Shape and Texture Recognition for Automated Analysis of Pathology Images

Violet Snell

Submitted for the degree of Doctor of Philosophy

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey

This research project is concerned with automated analysis of microscopic images used in clinical pathology for diagnosing disease. Application of computer vision methods can improve the accuracy, reliability and availability of tests, reduce the associated costs and ultimately improve patient outcomes.

Three different areas of pathology are covered:

- identification of clustered nuclei and detection of chromosomal abnormalities in DAPI-stained samples,
- diagnosis of auto-immune diseases from indirect immunofluorescence (IIF) images, and
- detection of dividing nuclei in H&E stained histopathology sections.

Despite the diversity of these application domains, the techniques used for their analysis are similar.

For cluster identification in DAPI images we focus on object shape and extend existing methods of shape analysis with novel measurements of the boundary profile which detect notches between overlapping nuclei in a cluster. For abnormality detection we focus on texture and develop a novel decision-tree dictionary for patch quantisation.

We continue to focus on texture for IIF images, developing suitable isotropic measurements as well as exploring the connections between classification of individual cells and whole patient samples.

Detection of dividing cells in tissue sections requires a combined assessment of shape, texture and colour in order to fully represent all relevant facets of the object. Here we develop a method for stain normalisation which efficiently compensates for batch variations in stain strength and proportions, followed by a full pipe-line of segmentation, feature extraction and classification, resolving issues of class imbalance implicit in detection of rare objects.

We develop an efficient and effective segmentation method, which is free of weight parameters and adaptable for use in different imaging modalities. We explore a variety of classifier types and ensemble structures, and suggest promising directions of future development in the broad application area of pathology image analysis.

Contents

1 Introduction **1**
- 1.1 Description of problem domains . . . 1
  - 1.1.1 DAPI . . . 1
  - 1.1.2 IIF HEp-2 pattern classification . . . 2
  - 1.1.3 Mitosis detection in H&E sections . . . 2
- 1.2 Motivation . . . 3
- 1.3 Existing methods . . . 3
  - 1.3.1 Segmentation . . . 4
  - 1.3.2 Shape analysis . . . 6
  - 1.3.3 Texture analysis . . . 8
  - 1.3.4 Machine Learning . . . 10
- 1.4 Outline and Areas of Contribution . . . 17

2 Abnormality Detection in DAPI images **21**
- 2.1 Application Domain . . . 21
- 2.2 Object type identification . . . 23
  - 2.2.1 Segmentation . . . 23
  - 2.2.2 Feature Extraction . . . 26
  - 2.2.3 Results . . . 30
  - 2.2.4 Discussion . . . 31
- 2.3 Abnormality detection . . . 32
  - 2.3.1 Methods . . . 33
  - 2.3.2 Experimental Results . . . 35
  - 2.3.3 Discussion . . . 36
- 2.4 Conclusions . . . 38

3 HEp-2 pattern classification **41**
- 3.1 State-of-the-art review . . . 42
  - 3.1.1 MIVIA Data Set . . . 43
  - 3.1.2 ICPR 2012 Contest . . . 44
  - 3.1.3 SNP HEp-2 Data Set . . . 46

3.2 Cell Experiments . . . 47
3.2.1 Evaluation protocols . . . 47
3.2.2 Feature sets . . . 48
3.2.3 Results . . . 50
3.3 Cell Distribution Experiments . . . 51
3.3.1 Normal distribution modelling . . . 53
3.3.2 Cumulative histogram modelling . . . 53
3.4 Discussion . . . 55
3.4.1 Analysis of experimental results . . . 55
3.4.2 Further work . . . 58
3.5 Conclusion . . . 59

4 Mitosis detection 61
4.1 Review of prior art . . . 63
4.1.1 ICPR 2012 Contest and MITOS dataset . . . 63
4.1.2 AMIDA Grand Challenge . . . 65
4.1.3 Gaussian Process Latent Variable Models . . . 67
4.2 Experimental methods . . . 68
4.2.1 Stain normalisation . . . 69
4.2.2 Detection of candidate locations . . . 72
4.2.3 Segmentation . . . 77
4.2.4 Feature extraction . . . 78
4.2.5 Classification . . . 79
4.2.6 GP-LVM detection of mitosis . . . 80
4.3 Results . . . 81
4.3.1 Extracted Features pipeline . . . 82
4.3.2 GP-LVM pipeline . . . 84
4.3.3 AMIDA Contest . . . 85
4.4 Discussion . . . 85

5 Reflection 89
5.1 Human vs Machine Learning . . . 90
5.2 Summary of Contributions . . . 91
5.3 Limitations . . . 93
5.4 Future directions . . . 95
5.5 Conclusions . . . 96

Acknowledgements 99

Bibliography 101

Chapter 1

Introduction

The subject of this research project is automated analysis of microscopic images of various human tissue samples, as used in clinical pathology for diagnosis and screening of various diseases. Improvements in their automatic assessment can greatly increase the accuracy, reliability, and/or availability of tests, reduce the associated costs and ultimately improve clinical outcomes. The majority of the work is concerned with the technical aspects of machine analysis and interpretation of digital images, but a small amount of biological and medical background is given in order to motivate the study.

1.1 Description of problem domains

The project covers automated analysis of images in three different areas of pathology, acquired using different stain types, but the techniques that are used for their analysis are broadly similar. The three staining methods are DAPI, ANA-IIF and H&E. DAPI (diamidino-2-phenylindole) is a fluorescent stain which bonds strongly to DNA-rich parts of the cell, allowing visualisation of the nucleus. Anti-nuclear antibodies (ANA) are used for diagnosis of auto-immune diseases, and are most commonly visualised through indirect immunofluorescence (IIF) with a substrate of HEp-2 cells. Hematoxylin and Eosin stain (H&E) is very widely used in *histopathology*, the study of whole tissues and their structures, as opposed to separate cells. Analysis of tissues can be considerably more complex than the relatively simple study of cells as individual objects.

1.1.1 DAPI

DAPI-stained cell nuclei in various human tissue samples can be used for diagnosis and screening of cancers and pre-cancerous conditions, and improvements in

their automatic assessment can greatly increase the accuracy and availability of tests, reduce the associated costs and ultimately improve clinical outcomes. Two tasks are investigated for this modality: identification of cell clusters that require further splitting, and the possibility of determining chromosomal abnormalities from nuclear appearance.

1.1.2 IIF HEp-2 pattern classification

A wide variety of auto-immune diseases affects different parts of the body, but are all associated with an immune reaction to, and an attack on, the person's own tissues. This reaction, known as anti-nuclear antibody (ANA), forms the most reliable basis for ascertaining the presence of, and establishing the specific type of auto-immune disease. The diagnosis is usually performed by highly trained physicians directly at the microscope, although better results can be obtained through digital imaging of the slides, as the fluorescence decays fairly rapidly. Both the overall brightness and the visual pattern of the fluorescence feed into the diagnostic decision, although many clinical settings will only use the brighter samples, known as *positive*, for identification of specific patterns. A large number of these visual patterns of fluorescence is described in the medical literature, and various groups or subsets of these have been targeted for automatic recognition by previous published works in the computer vision field.

1.1.3 Mitosis detection in H&E sections

H&E staining is the most widespread method of visualisation for histology slides, but this investigation is restricted to breast cancer biopsy specimens, and specifically to the task of mitosis detection within these. Mitosis is the process of cell division, and the proportion of cells within a tumour that are undergoing division gives an indication of the rate of growth of the tumour, and therefore its aggressiveness, the likely prognosis and the most appropriate treatments. The task is most challenging, for both human and machine, as the nucleus changes its shape and structure throughout the different stages of mitosis, and when this is combined with all the possible viewpoint orientations and exact slicing positions relative to the nucleus, the variety of resulting appearances is bewildering. Pathologists train for many years, and yet the inter-observer agreement on whether a nuclear figure is mitotic has been found to be only $\kappa = 0.38$ [1] (N=43, balanced sample, 7 observers). This creates considerable additional difficulty in training an automated process for mitosis detection, as the 'ground truth' is never entirely reliable.

1.2 Motivation

Automation of medical image analysis using computer vision methods can have a number of desirable benefits compared to manual techniques, which depend on the precise application area. At the simpler end, cell counting in homogeneous populations can improve accuracy, save time and consequently free up the skilled personnel for more complex and less repetitive tasks. The benefits come at the cost of having to digitise the pathology process, which so far remains largely optical for the majority of routine work, but is gradually being augmented by automated scanning technology, initially for archival and indexing, or remote consultation purposes [2,3]. Digital slides are also well suited for use in the training of new pathologists, and as their availability becomes more wide-spread and practitioner familiarity grows, the scope for automating their analysis increases.

Image analysis algorithms can provide a fast search for regions of particular interest within the slide, mechanise measurements of cell appearance that support the pathologist's decision making, and improve reliability of subjective evaluations. As growth of digital imaging in pathology accelerates, more data will become available for training and evaluating machine learning approaches, leading to improved performance which should, in turn, help to boost acceptance of these methods in clinical practice.

At its most advanced, automated image analysis can discover features of the tumorous tissue that have great prognostic value, but were previously not known by clinicians [4]. Discoveries such as these can then stimulate further research into the underlying biological processes that generate a particular tissue appearance, advancing the understanding of factors which affect the disease progression.

1.3 Existing methods

In its broadest sense, the subject of this work is *computer vision*, that is, an attempt to devise an algorithm capable of ascribing meaning to an image. Modern computer vision is largely based on machine learning methods, in their majority statistical ones, and several that are particularly applicable for image classification will be reviewed here. As the most comprehensive current review of image analysis in histopathology indicates [5], specific properties of pathology images, such as high data volumes or staining techniques, create a different set of requirements in their analysis as compared to the more general imagery of everyday objects commonly addressed by computer vision, and even to the more established computer-aided diagnosis in radiology. Much of this work has been concerned with identifying these distinguishing properties of microscopic pathology images, and analysing how they are connected to the suitability or otherwise of particular

algorithms. These differ between cytology, which is concerned with isolated cells or small clusters, and histology, which preserves the tissue structure and therefore more high-level information, but presents a considerably more complex analysis challenge as a result.

As a newcomer to the vision field, I expected this connection between image properties and the relevant algorithms to be well understood, but have found, with some dismay, a plethora of works evaluating a particular algorithm on a particular dataset (or a few at best) but offering no guidance as to its likely performance on a different set of images. Further exploration uncovered an ongoing tug-of-war between the hand-crafted algorithms, carefully designed and optimised for a very specific task, and the more generic vision architectures, which reach for the AI ideal of matching the human visual abilities, but nonetheless rely on a number of system parameters which have to be optimised by brute force in order to achieve acceptable performance.

The structure of the following sections tracks the pipeline of analysis undergone by an image: segmentation into objects of interest is followed by feature extraction, with features broadly sub-divided into those relating to the object's shape and to its internal texture; finally, the extracted features are combined by a classification algorithm to reach a decision.

1.3.1 Segmentation

The range of image segmentation methods is extremely broad. They can be based on thresholding (at the simplest end), clustering, histograms, edges, contour energy, watersheds, graph partitioning - the list goes on. Not only is there very little solid evidence as to the suitability of a particular method for use on a particular image type, but the very notion of segmentation performance or quality is highly ambiguous [6].

One option for evaluating segmentation performance is to compare the automatically segmented boundary with a manually drawn one, and measure the number or proportion of misclassified pixels or some form of distance between the two boundaries. This has two drawbacks: the necessity of procuring a manual annotation of the image, which quickly becomes very resource-intensive for a large dataset, and the potential inaccuracy of the manual contour, which is taken as 'ground truth'. In attempting to delineate the boundary of an object in an image with a finite sampling grid, no answer that is limited to integer pixel positions can be better than half a pixel wrong on average. Much more importantly, the manual boundary is subjective, with potentially large areas of ambiguity occurring in noisy real images. So any comparison to this subjective 'truth' is potentially misleading as to the quality of segmentation.

The alternative approach is one of qualitative assessment of segmentation

results, which suffers from limitations on the number of examples that can be evaluated in a reasonable time, or included in published work, as well as the obvious subjectivity of such evaluations. In cases of large differences between the candidate methods such comparisons can nonetheless be quite illuminating. Since segmentation is rarely the final objective, but more likely an early step in a processing chain involving measurements of the segmented objects, it is the effect of different segmentation methods on these measurements that is of greatest import, and can be the ultimate arbiter of the suitability of a particular segmentation process to the task. Such evaluation of segmentation quality is, of course, only applicable in that specific context, and does not provide further insight into its performance under other conditions.

The range of segmentation methods that are used on pathology images is much narrower, due to the particular nature of the images, and the choice is further constrained for each imaging modality and sample type. Since the earliest days of segmentation of cellular images, thresholding has been used extensively [7–9], and with remarkable success. This success can be attributed to the simple fact that many types of cell image contain objects with a fairly uniform interior and a reasonably strong contrast to the background. As long as a good threshold value is applied, the resulting contour will often match the visual boundary very closely, without recourse to more complex segmentation methods. The remaining question is how to select the best threshold, and it is here that the specifics of different cell and staining types come into play, creating a diversity of algorithms.

The one threshold selection method that is often unsuitable for use on cellular imagery is the first one in the textbook: Otsu's method [10]. This is based on the objective of minimising intra-class variance, or equivalently maximising the inter-class variance, which can be interpreted as finding a clustering for the foreground and background that is tight within each cluster and has well-separated centres. Unfortunately, the variances of background and foreground in cytopathology images are not comparable: the background is usually extremely flat, almost uniformly black in fluorescent images, while the foreground will exhibit some texture and therefore much higher variance. This difference of variances creates a strong bias, explored in depth by [11], resulting in thresholds that are much higher than optimal, and consequently losing outer edges of cells and producing extremely meandering contours for even the smoothest and rounded objects. Regrettably, this method does still get used in practice, as it is built into many libraries and toolkits, and included in most introductory courses and textbooks.

Based on a much weaker assumption of normal distributions, but explicitly allowing for differences in spread, Kittler and Illingworth's 1986 method of threshold selection based on minimum error (MET, [12]) does not suffer from such bias, and was still top of the table in a broad and robust 2004 survey

of thresholding techniques [13], nearly 20 years later. Other criteria that have been used for threshold selection in segmentation of cytology images include gradient measurements, originally suggested by Kohler in 1981 under the term 'contrast of edges' [14], and essentially looking for a threshold that would create contours with the greatest average difference between pixel values on opposite sides. This is distinct from using edge detection as basis of segmentation, which has to involve some form of merging decisions to construct a complete outline. A similar approach is employed by [15], although they prioritise the number of edge points that have high gradients, rather than a high average, and follow the thresholding with a number of heuristic post-processing steps to arrive at the final segmentation.

More sophisticated segmentation techniques, which go beyond thresholding, have also been used for cell images, in particular active shape contours, or 'snakes'. The 'balloon' form is especially suitable here [16], provided a good initialisation can be found. Different formulations of the energy function have been proposed for particular specimen types and imaging modalities, for example [17], but all have a high computational cost due to the large search space and iterative nature of the algorithm. Such models generally cope better with overlapping cells than thresholding does, as long as the amount of overlap is moderate, although in general segmentation of overlapping cells in low-contrast modalities, such as phase-contrast microscopy, remains an active research topic [18, 19].

Most of the advanced, complex segmentation methods remain tied to very specific image types and are not readily transferable to another domain [20]. In this work both DAPI and H&E cellular images have sufficient contrast to successfully employ adaptive thresholding. Our concern with clusters of overlapping cells is limited to identifying them, rather their segmentation into constituent parts, although there is an extensive body of literature addressing this related topic, most recent ones approaching detection of boundary concavities by looking for changes in the direction of the normal [21, 22]. Identification of clusters is largely based on analysing the shape of the segmented object, which is the subject of the next section.

1.3.2 Shape analysis

The breadth of methods for shape analysis proposed in the general computer vision literature is similarly enormous to that of segmentation. The most recent comprehensive review, carried out in 2004 [23], classifies the methods as either contour- or region-based, a distinction that is largely not relevant for cellular objects, as they do not contain holes. The other division is between global representations that analyse the shape as a whole, and structural, or local, representations, which break it down into a series of segments or primitives of some

form. The strengths of local approaches are partial matching and handling of occlusions, which is not a major issue in tissue images, provided that the segmentation has been done well. So the majority of relevant methods are global, from the simplest single features such as circularity or eccentricity, to more complex transform techniques such as Fourier or wavelet descriptors. Finally, a distinction is often drawn between information-preserving representations, which are invertible, and non-information preserving descriptions, which lead to ambiguity in reconstruction. As reconstruction is not a requirement for any of the applications we are considering here, this distinction is not relevant.

Another very thorough survey of shape features [24] describes and examines a rich menagerie of methods and attempts to characterise each one according to its invariance qualities as well as its resistance to noise, occlusions and non-rigid deformations. Some measure of the computational complexity involved is also given. The authors acknowledge that such judgements are necessarily approximate, but also depend on the 'type of shapes' under consideration, although no further guidance is offered on which might be most suitable for a particular combination of shape type (however that may be defined) and required task.

A more recent exploration of shape analysis in the specific context of biological images is given in the introductory chapters of [25], with a focus on statistical models of deformation. In the broadest terms, a statistical shape model is obtained from data and characterises the variations in shape that are present in the subject domain. Successful application of such models depends strongly on the choice of a suitable underlying shape representation. Principal component or eigenvector analysis can then be used to derive a statistical model of variation within the shapes. The representation chosen for the remainder of [25], based on landmark points, is not appropriate for the relatively simple, but much more variable, shapes of cellular and nuclear objects, and therefore not relevant here.

As confirmed by [5], the shape features most commonly used in histopathology automation are restricted to global, single-value measurements such as aspect ratio or solidity, and boundary transforms. Many of these are formulated to be invariant to scale and orientation changes, which is important for pathology images, while others require additional normalisation steps to ensure these invariance properties. They are also among the more computationally efficient shape representations [24], which becomes very valuable when the number of cells is large. One very well-established boundary transform method is Fourier Descriptors (FD), which applies Fourier frequency transform to the radial boundary profile. Among the advantages of FDs listed by [23] are their relative simplicity, in both computation, and in interpretability, or connection to the visually recognisable properties of the shape. A number of variants of Fourier Descriptors have been proposed in the literature, which were compared by [26] to find that centroid distance outperformed complex coordinate, angular and curvature

representations in retrieval applications.

The relative merits of different shape representations can only be evaluated in the context of a particular application, which generally requires some distance metric to compare the degree of similarity between two shapes. There is no objective universal measure of shape similarity, so a metric has to be chosen to reflect some relevant notion of shape difference, but also one that is appropriate for the particular shape representation. This point is explored little in recent literature, with majority of applications employing Euclidean distance regardless of the properties of their chosen shape features.

The importance of shape features in classification depends enormously on the nature of the images: cells that are uniformly elliptical cannot be differentiated by shape features. There are cases where they can be better distinguished according to their texture, which is the subject of the next section.

1.3.3 Texture analysis

If the number of shape analysis methods is staggering, that of texture is simply vast [27]. This is partly due to the fact that texture is a very ill-defined property, as compared to shape, and can be understood to mean quite different things depending on context. One fundamental division is between physical texture of the object and its appearance in an image, which depends on the imaging conditions such as lighting or distance. In the context of histopathology images, we are concerned primarily with appearance, as the physical objects are not measurable in any other way. The imaging conditions that most affect the appearance in this case are variations of stain distribution and microscope focus.

Image texture is fundamentally concerned with dependencies between pixel values at different spatial offsets and scales. As such, the simplest measure of texture is grey-level variance, either global across the whole object of interest, or local in a defined neighbourhood of each pixel. This measures the *amount* of variation between pixels, but takes no account of their relative position, which does however grant it automatic rotational invariance. Other early texture measures concentrate on edge [28] and gradient statistics [29] as salient factors in texture discrimination. Similar observations motivate the use of Gabor filter responses at different orientations and widths as texture representation [30], and, by extension, that of any other bank of filters, for example Laws' masks [31, 32] and MR8 filter bank [33].

By the far the most commonly used texture quantification technique is the second-order Grey-Level Co-occurrence Matrix (GLCM) [34], sometimes referred to as 'Haralick features'. These features are derived as various statistics of the joint distribution over grey-level value combinations for pairs of pixels that are a certain offset apart. The rotationally invariant version is typically used for

cellular analysis [35], as orientation of a cell should not impact on its assessment. A large number of statistics can be defined based on the co-occurrence matrix, but many of them are strongly correlated with each other, so no additional information is obtained from using more than two or three measures from each matrix. Of similar vintage is the Grey-Level Run Length matrix (GLRLM) [36], which operates on the distribution of lengths of 'runs' of identical or similar pixel values. Careful selection of quantisation step size is necessary in order to get the best performance from either method.

A more localised model of explicit higher-order correlation for analysing texture was proposed by Kurita and Otsu [37], but has not found wide-spread acceptance outside Japan. Instead, a much coarser model that only considers whether a pixel is brighter or darker than its neighbour has been spectacularly successful in texture-based segmentation and has spawned a multitude of variants and extensions [38–42]. Local Binary Patterns (LBP), introduced in the late 1990s by Ojala et al. [43], turn their very coarseness into a strength, as they are extremely robust to monotonic transformations of the grey-scale and computationally efficient. The local neighbourhood analysed by LBP and assigned a pattern code can be regarded as a micro-texton. Introduced by Varma and Zisserman [44], texton texture recognition has, as yet, seen little application to nuclear analysis. Similarly to a Bag-of-Words model, all patches of a certain size are extracted from the image and quantised according to a dictionary, or code-book, which is obtained by clustering patches from the training examples. Both dictionary and LBP methods compare objects or larger image regions based on histograms, i.e. empirical estimates of the distribution of particular textural patterns within the image, with χ^2 histogram distance most commonly employed for classification.

Transform approaches to texture analysis include Fourier spectrum [29] and wavelet analysis. The Fourier power spectrum is both straight-forward to compute and easily understandable in terms of frequency content of the image. Wavelet analysis [45–47] can provide not only the strength of different frequencies in the image, but also a measure of their spatial non-uniformity, which may be a discriminating factor in some applications.

Fractal dimension has found widespread use as a texture measure [48], including applications in cytology [49]. Its computation from images of limited resolution can be problematic, and the validity of some formulations has been questioned [50], as they are not invariant to linear intensity transformations.

Morphological approaches to texture divide into those of successive application of a structuring element to the image (granulometry) [51], and of quantifying morphological properties of connected elements in image slices at a range of thresholds [52, 53].

Some commonly desirable properties of textural representations are invariance to perturbations in scale, orientation or illumination, and many of the techniques

described above have variants or extensions that cater for these needs. In the context of pathology images, rotation invariance is essential, but is sometimes confused with the isotropic nature of most cellular textures, which allow simplified, lower-dimensional feature formulations. Invariance, or at least insensitivity, to blurring caused by variations in focus quality can be a very valuable property in quantification of microscopic texture, and current methods based on local phase quantisation can outperform LBP and Gabor filter banks [54].

Multi-scale representations capture additional information and often improve performance compared to a single scale chosen arbitrarily. Many texture representations have multi-scale extensions, for example patches of different sizes in a dictionary system, different offsets for co-occurrence, or sizes of rings for local binary patterns. Others have built-in multi-scale capability, for example wavelet or fractal techniques.

Another issue which can greatly complicate texture analysis is colour. Most of the images we are concerned with are either grey-scale or have a single dominant colour component and can be easily converted to grey-scale. Some extensions necessary to make full use of coloured or multi-spectral texture are suggested in [27, 55].

Despite the large number of works addressing the use of various texture analysis techniques, their relative applicability to a particular practical problem remains poorly understood, with little indication of the limitations and trade-offs that could guide the selection of most suitable method. What is shown to be 'best' actually depends entirely on the application examined by each particular review or comparison, and tend to be highly specific to the particular imaging conditions, including tissue type, staining method and magnification. One general trend which emerges from the application studies is that a combination of several different texture measurements often provides better classification performance than any individual method on its own, and the most suitable way of combining their contributions is one of the subjects covered in the next section.

