend in pupil diameter (F(1797)=229.88, p<.001; F(713)=34.98, p<.001; F(1818)=55.06, p<.001; F(1160)=139.96, p<.001; F(1229)=122.77, p<.001; F(1577)=246.79, p<.001; F(1123)=17.25, p<.001). The final case showed a significant difference for participant eight, but for this participant, their pupil size increased throughout the trial. This difference was not quite as marked as those showing decreases in pupil size but remains unique within our set (F(1989)=5.44, p<.05).

In Table 6.41 we can see the collected pupil diameter and fixation descriptive statistics for each individual who took part in the first trial. It should be noted that the fixation numbers listed here are for the whole recording and are produced automatically by the eye tracker. This is higher than the number of fixations used during the pupil diameter significance testing as it included initial calibration screens and end screens. These were removed from the significance testing so that only the experimental part of the image presentation has been assessed.

Table 6.41 Pupil diameter and Fixation Statistics

Partici-pant	Exp. Level	Trial Duration (mins)	Total Fixs	Mean Fix Duration (secs)	Fix Freq. (fix/s)	Pupil Diam. Change	Pupil Diam Sig
1	High	10.01	1145	0.307	1.9	none	.303
2	Low	13.08	1868	0.358	2.37	decrease	.000
3	High	10.18	772	0.340	1.25	decrease	.000
4	Low	16.58	1834	0.283	1.8	decrease	.000
5	High	8.29	1186	0.350	2.33	decrease	.000
6	Low	10.33	1284	0.388	2.03	decrease	.000
7	Low	15.24	1689	0.471	1.83	decrease	.000
8	Low	15.22	2045	0.417	2.22	increase	.020
9	High	8.56	1141	0.432	2.13	decrease	.000
10	High	12:05	1770	0.342	2.44	decrease	.037

In Figures 6.31 and 6.32 we can see two examples of the pupil diameter (reported in eye tracker units) plotted against fixations. One of the interesting things to note here is the spike in pupil size that occurs in 6.31 at the beginning of the presentation. This spike noticeably occurred in half of the pupil diameter recordings.

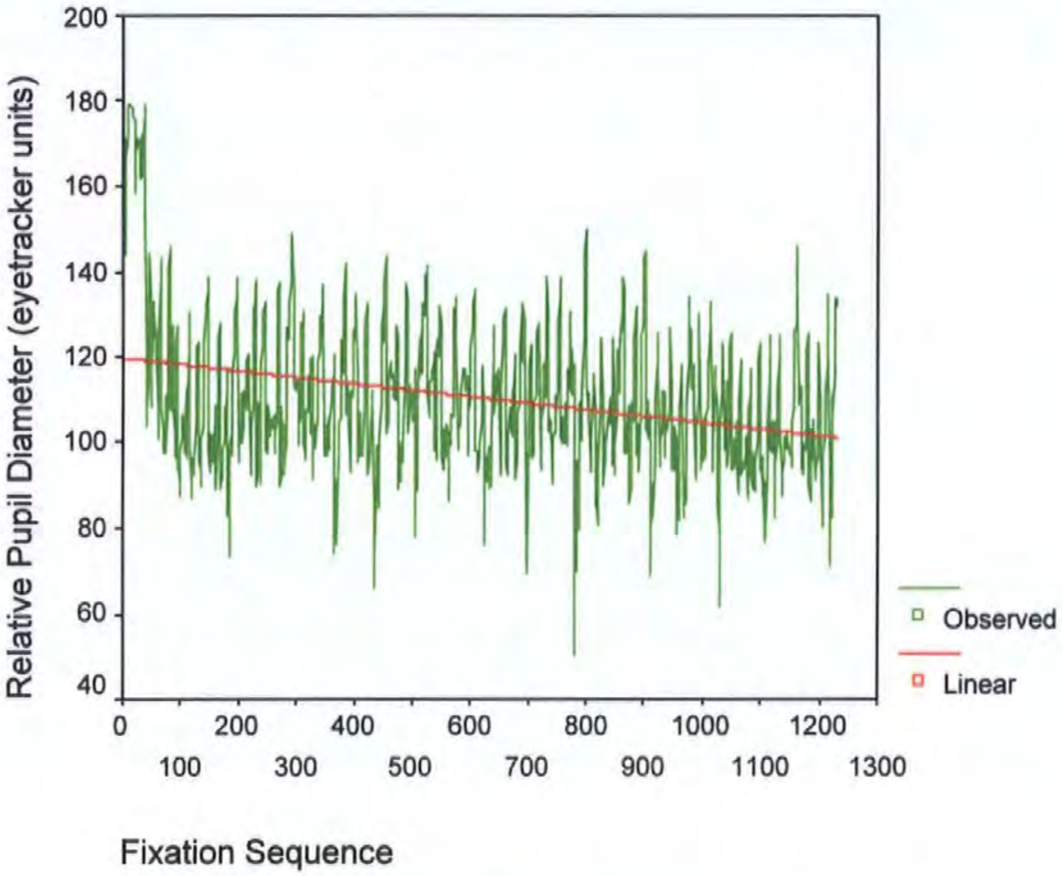


Figure 6.32 Shows a drop in pupil diameter for participant six throughout the experiment (F(1229)=122.77, p<.001)

The fixation statistics also show some interesting trends with respect to levels of experience. Shown in Table 6.42 are the mean averages for overall time taken, number of fixations, fixation duration and fixation frequency. This would seem to confirm that experienced screeners use different techniques than those who are less experienced. While fixation frequency and duration are similar figures, the small differences shown here are compounded throughout the image presentation. The length of time spent examining the images and deciding upon a classification shows an average difference of over four minutes. As we have seen, the experienced group showed better performance scores than the less experienced and this confirms that they also took less time and total fixations to make their decisions.

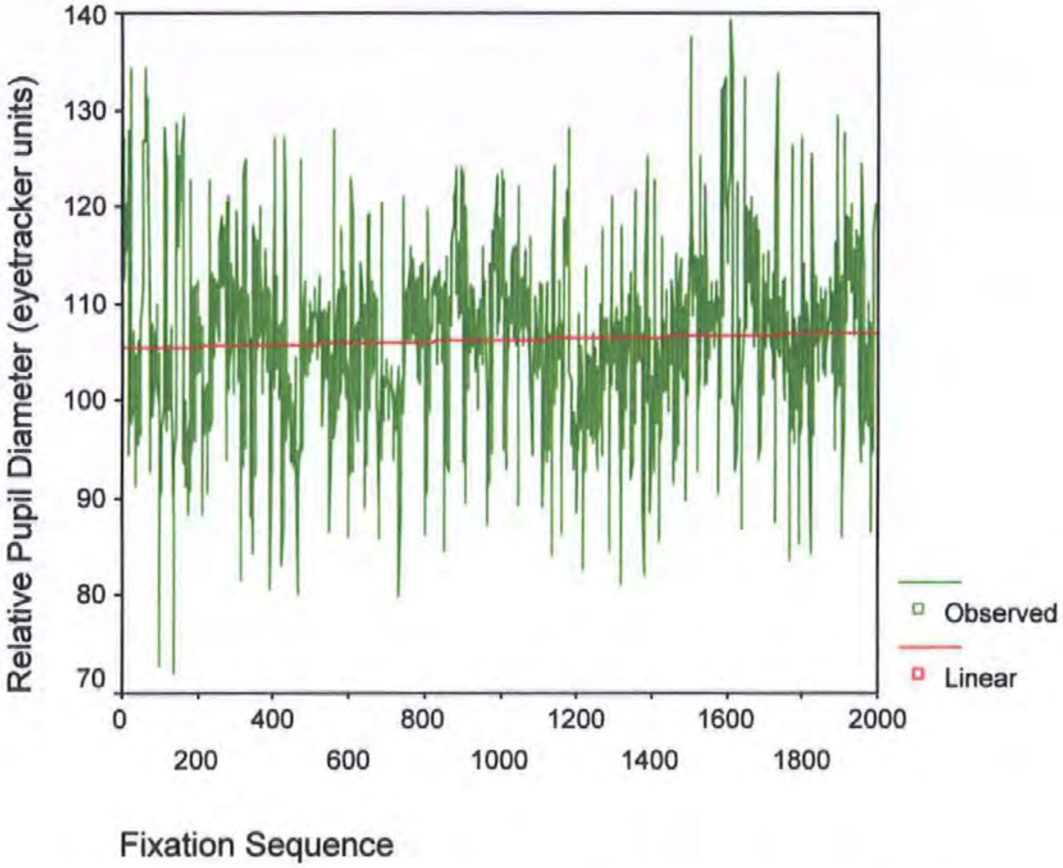


Figure 6.33 Shows an uncharacteristic increase in pupil diameter for participant eight throughout the experiment (F(1989)=5.44, p<.05)

Table 6.42 fixation means for most and least experienced groups

	Mean Time (secs)	Fixations	Fixation Duration	Fixation Frequency
Most Experienced	597.8	1202.8	0.3542	2.01
Least Experienced	857	1744	0.3834	2.05

It is also shown that there is a general trend for pupil diameter to shrink as the experiment proceeded. It is believed that the pupil slowly fluctuates completing a cycle in 25 – 50 seconds when fatigue is present with a general trend to shrink as fatigue sets in. A study by Yoss, Moyer, and Hollenhorst (1970) examined this fluctuation in airline pilots. Only 12% of those who had been well rested before the study showed this type of fluctuation, compared to 50% who had been poorly rested. We have shown that there was a shrinking of the pupil diameter through the experiment (taking approximately between 8 and 16 minutes to complete depending on the participant). This is a surprising finding given that the trial was so short, and that all but three people showed a similar significant downward trend in pupil size. Of those three participants one shows a significant upward and the other two show no significant differences. This data needs further exploration to see if there was a general fluctuation throughout the trial of the type indicated by Yoss *et al.*, however a fatigue based explanation would seem applicable. The initial spike seen in some of the recordings could be explained in terms of a reaction to taking part in the experiment. As each participant becomes more comfortable with the task they are undertaking, it is possible that they would relax and previous pre-experimental fatigue levels would be restored.

6.6 Summary of Results

The experiment produced three data types, which were interpreted using the basic model shown in Figure 6.8. Individual performance profiles were extracted revealing screener sensitivity, image coverage, abnormality coverage and saliency coverage. This also allowed the creation of an index based on saliency with which the machine colour texture analysis of images could be correlated using a data verification technique.

The main results from this analysis are as follows:

- Sensitivity performances for all conditions were at a level expected for qualified screeners

- Image coverage was generally less for more experienced screeners. There were no significant differences between the conditions

- Saliency coverage results show larger differences with ThinPrep images than for Papanicolaou images, and a general non significant trend for more experienced screeners to view less of the images salient areas. This trend reversed only for the correctly classified Papanicolaou images

- Abnormality coverage showed a significant difference between experience levels for moderate+ incorrectly classified images. A general non-significant trend shows that those with more experience view less of the abnormalities.

- A comparison of colour texture maxima density with saliency index 2 shows significant and highly significant correlations between density and saliency for the thirteen element vector's hue component. These differences are evident throughout a number of conditions. Differences between classification conditions are reported, though these are not consistent enough to draw strong conclusions.

- There is no predictive capability for group classification based on the machine colour texture canonical discriminant functions

- Classification data shows that performance levels improve with experience, suggesting that the methodology tests for an appropriate latent trait

- Pupil diameter data showed a significant decrease for eight participants, no change for one participant and an increase for one participant.