1.3.4 Machine Learning

Machine Learning is usually described as the study of systems that learn from data, in other words that can predict outcomes for new data points based on previously examined examples. Its fuzzy boundaries with Pattern Recognition, Artificial Intelligence and Statistics continue to be argued over by practitioners on all sides, although it is not clear to what end. Some of the distinctions are historical, such as Pattern Recognition's roots in statistical data analysis or AI's background in emulation of human cognitive function; others are a matter of emphasis or perspective: is the subject about problems and how to solve them, or tools and which way to use them? My personal view places statistics and

computational theory as foundations for the strongly overlapping fields of PR and ML within the much broader scope of AI, but useful insights from any area are all gratefully received.

All the applications we are concerned with in this work fall into the category of *supervised* learning, where some set of labels is attached to each of the known training examples, and the system is attempting to predict the labels for new points. This construction suffers from the fundamental problem of relying on the correctness of training labels, or 'ground truth', which in practice can be quite poor in the subjective area of medical opinion [56]. Limited work has been done to address this from a ML perspective [57]. On the positive front, it is easy to see how to evaluate the performance of a supervised learner by testing it on unseen material, whereas the performance of unsupervised tasks, such as clustering, is not so easily measured.

One aspect of performance evaluation which can, nonetheless, trip up the unwary is *class imbalance*, that is a large disparity in the frequency of occurrence of different labels, which is a common scenario in diagnostic systems where an overwhelming proportion of presented examples is disease-free. A naive measure of accuracy such as the percentage of predictions that are correct would fail to spot the problem with a system that missed most of the ill patients, if they were but a small proportion of the total. This issue is very clearly understood in medical statistics, although computer vision sometimes takes a rather laxer approach. Class imbalance also presents problems for the learning process itself: a minority class may be represented by so few examples that it becomes very hard to make generalisations; also, some algorithms are inherently vulnerable to imbalance, and effectively mistake rarity for unimportance [58–60].

This is but one illustration of the lack of statistical underpinnings in many machine learning methods. The introduction of Neil Thacker's attempt to define probability for scientists [61] describes this bleak state of affairs as follows:

> 'The dominant attitude to statistical methods being that we can largely pick various measures out of thin air and worry about how they behave on data afterwards, rather than deriving techniques from principles based upon the characteristics of the data.'

His assessment reflects my own view of much of the field, especially in computer vision applications: assumptions are rarely understood, let alone questioned, and the link between data characteristics and applicable methods is rarely explicit.

The typical pipeline in computer vision consists of feature extraction, which converts the huge mass of pixel values into a more compact representation, followed by a classifier. In this pipeline, the enormous choice of possible features, which has been explored in the previous two sections, albeit non-exhaustively, interacts with the choice of many different types of classifier, and performance

can only be evaluated for the whole system. Both features and classifiers have tunable parameters, and several feature types may be needed to comprehensively cover all relevant aspects of the image, resulting in a combinatorial explosion. The deluge of published papers describing a particular combination of attributes and classifier, with a tweak of some aspect of either or both, tested on a particular data set or two, is a reflection of the random stumble around algorithm space that is the outcome of this lack of systematic understanding.

Highlighted as an issue by Haralick over two decades ago [62], performance measurement of computer vision systems has hardly improved since, and certainly not reached the stage of his suggested systematic assessment of the effects of both parameter variation and input noise. The subject resembles biology at the stage of collecting beetles, rather than constructing evolutionary theories. One way that current research attempts to address this is the considerable effort directed towards the possibility of learning the suitable features, or *representation learning*. This is usually done in unsupervised contexts, and is aimed at discovery of features that are inherently good at representing the domain data by capturing the underlying factors causing variation [63]. The space of features that are learnable by these systems is frequently constrained to a particular architecture, e.g. auto-encoders or Restricted Boltzmann Machines, so they are parametric, albeit non-linear. This restriction is, perhaps, one of the reasons that representation learning is often done in combination with *deep learning* [64], which builds a hierarchy of representations from low-level ones that are closely related to the image pixel values, up to higher levels of abstraction that are closer to the target concepts of visual understanding. Multiple layers of Convolutional Neural Networks (CNN) have been applied to a range of challenging visual recognition tasks, including mitosis detection, with very positive results [65]. The same approach of stacking multiple layers of feature extraction can be successfully combined with non-parametric models such as Gaussian Processes (GP) and associated latent variable models (LVM) [66].

Latent variables are a fundamental concept in dimensionality reduction [67] and the closely related sub-field of manifold learning [68]. The input image space, where each pixel's colour corresponds to at least one dimension (more if the colour representation is richer than grey-scale), is very high-dimensional and creates enormous challenges for learning algorithms, for the simple reason that it is practically impossible to populate such a vast space with a sufficient number of training examples - the requirement grows exponentially with each additional pixel. It is consequently necessary to find a smaller number of underlying variables that control the correlations between pixel values to create the overall picture, and subsequently use these latent (i.e. unobserved directly) variables as the basis of classification. Many linear and non-linear methods of modelling such variables have been proposed, starting with the basic statistical technique

of Principal Component Analysis (PCA) and extending to a kaleidoscope of kernel, spectral, adaptive, probabilistic and information-theoretic justifications for a particular choice of embedding. Van der Maaten concludes in his comparative review [67] that the artificial datasets presented to illustrate the advantages of many non-linear methods are too specific, and the purported advantages fail to generalize to real data, where simple PCA continues to perform robustly. Although barely given a mention in the review, Gaussian Process LVMs, discussed in greater depth in Section 4.1.3, have made significant advances in recent years [69], and offer the additional bonus of Automatic Relevance Determination (ARD), which is the ability to determine how many latent dimensions are actually needed to represent all the salient variations of the data, a question that is left unanswered by most other types of dimensionality reduction.

The original aim of dimensionality reduction methods was visualisation, which partly explains their propensity for fixed target dimensionality and also for their basis in the concept of distance preservation, for a particular interpretation of 'distance'. This connection to distance metric learning, and by extension to one of the earliest types of classification - k Nearest Neighbours (kNN), which relies on a suitable measure of distance between pairs of examples - potentially reduces classification based on latent positions to kNN with a learnt distance metric.

Another conceptual difficulty with such an approach is that even latent variables that perfectly explain the variability in the data are not necessarily optimal for discriminative classification tasks: many variations will occur unconditionally of class, and therefore will only distract the classifier - these are sometimes known as *nuisance* variables. Imaging conditions are a classic example of such distracting variation in the context of image classification. Some notable attempts have been made to improve the discriminative value of learnt features in a supervised setting. Snoek et al. in [70] use a GP-LVM to ensure that it is possible to construct a smooth mapping from their trained auto-encoder representation to the class labels, without restricting the form of this mapping in any way, i.e. giving the auto-encoder non-parametric guidance towards a more discriminative representation. Although the results are clearly superior to those from a normal single-layer auto-encoder, the method cannot outperform a deep convolutional network, and requires tuning of a weight between representation accuracy and discriminative power in order to achieve best results. Urtasun & Darrell in [71] take a more direct approach of penalising examples that come from different classes but lie close in latent space during the optimisation. Their method also involves a trade-off parameter between its discriminative and generalisation abilities. Finally, a fully probabilistic treatment of supervised learning with GP-LVM is offered in [72], utilising the conditional independence of observed examples and their labels given the latent variables. So far, GP methods have proved good at generalising from a very small number of samples in a high-dimensional space,

but suffer from scalability problems if the number of samples is large, as their training is cubic in the number of samples. The use of GP-LVMs in classification is also problematic because they are formulated as forward mappings from latent space to observed values, so to determine the latent position of a new test point requires a slow numerical optimisation. However, some recent advances in sparse GP modelling are showing potential ways to overcome this limitation [73].

Of particular relevance to our application area, texture descriptors (introduced in Section 1.3.3) tend to be high-dimensional, and generic feature *selection* techniques are sometimes used to reduce their dimensionality by identifying a subset of the features that is most valuable for classification (although this subset may be specific to a particular type of classifier). As an intermediate step between this simplistic scheme and the fully flexible mapping of latent variables, Nielsen et.al. in [74] attempt to reduce dimensionality of textural features by taking a weighted sum of the ones that exhibit strong differences between classes. While correctly pointing out that as the number of features becomes larger than the number of training examples, simple feature selection results may no longer generalise, the authors make some strong assumptions, such as independence of input features, and some arbitrary choices, such as restricting the target space to 2 dimensions. Their potentially promising experimental results are not supported by a sufficiently robust analysis to show that the comparison with existing ad-hoc features passed to one classifier type is a valid one.

Another texture recognition technique, described in Section 1.3.3 as textons, is an example of the *dictionary learning* branch of ML. The basic premise of dictionary learning, which originates in the unsupervised context of encoding and compression of signals, is to represent each sample point by an approximation chosen from a set of predefined 'words', which together form the dictionary. The encoding is then formed listing the code, or identifier, of each word, the bit-length of which depends only on the size of the dictionary and not on the size of the original sample. In classification, it is the frequency of occurrence of each 'word', collected in a histogram, that is used as the discriminating basis. Two major questions arise from this configuration: firstly, how to arrange the dictionary entries in the original feature space so as to collect the most useful information in the histogram, and, secondly, what is the optimal measurement of difference between histograms? The first part also comes with attendant queries of how big should the dictionary be and how do we determine which dictionary entry is nearest, in other words what distance measure is most appropriate for our samples? The most common set of answers to these questions is 'k-means' with an empirically optimised value of k, Euclidean distance for the samples, and χ^2 distance for the histograms. As k-means (and similarly k-medoids) is an unsupervised clustering method, which aims to minimise the overall distortion of the representation under the Euclidean distance metric [75], it has no funda-

mental connection to the discriminative information contained in the resulting histogram. Similarly, the χ^2 histogram distance is an entirely heuristic way of comparing two empirical distributions. Some attempts have been made to take into account the discriminative needs of a supervised problem in dictionary design: based on minimising loss of discriminative information, [76] essentially aims to align partition boundaries with class density changes, making each partition more pure. The method is limited by hard assignment of each point to a partition, and is fundamentally concerned with predicting the label given a point, even though in dictionary-based classification the prediction is actually made from the distribution of (encoded) points, and it is this distribution that needs to be made more distinctive. An alternative approach to more discriminative use of dictionary learning, proposed in [77], makes a separate dictionary for each class, optimised for good reconstruction of samples belonging to that class, and then uses the reconstruction error as the discriminative feature. It makes underlying assumptions about comparable variability within each class, as an equal size of dictionary is employed for all classes, and therefore may not be applicable in some scenarios.

By far the most prevalent method of supervised classification in modern computer vision is Support Vector Machine (SVM). Introduced in mid-1990s, SVMs construct a dividing hyper-plane in feature space that separates the classes and gives the largest possible margin between the boundary and its nearest training points. SVMs are frequently, but not necessarily, combined with a kernel function, effectively boosting the dimensionality of the feature space and allowing the construction of far more complex boundaries. This extension of the method to address problems that are not linearly separable has accounted for much of its popularity, as does its deterministic behaviour and superior classification performance in many domains, particularly where the number of training samples is limited. The method does require optimisation of kernel and margin parameters, usually done by a hierarchical search for best performance on a validation set.

Gradually gathering momentum, especially in medical imaging applications, is a technique known as *Random Forests* [78]. Based on the earlier concept of decision trees, which make planar bisections of the feature space at each node, random forests add probabilistic and information-theoretic underpinnings to postulate that injecting randomness into the previously deterministic process of tree construction improves the robustness and generalisation capability of the overall system. Instead of seeking a single perfect answer (for example, *the* maximum margin separating hyper-plane of SVMs), random forests embody the collective wisdom of many different 'quite-good' opinions. The diversity, or randomness, of decisions implemented by individual trees is an essential component of the method's strength, and it usually relies on the plentiful supply of randomly generated features, which often consist of simple differences between pairs of pixels

or local integrals. By eliminating the need for a separate feature extraction step, which is never fully optimal, random forests unify many learning tasks, such as classification, regression and density estimation, into a single framework. They do require optimisation of parameters such as number of trees, their maximum depth and randomness, and generally perform better when a large quantity of training data is available. A notable recent proposal strengthens RF performance by giving higher weight to trees that are known to be better discriminators in combination with narrowing the pool of possible features to contain a selection of relatively strong ones [79].

The word 'learning' implies some change in the system, yet most formulations of supervised learning tasks in ML are actually one-shot: given a set of training data, construct a system capable of prediction of similar data, which will then be tested on a different set of data. Only the relatively small sub-fields of active learning and reinforcement learning address the questions of altering an existing system in light of new experience. Active learning approaches these from the specific view-point of the cost of labelling, and attempts to acquire labels only for those examples which will give it the greatest improvement in performance. Although these are very promising directions in the medical context, where acquisition of data can be particularly difficult [80, 81], they remain subjects of future work in our particular applications.

Despite its length, this overview of machine learning is far from exhaustive, giving only a cursory introduction to the topics most relevant to the applications of interest, and leaving out such large sections as genetic programming and evolutionary optimisation [82], sparse methods, independent component analysis, fuzzy logic and many more. Beyond the basic 'feature extractor plus classifier' pipeline lie questions of combining features of different origin and characteristics into a single prediction, either through optimisation of feature weights or more sophisticated kernel fusion [83]. Underlying all machine learning is the fundamental concept of distance between samples in the high dimensional measurement space, a notion of similarity or dissimilarity that is most appropriate for the categories that are being compared, and one that is potentially very different to the common Euclidean distance. Ultimately, feature extraction must preserve the information about this similarity in its manipulation of the pixels, as no clever classifier can compensate for its loss, but there is currently no universally accepted way to gauge the informational quality of a representation without pairing it with a specific classifier type.

1.4 Outline and Areas of Contribution

In the following three chapters of the thesis we examine in greater detail the three application domains described in Section 1.1, and their associated algorithms: Chapter 2 considers identification of clustered nuclei in DAPI-stained screening, and explores the scope for detecting chromosomal abnormalities from the appearance of single nuclei; Chapter 3 looks at diagnosis of auto-immune diseases through automated classification of immunofluorescence images; and Chapter 4 investigates methods of detecting mitotic nuclei in histopathology sections. Only the broad principles of the relevant existing methods have been reviewed here, and more detailed explorations of application-specific proposals are included in each of the chapters. Finally, Chapter 5 will summarise the contributions of this work and draw some conclusions, as well as offer suggestions for possible future directions.

The goals of this work in all three domains are highly challenging and open-ended. A major source of this challenge is the lack of clear understanding within the subject of the connection between image (or data) properties and the most suitable or promising method to apply. Although it is beyond the scope of this thesis to answer the broad question for all image types and all algorithms, we make some progress within the narrower scope of the specific pathology image categories.

In the initial task of Chapter 2, that of cluster identification, we focus on object shape as the primary characteristic of interest. Existing methods of shape analysis, reviewed in Section 1.3.2 above, are augmented with novel measurements of the boundary profile that are designed to detect notches between partially overlapping nuclei in a cluster. We then address the more ambitious undertaking of abnormality detection based on appearance, in the absence of definitive knowledge about visual features potentially associated with abnormality. Our focus here is texture, and we develop a novel dictionary construction, based on decision trees, to replace the more established models for texton quantisation in the texture analysis (see Section 1.3.3). This dictionary is extremely fast and, unlike the traditional methods, specifically targets discriminative power in different areas of feature space in order to improve final classification performance.

We continue to focus on textural features in Chapter 3, exploring and comparing a number of different approaches to measuring texture. Our main conclusion, however, is not about the relative merits of particular texture features, but the processes that should be used to evaluate and compare methods that assess individual cells when the diagnostic context is ultimately one of a whole patient sample. We investigate this subject in greater depth in Section 3.3, and compare methods that model the entire sample directly with those that view it as a collection of cells.

We raise the stakes once more in Chapter 4, by moving from cytopathology to histopathology, with the corresponding increase in scene complexity, and searching for objects with inherently variable appearance and ill-defined characteristics. Here we need to bring together aspects of shape, texture and colour in order to fully represent all relevant facets of the object, as well as address major complications on the classification side that arise from extreme levels of class imbalance implicit in detection of rare objects. We develop a unique adaptation of histogram matching suitable for transmission microscopy images in order to remove the effects of batch variations in stain strength and proportions on the colour profile, and hence on the texture and contrast measurements that we make. In a parallel to the conclusions of Chapter 3 we reflect here that the ultimate task of diagnostic relevance is not finding the individual objects, but assessing their average density, and this should form the basis for future evaluation of algorithms in this application.

Throughout the thesis, segmentation plays a prominent and vital role. The quality of segmentation results determines how accurately the downstream processes can measure shape, and how well the object's texture can be separated from that of the background. We propose an efficient and effective segmentation method, adaptable for use in different imaging modalities, and apply it to both DAPI-stained nuclei in Section 2.2.1 and H&E stained tissues in Section 4.2.3.

In every domain we examine, we have explored the use of a variety of classifier types as well as methods for combining them into ensembles. Although support vector machines (SVM) prove versatile and succeed in many of the learning tasks we consider, we have also explored the potential of random forests, k-nearest neighbour and Bayesian GP-LVM classifiers in application to our chosen domains. In the case of Random Forests we find that they can't compete with SVMs when given fixed manually constructed features as input, as they need the richer space of pixels and their combinations to randomise. Nearest neighbour methods are very dependent on the distance metric used, and are only applicable when the number of training samples is relatively small. GP-LVMs prove both computationally difficult and fundamentally more limited in their representational power than their formulation suggests, at least in their current stage of development.

Not all the experiments that I have tried out have made it into this document, just the major themes and threads of development. This is particularly true of feature extraction, which is the hardest part of the development process, the "black art", requiring intuition and ingenuity [84], and frequently resulting in disappointment. To avoid the disappointment and give a more solid basis to the belief that the features used are the best, or at least nearly the best, that they can be, the promising directions of research are those which draw features from a very large space of possible functions and evaluate them automatically, such

as the random combinations of attributes in random forests or linear kernels in multiple layers of deep CNNs.

Chapter 2

Abnormality Detection in DAPI images

This chapter covers automated analysis of DAPI-stained cell nuclei in human tissue samples used for diagnosis and screening of cancers and pre-cancerous conditions. The two parts of the processing chain that we focus on are identification of cell clusters and detection of chromosomal abnormalities from DAPI appearance characteristics alone. The latter is very much a speculative endeavour, as such a task cannot be performed reliably by human experts.

Following a description of the application domain in Section 2.1, Section 2.2 will present an enhanced method of segmentation for fluorescence images of cells, and detail the features necessary for successful segregation of single cells from cell clusters and other fragments present on the slide. For the textural characteristics that are the chief identification criterion for chromosomal abnormalities, Section 2.3 presents a novel decision-tree dictionary for patch quantisation, before overall conclusions are drawn in Section 2.4.

The data sets used throughout this chapter have been provided by Ikonisys Inc., a supplier of pathology equipment and services, and are not publicly available.

2.1 Application Domain

DAPI (4',6-diamidino-2-phenylindole) is a fluorescent stain that binds to DNA, and has been used extensively for visualisation of the cell nucleus in fluorescence microscopy. The samples used in our study have been processed to break down the cells and spread them in a thin layer over the slide. At 20 times magnification, each field of view captured by a microscope (Fig. 2.1) contains hundreds of nuclei.

The first step in processing these images is to identify the location of all the

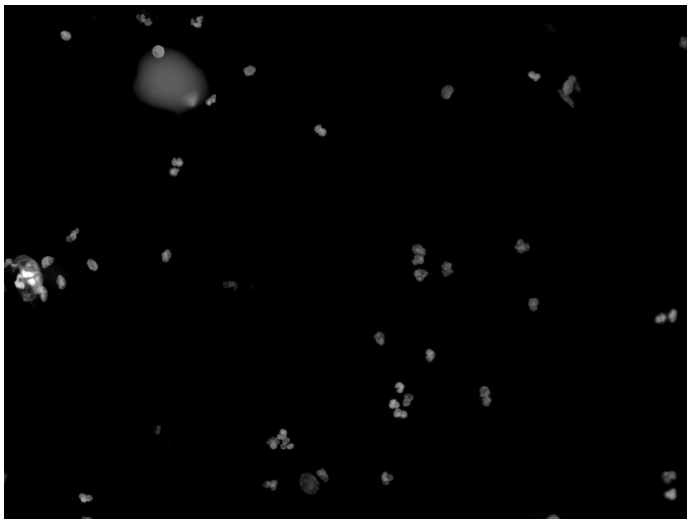


Figure 2.1: Example of 1600x1200 pixel field of DAPI-stained image captured by microscope at 20x magnification, showing an area approximately 0.25mm across

objects of interest, which is achieved through thresholding followed by connected component search. Objects below a certain area are rejected as noise or debris resulting from side-effects of the sample preparation process, while very large objects are excluded as they are likely to be very congested clumps of overlapping nuclei that are not possible to separate.

The objects that remain fall into one of three categories: single isolated nuclei, small clusters of two or three nuclei, or larger pieces of remaining debris. Classification of objects into one of these categories is the subject of Section 2.2. Single nuclei or split-up clusters can then be assessed for abnormal appearance.

Chromosomal abnormalities are known to cause certain changes in the appearance of the nucleus [85], which affect both its shape and texture. However these are not generally sufficient for establishing the abnormality status of a cell on their own. The reliable method of detecting chromosomal abnormalities is through a specific chromosome count based on FISH signals. FISH (Fluorescence in situ hybridization) is a technique for localisation of specific genetic sequences, which are marked by an attached fluorescent molecule. The FISH markers are imaged in a separate channel, which can be aligned with the DAPI images, and appear as small bright spots within the area of the nucleus. An excessive number of markers in a nucleus is a sign of abnormality, potentially related to oncogenesis and requiring further investigation. The question investigated in Section 2.3 is whether it is possible to predict abnormalities from DAPI appearance alone well enough for a first pre-screening stage, reducing the number of nuclei requiring more detailed and costly investigation with FISH.

2.2 Object type identification

In this supervised learning problem, DAPI images and segmented masks are provided in four categories: singles, doubles, triples and debris. Table 2.1 details the numbers of objects by type in the training and test partitions of the data set, and some examples of each category are shown in Fig. 2.2. The total number of labelled examples is over 14,000, and they include a large degree of variation in brightness, size and texture of nuclei, and arrangement of clusters. Unfortunately, exact details of the imaging conditions, equipment and the labelling protocol are not available.

Set	Single	Double	Triple	Debris
Training set	6676	2909	437	1503
Test set	700	685	519	675
Total	7376	3594	956	2178

Table 2.1: Numbers of objects in provided data sets

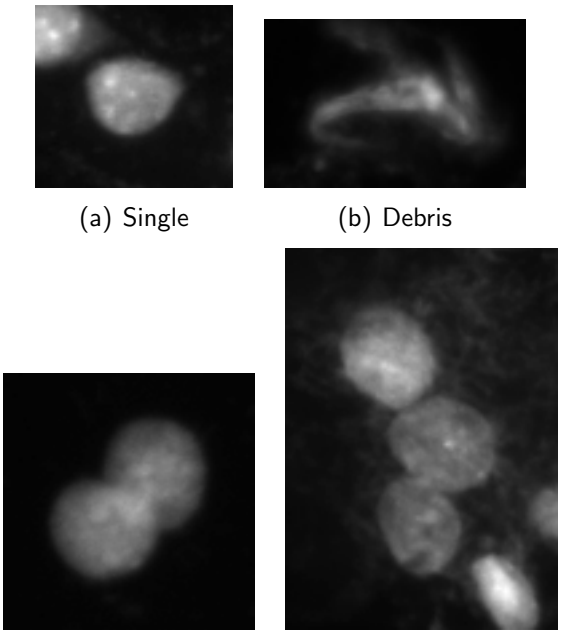
Although the data is labelled with 4 different categories, it is beyond the scope of this study to distinguish doubles from triples, although they are counted separately as it is sometimes informative to compare error rates and failure mechanisms between these two types. In light of this, for the remainder of this section the problem is defined as three-class, to sort the inputs into *singles, clusters* and *debris*.

Cluster identification relies significantly on a good segmentation to provide a reliable basis for shape assessment, and this step is discussed in Section 2.2.1. We then explore a number of attributes that are valuable for discrimination of clusters from single nuclei, as well as identification of debris objects, in Section 2.2.2. Experimental results are presented in Section 2.2.3.

2.2.1 Segmentation

Segmentation masks are supplied as part of the data set, but a significant proportion of them are poor, particularly among clusters. This will inevitably affect classification performance, as shape is the major discriminative factor.

Analysis of the masks shows that they are obtained by a simple thresholding operation, although a different threshold is chosen for each example object. In the most problematic cases the threshold used is 20 to 30 grey-levels above that which would produce a clear mask reasonably coincident with the object's visual boundary. The threshold is consistent with that computed by Otsu's method [10],



(a) Single (b) Debris

(c) Double (d) Triple

Figure 2.2: Contrast-boosted examples of each object type. Each nucleus is 30-40 pixels, or 5-7μm, in diameter.

the unsuitability of which for cellular images was explained in Section 1.3.1. We therefore devise an alternative method, still based on thresholding, but selecting the threshold based on different requirements, more aligned to the properties of DAPI images.

We search for the optimal threshold by measuring contour smoothness (sometimes referred to as circularity, defined as $S = A/P^2$ for area A and perimeter P) and the boundary gradient for each of the candidate threshold values, from the Otsu threshold down to 30 levels lower. The two measurements are combined in a weighted sum to produce a single quality metric:

$$T_{opt} = \arg\max_{t} \{G(t) + w_S \cdot \ S(t)\} \tag{2.1}$$

where G is the boundary gradient and S is the smoothness ratio at each threshold position t. The boundary gradient $G(t)$ is computed as the average of grey-level differences across the boundary, i.e. in the direction normal to the boundary at each point of its contour, with a base of 2 pixels either side of the contour position.

The weight w_S is calculated from the ratio of sample variances of the two

parameters across the image set as

$$w_S = \sqrt{\frac{\sum_i Var_t[G_i(t)]}{\sum_i Var_t[S_i(t)]}} \tag{2.2}$$

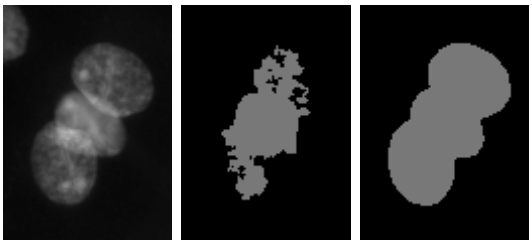
where i is the image index within the data set. To be clear, the variance is computed with respect to changes in threshold level for each image within the data set, and all such variances are averaged across all images in the data set. Such a measure provides a robust estimate of the range of each parameter in its sensitivity to threshold changes. Using the ratio w_S as a weight on $S(t)$ is equivalent to normalising each parameter by its range, as equation (2.1) can be reformulated as

$$T_{opt} = \arg\max_t \left\{ \frac{G(t)}{\sqrt{\sum_i Var_t[G_i(t)]}} + \frac{S(t)}{\sqrt{\sum_i Var_t[S_i(t)]}} \right\} \tag{2.3}$$

Thus the method gives an essentially equal weight to the two constituent measures, and provides a balance between searching for a rounded shape, but also matching the object's edge within the image.

(a) Single

(b) Double



(c) Triple

Figure 2.3: Examples of improved threshold selection, showing grey-scale source image, mask from Otsu's method, and the improved mask for each one.

The improvement in mask shape resulting from this process is consistent and visually apparent, with several examples given in Fig. 2.3, and is particularly pronounced for clusters where the constituent nuclei differ in brightness. The images

presented in Fig. 2.3 have been selected to demonstrate the most improvement over the baseline method, as they exhibit its most severe failure cases, generally caused by a high degree of brightness variation within the object.

The proposed method is subject to potential failure modes of a rather different kind: the selected threshold may break up a cluster whose nuclei only just touch, resulting in two separate single nuclei while the label indicates a cluster; no such cases occurred in this dataset as the selected threshold is generally lower than the original. Conversely, lowering the threshold may merge an object labelled as single with a nearby nearly-touching nucleus which was previously segmented separately (and is partially outside the region of interest). This eventuality has to be explicitly guarded against as part of the algorithm, detecting the threshold at which such a merge occurs from a sudden increase in the area of segmented object, and excluding the lower threshold values from the search. A few such cases are illustrated in Fig. 2.4.

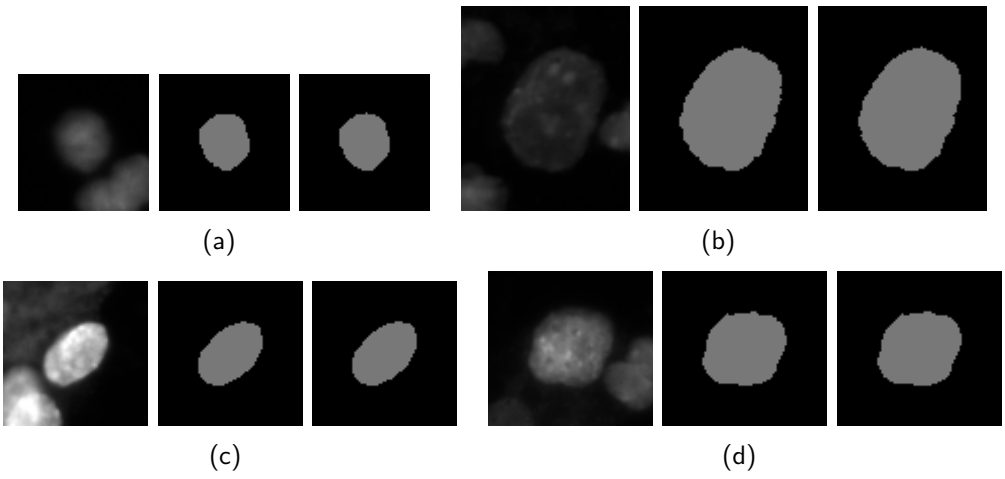


Figure 2.4: Examples of potential failure cases, showing grey-scale source image, mask from Otsu's method, and the proposed segmentation for each one, avoiding merge with nearby objects.

2.2.2 Feature Extraction

The measurements that we use to distinguish between the classes describe both the object's shape and its content.

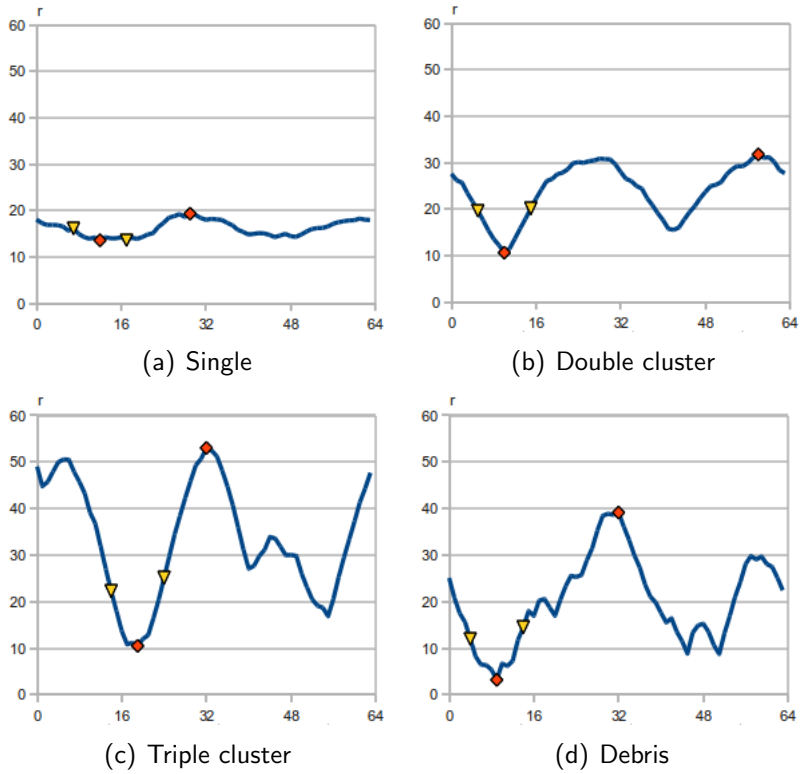
The geometric features, most of which were reviewed in Section 1.3.2, include the basic attributes of area, perimeter and circularity, which by themselves are sufficient to achieve around 90% accuracy. To help in separating singles

from clusters, we add a measure of concavity, defined as the area difference between the object and its convex hull, as proportion of the object area. Single nuclei are almost perfectly convex, and the addition of this measure improves the misclassification rate for this class.

We use the magnitudes of Fourier shape descriptor coefficients, rather than their complex components, to obtain explicit rotation invariance. We find that terms 1 through 6 carry the most useful information, with higher harmonics not bringing any improvement in classification. We choose radial profiles constructed as distance to centroid at equally spaced sections of arc length as the basis for the Fourier Descriptors, as they are robust to the highly convex configurations sometimes encountered among the cluster shapes where the centroid is outside the object.

The final morphological contributions come from direct analysis of these central distance profiles, some typical examples of which for the different classes are shown in Fig. 2.5. The characteristic feature of these profiles for most single nuclei is the small amount of variation, when compared to the other object types, and the very gentle slopes involved when the profile does vary. Most variation in single profiles comes from the elongated shape, which is smooth, whereas clusters tend to have much sharper notches and angles between the nuclei, which result in much deeper and sharper troughs in the profiles. Profiles of debris objects tend to be much noisier and generally less consistent in their shape.

Based on the observations above, two measurements are derived from the profiles: the first is a ratio between the minimum and the maximum of the profile, marked by red diamonds in Fig. 2.5, indicating the relative depth of the biggest trough within the profile. An inverse of this measurement (R_{max}/R_{min}) has been used in the literature to detect 'bulging' nuclei, but not for cluster detection [86]. The second set of measurements assesses the steepness of the sides of the lowest trough by taking gradients either side of the minimum. This is a different approach from the most works on notch (concavity) detection, which are usually based on the more complex analysis of changes in the tangent direction [21]. In this case we do not need to locate the positions of all significant concavities for cluster splitting, merely to establish whether they are present, so the simpler measurements of local gradients around the minimum are sufficient. The two gradients, from left and right, are sorted into the larger and the smaller, as no significance can be attached to the orientation. Both are normalised by the DC term of the Fourier transform, representing mean radius, to provide size invariance. To reduce the effect of noise, these are taken as differences from the minimum to values several points away from the minimum position, marked by yellow triangles in the Figure. It has been established experimentally that 5 points (out of the total of 64 used for Fourier analysis) is more robust to noise in the segmentation boundary and variation in cluster configurations than



(a) Single (b) Double cluster

(c) Triple cluster (d) Debris

Figure 2.5: Typical examples of central distance profiles for different object types, showing min and max positions (diamonds) and sides of lowest trough (triangles).

gradients taken at ± 1 or ± 3 points from the minimum, and provide a better overall classification performance. Together these features provide a very strong contribution to distinguishing single nuclei from other object types.

Explicit attempts to count the number of large indentations within the shape, either as large lumps in convex hull difference, or local minima in the central distance profile, have not proved fruitful. Neither did a measure which assessed the difference of the shape from a perfect ellipse with the same major and minor axes as the object under consideration.

As well as object shape, which provides the bulk of relevant cues for differentiation of single nuclei from clusters, as well as from debris, information derived from the object content is also beneficial. The first content feature is based on the observation that well-formed nuclei have a strong edge, while debris is smeary and blurred. This suggests that a measurement of image *gradient* in the direction normal to the boundary may allow better differentiation of debris from the other classes. The gradients are integrated around the boundary, and nor-

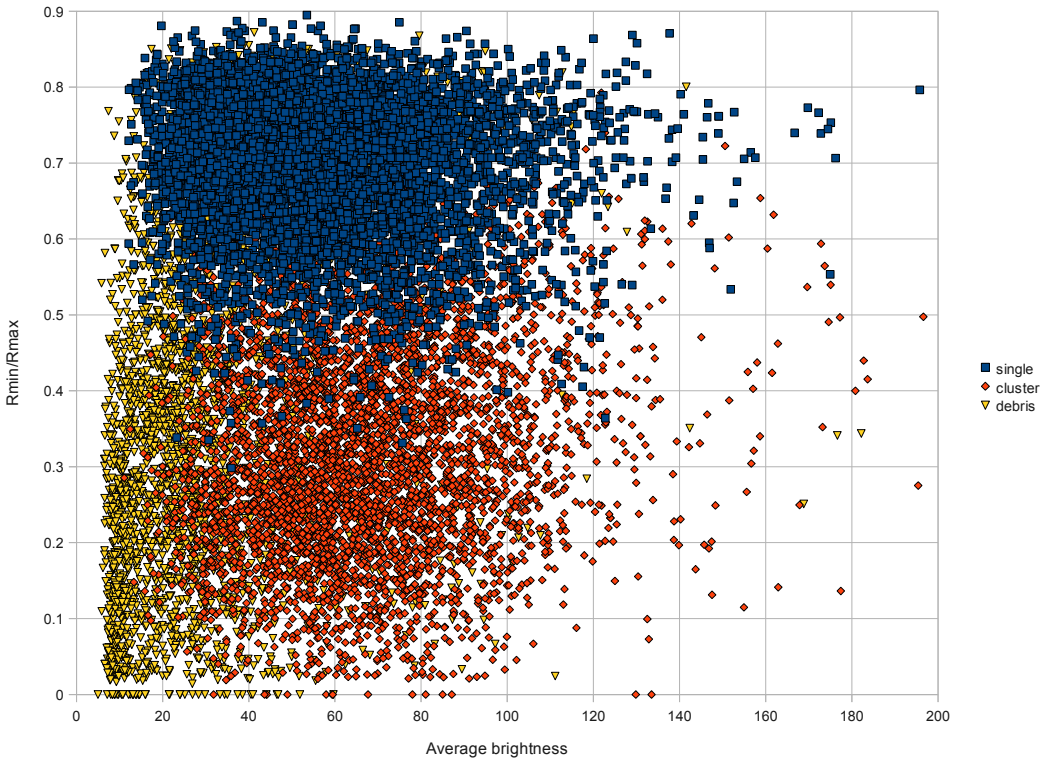


Figure 2.6: Average brightness provides separation between cluster and debris classes that are not well distinguished by morphology features.

malised by its length, as well as by the mean luminance of the object, to make the measurement independent of overall brightness of the object. Gradients taken at ± 1 pixel around the contour were found to be too noisy, while those ± 2 pixels either side of the contour were much more useful.

The *average brightness* within the object, which is used to normalise the gradients above, proves a surprisingly useful feature in its own right, largely due to its interactions with the morphological features, providing separation where the other features overlap. The scatter plot in Fig. 2.6 illustrates one such interaction, with the R_{max}/R_{min} feature, demonstrating additional separation of the cluster and debris classes which have overlap in the space of morphological parameters.

Another easily computed measurement is *standard deviation of luminance* within the object boundary. The variance is generally increased for singles and clusters by the presence of brighter and darker spots, and is even higher for clusters due to variation in brightness between the nuclei which comprise the cluster, and extra brightness in areas of overlap. Luminance variance in debris

tends to be lower, as it is largely amorphous and uniform.

A more specific assessment of this difference in texture between actual nuclei and smeary debris is supplied by a *spot filter*. This is a circular filter constructed as a difference of Gaussians of different widths ($\sigma_1 = 0.96, \sigma_2 = 1.55$), with total aperture of 7 pixels in each direction. The parameters are estimated from size of commonly encountered spot textures in the nuclei, and are used to compute the filter coefficients according to equation 2.4, where indices i and j have their origin at the centre of the filter. The filter is not separable.

$$DoG_{i,j} = exp\left\{-\frac{i^2+j^2}{(2*\sigma_1)^2}\right\} - exp\left\{-\frac{i^2+j^2}{(2*\sigma_2)^2}\right\} \tag{2.4}$$

The result of convolution with this filter is normalised by the average brightness of the object to ensure illumination invariance. This is followed by squaring the signal, to produce an energy measurement and to pick up darker as well as brighter spots. The energy signal is masked to exclude areas near the edge of the object's mask in order to avoid the filter's strong response to edges biasing the internal texture measurement. Finally we apply a heuristic threshold to reduce contributions from low-level noise (pixels with a low signal value are set to zero). The total of filter responses collected from the masked object area is used a feature for classification.

2.2.3 Results

We evaluate the performance of the proposed feature set with 10-fold cross-validation, using SVM with RBF kernel. The feature set consists of the following parameters, described in detail in the previous section:

- Area
- Perimeter
- Circularity
- Concavity (based on convex hull)
- Six Fourier Descriptor coefficients
- Distance profile R_{min}/R_{max}
- Sharpness of distance profile trough (ordered as smaller and larger)
- Average brightness
- Luminance standard deviation
- Luminance gradient across the boundary
- Spot-filter total energy

We compare this feature set with a base-line method based on complex Fourier Descriptors alone, comprising 10 terms each of real and imaginary parts

of the normalised Fourier coefficients, a total of 20 features. This well-established shape signature yields an overall error rate of 7.9% when applied to our image set. In contrast, the overall error rate of the proposed method is only 1.95%, nearly 4 times less than the base-line. The detailed confusion matrix of the proposed method is given in Table 2.2.

Classified as→ Actual class ↓	Single	Cluster	Debris
Single	7313 (99.13%)	52 (0.70%)	11 (0.15%)
Cluster	77 (1.69%)	4432 (97.43%)	41 (0.90%)
Debris	45 (2.07%)	52 (2.39%)	2083 (95.55%)

Table 2.2: Confusion matrix from 10-fold cross-validation using 17 features

As illustration of the efficacy of the distance profile features, when used on their own (three features: R_{min}/R_{max} and two trough gradients), these are able to distinguish single nuclei with an accuracy of 96.1%.

Classification of images segmented by the baseline Otsu's threshold, but applying the proposed set of features, yields an increased error rate of 2.22% ($\pm 0.37\%$).

2.2.4 Discussion

While the segmentation improvements described in Section 2.2.1 have a relatively small impact on classification performance (around 0.2%), the general approach is potentially useful in other segmentation applications. Although the computational cost of assessing each threshold is relatively high, the restriction of search space to one dimension allows a favourable overall cost comparison to two-dimensional segmentation methods which optimise some measure of a boundary's desirability, such as snakes [16]. The segmentation method is particularly suitable for this application because it disregards most of the textured object content, but concentrates on the resulting shape, which is known *a priori* to be smooth, and aligns it with highest edge contrast.

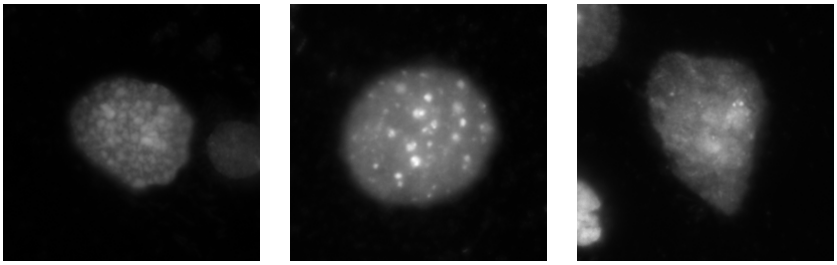
Among the features that have been evaluated for cluster detection, direct measurements on the central distance profile are notable for their novelty and efficacy. While Fourier analysis of these profiles is widely used for general shape matching, these measures are more tailored to the specific task of detecting notches between overlapping or touching nuclei within a cluster. Conversely, the rather general measurements such as luminance mean and standard deviation provide a surprisingly large amount of information to the classifier. Overall, it

is the diversity of the features that provides a strong basis for the classification performance.

2.3 Abnormality detection

As outlined in Section 2.1, DAPI images alone cannot provide a conclusive diagnosis of nuclear abnormality, but may supply enough clues to make an automated preliminary priority judgement. Therefore we do not expect to obtain very high accuracy rates, but are exploring the possibilities that computer vision techniques may offer as part of the bigger system.

The changes in nuclear appearance caused by abnormalities can be divided into differences of *texture* and of *shape*. Characteristic textural changes, induced by oncogene activation which alters the protein composition of the nuclear matrix, include very bright spots or a 'spongy' surface, as illustrated in Fig. 2.7, and the shape of abnormal nuclei can significantly deviate from the usual elliptical form [85]. It is important to note however, that there is a significant proportion of nuclei which are known to be abnormal (from FISH counts) but do not exhibit any of these signs in their DAPI images. Conversely, there is considerable natural variation of both texture and shape in normal healthy cells, some of which is similar to the signs of abnormality, for example a degree of brighter spots within the texture.



(a) Spongy texture (b) Bright spots (c) Irregular shape

Figure 2.7: Examples of abnormal nuclei showing various typical changes

The following sections describe an evaluation of a number of texture classification methods applied to this challenging problem, analyse the results and discuss potential ways to improve them further.

2.3.1 Methods

For developing abnormality detection we use a separate data set, consisting of 729 examples of (single) normal nuclei and 836 abnormal ones, all at 20 times magnification. Clearly this class ratio does not represent the general incidence of abnormality, but rather an approximately balanced set which provides a rich variety of examples of abnormality. Masks identifying the object boundary within the image are provided, but are drawn by hand and are very crude (polygonal), so can't be used for automatic shape measurements. Only pixels within the masked object area contribute to the texture measurements. The images vary in size from 20 to around 70 pixels across, so the number of pixels that can be used for texture assessment within each image is between a few hundred and several thousand. Information on the precise preparation methods and imaging conditions is, unfortunately, not available.

From the review of texture comparison methods given in Section 1.3.3, we choose two candidate methods for evaluation on our dataset: co-occurrence matrices and patch statistics. Simplicity and long-established use of GLCM (grey-level co-occurrence matrix) measures gives a base-line for the comparisons. The patch method provides a very general representation of texture, and as a relatively new development supplies an opportunity to advance performance of nuclear abnormality detection. We also test a specialist method designed specifically to assess texture of nuclear chromatin structures, called *contour complexity* [87]. It analyses changes in the length of outline obtained from different threshold levels, and claims to be much more sensitive than fractal dimension for detection of nuclear abnormality.

Six co-occurrence matrices were calculated for each image, with each matrix containing co-occurrence totals from four directions, spaced at 90°, as the textures of interest are isotropic. The offsets used are ± 1, ± 2 and ± 3 pixels, separately for axially aligned and diagonal (45°) orientations, as the diagonals are longer by a factor of $\sqrt{2}$. Contrast, mean and correlation statistics are computed for each matrix, resulting in a feature set of 18 attributes: 3 statistics for 3 pixel offsets at 2 orientations (axial and diagonal).

For the patch representation, we compare two alternative constructions of the dictionary: the widely used k-means clustering, as implemented by the Linde–Buzo–Gray (LBG) algorithm [88], and a novel decision-tree partitioning of feature space based on degree of overlap between classes, explicitly dividing the feature space into areas of different discriminating ability. To learn the decision tree, each dimension of the feature vector is assessed to determine the value of splitting feature space with a threshold in that dimension. The test threshold is placed half-way between the class means in the relevant dimension, calculated for the subset of points belonging to each leaf of the currently constructed de-

cision tree. The assessment compares proportions of points of each class on either side of the proposed boundary, and whether they differ significantly from the parent leaf's population. Significance is judged by the one-proportion z-test, defined as

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{2.5}$$

where p_0 is the proportion of abnormal patches in the parent population, and $\hat{p}$ is the proportion of abnormal patches in the putative subset, which has a total of n points. The denominator represents an estimate of the sample standard deviation based on the null hypothesis that the same binomial distribution with parent proportions applies on both sides of the proposed split. The z-test estimates the number of standard deviations by which the actual proportion deviates from the null hypothesis, and we reject the null hypothesis if $|z| > 2$ on one or both sides, corresponding to one-sided p-value of around 5%. If the proposed split is not found to be significant, no decision branches are added to the tree at this point, and we proceed to consider other dimensions. The process is repeated for all attribute dimensions until no further significant splits can be added. The z-statistic assumes that the sample is large enough for the binomial distribution to approximate the Gaussian distribution, placing a lower limit on the number of samples that are needed to make a valid assessment. At least 10 points of each class have to be present on each side of the boundary to satisfy the validity conditions, providing a natural stopping criterion to the growth of the decision tree. Each leaf of the final dictionary tree corresponds to a codeword. All patches wholly contained within the object boundary are converted to codewords, and the normalised histogram of codeword occurrence is used as attribute vector for classification of the image.