6.7 Conclusions

The experimental work in this thesis presents evidence to support the method of data verification. It shows that the basic methodology is capable of providing a framework for objective testing of the colour texture image analysis. While there is no evidence to suggest either the four or thirteen element vectors could provide a predictive capability based on colour texture, maxima density for the thirteen element hue component was highly significantly correlated with salient areas of each image. A refined version of the existing image analysis based on maxima density could provide a stronger predictive measure capable of locating salient areas on novel slide images. There are many avenues for further work to explore, and these will be discussed in the following chapter along with the implications this thesis has for existing work in the field.

Chapter 7

Discussion and Conclusions

7 - Discussion and Conclusions

7.1 Introduction

This thesis has presented a model for a novel approach to quality assurance of human expert cytological slide inspection. Relevant literature has been reviewed before the rationale for the experimental work that has taken place. The experimental work has been described and the results interpreted. In this chapter, the progress against project aims is reviewed, before the work is discussed in relation to the existing cervical screening programme and the current scientific literature. Finally, future possibilities for further work will be discussed before a conclusion is reached.

7.2 Review of Project Aims

This project has produced mixed results but there have been a number of successes. Before examining the implications from the evidence that has been presented, the project aims will each be examined.

Aim 1 - To provide a training tool for quality assurance assessment using gold standard images for use by histopathology laboratories.

The image set that has been used in this study constitutes a gold standard. They provide a reference set with which screener performance can be assessed on a number of measures. Image coverage, saliency coverage and abnormality coverage can all be evaluated, along with the modified sensitivity measures and fatigue. While the trends that can be seen in the data indicate that each of these measures is being successfully recorded, the statistical analysis has been inconclusive in many areas. It is for this reason that the novel approach to assessment that has been presented could not be used as a training tool until further fixation recordings have been analysed. This would strengthen the statistical analysis and, if the trends in the data are consistent, provide the statistical significances that are currently lacking. The experimental methodology could then be considered an effective training tool and this aim would be achieved

Aim 2 - To provide routine performance measurement assessment of cervical cytology screening using gold standard images.

The gold standard images have been used in an eye tracker based classification exercise and the results show consistently high performance levels based on the modified sensitivity and sensitivity + calculations. This would indicate that the methodology could provide routine measurement assessment. However, statistical analysis has not provided the support that would be required to recommend the methodology for routine assessment. As with the first aim, if the trends in the data remain as more participant data is recorded, the statistical basis that is essential would be provided.

Aim 3 - To provide online performance measurement and assessment of cervical cytology screening using images that are not gold standard.

To achieve this aim, the model would need to be extended to allow use of non-gold standard image presentation, through data verified using a new objective measure of salient areas. While it was not possible to provide performance measures based on non gold standard images that would indicate when salient areas have not been viewed, the principle behind achieving this has been demonstrated. The machine analysis of images provided mixed results, with the colour texture analysis showing no better predictive ability than chance alone. However, maxima density on the hue element of the thirteen element vector is highly significantly correlated across a large number of differing conditions. Unfortunately, unless the standard performance measures described in the first two have strong statistical basis, it would be difficult to justify basing a performance measure on maxima density. This is because the evidence based on the eye tracker fixations generally shows non-significant trends. Without being statistically certain of what is being measured in the first place, the significant correlations could relate to another variable other than saliency.

7.3 Discussion

The work presented in this thesis has shown that a new approach to quality assurance of human cervical cytology screening is feasible. Evidence has been presented that shows classification levels were at levels that would, under current NHS guidelines, be considered acceptable. A number of further measures have shown interesting trends without being statistically significant. The motivation for this has been to improve the quality assurance of cervical cytology screening by reducing the levels of intra- and inter-observer variation, both within and between laboratories.

The key to reduce the levels of variation that manifest during standard screening is to provide feedback to the individuals who carry out this task, as demonstrated in the study by Jones, Thomas, and Williamson (1996). This suggested that supplying feedback could reduce these levels of variation, and that simply discussing the way in which classifications are reached by screeners can be beneficial. The methodology described in this thesis, which uses data from a number of sources to provide information about performance, would allow performance feedback to be given to screeners viewing the gold standard image presentation while wearing an eye tracker. With a stronger statistical basis, the level of information that could be given to a screener about their performance is far beyond that provided by the existing quality assurance measures. For histopathology this could be very beneficial as it would allow examination of the classification sensitivity level and information about the general approach taken by each screener. In the field of cervical cytology quality assurance, this would represent the first time that eye tracking has been applied for this purpose. Providing information that is detailed enough to tell a screener not only their sensitivity levels, but also which specific aspects of their screening ability is falling below the level of other screeners would be a very positive addition to current quality assurance practices.

Given that this type of screener profiling appears to be a possibility, the next question is exactly how it could be implemented in histopathology laboratories. Naturally, there are a number of considerations that have to be addressed. These relate to ethics, time, cost and training. Before any new technology is introduced into a medical field, it has to satisfy a number of criteria, particularly if the advance in medical care that it may make is a small one. In the UK in particular, where the National Health Service provides the large majority of healthcare, any new technology has to be easy to implement. A piece of equipment that is difficult or complex to use is likely to take a lot of time and money away from other applications and this has been one of the fundamental reasons that some automated systems have failed (Broadstock, 2001). Time is perhaps the most valuable resource that cervical cytology screeners have, as they are limited in the amount of screening they can do in each day, so the idea that a person might have to retrain to use a new piece of equipment is always going to be problematic if the advantages are not immediate. Indeed, it is acknowledged in the LBC technology appraisal that introducing LBC will not drastically differ in quality of service from using Papanicolaou slides. It is the suggestion that fewer inadequate slides might save time in the long run that is the main reason for its acceptance (NICE, 2002).

One of the benefits of the methods and profiling presented by this thesis is that it would not disrupt the day-to-day activities of a busy laboratory. In its present form, the software allows limited self-testing based on eye tracker data and gold standard images. This process would take no longer to complete than screening the test slides that are part of the existing quality assurance program. Furthermore, the straightforward operation of the analytical software makes it easy to train people who might use it. After viewing the presentation, the file of fixations produced by the eye tracker is analysed with just a few clicks. Operating the software that performs this analysis is easy and intuitive, so minimal training would be required. There is also an advantage in running in parallel with the existing system that should not be overlooked. All of the analysis carried out so far has been aimed at enhancing the skills of human screeners rather than replacing them. As such, any use of the methods developed in this thesis would not present a problem should they be introduced into a laboratory.

Aside from the practicalities of introducing the type of quality assurance proposed in this thesis, there are also scientific implications that have to be considered. One of the more interesting observations from this work has been in the different trends seen between screeners with different levels of experience. More experienced screeners covered more of an image's area with fewer fixations suggesting that there may be differences in the way images are actually viewed. Less experienced screeners made more fixations on images, but these tended to be closer together and covered less of the image area at higher visual angles. While this could suggest that screeners with more expertise are using peripheral and parafoveal information from each fixation to find the location and direction of their next saccade and subsequent fixation more effectively, further fixation data would provide a definitive answer. Should the trends that have been shown in the data persist, the differences between those with different levels of experience would become significant.

Although this is a non-significant trend, it is worth considering the possible reasons as to why this trend was seen in the data. The first explanation would naturally be that this is an anomaly that will disappear as further experiments recorded more data. If this is not the case then there are two further possible explanations for this trend. Either more experienced screeners use this information, when less experienced do not, or both groups process the same information but the decision regarding the next fixation location is improved as experience is gained. Certainly the second of these two options is far more plausible as studies of expert judgement have shown. It is not the amount of information that experts have, as they are viewing the same images as less expert screeners. The difference is in how that information is used. It is likely that the experienced screeners are fixating on more relevant areas of the images in order to make a diagnosis decision and selection of these is based on their previous experience of screening. Less experienced screeners appear to spend more time searching an image for the relevant diagnosis information it can give them. This finding is in line with Yarbus's (1967) study of eye movements while evaluating paintings. Participants showed systematic preferences to repeatedly view the areas of the paintings that could most help them evaluate the pictures contents. When they were asked for information relating to different areas of the paintings, their searches were adjusted accordingly. In the context of cervical screening, this would suggest that the screeners are all viewing the information that they consider to be more relevant, with a better choice of which areas to inspect being made by the experienced screeners. While primary screening strategies are designed to cover the entire slide area, the implication for rapid review screening would be that more experienced screeners would pick areas to review that provide more information than less experienced screeners.

The image analysis that was used as part of this thesis was also rather unique for several reasons. The images were a lot larger than are used in the majority of image analysis studies at a resolution of 2048x1536x24b, with the file size for each image being 9,217kb. These images were also not part of a controlled set for the purpose of testing computer vision analysis. As such they represented a challenge for automatic analysis techniques. As we have seen, these issues were overcome with limited success. While the colour texture analysis could not predict saliency based groups, highly significant correlations were shown between eye tracker fixation data and the density of maxima surrounding salient areas for the thirteen element hue component of the images. To some extent this helps to validate the method of image analysis and demonstrates the ability of the analysis to handle such large image sizes. Very few systems use combined colour and texture information, and often when they do they treat colour and texture separately before combining the information (Van de Wouwer, Scheunders, Livens, Van Dyck, 1999). Recent attempts to combine both colour and texture information for the purpose of feature extraction have had varying success, and largely take one of three different approaches. Some derive textural information from the luminance plane along with pure chromatic features (Carson, Belongie, Greenspan, Malik, 2002), while others derive textural information from chromatic bands, extracting correlation information across different bands (Paschos, 2000; Mirmehdi and Petrou, 2000). The final group are those that process each colour band separately by applying monochromatic texture analysis techniques (Caelli, Reye, 1993; Thai, Healey, 1998). Caelli and Reye pointed out that the correlations existing between different colour channels over space determine the regions, textures and colours perceived by humans.

The approach to image analysis described in this thesis is most similar to this third group, as each image was transformed into its HSV components before a low-resolution multiscale analysis was performed. However, the analyses of cervical images were distinct because of the way in which each of the image's components was handled. While Hue was analysed separately, the Saturation and Value dimensions were combined. This was because Hue was expected to be a stronger measure, and although Saturation and Value were combined, the information in the combined measure encapsulated information from both separate measures. Although the method shares similarities to other studies, it is distinct within the image analysis literature.

7.4 Future Work

There are a great number of avenues that further work may take. Perhaps the most pressing of these is to collect further fixation data in order to verify or discount some of the trends that have been shown. During the course of this work, a large number of hospitals were approached to take part in this study to provide the validity that comes with testing on more than one site. While two further hospitals agreed in principle to be involved, this agreement came at a point where it was too late to include in this work. It is hoped that the next extension of this work will be carried out on more sites and with larger numbers of people. The work presented in this thesis, although not conclusive, provides enough evidence to support for further experimentation. This will be aided by the advancements in eye tracker technology specifically aimed at tracking the gaze on a computer screen. While this technology is primarily marketed towards researching website users, it will be of huge benefit when recording new participants. It will not only allow quicker easier testing, but also mean that more images can be added to the presentations giving even more data to work with. It will circumvent the need to use a bite bar to keep the head relatively still that has restricted the length of our trials to less than 15 minutes for the sake of comfort. Instead, more images could be presented over a longer period of time allowing more salient or abnormal features to be identified.

Further work also needs to be carried out to improve the ability of the colour texture analysis to predict saliency on non gold standard images. The thirteen element vector used in the current thesis demonstrated that it may be possible to predict saliency based on maxima density, but unless further fixation data is added, it cannot be absolutely certain that the measure is reliable. This is because there is not enough evidence to indicate the eye tracker data that has been recorded accurately shows salient areas. As such, the highly significant correlations may be correlated with a different aspect to that which the test is aimed at. If we consider for a moment that this correlation does allow the prediction of salient areas, further exploration of the number of vectors being used could also be beneficial. For example, the few negative correlations that are seen for the combined saturation and value measure could reveal more consistent negative correlations with fewer vectors. This could enhance the information provided by the hue component.

While much of the eye tracker's recorded information was used during the experimental work, there are three particular attributes that are recorded and are currently not included in the automated analysis that the software provides. Perhaps the most important of these is fixation duration. It is reasonable to suggest that a fixation lasting for a few seconds is more important in the decision process than a fixation lasting for a fraction of a second. Currently, there is no differentiation between the two types of fixations but adding this to the analysis should increase the sensitivity of locating salient areas. In addition, the interfixation degree and interfixation duration may also hold some valuable information regarding strategy. Interfixation degree refers to the calculated angle of each saccade from one fixation to the next. Interfixation duration refers to the time taken to move from one fixation to the next. This thesis has shown that there may be differences between experienced and inexperienced screeners in their choice of fixation and the distance the eyes travel to that point. Using these two additional attributes, a full scientific investigation of these differences can take place. In addition to these unanalysed attributes, the recording of the pupil diameter is not automatically analysed for significance. An interpretation and graphical representation of this should be included in the descriptive statistics that are produced for each screener as it can indicate fatigue levels. It needs to be easy enough for the cytology screeners to interpret and scientific enough to be meaningful. The upgrade of eye tracking equipment may also provide recordings on other attributes that have not been available at this time.

One interesting possible long-term aim is the possible use of eye trackers in microscopes. Existing technology is capable of adding an eye tracking element into such things as digital cameras. It is not hard to imagine that the same technology may be able to be placed into a screening microscope and, while a screener views a slide, a computer system can analyse their fixations. This would allow a screener to perform in their normal manner while a computer records and analyses the information from each fixation. This concept might also allow for a centralised database that contains information on thousands of fixations and their likely saliency. Should automated systems be introduced and human screening phased out, this would provide a valuable tool for development of cytology analysis systems.

Another area that will also provide valuable information is that of Liquid Based Cytology. While it has been decided in the UK to introduce LBC, there has not been a decision made on exactly which method to use (NICE, 2003). The methods described in this thesis can benefit this decision in two ways. Firstly it can compare different performances across the different preparations being considered in the same way that Papanicolaou and ThinPrep images have been compared. The second way it can be beneficial is in indicating when a screener reaches a performance level on LBC preparations that is comparable to those being achieved on Papanicolaou slides. It can provide a threshold for conversion once a predefined performance level has been reached.

For image analysis, thinlayer and monolayer preparations are of great interest as they provide an easier vision problem to solve than the cluttered Papanicolaou preparations. Currently, when screening a Papanicolaou slide, a screener will focus up and down, reflecting the three-dimensional nature of the cells. When there is only one layer, there is no need to keep refocusing like this as all the cells are on the same plane. As such they represent an easier problem for computer vision techniques to solve.

Finally, the software developed to automate the statistical analysis could be further improved. During the software development process, improvements were made to the programming language and the capabilities of the platform and operating system on which it runs. Because the process of creating the application had already started, it would have been costly and time consuming to convert to these newer versions. However, it would have undoubtedly provided more stability and speed. While the existing software is complete and capable of many powerful analyses, this is just one example of a way in which the overall package could be improved in the future. As computers become more faster and more capable there will be a noticeable difference to the performance of the analytical software as the areas that are currently computationally heavy will also become quicker at making the necessary calculations.

One final point to consider is that the methods described here may also be of use to other areas of research. Where an expert is involved in the task of making a visual classification, the method of deriving salient features could also be applicable. Areas that might benefit from this work include medical judgements based on images, such as X-ray or Electrocardiograms. Biological species taxonomy might also benefit, as might image retrieval databases by using the data verification method to discover what the most important features for each of the species. This could be extended to uses such as image retrieval from databases by examining which elements are important to humans when searching for images. Furthermore, this would allow objective testing of automatic computer analysis performance in any domain.

7.5 Conclusions

In conclusion, extensive work has been reported in the development of a computer system capable of predicting the areas of cytological images that are salient to the human screener who makes the cytological diagnosis. The concept of analysing images for salient features rather than abnormal ones has been introduced with evidence supporting this approach. The eye tracker has been demonstrated as a viable research tool for research in this area with the aim of providing salient areas and features based on a screener's fixations across cytological images. A feature marking exercise has been introduced as a method of classifying fixations. Evidence has been provided to show that a combination of these data types can be used for the analysis of cytological images, and that the image analysis methodology may be applicable to these types of images. Mixed results have been presented that support further investigation of both the methodology and analytical processes used throughout this thesis. The work presented in this thesis not only provides a research basis for further work, but also provides a methodology and analysis software aimed at a real world application. It also represents a multi-disciplinary solution to a complex image analysis problem.

References

American Society for Clinical Pathology (n.d.) Retrieved August 1St 2004, from ***http://www.ascp.org/general/about/pioneers/papanicolaou.asp***

Austoker, J. and McPherson, A. (1992) ***Cervical Screening (2nd Ed.)*** Oxford : Oxford University Press.

BBC.com (2001) Retrieved January 12th 2004, from ***http://news.bbc.co.uk/1/hi/health/1310684.stm***

Bijaoui, E. Starck, J-L. and Murtagh, F. (1994) Restauration des images multiechelles par l'algorithme a trous. ***Traitment du Signal, 11,*** 229-243.

Bishop, J.W., Cheuvront, D.A, & Sims, K.L. (2000) Evaluation of the AutoCyte SCREEN system in a clinical cytopathology laboratory. ***Acta Cytologica, 44,*** 128-136.

Brainard, D. H. (2001). Color vision theory. In ***International Encyclopedia of the Social and Behavioral Sciences***, N. J. Smelser & P. B. Baltes (Eds.), (4), 2256-63. Oxford : Elsevier.

Branca, M., Duca, P.G., Riti, M.G., Rossi, E., Leoncini, L., Turolla, E., Morosini, P.L, and the National Working Group for External Quality Control in Cervical Screening (1996) Reliability and accuracy of reporting cervical intraepithelial neoplasia (CIN) in 15 laboratories throughout Italy: phase 1 of a national programme of external quality control in cervical screening ***Cytopathology, (7) 3***,159-172.

Broadstock,M. (2001) Effectiveness and cost effectiveness of automated and semi-automated cervical screening devices: a systematic review of the literature. ***New Zealand Medical Journal, 114(1135)***:311-313.

Brotzman, G.L., Kretchmar, S., Ferguson, D., Gottlieb, M., and Stowe, C. (1999) Costs and outcomes of Papnet secondary screening technology for cervical cytological evaluation: a community hospital's experience. ***Archives of Family Medicine, 8,*** 52-55.

Brown, A.D., and Garber, A.M. (1999) Cost effectiveness of 3 methods to enhance the sensitivity of Papanicolaou Testing. ***Journal of the American Medical Association, 281,*** 347-353.

Brugal G, Garbay C, Giroud F, Adelh D (1979) A double scanning microphotometer for image analysis: Hardware, software and biochemical applications. ***Journal of Histochemistry and Cytochemistry, 27,*** 144 152.

Caelli, T., and Reye, D. (1993) On the classification of image regions by color, texture and shape. ***Pattern Recognition, 26.*** 461-470.

Campbell, F. W., and Robson, J. G. (1968) Application of Fourier analysis to the visibility of gratings. ***Journal of Physiology, 197***, 551-566.

Carson, C., Belongie, S., Greenspan, H., Malik, J. (2002) Blobworld: Color and texture-based image segmentation using em and its application to content-based image retrieval. ***IEEE Transactions on Pattern Analysis and Machine Intelligence, 24***. 1026-1038.

Chamberlain, J. (1986) Reasons that some screening programmes fail to control cervical cancer. In, ***Screening for cancer of the uterine cervix.*** M. Hakama, A.B. Miller & N.E. Day (Eds.). 161-168. Lyons : International Agency for Research on Cancer.

Colquhoun, P. (1976). Psychological and Psychophysiological Aspects of Work and Fatigue. ***Activitas Nervosa Superior, 18,*** 257-263.

Culverhouse, P.F., Williams, R., Reguera, B., Ellis, R., and Parisini, T. (1996) Automatic classification of 23 species of Dinoflagellate by artificial neural network. ***Marine Ecology – Progress Series, 139 (1-3)***, 281-287.

Dement, W.A., (1999) ***The Promise of Sleep.*** NY,US: Delacorte Press

Denaro, T.J., Herriman, J.M., and Shapira, O. (1997) Papnet testing system: Technical update. ***Acta Cytologica, 41***, 65-73.

Doekler, M. and Morris, J.A. (2003) How accurate are subjective judgements of a continuum? ***Histopathology, 42,.*** 227-232.

Drimbarean, A., and Whelan, P.F. (2001) Experiments in colour texture analysis. ***Pattern Recognition Letters, 22,*** 1161-1167.

Dyer, C. (1999a) Three women win in cervical cancer screening case. ***British Medical Journal, 318***. 484

Dyer, C. (1999b) Health authority loses cervical cancer smear appeal. ***British Medical Journal, 319***. 1391

Ericsson, K. A., and Lehmann, A. C. (1996) Expert and exceptional performance: Evidence of maximal adaptation to task constraints. ***Annual Review of Psychology, 47.*** 273-305

Fetterman, B., Pawlick, G., Koo, H., Hartinger, J., Gilbert, C., Connell S., (1999) Determining the utility and effectiveness of the NeoPath AutoPath 300 QC system used routinely. ***Acta Cytologica, 43***, 13-22.

Findlay J.M. (1997). Saccade target selection in visual search. **Vision Research, 37,** 617-631.

Findlay, J. M., Brown, V. and Gilchrist I. D. (2001). Saccade target selection in visual search: the effect of information from the previous fixation. ***Vision Research 41,*** 87-95.

Gatscha RM, Abadi M, Babore S, Chhieng D, Miller MJ, and Saigo PE. (2001) Smears diagnosed as ASCUS: interobserver variation and follow-up. ***Diagnostic Cytopathology, 25(2)***, 138-40.

Goodman, N. (1972) Seven structures on similarity. In N. Goodman, (Ed), ***Problems and Projects***, 437--447. NY, US: Bobbs-Merril.

Granka, L., Joachims, T., and Gay, G. (2004). Eye-Tracking Analysis of User Behavior in WWW Search. In ***Proceedings of 28th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '04)***, Sheffield, UK.

Hope, J. (2002, November 20) Cervical Cancer Vaccine Within 5 Years. ***Daily Mail (UK edition),*** pp 1,4.

Hubel, D. H. and Weisel, T. N. (1968) Receptive fields and functional architecture of monkey striate cortex. ***Journal of Physiology,*** **195**, 215-243.

Hubel, D.H., and Wiesel, T. N. (1962) Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. ***Journal of Physiology, 160,*** 106-154.

Jones, S., Thomas, G.D., and Williamson, P. (1996) Observer variation in the assessment of adequacy and neoplasia in cervical cytology. ***Acta Cytologica, 40(2),*** 226-34.

Komatsu, L. K. (1992) Recent views of conceptual structure. ***Psychological Bulletin, 112,*** 500-526.

Koss, L.G., Lin, E., Schreiber, K., Elgert, P., and Mango, L. (1994) Evaluation of the PAPNET cytologic screening system for quality control of cervical smears. ***American Journal of Clinical Pathology, 101***, 220-229.