Finally, we consider the option of combining results from multiple classifiers to improve overall accuracy. We use two classifiers with quite different feature sets to support each other in obtaining a better decision: the confidence or predicted probability output of the first classifier guides selective application of the second. For samples whose GLCM-based confidence is above a certain threshold T_c, the predicted class is used directly, without further assessment by the patch classifier. For samples with lower confidence, patch classifier is used instead; it is important that this second-stage classifier is trained exclusively on examples which have low confidence in GLCM decision. The optimal value of T_c=0.85 is established experimentally by cross-validation on the training set. This cascade construction can be seen as a very simple form of boosting, which is nonetheless adaptive to the variable accuracy of the two constituent classifiers in different regions of feature space, as well as focusing the downstream classifier's learning on the regions where the upstream one performs less well.

Method	Error rate (%)	ROC Area
GLCM	18.1 (σ=2.3%)	0.913
Contour complexity	27.3 (σ=3.6%)	0.811
LBG Dictionary	18.6 (σ=3.0%)	0.905
Discriminative Tree Dictionary	16.8 (σ=3.4%)	0.914
Cascade	14.6	(n/a)

Table 2.3: Results summary of all texture classification methods

2.3.2 Experimental Results

All experiments use an SVM with RBF (radial basis function) kernel for classification, and we report error rates with their standard deviations based on 10-fold cross-validation, summarised in Table 2.3.

The contour complexity measure proved rather disappointing: the lowest error rate, obtained in combination with average luminance and luminance standard deviation, is 27.3% (σ=3.6%).

GLCM was able to extract rather more information: the full feature set of 18 attributes can predict abnormality with error of 19.9% (σ=2.2%). However, evaluation of attribute information gain highlights the relatively low value of contrast features, and removing them from the feature set can actually drop errors to 18.1% (σ=2.3%).

Quantisation using a discriminative decision tree shows promise, improving the texture classification error rate from 18.6% (σ=3.0%) for LBG dictionary of 32 clusters, to 16.8% (σ=3.4%) for a tree of 1483 leaf nodes, using same patch features.

Receiver Operating Curves presented in Fig. 2.8 confirm the relative parity of GLCM and the two dictionary-based methods in this case, while the contour complexity measure is clearly inferior. This is also supported by area under the curve figures included in the last column of Table 2.3.

As success of our ensemble arrangement depends on the two individual classifiers making mistakes in different parts of input space, we assess the degree of correlation between the error instances of the two methods. We find that of 280 mistakes made by the GLCM-based classifier and the 257 mistakes of the patch-based classifier, there are 113 shared errors, giving a lower bound of 7.2% for error rate of the combined system. This would only be achievable if in each case of diverging classifier opinions we could perfectly predict which one is actually correct. The achieved combined error rate is 14.6%, significantly improving on both the constituent methods (18.1% for GLCM and 16.8% for patches).

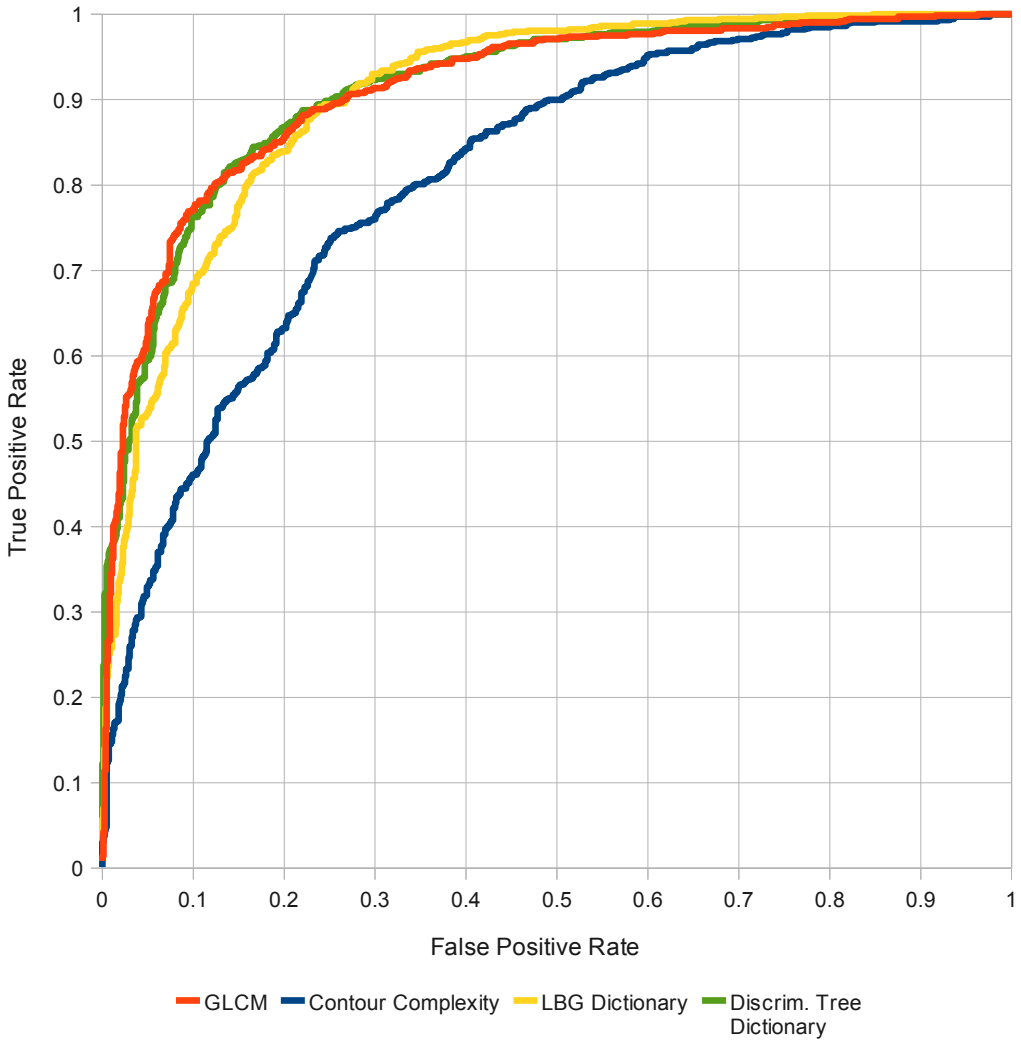


Figure 2.8: ROCs for the various abnormality classification methods

2.3.3 Discussion

It is very difficult to pick the most suitable method from the plethora of texture classification algorithms on offer. Trial and error is the prevalent selection mechanism, but the computational and development costs of trying all the potential combinations is prohibitive. While accuracy is a very important characteristic of performance, computational efficiency, robustness and transparency are also relevant. Specific aspects of the application may also influence the choice of method, for example the degree of similarity and overlap between the texture classes, which is very high in our case. We are additionally hampered by the relatively low magnification (20x) of the available images, which may not capture

sufficient detail for accurate recognition; and the crude manual segmentation, which precludes the use of shape attributes even though these are known to be related to malignancy. Our exploration is therefore quite limited, but does demonstrate potential for automation in this application.

Contour complexity has been claimed to be much more sensitive than fractal dimension for detection of nuclear abnormality [87]. However, this claim is based on a Kolmogorov-Smirnov test of distribution differences, not actual classification performance. While the distributions of contour complexity values are indeed different between benign and malign cells, a large proportion of both normal and abnormal nuclei have very low values of contour complexity, which therefore provides no discriminating information in these cases. So only a minority of malign cells, those with a high contour complexity, can be differentiated from the benign in a classification context, severely limiting the utility of this measure in a practical application. An additional limitation of this measure is its low dimensionality, fundamentally reducing its classification potential. This limitation is confirmed by experimental data on our own image set.

Co-occurrence matrices prove their enduring worth as fundamental representations of the statistical correlations underlying the concept of texture. It is particularly important to include multiple offsets to cover a broader range of texture scales, and to correctly handle the need for rotational invariance by combining contributions from different rotations of the image into a single matrix, where they can robustly reinforce each other, rather than compute separate statistics for each orientation to produce a larger, but noisier, feature vector.

The texton distribution comparison is not able to improve on GLCM results, although applying the χ^2 kernel for histogram comparison in the classifier could be potentially beneficial. The accuracy of the representation is limited by hard assignment of a patch to a single dictionary entry, especially as the dictionary size is constrained by computational cost of the k-means algorithm. Our novel discriminative dictionary constructs a much larger representation at a negligible computational cost, and is able to better separate the classes in this bigger feature space. It also requires no distance metric between patches, so there are no issues of finding an appropriate one. As the tree's principal objective is finding areas of relatively high discriminative ability in feature space, it is robust to issues of overall class imbalance, unlike dictionaries based on clustering which represent samples of high-density dominant class much more accurately than those of the minority class. The tree could be further improved at a small computational cost by taking account of the measured variances of each class within the parent population when calculating the proposed boundary position, rather than the current implicit assumption of equal variance.

The combined cascade classifier is able to implement a more complex decision boundary than either of the constituent methods. Although our assessment of

divergence between the error instances of the two methods is encouraging, we do not achieve the best possible ensemble accuracy as confidence of the first classifier is an imperfect predictor of error cases. The dependency between the two stages also creates difficulty in optimisation of the decision thresholds of the two classifiers.

Some preliminary examinations of and experiments with an augmented dataset containing additional focus planes either side of the images included in our abnormality data suggest that a significant proportion of the images we have used are not optimally focused, which would have strongly affected their textural properties and measurements. Although in some cases the shift of focus one step closer or further brings a distinct sharpening of the whole image, other cells exhibit differentiated blurring at opposite sides of the cell, caused by the cell's skewed orientation in relation to the plane of the slide. In these situations, a simple choice of optimal focus plane based on some global measure of sharpness would not be sufficient, and a more sophisticated adaptive focus method would have to be employed in order to provide the most accurate texture measurements.

The differences in accuracy between several of the examined methods are small, and do not provide a full picture of their relative merits. The true utility of each method within a larger system would depend on actual incidence of abnormality, which is not available for our study, and the desired balance between type I and type II errors when adjusted for that incidence. The precise distribution of errors will also interact to produce different effects on classification of a whole sample based on the cells it contains, a subject which is explored in much greater depth in Chapter 3.

2.4 Conclusions

In Section 2.2 we demonstrate a system able to distinguish single nuclei from nuclear clusters and also from fragmentary debris objects, which can be optimised to generate errors in less than 2.0% of cases. This is a very promising result in an application which could lead to major advances in accuracy and availability of early detection of cancers and pre-cancerous conditions, and it derives its strength from the richness of the feature vector used to describe the object. Development of specific measurements that target the characteristics of interest, such as the sharpness of notches in the radial profile, boosts the robustness as well as transparency of the overall design.

The abnormality detection task described in Section 2.3 is a difficult instance of texture recognition as the two classes are so similar and even ambiguous. We have tried to adapt the methods to squeeze the most discriminating information out of the available data. The discriminating dictionary tree developed for this

purpose can be applied to many other areas of computer vision which employ a bag-of-words approach. It is considerably faster than normal clustering methods in both training and assignment, and does not require an explicit distance metric for feature vectors. It is also much faster to train and has a more flexible structure than similar tree-structured descriptors that aim to directly optimise the classification performance [89], and is independent of the specific type of downstream classifier. Similarly, the combination of classifiers based on the confidence output of the first stage is potentially applicable to many other scenarios.

Chapter 3

HEp-2 pattern classification

This chapter is concerned with automated analysis of indirect immunofluorescence images (IIF) introduced in Section 1.1.2. As previously outlined, the particular type of images, which visualise the anti-nuclear auto-antibody (ANA) reaction using the HEp-2 cell line, are routinely used by pathology laboratories as the most reliable basis for diagnosis of auto-immune diseases. The diagnosis is usually performed by laboratory staff directly at the microscope, determining the specific type of auto-immune disease from the visual pattern of the fluorescence. A large number of these patterns is described in the medical research literature, although only a smaller subset of the more common ones is routinely differentiated by clinical laboratories.

Recent years have seen increasing interest in automation of parts of this diagnostic work-flow, both to reduce the pressure on overstretched pathology specialists and to provide a more objective and repeatable mechanism of image interpretation. The precise nomenclature of staining patterns continues to be a matter of debate within the medical community [90], with additional causes of variability in results stemming from differences in laboratory processes and natural variability of the reagents involved [91, 92], as well as the inter-observer variability in interpretation of the visual patterns, which is estimated to have only 76% agreement [93]. The photo-bleaching effect offers an additional difficulty for manual pattern interpretation, as the rapid fading of the fluorescence limits the time window for accurate analysis.

We review the latest research in this application area in Section 3.1, and analyse some of its short-comings in the context of the larger diagnostic system. We describe some of our own experiments in Section 3.2, and relate their results to the underlying search for connections between image characteristics and the best ways of measuring them as attributes suitable for classification. In particular, we suggest several measurements suitable for assessment of isotropic texture, which is frequently seen in biological objects, that are preferable to simple application

of full 2D analysis as they concentrate the relevant information in fewer features with less noise. A major theme of this chapter is the difference between cell and sample classification, so in Section 3.3 we depart from the late fusion approach and analyse each sample in its entirety, either by modelling the distribution of cell parameters within the sample, or by pooling contributions from all image regions that represent internal cell content into a single histogram of textural properties. The implications of both sets of experiments are analysed in Section 3.4, together with some promising suggestions for further improvements in this application area.

3.1 State-of-the-art review

The subject of automatic analysis of HEp-2 fluorescence patterns has been studied for around 15 years [94], and has gradually developed to cover automated detection of cells within the captured slide images, their segmentation [95], classification of the overall staining intensity as *positive* or *intermediate* [96–98], and finally identification of specific staining patterns associated with particular diseases [96,97,99–102]. A very recent comparative study of commercially available systems reported pattern recognition accuracies between 52% and 79% [103], although the manufacturer's own findings are often much more favourable. Meta-analysis of the various studies is highly challenging due to the variety of chosen class definitions, as well as the differing quality and quantity of images used in each one. Some studies only consider positive cells as part of the dataset, which allows a higher recognition accuracy than inclusion of the fainter, and therefore more difficult, intermediate intensity cells.

The most recent flurry of activity in the field has been prompted by introduction of public data sets, associated with contests or challenges, and aimed at attracting new researchers to the area. The HEp-2 Cells Classification contest at ICPR 2012 recognised the difficulty in comparing earlier works, and provided a single dataset for "the comparison of systems able to automatically recognize the pattern of cells within IIF images [..] on a large and significant set of real data" [104]. Unfortunately, the design of the dataset labelling was flawed, and its inaccurate description led to massive discrepancies between cross-validation and test performance, rendering the comparison ineffective.

The follow-up competition at ICIP 2013 reflected the statistical structure of the problem much better, and contained a much larger volume of cells, although it changed the definition of classes from that used in the earlier dataset, making cross-comparisons very difficult [105]. So far, the only work using this dataset that has reached publication performs a very broad-base comparison of feature and classifier types, and concludes that texture is the most relevant discrimi-

nant, and (of those tested) Laws' masks combined with RBF-SVM produce the best performance [106]. This dataset, as all the previous ones, contains images produced by a single laboratory using a single set of equipment, which limits its ability to fully test the generalisation capabilities of any proposed methods. The images contained in the two datasets described above exhibit significant differences of colour, resolution and focus quality, so it would not be possible to reliably predict the performance of a method on one dataset based on its performance on the other.

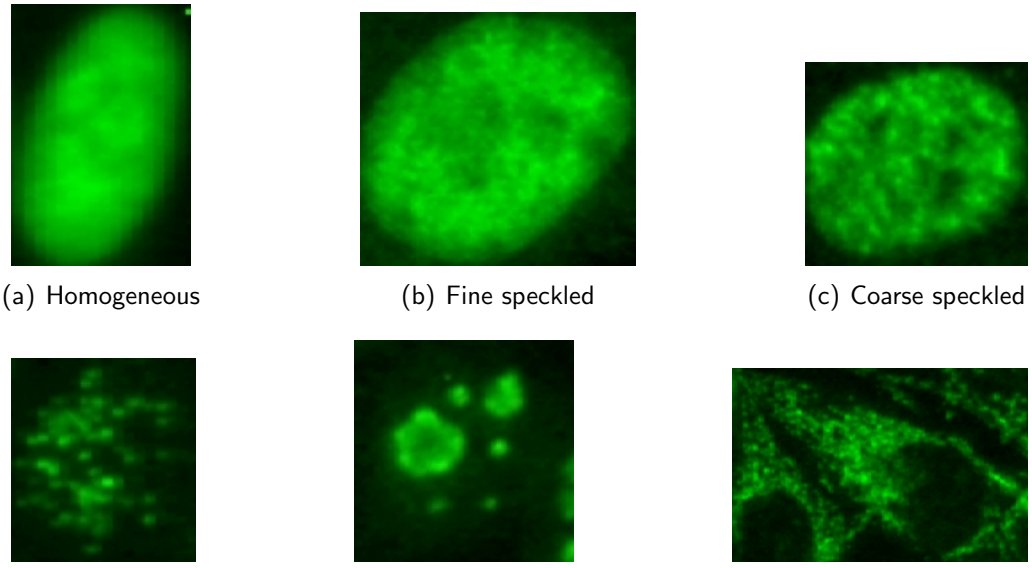
In addition to the two challenge datasets, a medium-size public dataset has been released recently by the same laboratory that prepared the ICIP 2013 contest data, Sullivan Nicolaides Pathology in Australia, under the name SNPHEp-2 [107]. The dataset, and a few of the corresponding published works, are described in detail in Section 3.1.3, following an in-depth review of the more numerous studies based on the earlier MIVIA dataset in Section 3.1.1. All are based on some combination of the computer vision and machine learning techniques described in Section 1.3.

3.1.1 MIVIA Data Set

The data consists of 1457 IIF images of individual cells, each having an associated binary mask (removing issues of segmentation from any comparison), an intensity label (*positive* or *intermediate*), and a ground-truth class label from one of 6 classes. The classes are as follows:

- **Homogeneous**: a diffuse pattern, fairly uniform across the whole nucleus.
- **Fine speckled**: a very fine-grained isotropic texture, not dissimilar to white noise.
- **Coarse speckled**: an isotropic texture of somewhat larger specks.
- **Centromere**: this class is characterised by large numbers of strong bright spots on a darker background. These are 2-3 pixels across, and 40-60 are supposed to be present, although in a number of intermediate intensity examples of this class none are visible to the eye, even after contrast normalisation.
- **Nucleolar**: a small number (less than 6) of larger bright areas within the nucleus.
- **Cytoplasmatic**: these nuclei are characterised by a strongly irregular shape, as compared to the generally elliptic nature of all other classes. The texture is equally irregular.

Examples of each class are given in Fig. 3.1, contrast boosted to make their detail more visible. Typical contrast range for positive examples is around 120 grey-levels, but can be as low as 25 levels for cells in intermediate samples, greatly



(a) Homogeneous (b) Fine speckled (c) Coarse speckled

(d) Centromere (e) Nucleolar (f) Cytoplasmatic

Figure 3.1: Positive examples of each class

exacerbating the effect of sensor noise. Image sizes range from 45 to 130 pixels across.

The images suffer from a number of artefacts: sensor impulse noise affecting groups of between 4 and 8 pixels across appears in a number of locations; 4-pixel wide vertical banding is visible in areas of high gradients, and probably originates from a crude up-sampling algorithm in the scanning device. Finally, a variation in focus precision, which can affect textural measurements, is present within the image set.

3.1.2 ICPR 2012 Contest

The HEp-2 Cells Classification contest held at ICPR 2012 attracted the participation of 28 groups from across the world, evaluating a rich variety of algorithms. Sadly, due to a flaw in the experimental design of the contest, its results, now published in [108], shed little light on the actual usefulness or otherwise of any of the methods. The contest description misrepresented the labelling procedure, suggesting that each cell was independently labelled. The submitted algorithms, designed on the basis of this statistical independence of the supplied data points, performed very well in cross-validation on the training set (error rates of around 5% were reported by a number of participants) but failed across the board when exposed to the test set, which consisted of genuinely independent samples from

other patients. Rather than over 700 training samples across 6 classes, the training set only contained 14 independently labelled whole-well images for each class, which had then been split up into cells with inherited, propagated labels.

Despite the contest's failure to give definitive answers about the relative merits of different feature sets for this application, it is nonetheless informative to consider the range of features employed by the participants. The summary report for participants includes method descriptions of all submissions, and a number of more detailed papers were included in ICPR proceedings [109–116]. Most algorithms included some form of texture measurement, and often a combination of several; co-occurrence matrices, Local Binary Patterns, and various extensions thereof, as well as gradients, frequency transforms and Gabor wavelets were represented. Morphological and granulometry features form another strand, recognised by several researchers as relevant to these image patterns. Convolutional, dictionary learning, randomized and evolutionary feature extraction methods were also evaluated.

Following identification of the cause of the gap in performance between training and test conditions, additional labelling information was released, which relates each cell image to its parent sample image. This allows cross-validation procedures to take account of the dependencies between cells originating from the same sample, and structure cross-validation folds in the same manner as the training-test split, i.e. always evaluating performance on cells that have come from a previously unseen patient, not just previously unseen cells. A special issue of the Pattern Recognition journal, containing follow-up studies based on this augmented dataset [117], is currently still 'in press'. The protocols prescribed for the special issue articles allow comparison of both cell-level performance and the accuracy of sample-level predictions, assuming an independent prediction for each cell and a vote for most frequently selected class. Given that human experts examine the entire sample before making their decision, we find this approach too limiting, and discuss potential improvements in Section 3.2.

Additionally, it is worth bearing in mind that some classes are not fully characterised by the appearance of their inter-phase nuclei as described earlier, but can only be distinguished if mitotic (dividing) cells are present within the sample and exhibit the characteristic features of a particular class. The appearance of these nuclei is very different from the rest of the sample, and they had been explicitly manually excluded from the original contest dataset. Only a few mitotic cells are present in any one sample (and sometimes none at all), making a proper statistical treatment of the connection between mitotic and inter-phase appearance very difficult.

Several of the special issue papers are of particular interest: the best overall result at the sample level is reported in [118], with accuracy of over 95%, which is considerably higher than any other reported results. The method is based on

subclass discriminant analysis (SDA, [119]), and combines morphological features at 7 different thresholds with both global and local textural statistics. It remains a matter of concern, however, that the number of sub-classes is in this case very close to the number of samples of each class (which are very few in this dataset), and the sub-classes may be learning to model each sample rather than sub-classes as such. Separate feature selection and classifiers for positive and intermediate samples are potentially beneficial due to their different characteristics, but discovery of sub-classes that is inherent in SDA is perhaps less justified.

Winners of the original ICPR 2012 contest, Nosaka et al. improve on their own result by introducing explicit rotational invariance of the co-occurring binary patterns, but also find it necessary to include rotated versions of training images in their training set, which is counter-intuitive [120]. A promising direction of optimising a dictionary for more discriminative representation gives disappointing results in [121], reaching only 71.4% sample level accuracy with leave-one-out protocol, perhaps due to the inherent limitations of using reconstruction error as the basis of classification, and its consequent sensitivity to noise, which is high in this dataset. Finally, Theodorakopoulos et al. [122] propose a dissimilarity representation, based on a combination of local gradients and a rotationally-invariant version of CoALBP, and although their consideration is limited to cell, rather than sample, similarity, they do remark on the lack of coherent block structure within their dissimilarity matrix, something which we explore further in Section 3.2.

3.1.3 SNP HEp-2 Data Set

The images from 40 patients contain 1884 cells in 5 different classes [123], which match those in the MIVIA dataset, except for the omission of *cytoplasmatic*, which is a reasonable exclusion given that this class is easily distinguishable by shape, and only dilutes evaluation of texture recognition performance when mixed with the others. This dataset exhibits a much greater variation in focus precision than in the earlier images, and wide variation of exposure, including strong oversaturation in some cases, but has lower levels of white noise due to better cooling of the camera. No mitotic nuclei are differentiated in the labelling of this dataset.