Krauzlis, R.J. (2004) Recasting the smooth pursuit eye movement system, ***Journal of Neurophysiology, 91,*** 591-603.

Krauzlis, R.J. and Adler, S.A. (2001) Effects of directional expectations on motion perception and pursuit eye movements. ***Visual Neuroscience, 18,*** 365-376, 2001

Lee, K.R., Ashfaq, R., Birdsong, G.G., Corkill, M.E., McIntosh, K.M., & Inhorn, S.C. (1997) Comparison of conventional Papanicolaou smears and a fluid based, thin-layer system for cervical cancer screening. ***Obstetrics and Gynecology, 90,*** 278-284.

Li, A.. and Lennie, P. (2001) Importance of color in the segmentation of variegated surfaces. ***Journal of the Optical Society of America, 18 (6).*** 1240-1251.

McCrory, D. C., Matchar, D. B., Bastian, L., Datta, S., Hasselblad, V., Hickey, J., Myers, E. *et al.* (1999). Evaluation of cervical cytology: Evidence Report/Technology Assessment Number 5. (Prepared by Duke University under Contract No. 290-97-0014). ***AHCPR Publication No. 99-E010.*** Rockville, MD: Agency for Health Care Policy and Research (AHCPR).

Miller G.A. (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. ***Psychological Review, 63***. 81-97.

Mirmehdi, M., Petrou, M. (2000) Segmentation of color textures. ***IEEE Transactions on Pattern Analysis and Machine Intelligence, 22.*** 142-159.

Moss, SM, Gray, A, Legood, R, & Henstock, E (2003) Evaluation of HPV/LBC Cervical Screening Pilot Studies. ***First report to the Department of Health on evaluation of LBC.*** Sutton: Institute of Cancer Research

Nanda, K., McCrory, D. C., Myers, E. R., Bastian, L. A., Hasselblad, V., Hickey, J. D., & Matchar, D. B. (2000). Accuracy of the papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. ***Annals of Internal Medicine, 132,*** 810-9.

National Audit Office (1998) The performance of the NHS Cervical Screening Programme in England. ***Report by The Controller and Auditor General, NAO***. London : HMSO.

National Heath Service Cancer Screening Programme, (n.d), Retrieved April 1St 2005, from ***http://www.cancerscreening.nhs.uk/cervical/risks.html***

National Heath Service Cancer Screening Programme (2000) ***Achievable standards, benchmarks for reporting and criteria for evaluating cervical cytopathology.*** Sheffield: NHSCSP Publications (Publication no. 1).

National Institute for Clinical Excellence (2002) Guidance on the Use of Liquid Based Cytology for Cervical Screening. ***Technology Appraisal Guidance No. 5.*** London: NICE; 2002.

National Institute for Clinical Excellence (2003) Guidance for the use of liquid-based cytology for cervical screening. ***Technology Appraisal No. 69.*** NICE:london

O'Leary, T., Tellado, M., Buckner, S-B., Ali, I., Stevens, A., and Ollayas, C. (1998) Papnet assisted rescreening of cervical smears: cost and accuracy compared with 100% manual rescreening strategy. ***Journal of the American Medical Association, 279,*** 235-237.

Office of National Statistics, The (1998) ***Estimates of newly diagnosed cases of cancer, England and Wales 1993-1997.*** London, UK: ONS

O'Sullivan, J.P. (1998) Observer variation in gynaecological cytopathology. ***Cytopathology, 9 (1)***, 6-14.

O'Sullivan, J.P., Ismail, S.M., Barnes, W.S., Deery, A.R., Gradwell, E., Harvey, J.A., Husain, O.A., Kocjan, G., McKee, G., Olafsdottir, R., Ratcliffe, N.A., and Newcombe, R.G. (1996) Inter- and intra-observer variation in the reporting of cervical smears: specialist cytopathologists versus histopathologists. ***Cytopathology, 7,*** 78-89.

Palm, C., Keysers, D., Lehmann, T., and Spitzer, K. (2000) Gabor Filtering of Complex Hue/Saturation Images for Color Texture Classification.

Proceedings of the 5th Joint Conference on Information Science (JCIS2000) 2, Atlantic City, USA, 45-49.

Paschos, G.: (2000) Fast color texture recognition using chromacity moments. ***Pattern Recognition Letters, 21.*** 837-841.

Pattanaik, S. N., Fairchild, M.D., Ferwerda, J.A., & Greenberg, D.P. (1998) Multiscale model of Adaptation, Spatial Vision and Color Appearance. ***Proceedings of IS&T/SID's 6th Color Conference***, Arizona, U.S.

Pollatsek, A., Rayner, K., & Collins, W. E. (1984). Integrating pictorial information across eye movements. ***Journal of Experimental Psychology: General, 113***, 426-442.

Potter, T. (1999) Assessing the skill of an expert cytologist engaged in cervical smear categorisation tasks. ***MSc Computational Intelligence Thesis.*** Plymouth: University of Plymouth.

Rayner, K.(1998). Eye movements in reading and information processing: 20 years of research. ***Psychological Bulletin, 124(3)***, 372-422.

Rips (1989) Similarity, typicality and categorization. In Vosnidau, S, Ortony, A. (Eds.) ***Similarity and Analogical reasoning.*** Cambridge: Cambridge University Press.

Rips, L. (1990) Reasoning. ***Annual Review of Psychology, 41,*** 321 – 353.

Rosch, E. (1975) Cognitive representations of semantic categories. ***Journal of Experimental Psychology: General 104***,192-232.

Sanders, A.F. (1993) Processing information in the functional visual field. In G. d'Ydewalle & J. Van Rensbergen (Eds.), ***Perception and cognition. Advances in eye-movement research*** (pp. 3–22). Amsterdam: North-Holland.

Shanteau, J. (1992) How much information does an expert use? Is it relevant? ***Acta Psychologica, 81***, 75-86.

Shanteau, J., and Stewart, T.R. (1992) Why Study Experts? Some Historical Perspectives and Comments. ***Organizational Behavior and Human Decision Processes, 53,*** 95-106.

Sokal, R.R. (1974) Classification: Purposes, Principles, Progress, Prospects. ***Science, 185 (4157),*** 1115-1123.

Solomon D, Davey D, Kurman R, Moriarty A, O'Connor D, Prey M, Raab S, Sherman M, Wilbur D, Wright T Jr, Young N; Forum Group Members; Bethesda 2001 Workshop (2002) The 2001 Bethesda System: terminology for reporting results of cervical cytology. ***Journal of the American Medical Association, 287(16),*** 2114-2119.

Tezuka, F., Oikawa, H., Shuki, H., and Higashiiwai, H. (1996) Diagnostic efficacy and validity of the ThinPrep method in cervical cytology. ***Acta Cytologica, 40,*** 513–518.

Thai, B., and Healey, G. (1998) Modelling and classifying symmetries using a multiscale opponent colour representation. ***IEEE Transactions on Pattern Analysis and Machine Intelligence, 20.*** 1224-1235.

Toth, L. and Culverhouse, P.F. (1999) Three dimensional object categorisation from static 2D views using multiple coarse channels, ***Image and Vision and Computing, 17***. 845-858.

Tsao, Y-C., Drury, C.G. and Morawaski, T.B. (1979) Human performance in sampling inspection. ***Human Factors, 21,*** 99-105.

Tucker, J.H., and Shippey, G. (1983) Basic performance tests on the CERVFIP linear array prescreener. ***Analytical and Quantitative Cytology and Histology, 5,129***-37.

Tversky, A., and Kahneman, D. (1974) Judgement under uncertainty: Heuristics and biases. ***Science, 185,*** 1124-1131.

Van de Wouwer, G., Livens, S., Scheunders, P., and Van Dyck, D. (1999) ***Wavelet Correlation signatures for Color Texture Characterization,*** Pattern Recognition (Special issue on Color and Texture Analysis), 32(3). 443-451.

Veneti, S., Papaefthimiou, M., Symiakaki, H., & Ioannida-Mouzaka, L. (1999) Papnet for cervical cytology screening: Experience in Greece. ***Acta Cytologica, 43,*** 30-33.

Wang, H. & Culverhouse, P.F. (2004) The Categorisation of Similar Non-rigid Biological Objects by Clustering Local Appearance Patches. ***Lecture Notes in Computer Science Intelligent Data Engineering and Automated Learning - IDEAL 2004: 5th International Conference, Exeter, UK.*** 65-72.

Warm, J.S.(Ed) (1984) ***Sustained attention in human performance.*** John Wiley & Sons, NY, US.

White, C.W. (1976) Visual masking during pursuit eye movements. ***Journal of Experimental Psychology: Human Perception and Performance, 2,*** 469-478.

Wied, G.L., Bahr, G.F, and Bartels, P.H. (1970) Automated analysis of cell images by TICAS. In ***Automated Cell Identification and Cell Sorting,*** G.L. Wied, and G.F. Bahr, (Eds). 195-360. NY,US:Academic Press.

Wied, G.L., Bartels, P.H., Bahr, G.F., and Oldfiled. D.C. (1968) Taxonomic intracellular analytic system (TICAS) for cell identification. ***Acta Cytologica, 12,***180-204.

Wilson, H.R. (1991) Psychophysical models of spatial vision and hyperacuity. In ***Spatial Vision***, Vol 10, Vision and Visual Dysfunction, D. Regan (Ed) Boca Raton, FL, US: CRC Press, 64-81.

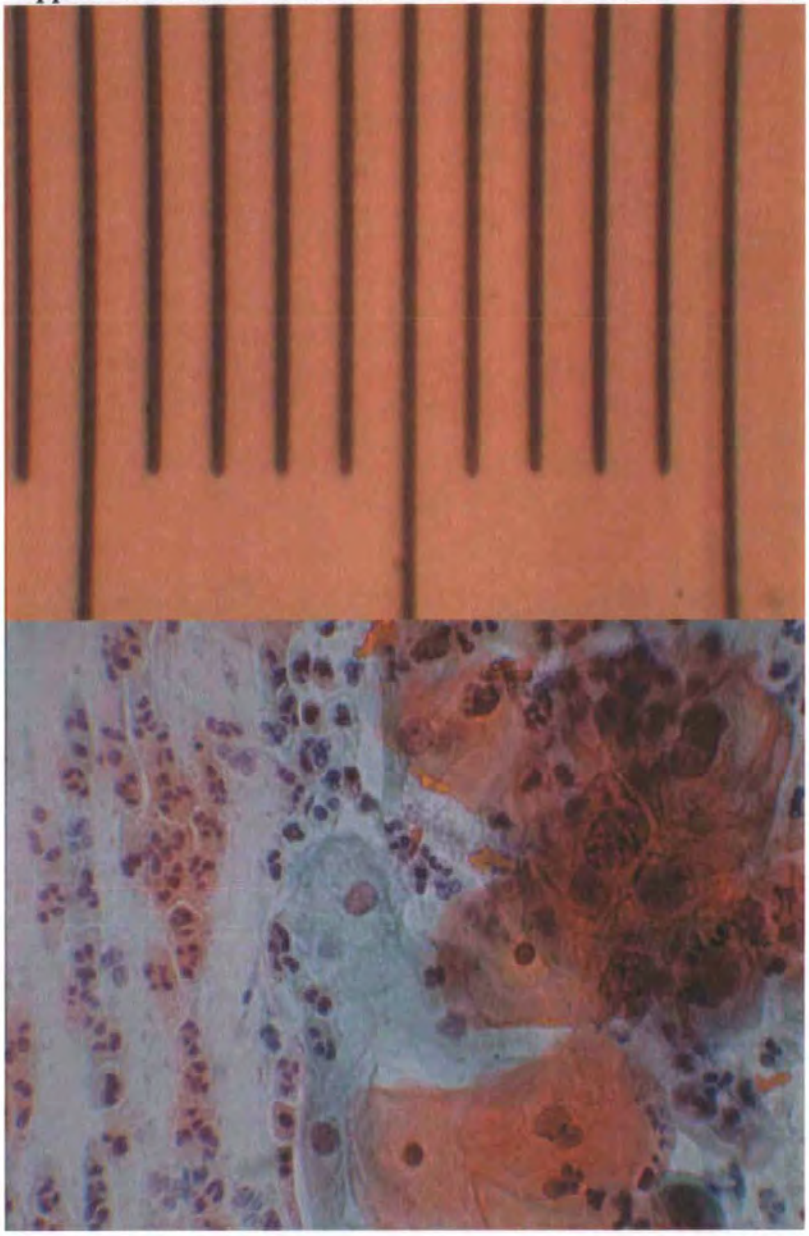
Yarbus, A.F. (1967) ***Eye Movements and Vision.*** NY, US: Plenum Press.

Yobs AR, Plott AE, Hicklin MD, Coleman SA, Johnston WW, Ashton PR, Rube IF, Watts JC, Naib ZM, Wood RJ (1987) Retrospective evaluation of gynecologic cytodiagnosis. II. Interlaboratory reproducibility as shown in rescreening large consecutive samples of reported cases. ***Acta Cytologica, 31,*** 900-910.

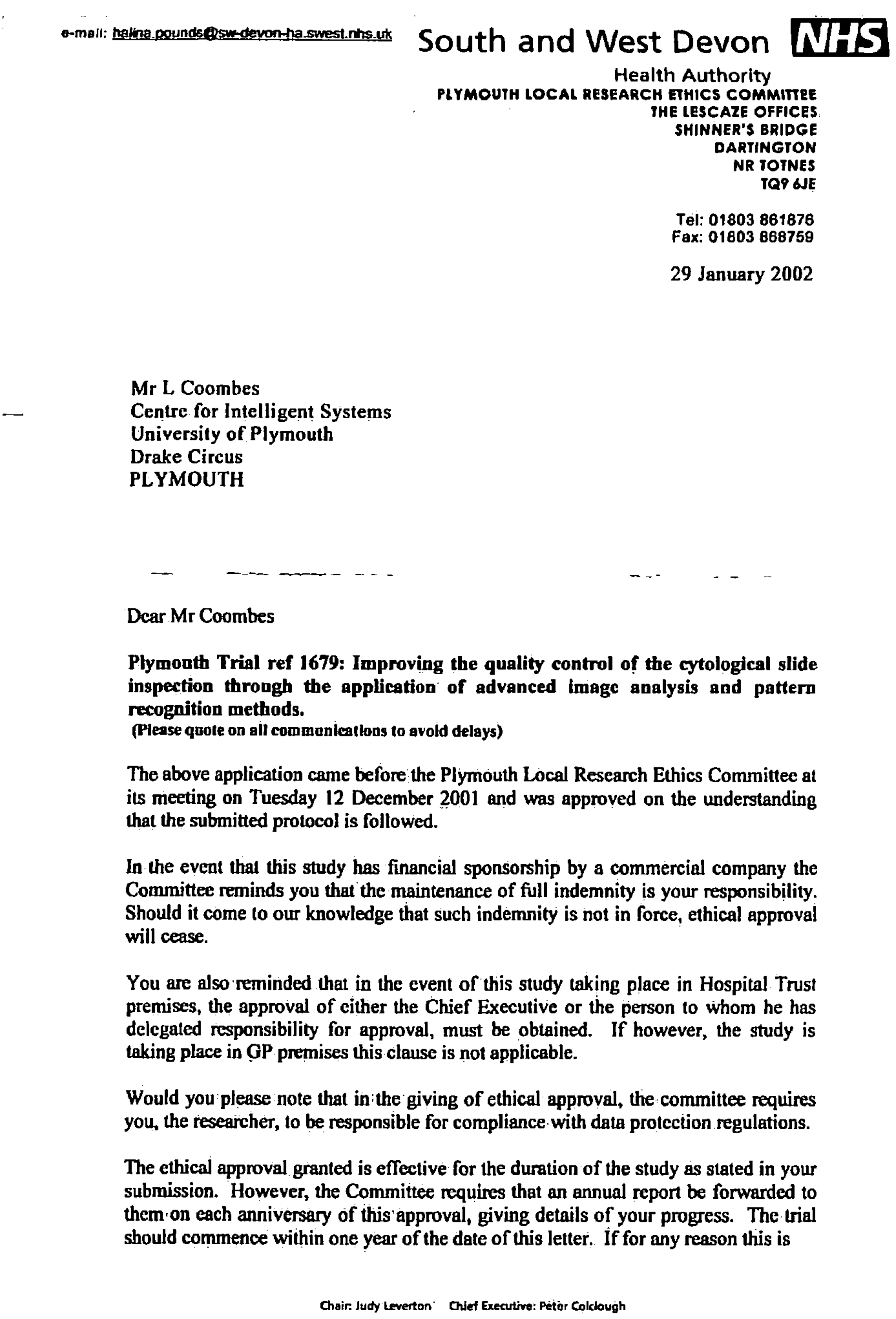
Yoss, R.E., Moyer, N.J., Hollenhorst, R.W. (1970) Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. ***Neurology, 20,*** 545-554.

APPENDIX A – Leica Calibration at x40 magnification

Appendix Leica calibration slides at x 40 each unit = 0.01cm



APPENDIX B – Ethics Approval Letter



e-mail: halina.pounds@sw-devon-ha.swest.nhs.uk

South and West Devon NHS
Health Authority
PLYMOUTH LOCAL RESEARCH ETHICS COMMITTEE
THE LESCAZE OFFICES
SHINNER'S BRIDGE
DARTINGTON
NR TOTNES
TQ9 6JE

Tel: 01803 861876
Fax: 01803 868759

29 January 2002

Mr L Coombes
Centre for Intelligent Systems
University of Plymouth
Drake Circus
PLYMOUTH

Dear Mr Coombes

Plymouth Trial ref 1679: Improving the quality control of the cytological slide inspection through the application of advanced image analysis and pattern recognition methods.
(Please quote on all communications to avoid delays)

The above application came before the Plymouth Local Research Ethics Committee at its meeting on Tuesday 12 December 2001 and was approved on the understanding that the submitted protocol is followed.

In the event that this study has financial sponsorship by a commercial company the Committee reminds you that the maintenance of full indemnity is your responsibility. Should it come to our knowledge that such indemnity is not in force, ethical approval will cease.

You are also reminded that in the event of this study taking place in Hospital Trust premises, the approval of either the Chief Executive or the person to whom he has delegated responsibility for approval, must be obtained. If however, the study is taking place in GP premises this clause is not applicable.

Would you please note that in the giving of ethical approval, the committee requires you, the researcher, to be responsible for compliance with data protection regulations.

The ethical approval granted is effective for the duration of the study as stated in your submission. However, the Committee requires that an annual report be forwarded to them on each anniversary of this approval, giving details of your progress. The trial should commence within one year of the date of this letter. If for any reason this is

Chair: Judy Leverton Chief Executive: Peter Colclough not possible, confirmation should be sought from this committee that the ethical approval is still applicable. Notification of the trial's completion, suspension or premature termination must be advised in writing with a copy of the final summary or full report where possible. Notification is also required in respect of any relevant Serious Adverse Events or Adverse Drug Reactions and any Protocol Amendments.

Yours sincerely

Halina Pounds

HALINA W POUNDS (MRS)
ADMINISTRATOR
Plymouth Local Research Ethics Committee
(signed by the Administrator on behalf of the Chairman in accordance with the Committee's Operating Procedures)

APPENDIX C – Experiment Briefing

Information Sheet for Participants

Research Title : **Improving the quality control of human expert cytological slide inspection through the application of advanced image analysis and pattern recognition methods**

Main Researcher :......... Mr Lee Coombes
Supervisor : ....Dr Phil Culverhouse

Thank you for offering to be a part of the experimental research taking place here today. This is a brief introduction to the specific nature of our research followed by the specific details of the experiments.

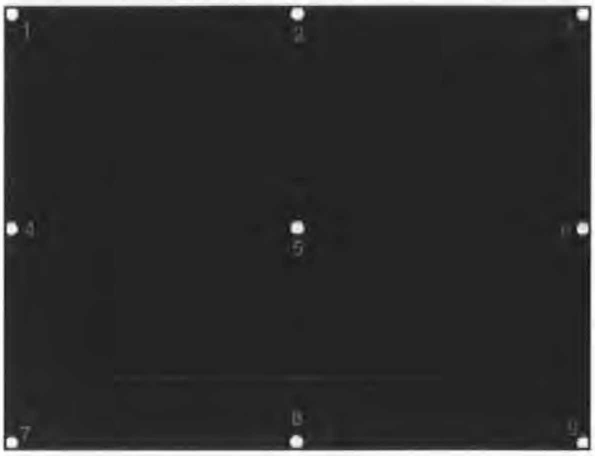
We are currently investigating the viability of developing a computer system that would be able to assess the skill level of a cytological screener. In order to do this we first have to look at how an expert examines a slide before deciding on its classification. When we have this data we can then compare it to our own method of finding these important features.

For this purpose, there are two straightforward tasks we would like you to complete today.

Eyetracking

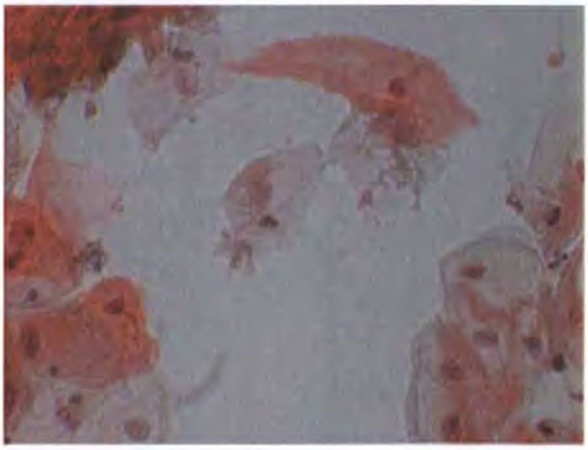
The first involves viewing a number of images and then deciding on a classification based only on the data contained in each one. This task is to be completed while wearing an eyetracking device. Because of the practical difficulties of getting data using real slide examinations, high-quality images will be used instead and presented to you on a computer screen. We can then monitor your eye movements while each image is being viewed. Each image will require a decision to be made regarding the level grade of abnormality. Because of the sensitive nature of the eyetrackers measurements it is important that the head of the person wearing the equipment is kept still. To do this we will need to mould a bite-bar to your teeth in the same way as a gum-shield might be fitted. The bite-bar keeps the head relatively motionless allowing the most accurate measurements.

The screens used during the trial are as follows..



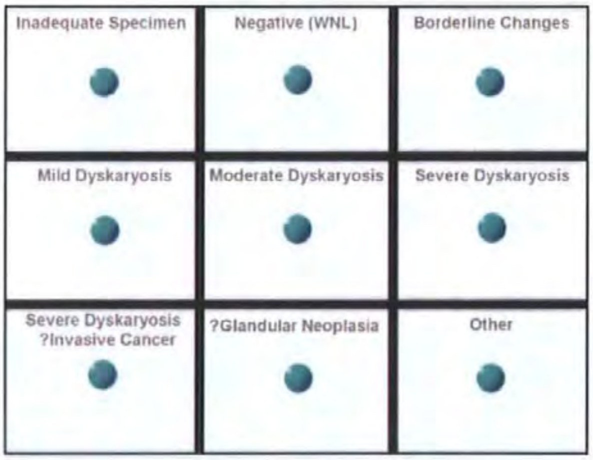
A 9-point calibration screen.
This is shown at the beginning of the image set to allow us to calibrate the eyetracker to your visual field. Once you are wearing the eyetracker and the experiment begins this will be the first screen you see. You will be asked to look at each point in turn before continuing through the images. This screen is only displayed once.