So far, relatively few works evaluating methods on this dataset have reached publication; of these, Faraki et al. is notable for the explicit consideration of how robustly a method maintains its performance when applied to a different data set than it was trained on [124]. The theme of regional codebooks, giving separate consideration to cell edges and their interiors, first suggested in the original paper describing the dataset [123], continues in [125]. The improvement obtained by additional application of multiple kernel learning is variable, depending on

which dataset it is evaluated on. An approach based on so called Spontaneous Activity Patterns (SAP), evaluated in [126], shows better cell-level accuracy than some of the other proposals, but, as we demonstrate through our experiments in the following section, cell-level accuracy is not a reliable guide to sample-level performance, which is the ultimate goal of this application.

3.2 Cell Experiments

It is evident from the review in the previous section that the majority of current work concentrates on improving accuracy of recognition for individual cells, with an implicit assumption that this will automatically translate into improved sample decisions downstream. In this section we demonstrate through comparative experiments (published in [127]) that this is not necessarily the case, and that therefore a richer statistical model of connections between cell appearance and sample class is needed. We propose the use of a distance metric for *sets* of cells, which can take into account the full set of measurements from a patient's sample, instead of narrowing the cell information down to a hard class decision before allowing it to be combined with information from other cells within the sample. We show that this approach has a stronger connection to the ultimate goal of performing a clinical diagnosis, and provides the researcher with a richer insight into the causes of confusion.

We describe the different protocols that are used to compare features in Section 3.2.1, and detail the evaluated descriptors in Section 3.2.2. Experimental results for each combination of feature set and evaluation protocol are summarised in Section 3.2.3.

3.2.1 Evaluation protocols

Two experimental protocols are compared: the original ICPR contest protocol and a sample-based cross-validation procedure. For each protocol, we report both the accuracy of prediction for individual cells, and the sample-level predictions made by highest vote share. Each protocol is applied to a number of different descriptors in order to evaluate the correlation between cell- and sample-based performance.

Contest protocol

The original contest data set was split into training and withheld test portions, with separate patient samples used for each portion. The training set contained 2-3 labelled samples from each class, but cells from all the samples were mixed

together, with no information on which cell came from which sample. Such a data set only allows for cell-based cross-validation within the training portion, resulting in folds which contain cells from the same sample in both training and validation sections. Under these conditions, methods that are sensitive to a sample's specific imaging characteristics, such as focus or contrast, rather than broader class characteristics, can provide a very accurate prediction for the validation set, but may not generalise well to the held-out test. For this protocol, we report both cell and sample-level accuracy for the test set, and compare the former to the average accuracy obtained by cross-validation within the training set in order to assess the generalisation performance.

Sample-based cross-validation

The second protocol addresses the problems of the contest protocol by using additional information about the source sample for each cell. It is a leave-one-out procedure, where all cells from a single sample are held out as validation set for each fold. This gives a much fairer assessment of the expected real-life performance of a classification method, and has the additional benefit of a much larger training set, with 4-6 independent samples of each class available for training in each fold. Prediction of class label is still made independently for each cell, without making use of any information from other cells within the same sample.

3.2.2 Feature sets

As the class descriptions in Section 3.1.1 make clear, most of the distinctions between HEp-2 patterns are based on textures. With this in mind, we compare a number of different approaches to texture measurement against each other. As the cytoplasmatic class is also characterised by shape, we include circularity of the mask (calculated as area divided by square of the perimeter) as shape descriptor in every feature set. All feature vectors also include the basic measurements of pixel value average within the cell mask, and their standard deviation normalised by min-max contrast range of the entire cell image. We also note that all the textures are completely isotropic, allowing simplified formulations compared to the general case. As the fluorescence is monochromatic, we further simplify texture assessment by only using the dominant green component of the images.

DCT based descriptor

We note from the class descriptions that their distinctions are often ones of *scale*, rather than a specific textural pattern. This is most apparent in fine vs.

coarse speckled cases, but also continues to larger spots in centromere, and even larger bright areas in nucleolar. We therefore use the power spectrum to capture the scale at which textural variation is strongest, as described in greater detail in [113].

The frequency analysis is performed as a 32-point DCT of line sections from inside the segmented mask boundaries. As the texture is isotropic, a 1-dimensional transform is sufficient to establish its frequency distribution. Transforms from all the qualifying lines within a cell image are averaged to reduce variability and noise, and intensity normalised by min-to-max range of the image. The higher frequencies of the transform are dominated by noise, so it is found beneficial to use only the lower 16 of the resulting coefficients for classification.

Pixel differences

Pixel difference statistics at different scales are another way to capture the variation of textural energies. Basic average absolute difference between nearby pixels (horizontal and vertical offsets combined), is defined in Eq. 3.1, with pixel intensity at position (i, j) denoted $I_{i,j}$ and the summation covering only those pixels that are within the segmentation mask C of the cell.

$$D(\delta) = \frac{1}{|C|} \sum_{(i,j)\in C} |I_{i+\delta,j} - I_{i,j}| + |I_{i,j+\delta} - I_{i,j}| \tag{3.1}$$

When offset $\delta = 1$, the difference is highest for fine speckled and homogeneous classes, whereas differences from 2 pixels apart ($\delta = 2$) are increased for coarse speckled and centromere. Subsampling the image by a factor of 2 in each direction (following a suitable low-pass filter to avoid aliasing) and applying the pixel-difference operator again creates a textural signature at a coarser scale. The subsampling smooths out most of the finer textures, but brings the stronger gradients of centromere and nucleolar classes to pixel-level scale. Further levels of subsampling are not useful in this particular application, as resulting images are too small to retain any relevant information, but could be used in the general case to create a multi-scale representation of the texture.

The difference averages at the various scales are strongly linearly correlated with each other, but at characteristically different slopes for each class. We therefore derive the most classification benefit by taking pairwise ratios between measurements at different scales, and including them in the feature vector. The ratios are also independent of overall brightness and contrast of the image, aside from quantisation effects.

Morphology features

Another way of comparing these textures, used by a number of contest participants [114, 115, 128], is granulometry or morphological measurements of image slices at different thresholds. Similarly to [128], we consider 7 thresholds equally spaced between the extremes of intensity within each image, and compute 3 parameters from the connected objects produced at each threshold:

- mean area of each object relative to the area of the nucleus mask
- variance of all object relative areas
- average circularity of all the objects

Again following [128], we filter out objects below a certain size (1% of the average object area) as noise. The resulting descriptor has $7 * 3 = 21$ features.

Co-occurrence features

Another well-established and common method of quantifying texture characteristics is the grey-level co-occurrence matrix (GLCM). It was used by a great many of the contestants as part of larger feature vectors, and so it is useful to compare its contribution. As the textures in question are isotropic, it is not necessary to consider different orientations separately, but offsets of different length can provide extra information about different scales of texture, so we include contrast, energy and correlation for matrices at $d = 2$ and $d = 4$ in the descriptor.

3.2.3 Results

A summary of the results, comparing the feature sets to each other, is given in Table 3.1. "Contest:cells" gives the cell-level accuracy on the test partition of the contest dataset for methods trained on the training partition, and "contest:samples" gives the sample-level accuracy using the same train-test split. "Leave-1-out:cells" row lists the cell-level accuracy using the sample-based leave-one-out cross-validation protocol, and finally "Leave-1-out:samples" lists the sample-level accuracies from the same protocol. All experiments in this section use multi-class SVM with RBF kernel, provided by LibSVM library [129], whose hyper-parameters are determined by a cross-validation grid search.

Tables 3.2 through 3.5 present the detailed cell-level confusion matrices for the leave-one-out protocol, to allow analysis of the suitability of each feature set to identification of particular classes.

Additional experiments using a combined feature set containing both DCT coefficients and difference statistics, as well as the shape parameter of circularity,

Evaluation	DCT	Pixel Diffs	Morphology	GLCM
Training	90.9%	95.3%	87.8%	91.1%
Contest:cells	52.3%	56.5%	52.2%	35.3%
Contest:samples	71.4%	71.4%	64.3%	35.7%
Leave-1-out:cells	53.5%	53.7%	50.6%	39.4%
(positive)	60.7%	62.8%	61.4%	45.6%
(intermediate)	48.1%	38.7%	41.4%	26.2%
Leave-1-out:samples	64.3%	64.3%	**71.4%**	60.7%

Table 3.1: Summary of accuracy rates for the different feature sets and forms of evaluation, highlighting overall best in bold type.

True Class	Centr	Homog	Nucl	Coarse	Fine	Cytopl	FNR
Centromere	78.7%	0.8%	5.3%	5.0%	10.1%	0.0%	21.3%
Homogen	0.0%	53.2%	24.9%	1.8%	18.8%	1.2%	46.8%
Nucleolar	21.6%	46.5%	16.6%	8.3%	7.1%	0.0%	83.4%
Coarse	7.6%	1.9%	7.6%	61.4%	21.4%	0.0%	38.6%
Fine	19.7%	34.1%	13.5%	7.2%	25.5%	0.0%	74.5%
Cytoplasm	0.0%	4.5%	0.0%	1.8%	1.8%	91.8%	8.2%

Table 3.2: Cell-level confusion matrix for leave-one-out protocol using DCT features, expressed as percentages of number of cells of true class in the test set

which is very similar to the method in [113], show an improvement over the constituent features taken on their own: the accuracies for the leave-one-out protocol are 56.2% for cells, and 67.9% for samples. This suggests that the two feature types provide some complementary information and can support each other in different sections of the dataset, but cannot surpass the overall accuracy of the morphological features.

3.3 Cell Distribution Experiments

Human experts assessing a sample take account of the appearance of all its cells together, and assign a single class label to the entire image. Experiments placing a human expert into the same conditions as the cell-based protocols presented in the previous section, i.e. only able to examine a single cell at any one time, show that their performance on the MIVIA dataset drops to 73% cell-level accuracy [108]. We therefore explore the possibilities of basing the overall sample decision on a model directly representing the properties of the sample as a whole, as opposed to voting by independently examined cells.

True Class	Centr	Homog	Nucl	Coarse	Fine	Cytopl	FNR
Centr	70.6%	0.8%	7.0%	6.2%	14.3%	1.1%	29.4%
Homog	1.5%	53.0%	16.4%	4.2%	23.0%	1.8%	47.0%
Nucl	31.1%	29.9%	29.5%	6.6%	2.9%	0.0%	70.5%
Coarse	8.6%	1.9%	8.6%	64.3%	16.2%	0.5%	35.7%
Fine	28.8%	27.4%	1.4%	8.7%	33.2%	0.5%	66.8%
Cytopl	5.5%	8.2%	0.0%	5.5%	8.2%	72.7%	27.3%

Table 3.3: Cell-level confusion matrix for leave-one-out protocol using pixel difference features, expressed as percentages of number of cells of true class in the test set

True Class	Centr	Homog	Nucl	Coarse	Fine	Cytopl	FNR
Centr	66.9%	5.9%	12.0%	7.6%	6.4%	1.1%	33.1%
Homog	11.2%	35.5%	5.8%	10.0%	37.3%	0.3%	64.5%
Nucl	37.8%	16.6%	35.3%	5.0%	4.1%	1.2%	64.7%
Coarse	7.6%	10.0%	6.7%	61.9%	13.8%	0.0%	38.1%
Fine	7.2%	44.7%	1.4%	7.2%	39.4%	0.0%	60.6%
Cytopl	5.5%	4.5%	10.9%	2.7%	0.0%	76.4%	23.6%

Table 3.4: Cell-level confusion matrix for leave-one-out protocol using morphological features, expressed as percentages of number of cells of true class in the test set

True Class	Centr	Homog	Nucl	Coarse	Fine	Cytopl	FNR
Centr	62.7%	10.4%	7.3%	5.6%	5.9%	8.1%	37.3%
Homog	16.1%	31.8%	11.8%	9.4%	21.5%	9.4%	68.2%
Nucl	22.4%	58.5%	2.9%	7.9%	6.2%	2.1%	97.1%
Coarse	10.0%	11.4%	3.8%	50.5%	18.1%	6.2%	49.5%
Fine	8.2%	31.7%	5.8%	11.5%	42.3%	0.5%	57.7%
Cytopl	17.3%	19.1%	3.6%	16.4%	3.6%	40.0%	60.0%

Table 3.5: Cell-level confusion matrix for leave-one-out using co-occurrence features, as percentages of number of cells of true class in the test set

3.3.1 Normal distribution modelling

One approach to modelling the overall properties of a sample is to estimate the distribution of cell parameters within it, and compare distribution overlap for samples of the same class and from different classes. A feature extraction process which is invariant to the 'distractor' variables, such as differences in overall sample intensity or focus, while being sensitive to the true class-dependent characteristics of the image, should produce distributions that overlap strongly with those from the same class, while being well separated from distributions of other classes. We use the Bhattacharyya distance (denoted D_B) for multivariate Normal distributions, based on sample mean and full co-variance, calculated according to Eq. 3.2, where μ_1 and μ_2 are sample means of the two distributions, Σ_1 and Σ_2 are the corresponding co-variances, and the combined co-variance is $\Sigma = (\Sigma_1 + \Sigma_2)/2$.

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)\Sigma^{-1}(\mu_1 - \mu_2)^T + \frac{1}{2}\log(\frac{|\Sigma|}{\sqrt{|\Sigma_1||\Sigma_2|}}) \tag{3.2}$$

The normality assumption holds better for some feature sets than for others. We produce distance maps for a variety of feature sets to illustrate their strengths and weaknesses.

The distribution overlap data is presented as distance maps, with dark points corresponding to closely overlapping distributions, and brighter ones being more separated. The samples are grouped by class and also by intensity, so that the first 3 samples are centromere and positive, the next 3 are centromere and intermediate, followed by the 5 homogeneous samples, similarly split by intensity, etc. This arrangement allows for easy visualisation of the expected performance of a feature set, based on the degree of block-diagonality within the distance map.

The distance matrix for the distributions of cells using DCT features is visualised in Fig. 3.2, that of the pixel difference feature set in Fig. 3.3, and GLCM in Fig. 3.4.

3.3.2 Cumulative histogram modelling

Another approach to whole-sample representation works better for features that are based on histograms, i.e. empirical estimates of probability distributions, Local Binary Patterns (LBP) being a typical example for textural properties. In this section we compare the decisions based on LBP for individual cells with those accumulated from all the cells in a whole sample. We choose a rotation-invariant form of uniform 8-point LBP, applied at multiple scales to provide information about a range of sizes of textural feature. Histograms are compared using the

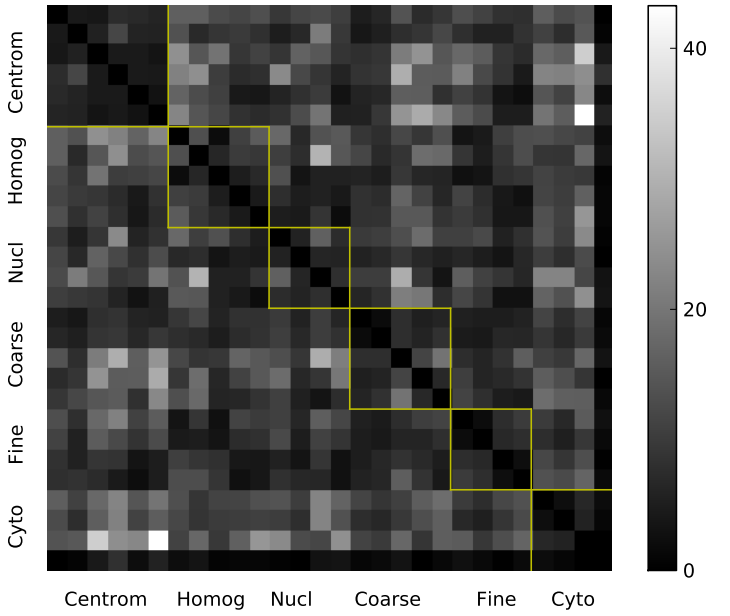


Figure 3.2: Distance map using DCT features

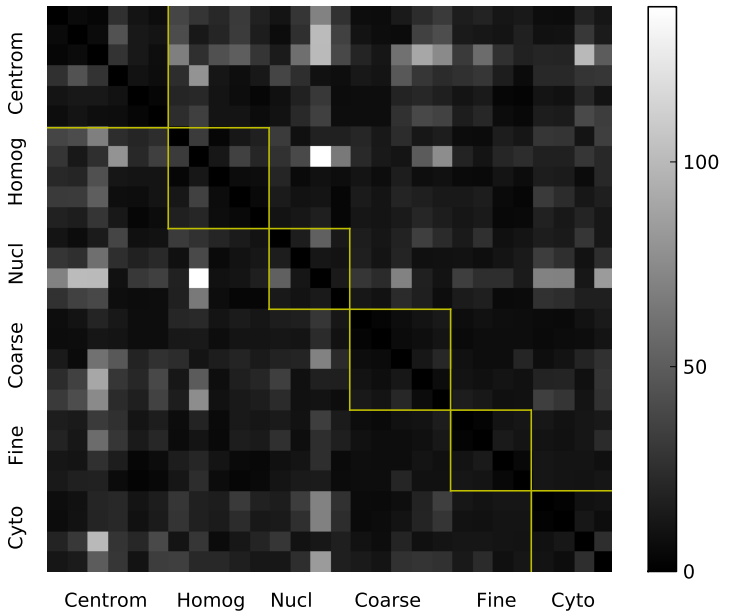


Figure 3.3: Distance map using pixel difference features

Bhattacharyya distance, calculated according to equation 3.3, where p and q are the normalised LBP histograms of the two cells or samples being compared.

$$D_{LBP}(p,q) = -\ln \sum_i \sqrt{p_i \cdot q_i} \tag{3.3}$$

Two additional coarser scales are produced by Gaussian filtering and down-sampling, followed by application of the same 8-point uniform LBP operator, resulting in a total of 30 attributes, of 10 LBP bins at 3 scales. Cytoplasmatic class is excluded throughout these experiments as indications of its distinctive shape cannot be easily included in a textural histogram vector.

We use the leave-one-out protocol to obtain classification accuracy estimates for k Nearest Neighbours algorithm with optimal value of $k = 1$ determined by cross-validation. All cells from a single sample are excluded from training and used as validation set in both cell and sample predictions. The majority vote by independent cells correctly predicts 62.5% of samples (15 out of 24), based on a cell-level accuracy of only 43.9%. Accumulated histograms for a whole sample produce correct predictions in 41.7% of cases (10 out of 24). A complete sample-level distance matrix is illustrated in Fig. 3.5.

3.4 Discussion

We have performed an experimental comparison of a number of different texture measures on a publicly available dataset of medical images. Our main goal was not necessarily to achieve the best possible result, which would be unrealistic given the absence of earlier publications providing a base-line, but to approach a better understanding of the intrinsic properties of this type of images and their class characteristics. We also examine and compare different approaches to combining information from individual cells comprising a sample in order to make a class prediction for the entire sample.

3.4.1 Analysis of experimental results

We have performed experiments comparing both different texture features for their efficacy in predicting the correct immunofluorescence pattern class, and between different methods of combining cell properties into a sample prediction. Due to the limitations of the leave-one-out procedure, no spread can be placed on the accuracy estimates, making it impossible to judge the significance of any difference in performance.

Of the feature sets tested, the morphological parameters clearly outperform the others on the ultimate measure of sample-level decisions in a leave-one-out

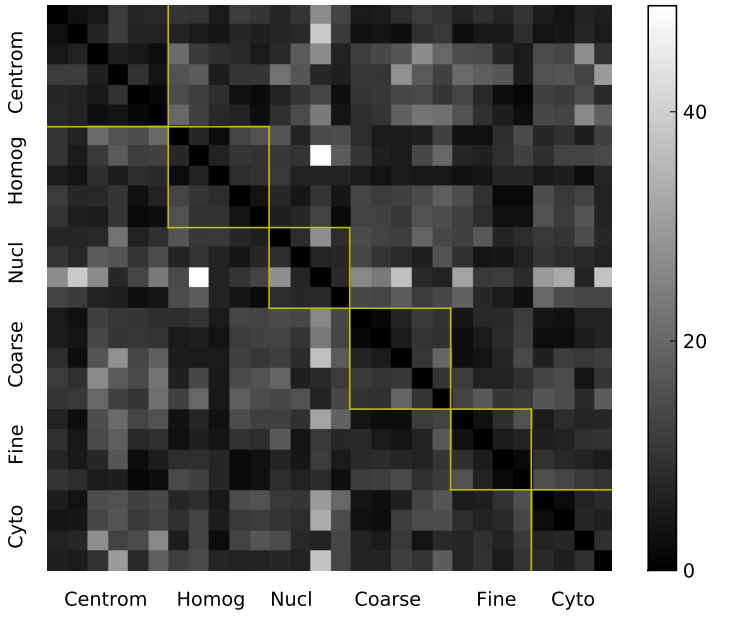


Figure 3.4: Distance map using co-occurrence features

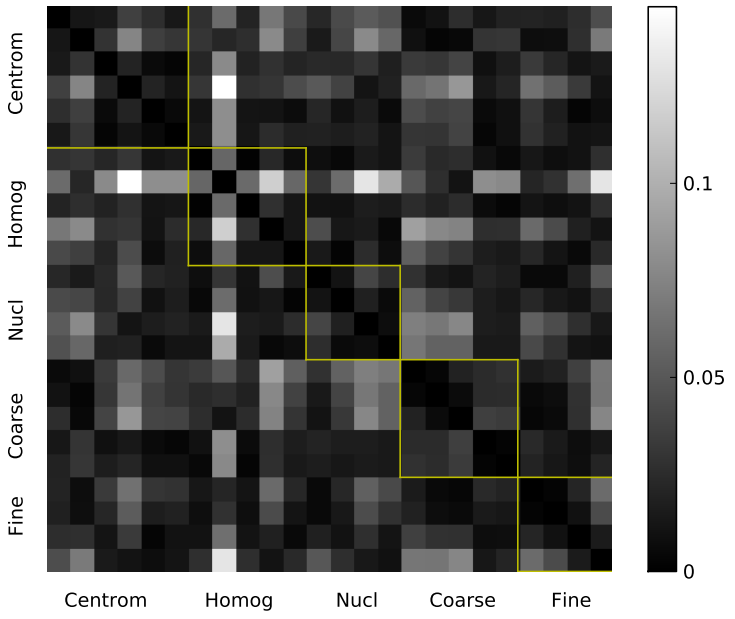


Figure 3.5: Distance map using cumulative LBP histograms

protocol (see Table 3.1). The important point to note is that this difference of performance could not be predicted from the cell-level performance of the various classifiers, as the morphological features actually perform slightly worse for that evaluation, and its training cross-validation results are actually the worst of all the features tested. It is possible that the improved performance is mostly due to the comparatively larger dimensionality of the morphological vector, rather than its intrinsically greater relevance to the class characteristics.

Examination of the detailed cell-level confusion matrices in Tables 3.2-3.5 shows that the DCT-based descriptor performs very poorly on nucleolar samples, but is perfectly accurate on cytoplasmatic ones. This is reflected in a strong block associated with the cytoplasmatic class in its distance matrix (bottom right in Fig. 3.2), as compared to a thin diagonal line for nucleolar. GLCM is the worst performing feature set across the board, but is also the smallest vector. It particularly struggles with the nucleolar class, suggesting that, in its current formulation, it does not extend to large enough scales, and its quantisation level may not be optimal for this application. Pixel-difference features seem most suited to the centromere class, but could also be used to separate the finer-grained classes (ie homogeneous and both speckled) from the rest.

Careful inspection of the predicted class for each sample (available in [127]) shows that some samples are predicted correctly with every feature set, while others are wrong in every case, suggesting that there may be an issue of variable image quality which is affecting the texture itself, however it is measured. Specific blur-tolerant texture descriptors may need to be deployed to combat this problem.

Distribution-overlap distance maps clearly visualise the fundamental problem: none of the feature sets is able to produce a block-diagonal matrix which would indicate reliable similarity within classes and differentiation between them. There is too much variability within classes which is not adequately represented by the few examples that are available. Even within just the positive samples of the homogeneous class there is a lack of class consistency that is apparent in every distance matrix: instead of a block, it is showing up as a diagonal cross, because the middle of the 3 available samples is very different from the other two.

Similarly to the contest report [108], we find that intermediate intensity patterns are much more prone to errors than positive ones. This discrepancy persists with all types of attribute, including local binary patterns, which are generally taken to be very robust to changes of intensity as they only take account of the sign of pixel differences, and not their magnitude. We speculate that in this case the differences between positive and intermediate images are more qualitative than mere scaling of intensity; that the intermediate images are so faint that pixel differences which were noticeable in their positive counterparts fall below the 8-bit quantisation step of the sensor, and therefore count towards a different LBP bin. This is supported by the distance map of whole-sample LBP

histograms, which in several classes shows a within-class block pattern of much greater similarity between samples of same intensity (see Fig. 3.5, visible as two smaller blocks instead of one large block covering the whole class).

While LBP accuracy rates are not directly comparable with the other results because of the exclusion of cytoplasmatic class, which affects the percentage represented by each sample in the dataset, the prediction accuracy of the accumulated sample histograms is disappointingly low. It is likely that the observations above, on the intrinsic differences between the intermediate and positive images due to quantisation effects, also come into play here, as the only truly similar samples within the dataset are those that have both the same class and the same intensity. For many classes there are only 2 examples in this dataset that share both these labels, which in a leave-one-out protocol results in a single comparable training example for many test samples. This is simply not enough to make a reliable prediction, especially when the number of classes is large.

3.4.2 Further work

The distribution distance matrix can be used as the basis of an ensemble combining distance information from several feature vectors. Other ways of combining two or more different feature sets through either early or late fusion should also be explored, as there are indications of complementary information represented by different texture measures. As some feature types are more suited to identification of certain classes, they could also be combined in a cascade which filters out each class based on its most favourable features.

Another potentially fruitful approach to addressing the shortage of labelled image data is the use of semi-supervised methods. Leverage of large numbers of unlabelled HEp-2 images could allow development of much better understanding of the effect of imaging conditions on the resulting image texture, and compensating for these common variations in classification, for example through the use of manifold learning or subspace methods [130].

If or when larger quantities of labelled training data become available, it would become possible to treat positive and intermediate samples of each pattern as separate sub-classes, which could be very beneficial to their recognition as we find their characteristics to be significantly different from each other. At least 3 or 4 samples of each combination of intensity and pattern class would have to be present in each of training and test sets to produce meaningful learning results in this case.

Mitotic cells, whose appearance allows the experts to differentiate between fine and coarse speckled patterns, must be brought into the analysis and combined with information from the rest of the sample. An extensive range of methods for combining evidence would need to be explored to determine which

is most appropriate for this situation.

The most recent competition on the subject of automated pattern recognition for IIF images, the I3A workshop at ICPR 2014[1] (held after the completion of this thesis), should make a significant step forward in the research, as it includes both cell-level and specimen-level classification, with no restrictions on methods. Its dataset includes the mitotic nuclei necessary for differentiation of certain classes, although they are not labelled, and matches the class definitions of the dataset used in the ICIP 2013 contest.

3.5 Conclusion

We conclude strongly that cell-level performance of a classifier offers little guidance to its performance in whole-sample decisions, even in a simplistic majority-vote setting. This is supported by the recently published detailed report of the ICPR 2012 contest findings [108], which shows great variability between method rankings by cell-level and by sample-level performance. Consideration of the sample as a whole, including complete measurements from all the relevant cells, allows the application of a much richer set of pattern recognition methods, and is a better match for the ultimate goal of replicating the diagnostic decision of a physician. Whilst considerable progress is being made in identifying likely methods for single cell classification, we feel that assessment of their suitability for use in a realistic clinical system requires a larger quantity of data that more fully covers the variability of cell appearance, such as the new SNP dataset.

[1] http://i3a2014.unisa.it

Chapter 4

Mitosis detection

The subject of this chapter is detection of mitotic figures in breast biopsy images, which are described in Section 1.1.3, and it is by far the most complex challenge of the three domains investigated in this thesis. There are no separate cells with plain background, but an intricate mosaic of multifarious shapes and textures. It takes many years of training to identify the different constituent elements within the sliced tissue, which varies in its appearance depending on a number of factors: the organ from which the tissue is taken, tumour type and stage, the preparation process undergone by the sample and individual patient characteristics.

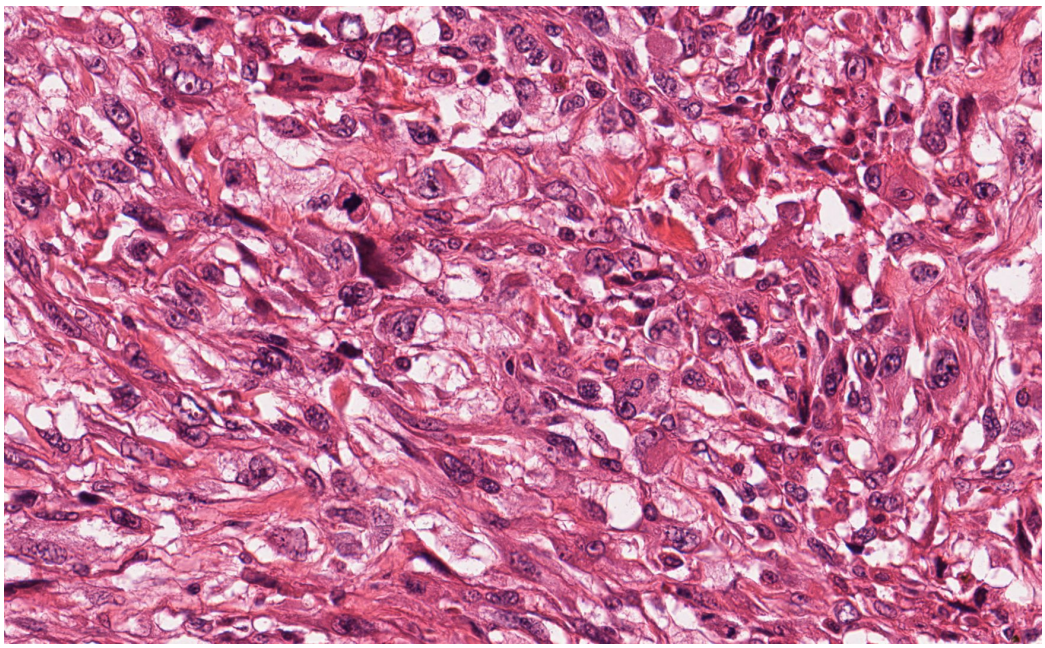
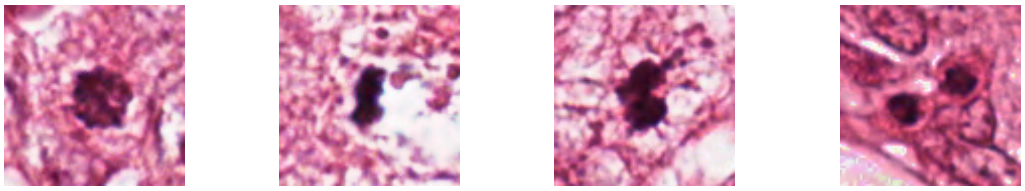


Figure 4.1: Example section of H&E stained breast biopsy image

Identification of cells which are dividing, or mitotic, is just one step on the way to grading the tumour: the frequency of division is correlated to aggressiveness of tumour growth, but there are other indications which also contribute to the grade, such as nuclear pleomorphism (marked variation in nucleus shape and/or size) and tubule formation, which are outside the scope of this study. Identification of dividing nuclei is itself an imprecise art, chiefly due to the extreme variability of the visual presentation of the nucleus, depending on the exact phase of the division process in conjunction with the nucleus orientation relative to the slice plane. Four major phases of mitosis are recognised:

- **Prophase**, in which the genetic material of the cell condenses, relative to its inter-phase diffuse state.
- **Metaphase**, in which the ball of nuclear material elongates.
- **Anaphase**, in which the nucleus splits into two parts. This phase is brief, and is therefore relatively rarely observed.
- **Telophase**, in which the two new cells move apart and gradually assume the normal inter-phase appearance. This phase can be particularly difficult to handle as there are two separate segmented objects present, which nonetheless have to be identified as a coherent pair and counted as a single mitotic figure.

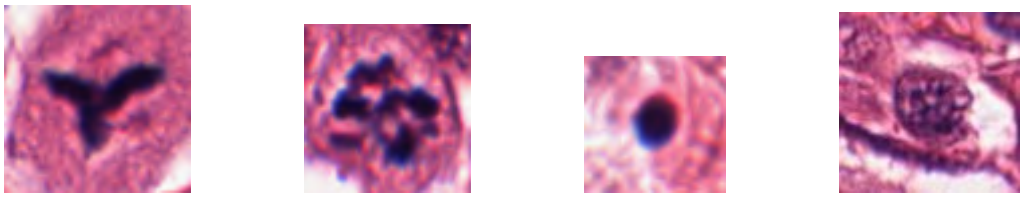


(a) Prophase (b) Metaphase (c) Anaphase (d) Telophase

Figure 4.2: Examples of each phase of mitosis

Typical appearance changes associated with each phase are illustrated in Fig. 4.2, although it is worth noting that these are subject to considerable variation, depending on viewing angle or due to abnormalities of the tumour cells. For example, instead of normal elongation in metaphase, one can observe cases of three or four-way splits, as shown in Fig. 4.3(a). The strongest distinguishing feature of mitotic nuclei is their dense staining with the hematoxylin (H) component of the stain, which binds to DNA-rich regions and colours them a deep purple (the eosin component produces the bright pink colouring of the majority of other cell structures). However, the colouring alone is insufficient for

accurate detection, due to the presence of distractor objects of similarly dense appearance, particularly cells that are *apoptotic*, or undergoing the process of cell death, see Fig. 4.3(c). Other highly confusing scenarios include metaphase nuclei that have been sliced across the mitotic spindle, rather than along it, and consequently appear as a collection of scattered dots, the end-on view of the stretched chromosomes, as in Fig. 4.3(b).



(a) 3-way division (b) Spindle slice (c) Apoptotic (d) Inter-phase

Figure 4.3: Examples of difficult presentation, and an inter-phase cell

We now review the published work on the specific subject of automated mitosis detection in H&E-stained breast biopsies, with reference to the image analysis and machine learning methods covered in Section 1.3. We then describe the detailed experiments carried for this project in Section 4.2, present the results in Section 4.3, and discuss the findings in Section 4.4. Our contributions cover stain profile normalisation, segmentation, feature extraction and imbalanced class learning, as well as particular adaptations in patch pre-processing which enable use of GP-LVM methods in this application.

4.1 Review of prior art

On the specific topic of mitosis detection in malignant breast tissue with H&E staining, two public datasets are available, each associated with a challenge or contest comparing the performance of a wide range of algorithms. We describe each dataset, and the methods submitted to the corresponding contest, in the following two sections. It is important to bear in mind that low agreement among experts makes it challenging to provide definitive ground-truth labels for any such dataset [5, 65].

4.1.1 ICPR 2012 Contest and MITOS dataset

'Mitosis detection in breast cancer histological images', a contest held at ICPR in 2012, provided a dataset of scanned biopsy slides from 5 patients, each one represented by 10 high-power fields (HPF) at 40x magnification [65]. Each HPF

of 512μm by 512μm was scanned by three different scanners, two of which came from different manufacturers but were of similar resolution, and a 10-band multi-spectral microscope, which also included 17 different focus planes (Z-stack) in 500nm steps. The contest submissions were evaluated separately for each scanner, with the majority of entrants only supplying predictions for the first scanner. The ground-truth annotations included not only the centroids, but a full pixel-wise segmentation of each mitotic figure, which in the case of telophase cells may not include the centroid itself.

The dataset is split into 70% training and 30% test, with 3 fields from each of the patients held back for testing. This split makes the contest task considerably easier than the realistic one of presenting an entirely unseen sample for testing, as the training portion contains images from the same patient samples as the test. Predictions from participant algorithms were evaluated and ranked by F_1-measure, with detections counting as a true positive if their centroid was within 8 μm of the ground-truth centroid. The original contest description suggested that finer distinctions between algorithms of same F-measure would be evaluated based on the extent of overlap between ground-truth and predicted segmentations, but in the event the spread of results was quite broad, and some of the entries did not include detailed segmentations.

Only 4 teams submitted results for the multi-spectral Z-stack dataset, and all the results were much poorer than for the normal 3-channel RGB images, despite the higher resolution of images and the additional information contained in the multiple focus planes and the additional spectral bands, which was clearly a disappointment. However, the drop in performance is easily explained by the lack of special treatment to extract the greatest possible amount of information from the additional data; of the few entries that described their multi-spectral and multi-focus algorithms during the contest workshop, all restricted themselves to picking a single best focus plane, and a single best channel, across the entire dataset. Better results may have been obtained by applying adaptive focus techniques and using a combination of channels to derive the optimal projection of the hematoxylin and eosin signals.

There were 14 submissions for the main image set, the best of which achieved an F_1-measure of 0.78, with a considerably higher precision of 0.89 but lower recall of 0.70. The balance between precision and recall varied quite strongly between methods.

Of the highest ranking entries, IDSIA's (Dalle Molle Institute for Artificial Intelligence, Switzerland) deep convolutional neural network [131], consists of 10 alternating layers of small-aperture convolutional linear filters and max-pooling sub-sampling operators, followed by a final fully-connected layer of neurons producing a 2-class prediction of the probability that the input window is centred on a mitotic figure. The network operates directly on the RGB values of the input

image pixels, and assesses a dense grid of window offsets. The training of such a deep network requires immense amounts of computation, and would take years without GPU acceleration, which brings it down to a matter of days. Similarly, test-time application of deep convolutional networks is quite computationally intensive, and can take several minutes per high-power field unless accelerated by a GPU implementation.

Most other entries followed the more traditional detector pipeline of identifying candidate or seed points, segmentation of the nucleus outline and computation of features based on the segmented mask, and finally classification as mitosis or not, based on the extracted features. Most of the candidate detections were based on simple thresholding, followed by mathematical morphology filters, but the entry from the University of Warwick included an additional step of high-level segmentation of the tissues to exclude non-tumour areas from the detection process [132, 133]. Although the segmentation masks of the detected nuclei were not used for the ranking comparison, and some of the entries did not even include detailed segmentations, the quality of segmentation would affect the accuracy of the extracted features in this type of pipeline, so it remains an important step. The most crucial choice, however, is the constitution of the feature vector, which varied considerably between methods; careful tuning of the classifier to cope with the potentially very high level of class imbalance (depending of the precise characteristics of the earlier candidate detection step) is also essential. Detailed descriptions of individual methods can be found in [134–139].

All samples in this dataset were of high-grade tumours, with little variation in mitosis density between patients. This choice was probably made on the grounds of providing the greatest number of mitotic figures for a given size of dataset, but it does limit how representative the dataset is of the range of cases typically presented for annotation.

4.1.2 AMIDA Grand Challenge

Following the findings of the 2012 contest, a new challenge, named 'Assessment of Mitosis Detection Algorithms' (AMIDA)[1], was held at the MICCAI (Medical Image Computing and Computer Assisted Intervention) 2013 conference. A new dataset was provided by the University Medical Center Utrecht, which included a much greater variety, as well as number, of patient samples, covering both high and low grade tumours. The experimental design was also improved, so that the training and test images never came from the same source sample; this is a much more realistic scenario than that of the ICPR 2012 contest, but does result in lower overall levels of accuracy. In all, 12 patient samples were available for

[1] http://amida13.isi.uu.nl

training, and a further 11 used for test.

Additional care was taken with ground-truth labelling of the images, using two independent annotators, and a panel of two further experts to adjudicate on any cases of disagreement between them. Only centroids of the agreed mitotic figures are listed as ground-truth, and any predictions within 7.5μm, or 30 pixels, of the labelled position are considered true positives. The density of mitotic figures varies between 0 and 13 per HPF, with at least 10 fields from the diagnostic region-of-interest of each sample, but as many as 60 for some samples. In total, the training set contains 550 examples of mitosis across 331 HPFs, including around 30 examples of telophase pairs. Evaluation by F_1-measure was performed both for the dataset overall, which gives a greater weight to patient samples with lots of mitotic figures, and for each patient sample separately, providing greater insight into the strengths and weaknesses of the submitted algorithms on different input material. As some (low-grade) samples contain no mitotic figures in any of their images, making conventional F-measure evaluation impossible (as the number of true positives is zero), the ranking of algorithms for these cases was performed based on the number of false positives alone.

A single device was used for digitization of the slides (with similar resolution to that of the ICPR 2012 dataset), but the length of time over which the samples had been collected meant that there was a lot of variation in stain strength and exact hues, as different batches of stain would have been in use in the laboratory at different times. In this respect the AMIDA challenge is again harder, but more representative, than the earlier contest, and many entries suffered drops in accuracy on samples with particularly strong or weak staining. Some ways of dealing with these variations have been suggested in the literature, which take into account the light transmission (as opposed to reflection) nature of bright-field microscopy and therefore operate in the logarithmic space of optical density, but they usually target the more challenging task of multiple stain separation and may therefore be unnecessarily complex for this application [140].

The challenge received results from 14 teams from around the world, and was again won by a deep convolutional neural network from IDSIA, by an even greater margin than previously (full competition results are listed in Table 4.3 and will be published in [141]). No method descriptions have as yet reached publication, but some teams were present at the workshop to give an overview of their algorithms. A common thread through several methods was the use of 'blue ratio' (BR) images as the colour pre-processing step that 'accentuates' the nuclear regions. First appearing in [142], and defined as $BR = \frac{100 \cdot B}{1+R+G} \cdot \frac{256}{1+B+R+G}$, this seems to be an entirely heuristic transformation, with no basis in measured stain properties, although perhaps an improvement on the entirely simplistic approach of using the Red channel on its own, chosen by some of the teams.

Second-best performance was achieved in both overall and per-patient rank-

ings by the team from Technical University of Denmark, whose method, titled 'Donut spatial pooling', was based on histograms of colour, gradient orientation and shape index, collected from soft concentric ring regions around the test position. The use of shape index, calculated as $s = \frac{2}{\pi} arctan(\frac{\kappa_1+\kappa_2}{\kappa_1-\kappa_2})$, where κ_1 and κ_2 are the two eigenvalues of the Hessian matrix $\nabla^2 L$ of the Gaussian scale-space $L = G * I$ of the image I, captures the principal local curvature of the image surface and was found to be the most significant component of the feature vector in terms of classification performance. The classifier chosen by this team, an RBF-kernel SVM, is probably not the most suitable for histogram features, particularly in terms of kernel, and even better results may have been achieved with greater optimisation of this part of the process, compensating for class imbalance and selecting a kernel targeted at distribution comparisons, such as χ^2. All features used in this method are inherently rotation-invariant, which is an important property in this application, but treating the three colour channels completely separately spreads the extracted information over a larger number of bins than necessary and holds back their effectiveness in recognition.

Below the second place the two rankings, those based on overall accuracy and those giving equal weighting to each patient case, regardless of how many mitoses are present in the corresponding images, differ considerably from each other. Neither form of evaluation can be considered the ultimate way to assess performance, particularly when using a relatively small dataset such as this one; the overall F-score gives excessive weight to accuracy on the high-grade cases with lots of mitoses, and per-case accuracy can vary enormously for any particular method, from completely wrong to total perfection, making their average somewhat meaningless. Many of the lower-ranked entries suffered from extreme imbalance between precision and recall, mostly in the direction of quite high recall but very poor precision. Of the methods presented at the workshop and reporting their cross-validation performance on the training set (and correctly implementing a patient-based cross-validation procedure), the difference between cross-validation and test results was around 0.1 on overall F-score, suggesting that even this larger training set did not cover sufficient variation of the input material, and the test set presented new challenges.

4.1.3 Gaussian Process Latent Variable Models

Although Section 1.3.4 provided a broad overview of machine learning methods applicable to histopathology image analysis, a more detailed exposition of one of the models which has been evaluated as part of this project is given here.

In statistics, a Gaussian Process is a stochastic process characterised by a

mean function m and a covariance or kernel function K, denoted as

$$X \sim GP(m, K) \tag{4.1}$$

and having a multivariate Normal distribution when sampled at any one point. In the context of dimensionality reduction, Gaussian Processes allow the extension of principal component analysis (PCA) with non-parametric non-linear mappings from the latent variables, while maintaining a fully probabilistic framework [143]. The probabilistic treatment allows estimation of the model's uncertainty at any point in the latent space, while the non-parametric nature of the mapping gives ultimate flexibility to cope with highly non-linear relationships to the high-dimensional image space. The main drawbacks of the method are non-convexity of the optimisations needed to find the best mapping for a particular configuration of observed data points, and the lack of a direct reverse mapping from the observed samples to their latent positions, necessitating further non-convex optimisation when assessing new test points. Extensions of the method using variational inference allow automatic determination of the number of latent dimensions necessary to describe the relationships inherent within the data [69], but these come at the expense of further computational complexity. Explicit modelling of noise as an additive component in the observed variables helps to further isolate the meaningful latent interactions. Finally, a variety of possible kernel functions allows representation of a rich set of potentially non-stationary processes, although the infinitely differentiable Gaussian kernel is by far the most common in practice.

4.2 Experimental methods

In this section we describe the full details of the algorithms submitted to the AMIDA challenge, one solely by the University of Surrey, and one in collaboration with the University of Sheffield's Institute for Translational Neuroscience (SITraN). The two submissions share the first blocks of the detection pipeline: colour normalisation (Section 4.2.1), detection of seed points (Section 4.2.2, greyscale conversion and segmentation (Section 4.2.3). The traditional supervised pipeline then proceeds with feature extraction (Section 4.2.4) and classification (Section 4.2.5), while the alternative GP-LVM approach (Section 4.2.6) works directly on the pixel values of the image patches. The GP-LVM modelling and simulations were carried out by SITraN in Sheffield based on labelled rotated image patches, and their results are included here for completeness.

Exactly the same methods are applicable for the MITOS dataset, and some of the sections cover our findings for both cases. The main novel contribution

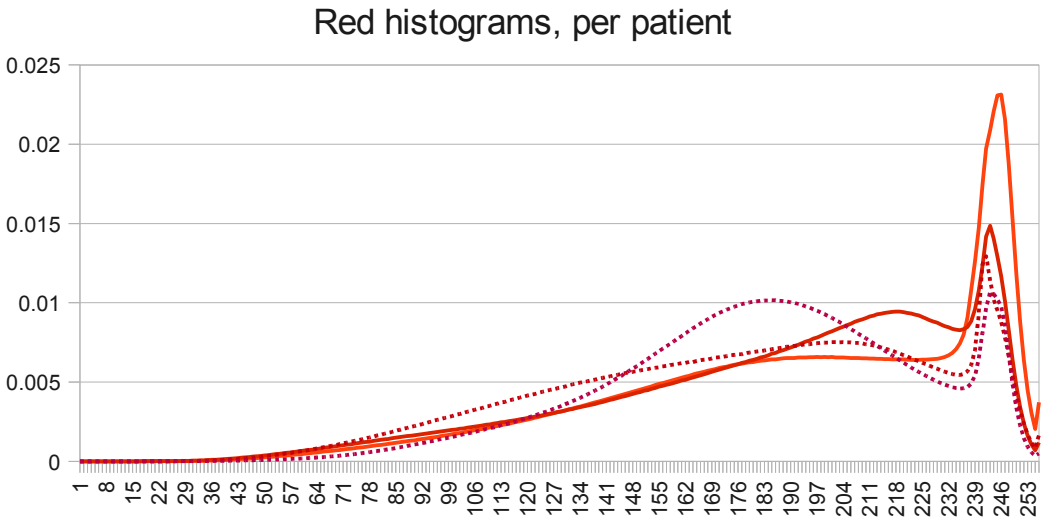


Figure 4.4: Red channel histograms for a selection of cases from the AMIDA database, each line corresponding to distribution of pixel values originating from a single patient sample

of this work to the detection pipeline is the stain normalisation process, which is presented in the next section.