A 1-point calibration screen.
This will be shown prior to each image you view. ***It is important that you fixate on the central white dot and pause briefly before moving on to the next screen***. This allows us to assess the calibration of the eyetracker throughout the presentation.



The Image

This is a sample of the type of image you will see. The image set has a variety of grades from both Conventional Pap and ThinPrep methods of slide preparation. Take as long as you need to inspect the image before moving onto the decision grid.



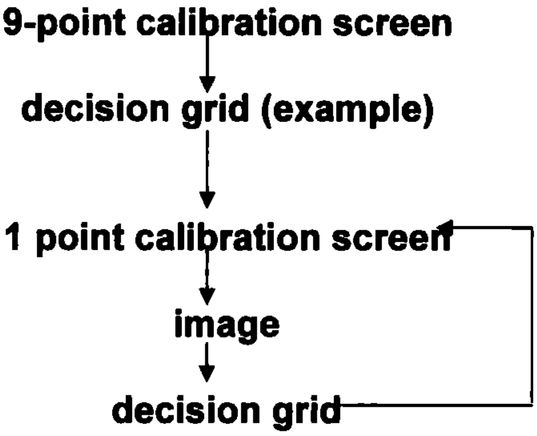
Inadequate Specimen ●	Negative (WNL) ●	Borderline Changes ●
Mild Dyskaryosis ●	Moderate Dyskaryosis ●	Severe Dyskaryosis ●
Severe Dyskaryosis ?Invasive Cancer ●	?Glandular Neoplasia ●	Other ●

The Decision Grid

This is presented after each image. Once you have inspected the image, this screen is presented which allows you to indicate which classification you think the image should have. ***It is important that you fixate on the central appropriate grade and pause briefly for a few seconds before moving on to the next screen***. When the data from the trial is analysed we will be able to tell what your classification would be from the fixations made on this screen.

Order of presentation

We will control the display up until the 9-point calibration screen when control will be given to you via the buttons attatched to the bar in front of you. Both buttons have the same function, to move the presentation on to the next screen, so only one button needs to be pressed at a time and pressed only once. The screen moves on when you release the button.



9-point calibration screen

decision grid (example)

1 point calibration screen

image

decision grid

Important points

When the experiment has begun, you should continue biting on the bar until the final screen is shown. This is because it is vital that your head remain as still as possible. A message will be displayed when the trial ends to let you know when you can stop biting.

This is not a reaction time experiment. You should take as much time as you need to analyse each image and only move on when you are confident that you can grade the image.

You will not be able to talk while the experiment is in progress. If you make a mistake, such as flicking past more than one image, it is important that you carry on regardless. The eyetracker data will show if there have been any problems and these can be dealt with accordingly.

Feature marking

The second of the two tasks involves manually marking abnormal features on the same set of images shown in the eyetracking task. This is so we can tell if the fixations made during the eyetracking trial are on normal or abnormal features. You will be given a selection of images and all you need to do is mark with a dot any and all of the abnormal features present.

The images will be presented in a piece of software which allows you to mark points on the image using the mouse. When all of the abnormal features for an image have been marked, you should select the next image in the menu on the left of the screen. You can also comment on the slides (ie if you think it is inadequate) by using the right mouse button to make a comment box appear. Where there are too many abnormal features to mark, mark the most important and leave a comment.

When all of the images have been marked the experiment will have ended

Finally....

Please remember that there are no right or wrong answers. We are not here to test you but to gather a pool of data based on your considerable expertise in this field. All your data will be held anonymously and confidentially. You have the right to withdraw your data from this study at any time.

You will be shown short trial versions of both tasks before completing them. If you have any further questions then please feel free to ask.

APPENDIX D – Experiment Debriefing

Debriefing

The data we are gathering here today is a vital part of the research we are carrying out. It is a very early stage in the development of a quality assurance system which we believe will enhance the current Quality Assurance measures. As a participant in this study you are welcome to get in touch with us to find out how your data is being used and to learn more about our work.

Contact Details

Main Researcher –	Mr Lee Coombes	L1Coombes@Plymouth.ac.uk
Supervisor –	Dr Phil Culverhouse	P.Culverhouse@Plymouth.ac.uk

Centre for Intelligent Systems
University of Plymouth
Plymouth
Devon
PL4 8AA

Thank you for your time and participation

APPENDIX E – Experimental Consent Form

PARTICIPATION CONSENT FORM

Name of Research Study...

Investigator...Supervisor..

NOTE: THIS FORM MUST BE SIGNED BEFORE YOU PARTICIPATE

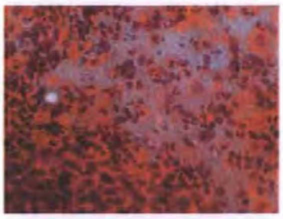
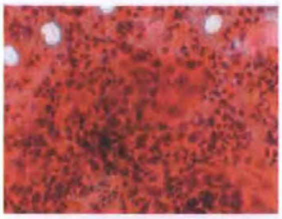
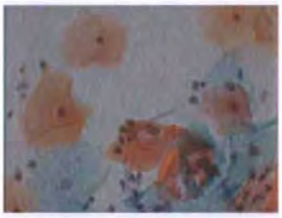
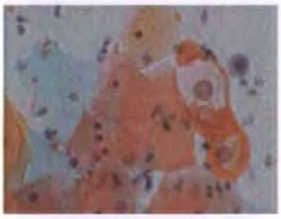
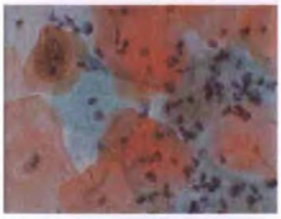
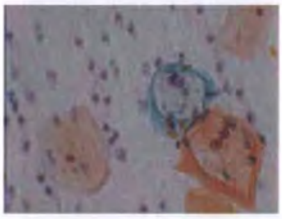
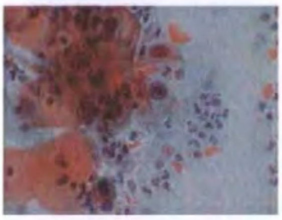
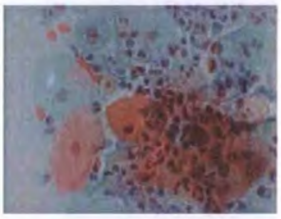
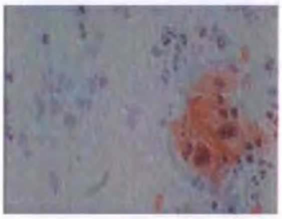
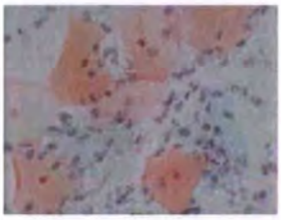
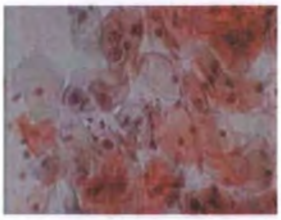
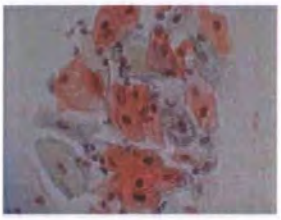
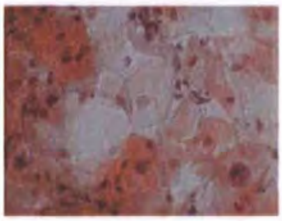
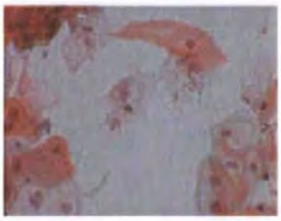
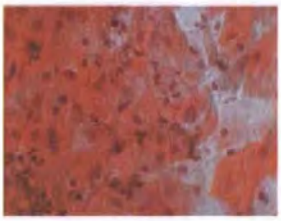
We, the undersigned, hereby consent to participate in the above research study. We give our consent having received satisfactory answers to our questions concerning the study, in the full knowledge that we have the right to refuse to participate and knowing that we may withdraw from the above study without penalty at any time. We also understand that every effort will be made to protect the anonymity of our responses.

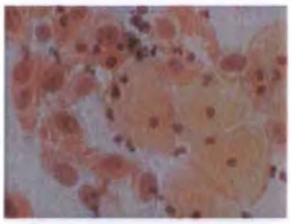
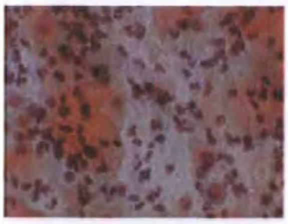
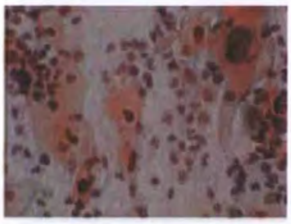
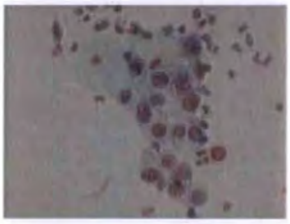
	DATE	NAME (please print)	SIGNATURE
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			

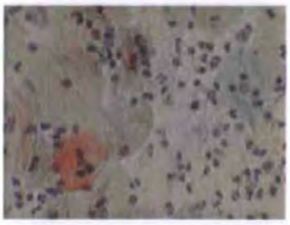
I certify that the names, dates and signatures on this sheet are authentic.

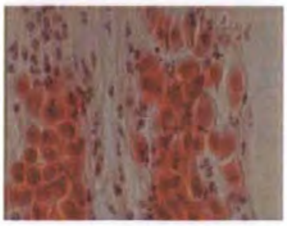
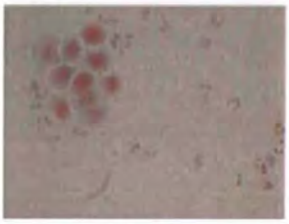
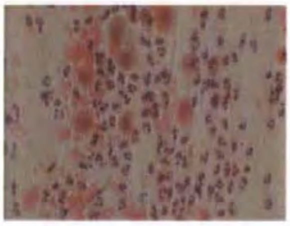
Signature of Investigator...Date.....................................