4.2.1 Stain normalisation

To compensate for variability of staining and preparation, the images are first aligned in colour space. Figs. 4.4 through 4.6 show examples of individual channel histograms from several samples in the AMIDA dataset, to illustrate the type and degree of variation between slides. As so much of the downstream processing is dependent on pixel colour, it would be unwise to feed images of such variable colour distributions directly into the next stage.

The only other contest entry to include explicit stain normalisation as part of their pre-processing stage was the team from University of Warwick [132, 144]. They cite an earlier work for details of the method used, which does not actually describe a method for stain normalisation [145]. Instead, the algorithm targets classification of each pixel as belonging to one of the two component stains, or to background, based on a combination of its full colour vector and a global 'context' vector derived from the image histogram. The most confident areas of this pixel-wise classification are then used as inputs for computing a full stain deconvolution matrix. Manual pixel labelling is required to train the classifier and the method does not produce images with a normalised stain profile, so cannot be considered comparable to our proposal.

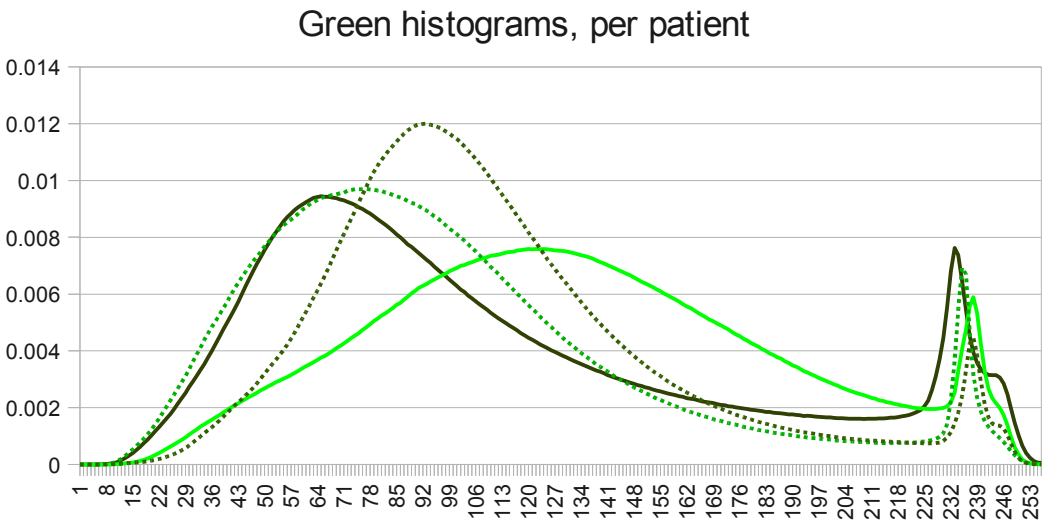


Figure 4.5: Green channel histograms for a selection of cases from the AMIDA database, each line corresponding to distribution of pixel values originating from a single patient sample

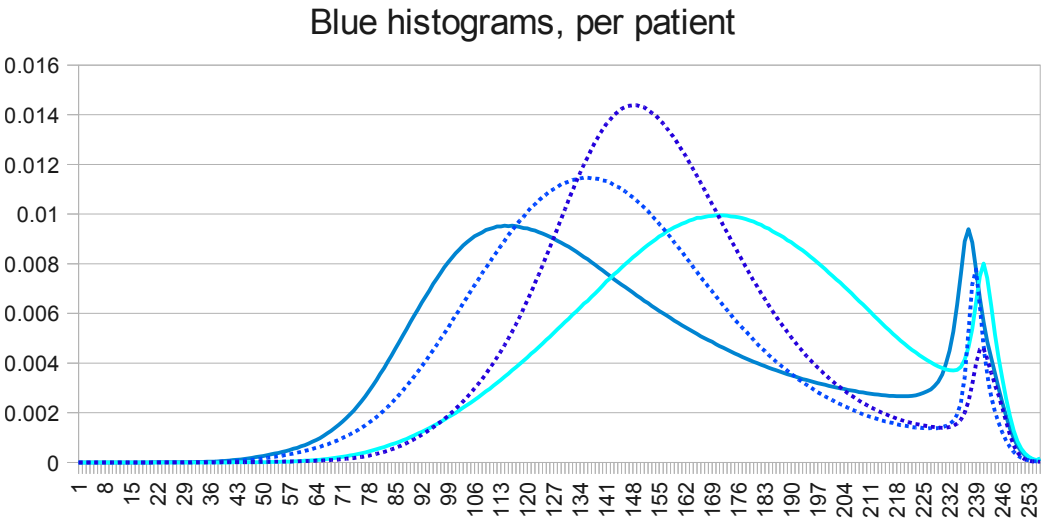
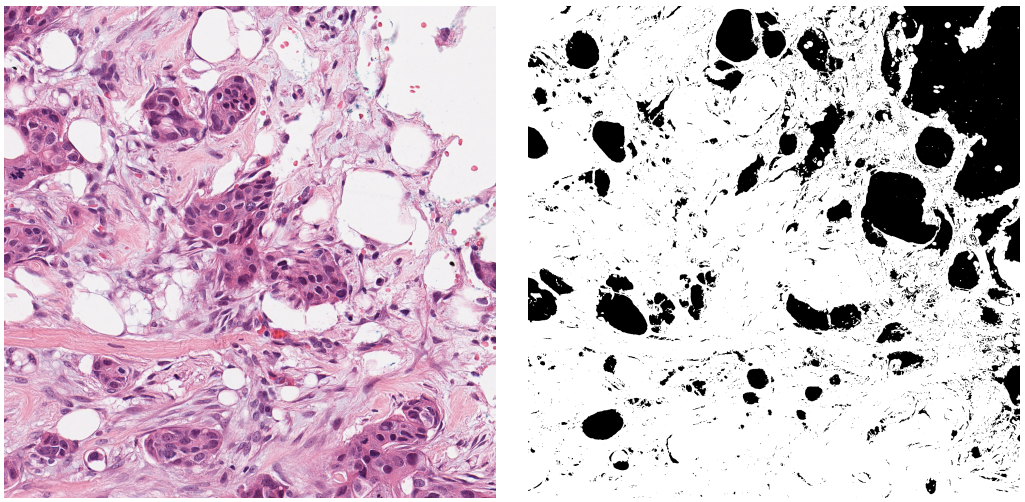


Figure 4.6: Blue channel histograms for a selection of cases from the AMIDA database, each line corresponding to distribution of pixel values originating from a single patient sample



(a) Full colour HPF (b) White exclusion mask

Figure 4.7: Example HPF covering tissue edge, and the corresponding mask excluding all white areas

The distributions in Figs. 4.4-4.6 all have two modes: a broad peak at lower pixel intensities (between 50 and 120 for the green channel) and a much sharper peak at the very top of the range. Blue channel histograms follow similar profiles, with a broad peak at an intermediate range of values, between 130 and 150. In the red channel, the lower-value broad peak is in some cases so close to the higher peak that they are no longer distinguishable as separate modes. The cause of the separate sharp peaks at the top (white) end of the range is the presence of variable amounts of white space (holes, tears or edges) within the tissue on the slide; an example is shown in Fig. 4.7(a). Variable amounts of adipose (fatty) tissue also contribute to the white peak, and therefore affect the whole distribution, but have no diagnostic bearing. We therefore seek to exclude these white, or near-white, areas from all further processing, and particularly from any adjustments of the colour profiles of different slides. We base the decision of which pixels are considered "white" on a threshold applied to the green channel, as the green channel histograms have the best separation of the two modes, due to the low green content of either staining dye. We select the threshold to be used for each 2K x 2K field separately, finding the lowest point between the two peaks in the field's green histogram. Selecting a single threshold for the entire slide would be sub-optimal as the proportion of white areas varies across different parts of the slide and the overall histogram is not as clearly split as those from single fields. An example mask resulting from such threshold selection is shown in Fig. 4.7(b), alongside its source image.

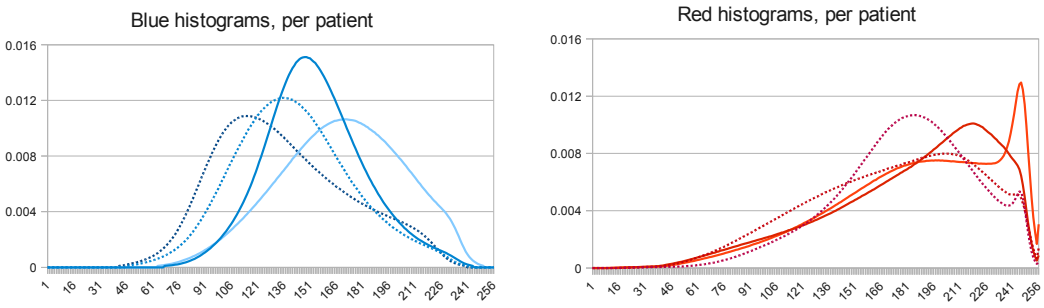


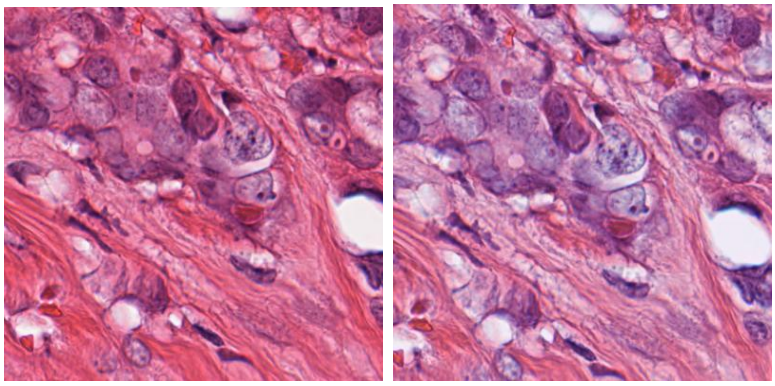
Figure 4.8: Histograms for a selection of cases from the AMIDA database, following exclusion of white areas

Histograms of the remaining pixels, following exclusion of white areas, are shown in Fig. 4.8. The second peak has disappeared completely from the blue channel, and for most of the red channel histograms the second peak has been greatly reduced, if not completely eliminated. It is now possible to apply histogram matching methods to adjust the colour profile of differently stained slides, without bias from the proportion of white areas present in each sample. Each colour channel is adjusted independently, with mean histogram from the whole training set used as the target distribution, and a histogram of each patient's images (taken together, not from each HPF) as the source. For reference, histogram matching is performed by comparing cumulative histograms of the source and destination profiles, and replacing the intensity value of each source pixel with the value that reaches the same level in the target cumulative histogram. The visual effect of the adjustment is illustrated in Fig. 4.9.

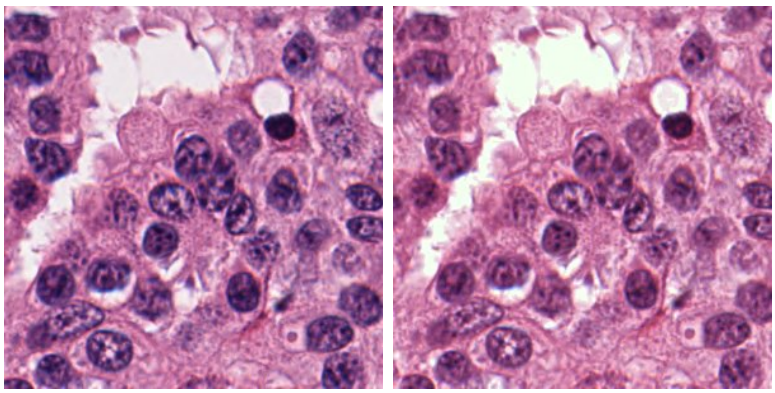
Following the histogram matching procedure, the images can be analysed based on pixel colour without bias from staining variation between different samples.

4.2.2 Detection of candidate locations

The next step in the detection pipeline is identification of locations which warrant more detailed investigation, sometimes referred to as "seed points". We base this initial pre-selection primarily on pixel colour, as mitotic figures are known to be characterised by condensed chromatin, manifesting as darker purple areas following the staining. The strong distinctions of pixel colour between mitotic figures and the rest of the image are demonstrated by Fig. 4.10, which shows separate histograms of the pixels labelled as mitotic in the ground-truth annotations of the MITOS dataset as dashed lines, and overall colour distribution in the same dataset as solid lines. The mitosis distributions are noisier, simply because there are relatively few pixels contributing to them, but occupy a much lower



(a) Source and adjusted images for excess of eosin staining



(b) Source and adjusted images with excess of hematoxylin staining

Figure 4.9: The effect of histogram-matching on varied stain strength

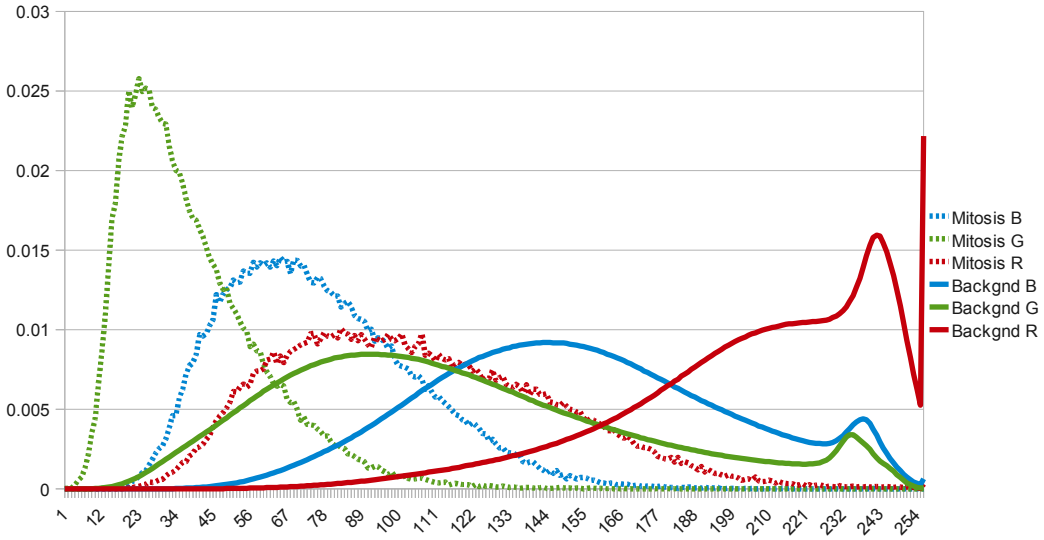
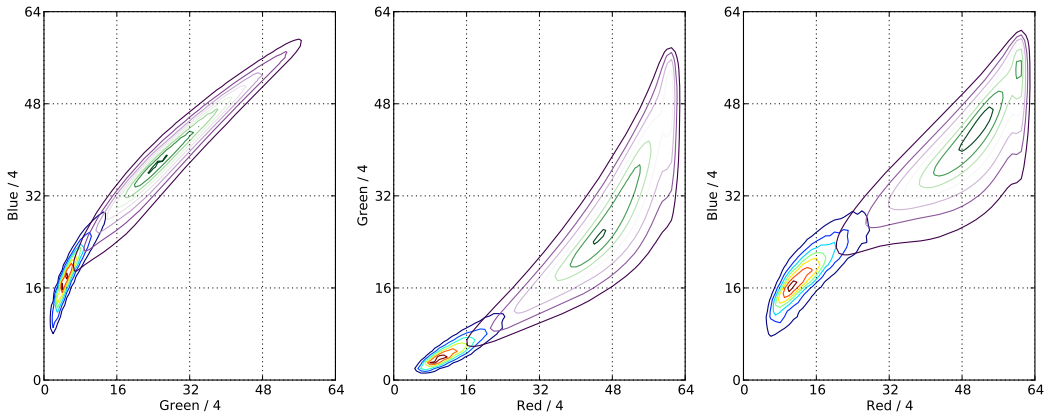


Figure 4.10: Histograms of pixels belonging to mitotic figures (dashed lines), against background pixels (solid lines) from the MITOS dataset

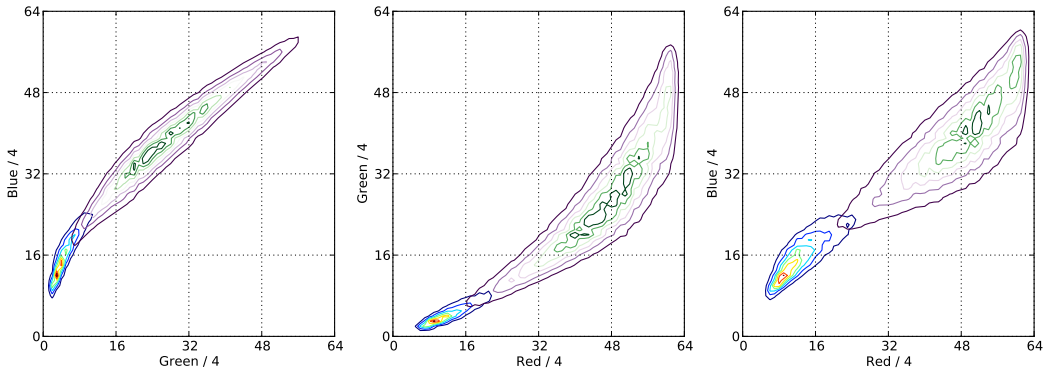
section of the intensity range.

The labelling of individual pixels as "mitotic" is somewhat more difficult for the AMIDA training set, as the annotations do not include a full segmentation, only a centroid of the mitotic figure. As an approximation, we take locations within 10 pixels of the labelled centroid as likely to be within the nucleus, and collect full 3-dimensional RGB histograms of these as the colour distribution of mitotic pixels. Although it is impossible to visualise the full histogram, we show 2-D projections along each of the axes, presented as contour plots, in Fig. 4.11. The histograms are collected in 64 bins along each axis, or 4 intensity levels per bin, in order to reduce the overall size of the histogram and to limit the effect of noise, particularly for the less numerous mitotic pixels. The extremely small amount of overlap that is seen between the mitotic and the all-pixels distributions in 3 dimensions demonstrates that, unlike the single-channel histograms in Fig. 4.10, the full RGB colour of a pixel can give very good guidance as to the likelihood of that pixel belonging to a mitotic nucleus. Fig. 4.11 also illustrates the beneficial effect of the earlier histogram matching step (described in the previous section) on the separability of mitotic nuclei based on pixel colour. The distributions in Fig. 4.11(b) are tighter and less overlapping with each other than those in Fig. 4.11(a). The Bhattacharya coefficient of overlap between the two distributions reduces from 0.57 to 0.55 as a result of the histogram matching process.

The two histograms, of mitotic pixel colours and of the overall colour distribution, are used to create a quantised mapping from pixel colour to likelihood



(a) Mitotic and overall colour distributions before Histogram Matching



(b) Mitotic and overall colour distributions after Histogram Matching

Figure 4.11: Projections of 3-D histograms of pixels close to mitosis centroids are shown in red-to-blue colour map, and background pixels in purple-to-green, before and after stain normalisation on the AMIDA dataset.

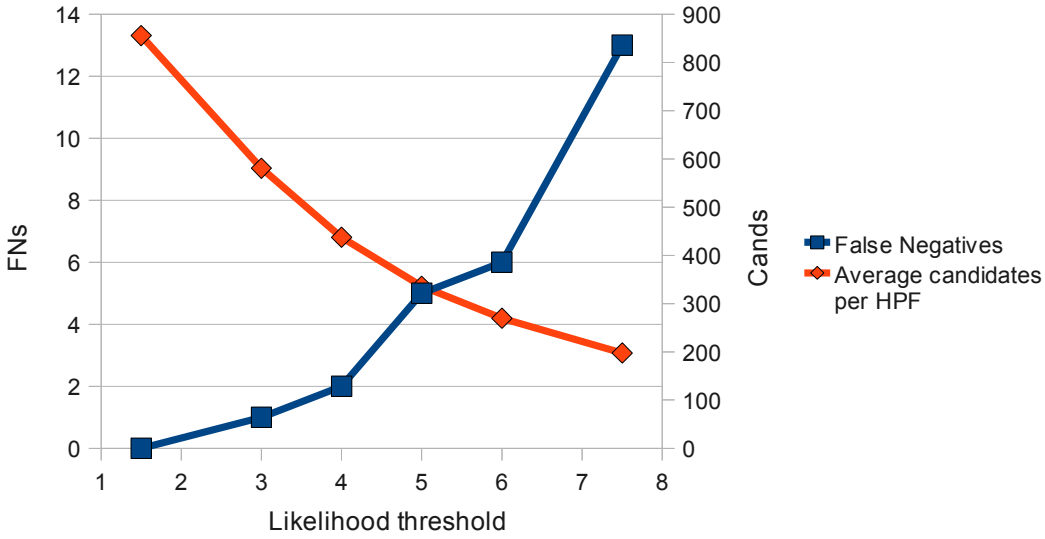


Figure 4.12: Trade-off between number of candidate locations and of false negatives in the first stage for the MITOS dataset

of mitosis, as likelihood ratio $P(R, G, B|mitosis)/P(R, G, B)$. The likelihood ratio is clipped to avoid extremes generated in areas of colour space where the denominator is low, and applied to every pixel to produce a map of likelihood across the image. This map is then filtered by a 5x5 box filter, to ensure that the detected locations come from a spatially coherent group of the likely pixels, not single dots of noise. Following the filter, a threshold is applied to generate a binary map, and the centres of contiguous objects within it are taken as candidate locations.

The choice of threshold level is crucial in controlling the trade-off between false negative rate and class imbalance (and therefore false positive rate): a low threshold will generate a massive number of false candidates requiring detailed analysis and create an extreme class imbalance to the true positive examples, while a high threshold will miss more ground-truth positions before they have a chance of more detailed consideration, creating an underlying bias for the false negative rate of the overall system. Fig. 4.12 shows the effect of this trade-off for the MITOS dataset, with a similar relationship exhibited by the AMIDA images. The size of the box filter aperture was optimised in conjunction with the likelihood threshold, as it also affects both the number of generated candidates and which of the ground-truth positions are missed.

At this stage in the pipeline, a large number of possible mitotic locations have been detected by a relatively fast pixel-based method. The remaining process is one of classic supervised classification, with features extracted from the patch around each candidate location, and used either for training or test.

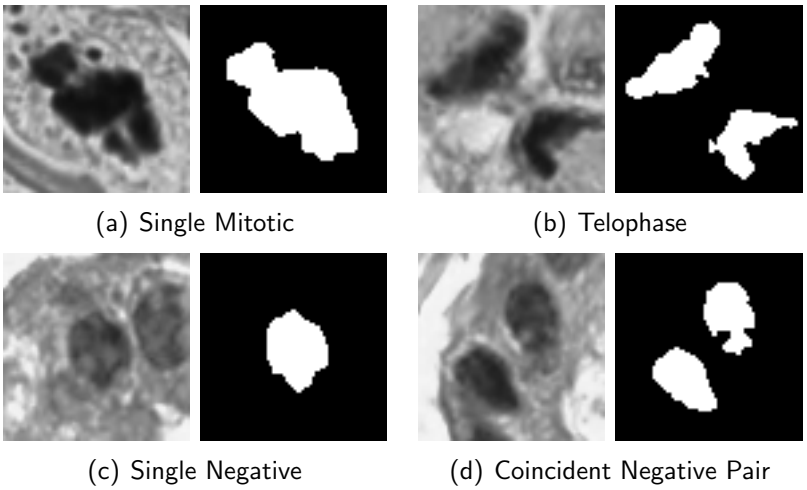
4.2.3 Segmentation

As many of the features used by specialists to distinguish mitosis are related to shape, the automated process requires a segmented outline of the object under analysis in order to measure and describe its shape. We use a 70 x 70 pixel square around each candidate location for the detailed assessment, as none of the ground-truth segmentations in the MITOS dataset extend outside this region. The image is first converted to greyscale, as segmentation in full colour would be much more complex, as well as unnecessary in this case: we know that a single stain is responsible for the colouring of nuclei, as evidenced by the narrow, near-linear, spread of mitotic pixel colours in Fig. 4.11(b). We use PCA of these pixels around labelled mitotic positions to calculate their dominant staining axis in colour space, and project onto this axis to obtain greyscale images.