APPENDIX F – Papanicolaou Image Set

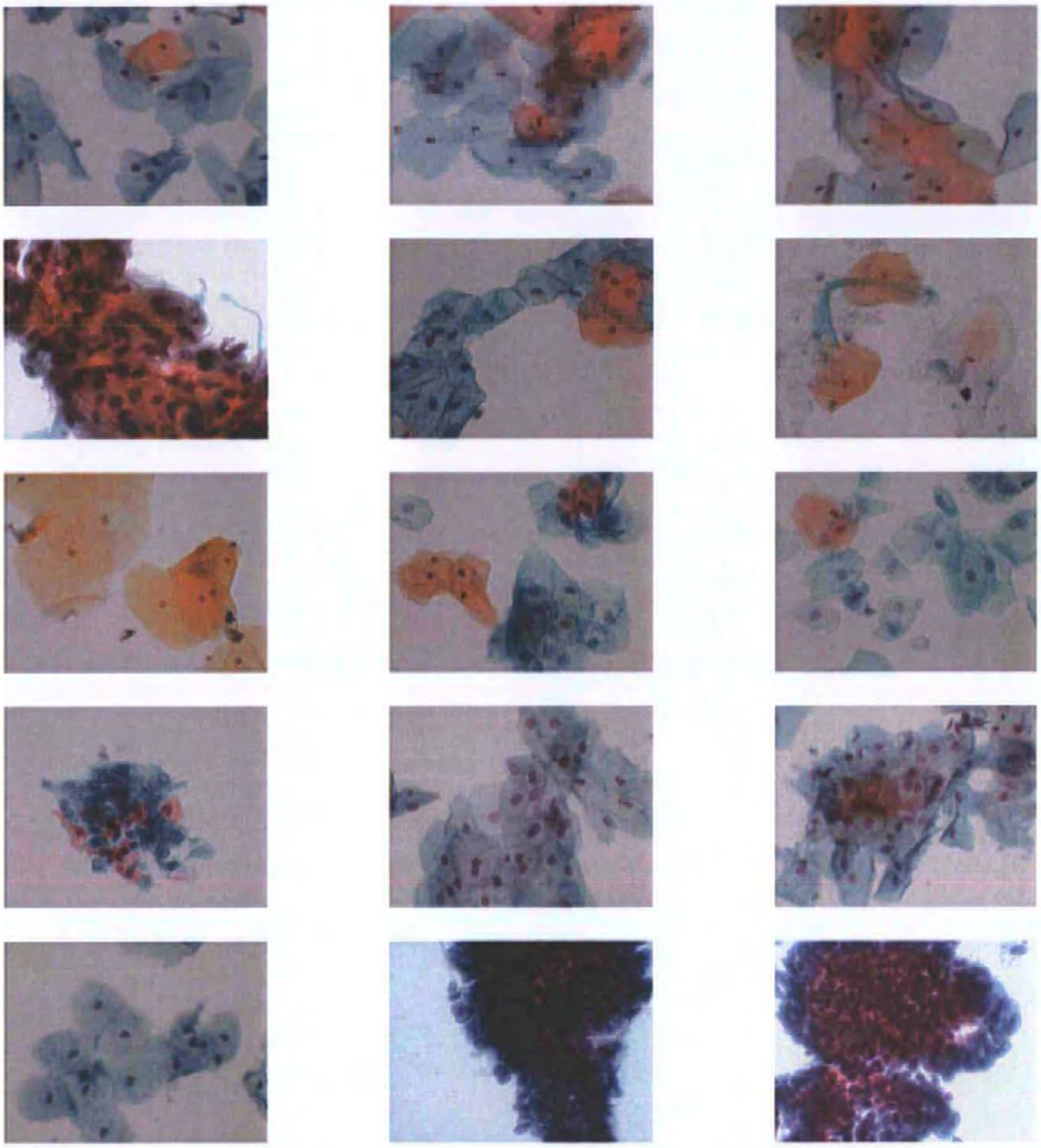
















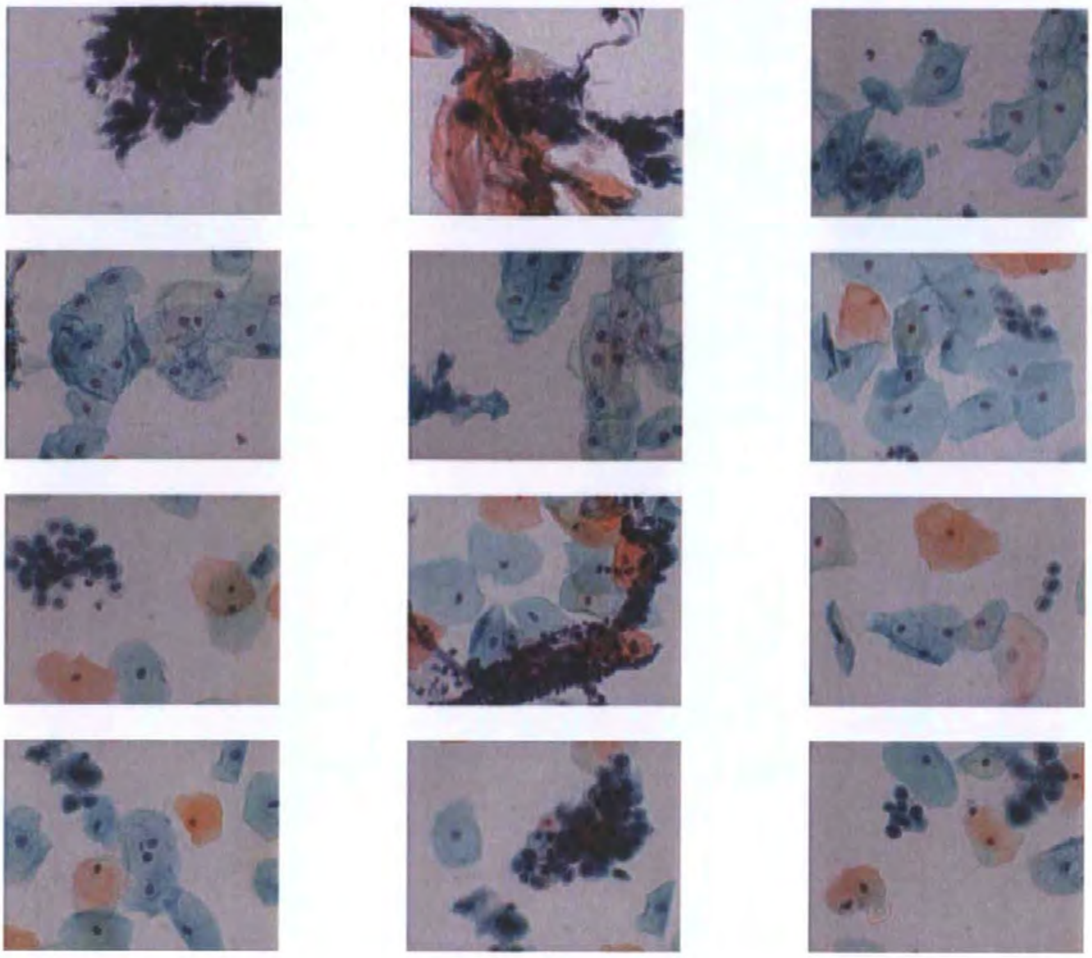







APPENDIX G – ThinPrep Image Set





APPENDIX H – DirectX Image Presentation Software

On Accompanying CD

APPENDIX I – Image Labelling Software

On Accompanying CD

APPENDIX J – EQA Statistical Analysis Software

On Accompanying CD

APPENDIX K – Raw Data

On Accompanying CD

APPENDIX L – Published Papers

Pattern Recognition in Cervical Cytological Slide Images

L. R. Coombes
Centre for Robotics and Intelligent Systems,
SoCCE, University of Plymouth,
Plymouth, Devon, England. PL4 8AA
Lee.Coombes@Plymouth.ac.uk

P. F. Culverhouse
Centre for Robotics and Intelligent Systems,
SoCCE, University of Plymouth,
Plymouth, Devon, England. PL4 8AA
Phil.Culverhouse@Plymouth.ac.uk

Abstract

We describe a novel model for quality assurance of the cervical cytological slide screening process. We offer evidence for our model from a combination of eye-tracked fixations of cervical cytology screeners and manually marked features that are used as a guide to maxima identified using an Atrous wavelet transformation. The results show that the distribution amongst the groups is not random, with Hue proving particularly valuable in describing the images. Further work will refine this model to improve the discrimination between the groups.

1. Introduction

The existing program within the UK for screening of cervical cancer has been established for a number of years and has seen a marked reduction in both incidence and mortality rates [11][13]. The National Health Service's Cervical Screening Program's 'Achievable Standards, Benchmarks for Reporting, and Criteria for Evaluating Cervical Cytopathology' [10] states that cervical smears must be competently obtained and interpreted at least every five years to prevent 80 – 90% of invasive cancers. While the screening program itself works largely with a cautious approach, with slides re-screened to confirm the diagnosis, there remain several sources of variation both within and between laboratories [2][7]. One way to reduce the variation is with further training of those staff involved in categorisation. This is shown in [7] which demonstrates that attending training courses or discussing the criteria through which slides were diagnosed reduced this variation. Unfortunately, training is expensive both in time and cost to the individual, the laboratory and the health authority. Attempts have been made to introduce automation within the screening process aimed at improving the overall performance levels but these have to be extensively tested and proven before they can be introduced as standard. It is shown that this in itself can be problematic as many of the studies lack the scientific rigour that is required to make an informed decision regarding their usefulness [3].

2. Our Approach

We propose a novel approach to reducing variation and improving the performance of cervical cytological slide screeners. This is based on the limited success of automated recognition devices where they have been in operation [3][5][8] but also addresses one of the main drawbacks with automating any part of the screening process. Because final classification of each slide is based on human judgement, there is not 100% agreement as to what properties each classification has. Guidelines are available [10] but there are not distinct boundaries between each possible classification. This means that any automatic recognition that takes place will at best be based on human definitions of categories that are flexible. The infinite variation of slide material means that we will never have 100% agreement between humans and recognition systems and this is a source of ethical issues regarding the judgements the systems make. We propose an intervention that we believe will be able to reduce variation between screeners while avoiding the ethical problems faced by replacing part of the screening process itself. By identifying the most salient features on each slide, it would be possible to judge whether these are being viewed when a slide is screened by using data from an eye-tracking piece located in a microscope. Even minimal feedback regarding performance would reduce variation as shown in [7] and that this could be provided back to the screeners in real time. The work presented here shows how we derive what constitutes an important feature using both machine analysis and data from human screeners. Our aim is to provide a strong theoretical basis for our approach from which further development can take place.

3. Experimental Work

In order to discover which features are most important to the classification decision process we recorded the skills of 10 trained and practicing cytological screeners. They were presented with images taken from cytological slides and had to make a decision on what classification they should give it. This was done while wearing an eyetracking device so that we could catalogue all of the eye fixations made during the decision process. They then completed a manual feature marking exercise on the same set of images to indicate where abnormalities were located. This experimental procedure was designed specifically with the analysis in mind. We begin the analysis with the data from the second task; the manual feature marking. This gives us information on where the participants believe there are abnormalities on the slides. Similar abnormalities marked by a lot of participants are assumed to be more salient than those marked by only one person. In other words, these areas are the most important to consider before making what is perceived to be the correct diagnosis for the slide. This allows us to classify each of the eyetracker fixations according to the proximity of abnormal features. From this data we have produced a 'saliency index' that ranks all of the eyetracking data points in order of their importance. It is this index that we use as a guide to the effectiveness of our machine analysis

4. Machine Analysis

To produce a list of features we first decompose the image into a hue and a saturation/value combined component. When a slide is viewed it is illuminated by the microscope's backlight and this is adjustable depending on personal preference. This means that hue is relatively stable while both saturation and value vary considerably dependant upon the amount of illumination used. We would therefore expect hue to be the superior measure. We use an Atrous wavelet transformation [1] to identify maxima that relate to features at various resolutions within the images used in this study. This method is employed in the recognition of marine microplankton from images of seawater, where successful categorisation of morphologically similar species has been demonstrated [4][14]. A study by [12] suggests that this methodology is also appropriate for cervical smear image analysis. Each of the maxima is checked against the co-ordinates of the eyetracker fixations to look for proximity to two types of feature – abnormal and normal. If no proximity is found then it enters a third group that we class as unimportant features as they have not been viewed by screeners while assessing the slides.

5. Results

The results reported here relate to an initial exploratory investigation of both the data and methodology involved. Because there are three complimentary types of data in use, Eye-tracking fixations, feature-marked points, and machine identified maxima, there are many ways of exploring relationships existing between the triad. In this instance we performed a Canonical Discriminant Analysis to provide a measure of predicted group membership based upon texture measures taken at four resolutions in either the Hue or Saturation/Value planes. The image data itself is also split according to the images contents depending upon the preparation method involved; the liquid-based ThinPrep method or the more common Papanicolaou method. Unlike the Pap method of preparation, ThinPrep allows for a single cellular layer on a slide and this should make it easier for the screener to see abnormalities. In our case this means that the colour and textures sampled from the images are a truer reflection of the cells being sampled. Pap slides often have cells clumped together and the density of these can cause changes in the colour and texture seen.

Figures 1, 2, 3 and 4 show the distributions of the two functions across preparation methods for both Hue and Sat/Val. We can see that the variance of points is particularly marked for the ThinPrep/Hue analysis, which we would expect from monolayer preparations such as ThinPrep. We can also from Tables 1 to 4 that it is the Pap/Hue that has the highest overall predictive value (highlighted for each group), however this can be misleading as it is not our aim to classify slides or predict groups. Instead we aim to judge the saliency of any given feature based on the texture measures taken.

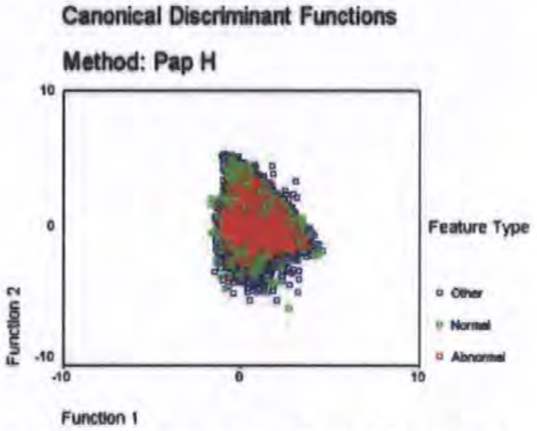


Figure 1: Distribution across the two functions for Hue/Papanicolaou

Table 1: Predicted Group membership for Hue/Papanicolaou (1=abnormal 2=normal 3=other)

			Predicted Group Membership			Total
			1	2	3	
Grp	Count	1	**316**	66	407	789
		2	1863	**509**	3242	5614
		3	7951	2128	**16779**	26858
Grp	%	1	**40.1**	8.4	51.6	100.0
		2	33.2	**9.1**	57.7	100.0
		3	29.6	7.9	**62.5**	100.0

Although the success rate of predictive group membership is low, we have the ability to take one of the recorded eye fixations and rate it for its importance in the diagnosis process. Currently there would be a strong bias towards the group of features that we have shown are not viewed during the decision process, as the recorded data is skewed in favour of this larger group. As the discriminant functions tend towards the centre of the cluster of points there will be less certainty that the feature being examined is irrelevant. In some of these cases there would be no doubt at all that the fixation refers to a feature that is important in making a diagnosis. In both of these instances, we would see the feature being rated more highly on our final saliency index.

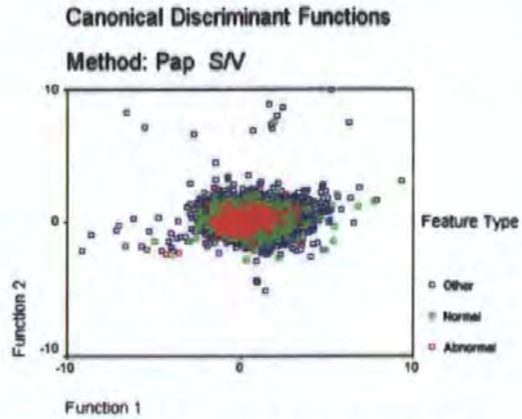


Figure 2: Distribution across the two functions for SatVal/Papanicolaou

Table 2: Predicted Group membership for SatVal/Papanicolaou (1=abnormal 2=normal 3=other)

			Predicted Group Membership			Total
			1	2	3	
Grp	Count	1	**423**	202	59	684
		2	2437	**1617**	410	4464
		3	12831	7590	**1941**	22362
Grp	%	1	**61.8**	29.5	8.6	100.0
		2	54.6	**36.2**	9.2	100.0
		3	57.4	33.9	**8.7**	100.0

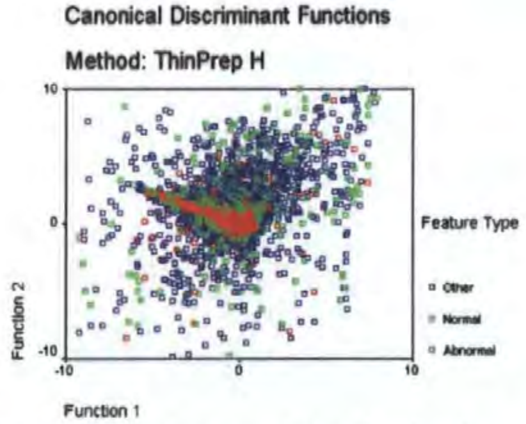


Figure 3: Distribution across the two functions for Hue/ThinPrep

Table 3: Predicted Group membership for Hue/ThinPrep (1=abnormal 2=normal 3=other)

			Predicted Group Membership			Total
			1	2	3	
Grp	Count	1	**271**	606	52	929
		2	1373	**3588**	233	5194
		3	8349	18018	**1287**	27654
Grp	%	1	**29.2**	65.2	5.6	100.0
		2	26.4	**69.1**	4.5	100.0
		3	30.2	65.2	**4.7**	100.0

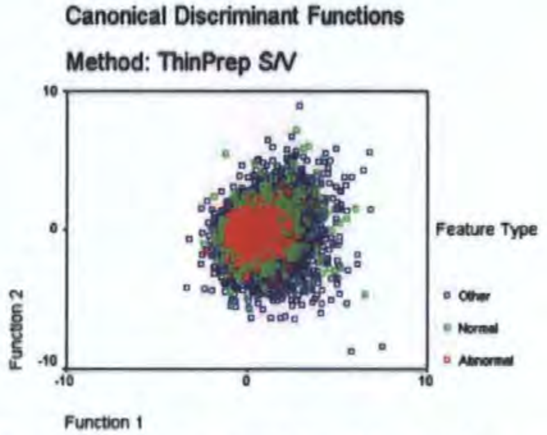


Figure 4: Distribution across the two functions for SatVal/ThinPrep

Table 4: Predicted Group membership for SatVal/ThinPrep (1=abnormal 2=normal 3=other)

			Predicted Group Membership			Total
			1	2	3	
Grp	Count	1	**404**	213	190	807
		2	1851	**1219**	1039	4109
		3	10550	5650	**6099**	22299
Grp	%	1	**50.1**	26.4	23.5	100.0
		2	45.0	**29.7**	25.3	100.0
		3	47.3	25.3	**27.4**	100.0

6. Discussion and Conclusions

The results show that the distributions of data are not random and give a good basis for further development of both the methodology and techniques used in this study. We have shown that there are significant differences between the methods of preparation with regard to their predictive abilities and that there is scope for further refinement. We expected that Hue would be a better reflection of the images contents and certainly for the Pap slides this seems to be the case. The work carried out so far shows that we have a good basis from which to further develop our model for quality assurance. We also believe that as we collect further data to add to the model that this will improve the discriminability of the groups. We currently have a great number of data points in the 'other' category that means that there is a bias towards this groups classification. As more data is collected, the model will become stronger and the groups more tightly defined. There are also ways of improving the existing data that need to be explored like tightening the criteria by which the original group is judged. We believe the model will become robust over time but at this first iteration it is still rather sensitive because of the limited number of abnormal and normal features indicated in the saliency index. We are currently using a simple texture measure as a means of testing our approach to this area and this has shown limited success. In these terms we have been successful in showing that the way we derive a salient feature is valid. We will be looking to further develop the measures used and increase the likelihood that any given feature can be rated accurately for its saliency in the classification process. Developments in our understanding of combined colour texture perception and analysis [6][9] are also of great interest because the recognition of abnormalities relies very much on both. There is a lot of scope for further development of our model but we have shown that the basic principles of how we derive our saliency index are good and can be used for further work. Increasing the data in the abnormal groups and further development of the way the maxima are derived should produce a good working model in the near future from which a system such as we describe in this paper can be created. This will allow us to judge whether the most salient areas of a slide have been examined while being screened providing a quality control measure which runs in parallel to the existing screening program. As such it would not suffer from the same problems being faced by alternate automated screening interventions.

References

[1] Bijaoui, E. Starck, J-L., Murtagh, F. Restauration des images multiechelles par l'algorithme a trous. *Traitment du Signal*, 11, 229-243, 1994.

[2] Branca, M., Duca, P.G., Riti, M.G., Rossi, E., Leoncini, L., Turolla, E., Morosini, P.L., and the National Working Group for External Quality Control in Cervical Screening. Reliability and accuracy of reporting cervical intraepithelial neoplasia (CIN) throughout Italy: Phase 1 of a national programme of external quality control in cervical screening. *Cytopathology*, 7, 159-172, 1995

[3] Broadstock, M. Effectiveness and cost effectiveness of automated and semi-automated cervical screening devices: a systematic review. NZHTA Report 2000, 3(1), 2000

[4] Culverhouse, P.F., Williams, R., Reguera, B., Ellis, R., and Parisini, T. Automatic classification of 23 species of Dinoflagellate by artificial neural network. *Marine Ecology – Progress Series*, 139 (1-3), 281-287, 1996

[5] Fetterman, B., Pawlick, G., Koo, H., Hartinger, J., Gilbert, C., Connell S., Determining the utility and effectiveness of the NeoPath AutoPath 300 QC system used routinely. *Acta Cytolo*gica, 43, 13-22, 1999

[6] Hoang, M. A. & Geusebroek. J. M. Measurement of color texture. *Proc. of the 2nd International Workshop on Texture Analysis and Synthesis*, 73-76, 2002

[7] Jones, S., Thomas, G.D.H., and Williamson, P. Observer Variation in the Assessment of Adequacy and Neoplasia in Cervical Cytology. *Acta Cytologica*, 40, 226-234, 1996

[8] Koss, L.G., Lin, E., Schrieber, K., Elgert, P., and Mango, L. Evaluation of the Papnet screening system for quality control of cervical smears. *Anatomic Pathology*, 101(2), 220-229, 1994

[9] Li, A. & Lennie, P. Importance of color in the segmentation of variegated surfaces. *Journal of the Optical Society of America A*. 18 (6), 1240-1251. 2001.

[10] National Health Service Cervical Screening Programme. Achievable Standards, Benchmarks for Reporting, and Criteria for Evaluating Cervical Cytopathology. NHSCSP Publications, Sheffield, 2000

[11] Office of National Statistics. *Estimates of newly diagnosed cases of cancer England and Wales 1993-97*. ONS Monitor MB1 98-2, 1998

[12] Potter, T. *Assessing the skill of an expert cytologist engaged in cervical smear categorisation tasks*. MSc Computational Intelligence Thesis. Plymouth: University of Plymouth. 1999

[13] Sasieni, P., Cuzick, J., Farmery, E. Accelerated decline in cervical cancer mortality in England and Wales. *Lancet*, 346: 1566-1567, 1995

[14] Toth, L. and Culverhouse, P.F. Three dimensional object categorisation from static 2D views using multiple coarse channels, *Image and Vision and Computing*, 17. 845-858, 1999