The basic segmentation algorithm is similar to that described in Section 2.2.1 for DAPI-stained nuclei, in that it selects a threshold which optimises a combination of two different aspects of the boundary. However, the object shapes in this application are more variable and not necessarily smooth or elliptical, so we base the threshold selection on a different combination of properties: the greatest gradient across the segmentation boundary together with the lowest variance of the pixels within the foreground object. This additional criterion favours internal solidity of the segmented object, which for many negative candidates - inter-phase nuclei whose chromatin is quite dispersed - results in highly irregular outlines.

Major complications arise in the segmentation of telophase pairs, as the patch contains not one, but two separate objects, and their joint centroid lies outside either of the objects. The two daughter nuclei trigger the pixel-level detector described in the previous section at two separate locations, yet need to be assessed as a unified pair. Any two objects within a certain distance of each other may, in fact, constitute a telophase pair, and have to be entered into the candidate list as a jointly centred and segmented patch, as well as separate objects in their own right. This further adds to the class imbalance problem, as many more coincidentally adjacent pairs are listed as candidates. To reduce this burden, two-object segmentation applies the same threshold across both objects, and rejects any patches that do not produce objects of a comparable size (within 30% of each other) and similar intensity (within 8% of pair average) to each other.

Centroid of the segmented object(s) is used as the new location of the detection, and duplicate removal is applied to filter out patches which started as separate hits of the pixel-colour detector, but converged onto the same position following segmentation and the positional refinement that it brings. Examples of segmentation results for various categories of objects are shown in Fig. 4.13.



(a) Single Mitotic (b) Telophase

(c) Single Negative (d) Coincident Negative Pair

Figure 4.13: Examples of segmentation for candidate patches from the AMIDA dataset.

To reduce the class imbalance, minimum limits are imposed on object area (50 pixels) and on contrast between foreground and background means (background at least 70% brighter than foreground), filtering out many small or faint negative candidates without any loss of true positives. This step produces a total of 100.5K negative patches for the AMIDA training set, of which 75K are single objects, and the rest pairs. For comparison, the training portion of this data set contains 550 positive examples, of which 30 are pairs.

4.2.4 Feature extraction

The following rotation-invariant features are calculated for each segmented object and the surrounding greyscale patch:

- Area, in number of pixels
- Circularity, calculated as perimeter squared over area
- Convex hull area as proportion of the object area
- Elongation of minimal-area fitted rectangle, calculated as major axis over minor axis
- Fourier Descriptors, based on radial profile of segmented shape, 64 points. The first 5 terms are used as individual attributes (normalised by the DC term), and the rest are added together as high-frequency total.
- Contrast ratio between background and foreground means, excluding white holes.

- Average "depth" of object relative to segmentation threshold, calculated as the difference between the threshold and the foreground mean.
- Morphology of slices at $1/3$ and $2/3$ of total depth (between segmentation threshold and minimum value inside mask). At each depth, we calculate the average area per contiguous object, as proportion of the overall object area. This gives an indication of how quickly the object breaks up internally with lowering threshold.
- Average gradient across the segmentation boundary, indicating the sharpness and contrast of the edge
- Average contrast-independent edge sharpness across the boundary, measured as ratio of ± 1 pixel gradient to ± 2 pixel gradient at each point
- Standard deviation inside the object
- Mean local variance inside the object, measured on densely sampled 7x7 patches that lie wholly within the object mask
- High-band energy inside the object, measured as $\|I - I_{low}\|_2$ where I_{low} is the output of 7x7 low-pass box filter and the norm is computed with a mask which has been eroded with a 3x3 structuring element in order to reduce edge effects
- Average local variance of background, measured on densely sampled local patches of size 5x5 that lie wholly outside the object mask
- Average low-band energy of background, $\|I_{low}\|_2$ for same 7x7 box filter as the object's high-band
- Ratio of high-band and low-band energy for background areas (outside a dilated object mask)

Of these, around a third relate to shape of the object, a third measure intensity parameters, and the last third describe certain aspects of the textures inside and outside the object. A total of 23 attributes is combined to represent all the relevant aspects of the object and its context in a single feature vector. Each feature is normalised to zero mean and unit variance prior to using it for classification.

4.2.5 Classification

The biggest challenge for classification in this application is class imbalance, closely followed by the sheer number of training points coming from the detection stage. For single objects the imbalance is 150:1, and for pairs it is 800:1 (AMIDA dataset). Inspired by the success of under-sampling methods proposed in [146], we address both imbalance and size of dataset for single objects by employing dominant (negative) class sub-sampling with model averaging. The negative part

of the training set is split into a number of partitions, and each one is combined with all of the positive examples to train an RBF SVM, with appropriate class weights to compensate for the remaining imbalance. The predictions from all the models are averaged to give the final estimate.

Model parameters, consisting of SVM hyper-parameters and the decision threshold applied to the average score, were selected using a cross-validation procedure, based on leaving out all cells from one patient as the validation set (leave-one-out). Experiments were also performed to assess the impact of changing the number of negative-example partitions, and therefore classifiers in the ensemble.

For pairs, each object is first assessed by the single-object classifier, and if at least one of the constituent parts has a high enough prediction, the pair is assessed further on its pair-specific features. This filtering step reduces class imbalance for pairs and produces a dataset which is sufficiently small that no model averaging is needed. In addition to object attributes described in Section 4.2.4, pairs are characterised by ratio and average of a subset of the parameters from each of the objects, which assess their compatibility as a pair: area, contrast, circularity, depth and elongation. In addition, the total of the two single-object prediction scores is used as an extra feature. A separate RBF SVM is then trained for performing identification of telophase pairs.

4.2.6 GP-LVM detection of mitosis

Our application of Gaussian Process Latent Variable Models, described in Section 4.1.3, to detection of mitotic figures follows the same initial steps of stain-normalisation (Section 4.2.1), and candidate detection (Section 4.2.2) as the traditional features-plus-classifier path. Pixel values of the resulting patches then serve as observed measurements of the model.

To improve the correlation between pixel values from different samples, the segmentation algorithm of Section 4.2.3 was augmented to include rotational alignment of the segmented object(s). The necessary angle of rotation is determined from PCA of pixel coordinates that make up the segmented object area, and a rotated patch is extracted from the full field image, although the segmentation masks themselves are not directly used in the modelling. Rotating all objects to spatially align along the same axis saves the need for complex modelling of the angle of rotation as one of the latent variables, and gives a more consistent meaning to the intensity of a pixel at a particular position: all elongating metaphase nuclei affect the values of the same set of pixels, reinforcing each other's contributions.

To further reinforce spatial connections between neighbouring pixels, which would ordinarily be treated by GP-LVM as completely independent dimensions of

the input space, we construct a scale-space pyramid of each patch and augment the input vector with pixel values from the higher levels of the pyramid. As each coarser pixel represents a combination of several pixels from the lower level, it provides additional information on the overall layout of a patch and correlations within it which would otherwise be inaccessible to the GP-LVM as it models each observed dimension (image pixel) independently.

The main challenges faced in applying GP-LVM to this problem are computational: a training set of 550 positive examples, of nearly 5'000 dimensions each, is already at the upper limit of what is currently possible to construct a GP-LVM for; the addition of over a hundred thousand negative examples puts it squarely into the realm of the impossible. Our solution was to build two separate models, one for the positive manifold, and another for a heavily sub-sampled selection of negative examples; the detector's decision is then based on whichever model predicts a higher likelihood, with weights to compensate for the original class imbalance.

The sheer number of test points generated by the candidate detection mechanism also presented a computational problem, as the latent position of each one had to be calculated by an iterative optimisation, for each of the positive and the negative models. It is not possible to sub-sample the test set as this would randomly miss a high proportion of the positive candidates. This severely limited the scope for experimentation with different options and settings of the algorithm, reducing the final performance.

The final thorn in applying GP-LVM to this challenging scenario is selection of the appropriate noise level within the model. For smaller datasets the most suitable noise level can be determined by a brute-force search across a certain range, to find one that gives the best converged model. However, on this large dataset exhibiting strong textural variations, this proved impossible as none of the noise settings within the normal range (10-30 dB) could produce a converging model. Both the positive and the negative models had to use the extremely high noise setting of 2 dB in their training.

4.3 Results

Most of the detailed results presented here are for the AMIDA dataset. For comparison, both test results reported in [65] and training accuracy from our own experiments on the MITOS dataset show much higher levels of mitotic figure recognition: above 70% and around 65% respectively. However, the significance of these results is compromised by the training-test split which uses the same patient samples across both partitions, and by the general homogeneity as well as low number of samples in the MITOS dataset.

Results for the traditional feature extraction pipeline, as described in sections 4.2.1 through to 4.2.5, are detailed in the following section, while those for the GP-LVM approach are given in Section 4.3.2.

4.3.1 Extracted Features pipeline

Patient-based leave-one-out cross-validation on the training set was performed to establish the effect (if any) of the number of negative partitions on the performance of the single-object classifier ensemble. A separate optimisation of SVM hyper-parameters and decision threshold was carried out for each of the sub-sampling ratios tested. As can be seen from Table 4.1, the effect is negligible up to ratios of 50:1, so as a compromise between performance and training speed, 30 classifiers were used. Please note that the percentages in this table are not directly comparable with those in Table 4.2, as they are measured in proportion to the number of single-object positive examples, not the overall total. Experiments were also performed to evaluate the effect of having additional classifiers in the ensemble, for the same level of random sub-sampling of the dominant class; these showed no discernible benefit.

Sub-sampling Ratio:	10:1	15:1	30:1	50:1	300:1
F-score	45.2%	45.1%	45.0%	44.9%	43.0%

Table 4.1: Effect of dominant class sub-sampling and model averaging on cross-validation accuracy for single object patches (AMIDA dataset).

Selection of parameters for the pairs classifier was done taking into account the numbers of true positives, false positives and false negatives produced by the single-object classifier, and looking for highest overall F-score. It was found that this resulted in the same parameter values as those optimised for best F-score among the pair samples only, although the two methods would not necessarily agree in the general case. Selection of threshold for the pair filter, applied to individual object scores in order to decide whether the pair is deserving of further assessment, was guided by balance between its impact on the number of false negatives (rejection of positive examples where one half is unusually small or faint) and on the class imbalance (letting through a great flood of coincidentally close negative examples). The final choice of threshold value of 0.4 results in missing 7 positive examples (out of a total of 550) and reduces class imbalance for pairs from over 800:1 to around 30:1 (depending on the exact configuration of the single-object classifier). As the number of telophase pairs in the training set in relatively small, such class imbalance does not result in a computationally difficult size of training set, but class weights are needed in the SVM to cope

with this degree of imbalance. The contribution of the pair classification stage to the overall error rates can only be evaluated on the training set, as ground-truth for the test set remains undisclosed; the cross-validation F-score for pairs alone is around 38%.

Entry	Training F-score	Test Precision	Test Recall	Test F-score
#1	45.2%	41.2%	26.5%	32.2%
#2	43.8%	38.2%	28.0%	32.3%
#3	44.1%	35.7%	33.2%	34.4%

Table 4.2: Summary of cross-validation and test results for the AMIDA challenge submissions.

In total, three separate submissions were made for the challenge, with slightly different optimisations of model parameters, and resulting in different test scores, which are summarised in Table 4.2. Submission #1 was optimised for best overall F-score, which gives much higher weight to performance on patient samples that contain a large number of mitotic figures, and therefore can perform quite poorly on the low-grade cases with few mitoses. In an attempt to even out the performance across different samples, submission #2 was optimised to give the best average score based on equal weight for each patient. Although this approach improves the balance between precision and recall on the test set, the overall F-score is unaffected. Both of these submissions show a considerable gap between precision and recall, which is not evident in the training cross-validation results, so the final submission deliberately favours recall at the expense of precision. At the operating point measured by cross-validation, an improvement of one in the number of false negatives can be offset by a deterioration of five in the false positives and maintain the same F-score, due to the specific formulation of F-measure as a function of precision and recall. Submission #3 therefore lowers the decision threshold, as compared to the optimal value for best cross-validation score, to allow five times more extra false positives than it loses from true positives; although the cross-validation precision drops to 37.2%, recall rises to 54.2%. Most importantly, the desired effect of better balance between precision and recall on the test set is achieved, and gives an overall boost for the test F-score.

An analysis of the total number of detections for each test patient case, and the corresponding density of mitoses per unit area, shows a 0.82 correlation to the true mitotic density, see Fig. 4.14.

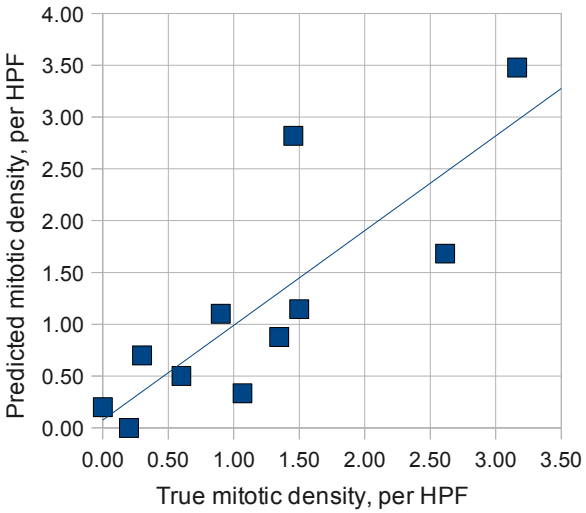
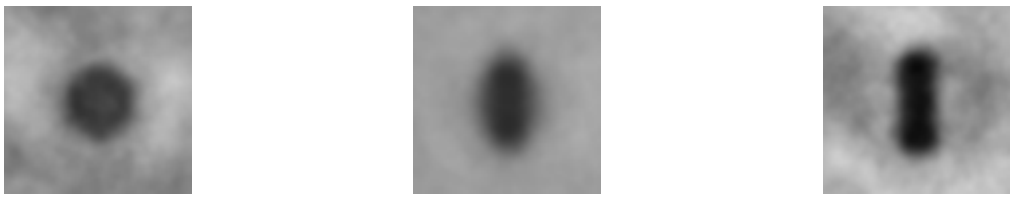


Figure 4.14: Correlation between true and predicted mitotic density for patients in AMIDA test set, and the linear regression line of the two densities.

4.3.2 GP-LVM pipeline

A single submission was made for this computationally expensive method, resulting in test precision of 11.9%, recall of 10.7%, and a combined F-score of 11.3%. Some illustrative examples of the latent axes found by the GP-LVM positive model are shown in Fig. 4.15. Other axes are related to object contrast, curvedness or the splitting into two symmetrical parts observed in telophase. These images are sampled from the generative model, and therefore not present in the training set.



(a) Large -ve (b) Zero (c) Large +ve

Figure 4.15: Images reconstructed by the positive GP-LVM at varying positions along latent dimension 8, clearly related to vertical elongation of the object.

4.3.3 AMIDA Contest

For contextual comparison, we include a summary of the AMIDA challenge results, published in full in [141].

Team name	Precision	Recall	F_1-Score
IDSIA	0.610	0.612	0.611
DTU	0.427	0.555	0.483
SURREY	0.357	0.332	0.344
ISIK	0.306	0.351	0.327
PANASONIC	0.336	0.310	0.322
CCIPD/MINDLAB	0.353	0.291	0.319
WARWICK	0.171	0.552	0.261
POLYTECH/UCLAN	0.186	0.263	0.218
MINES	0.139	0.490	0.217
SHEFFIELD/SURREY	0.119	0.107	0.113
NTUST	0.011	0.685	0.022

Table 4.3: Summary of all entries for the AMIDA challenge, giving precision, recall and F_1-score for overall numbers of detections.

The top two entries of Table 4.3 have been described in detail in sections 4.1.1 and 4.1.2, with our proposed method coming third. Many of the lower ranked methods used similar detection pipelines with initial candidate locations chosen on the basis of colour, followed by extraction of hand-picked features and some form of supervised learning, but none included either explicit stain normalisation or special treatment of telophase pairs.

4.4 Discussion

The cross-validation accuracies listed in Table 4.2 are over 10% higher than the corresponding test results, and similar gaps of 5-10% between validation and test were reported by other participants in the challenge. Such large discrepancies in performance estimation suggest that the training set does not fully represent all the variations present in the underlying data, as the test set offers additional, unforeseen, challenges. To overcome the difficulty of obtaining additional labelled images that would increase the representational coverage of the training set it may be possible to use semi-supervised methods to leverage large quantities of unlabelled images, which are more easily available, to build a richer model of possible tissue appearances.

The histogram adjustment stage of the pipeline has proved to be very valuable: as the only entry to have an explicit normalisation step, the method achieved a much higher ranking for certain cases with exceptionally light or heavy staining which caused difficulty for other methods. The candidate point detection is probably sub-optimal, and many participants reported less extreme class imbalance ratios for the training set through better seed-point selection methods. Alternative means of selecting promising patches for further detailed assessment, such as a suitable variant of the "objectness" measure, could improve both the class imbalance and the final accuracy of the system [147].

Although no data is available for comparison, it could well be the case that additional care to cope with the complexities of telophase pairs throughout the chain has given this method the edge over other algorithms of similar structure. One unfortunate drawback of the separate treatment of single objects and pairs has been the inability to provide a unified prediction score for every patch which could be subjected to a single threshold to produce the final decision. The challenge organisers allowed submission of probabilistic predictions for every candidate location, which would have enabled production of complete ROC curves, but only one threshold could be submitted for actual competition performance measurement for all cases, whereas our separate single and pair classifiers require separate decision thresholds.

The optimisation of SVM hyper-parameters provided an unexpected difficulty: the results are evaluated by F-measure, but the conventional search strategy of coarse-to-fine grid fails to find the global optimum, because the F-score surface, unlike pure error rate, is not convex. This is even more pronounced for the averaged per-patient scores, as a single different decision in a sample with very few or no true positives can massively alter the measured performance of the whole system. A compromise measure of performance, where each patient score is weighted by the number of HPF images available for that patient, as opposed to equal weight for each patient score, smooths out some of the more extreme sensitivities.

Predictive performance of the GP-LVM proved hugely disappointing. To circumvent the need for a manually crafted feature set of relevant attributes would have been a major gain for the application area, as well as a notable expansion of valuable applications for GP-LVMs. So far, the data has proven to be too noisy, as well as too large, for current GP-LVM implementations to successfully digest. It is also possible that this problem domain is inherently so complex as to require some form of deep learning, as even the most non-parametric of models cannot connect the latent variations to their image expressions in one step. Although the latent axes found by the current model look promising, and can be ascribed meanings related to nucleus shape or intensity, they entirely fail to deal with the textural aspects of the images.

To extend the work in the future, several improvements can be suggested: bigger patches would allow more information to be gathered from the surrounding context and also avoid cropping of some widely separated pairs. Full colour processing, or two channels for separate eosin and hematoxylin signals, could boost performance as texture of the eosin-stained protein in the surrounding area may be informative.

Our overall performance of 3rd place out of 14 entries shows the merits of this carefully constructed algorithm, but the significantly higher performance of the leading entries suggests that more unconventional measures are needed in order to solve this highly complex and ambiguous challenge. Deep learning, whether based on convolutional networks or other learners such as Gaussian Process, has shown a lot of potential in many recent works, and definitely removes the guesswork inherent in manual selection of the most relevant image features. In the case of this particular clinical application, the most important goal, however, is not the precise localisation of mitotic figures, but a measurement of their density. The density scatter in Fig. 4.14 is very similar to those obtained when comparing manual mitotic counts from light microscopy and from digital slides: there is greater spread at higher densities and tighter correlation at low grades [148]. The correlation figure of 0.82 confirms that even a relatively low F-score for detection of individual mitotic cells can be of great benefit in the diagnostic task of tumour grading.

Chapter 5

Reflection

The work described in this thesis has covered a range of applications for automated analysis of microscopic histopathology images. On first inspection, the very attempt is a little foolhardy, as, unlike in the majority of vision tasks, the human visual system itself struggles to deliver a consistent and unambiguous answer, thus depriving us of the luxury of reliable ground truth. But the reward for success is correspondingly greater, as it allows objective measurements to become the basis of diagnostic decisions in a repeatable mechanism.

Automation of histopathology image analysis is finally attracting the attention of researchers, and has enjoyed an increasing amount of coverage and cooperation between medical specialists and computer scientists [149–151]. The scope of these research projects is very diverse, and they cover all parts of the image-processing pipe-line needed to address the overall goal of diagnostic assistance: auto-focus to ensure acquisition of the best possible images [152], identification or segmentation of tissue types [133, 153, 154], segmentation of individual nuclei [155, 156], analysis of nuclear features and their connections to diagnostic or prognostic labels [80, 157], which can include grading of disease progression or severity [158]. They cover a wide range of microscopic imaging modalities and staining techniques, but often fall into the trap of blindly following the human procedures and steps towards a decision, instead of concentrating on the only solid evidence, which is outcome [159]. The outlook for the future is, however, very positive [160]:

> further advances in image analysis algorithms are warranted in order to fully realize the benefits of digital pathology in medical discovery and patient care. In coming decades, pathology image analysis will extend beyond the streamlining of diagnostic workflows and minimizing interobserver variability and will begin to provide diagnostic assistance, identify therapeutic targets, and predict patient outcomes and therapeutic responses.

We will first reflect on the underlying connections between machine learning paradigms and human learning and decision making, as there is much to be gained from cross-fertilisation between these areas. In the remainder of the chapter we will summarise the contributions made by this work, dissect its limitations and suggest possible ways around them, before drawing the final conclusions.

5.1 Human vs Machine Learning

As there is no such thing as a free lunch [161], assumptions have to be made in order to be able to learn inductively at all, even if very simple ones [162]. There are enormous parallels here with human learning and education systems. All education is social, and involves training with a teacher, who may not be a perfect oracle and therefore produces some label noise, but possesses the know-how of the subject as well as the correct answers which can be used as training examples. This knowledge of how the subject's processes should be conducted corresponds to the structural assumptions which must be made *a priori* to achieve machine learning. Their importance is evident in the drilling to 'show your workings' which is enforced on human students from an early age. Off-training set (OTS) error, the only unbiased estimate of an algorithm's performance corresponds to the unseen examination questions present in the vast majority of educational qualifications, and the variation in pass marks between subjects is quite similar to the range of accuracy figures achieved by state-of-the-art algorithms for problems of different degrees of difficulty.

'Study skills', or meta-learning about how to learn, are the subject of much fundamental research underlying machine learning. Amongst valuable study skills, the ability to distinguish essential information from irrelevant corresponds to robustness of an algorithm to nuisance variables, spotting patterns of similarity is the principal inductive basis of all machine learning, and the capacity to identify gaps in one's own learning translates to a boosting algorithm's increased attention to incorrectly classified examples. As a student's proficiency at study skills is judged by their mastery of a number of a different subjects, so the excellence of a learning algorithm is only evident from its successful application in a broad range of different learning tasks.

Recent growth of interest in learning methods which involve an element of randomness, such as random forests, could be justified by the necessity of sleep for generating truly creative solutions to complex problems, as classically exemplified by the periodic system of elements appearing to Mendeleyev in a dream. The random connections and combinations generated by the sleeping mind and perceived as dreams are, in their vast majority, not useful and are discarded, just as the bulk of randomly generated feature combinations and thresholds are

rejected during the construction of random decision tree, but the process does throw up some gems on occasion.

Classifier ensembles, and the manner in which they are improved by diversity of the constituent hypotheses, are echoes of every panel, council, committee or representative body ever convened by humans to pool their collective wisdom and experience of multiple individuals in order to arrive at a better decision.

Distinctions between specialist and more general skills are reflected in the restrictions placed on the probability distribution whose samples the algorithm is trying to predict. Distinctions between skill and knowledge also find a reflection in machine learning as differences between supervised training for prediction making in a specific task and discovery of more general structure through clustering or visualisation. Finally, human judgement, as precursor to decision and action, is always associated with a valuation of the potential outcomes, which is encoded in machine learning as a loss or utility function [163]. There are also connections between the improvement in long-term neuronal potentiation brought by the excitement and satisfaction of a successfully completed task, through the effect of dopamine and noradrenaline on synapse formation and growth, and machine learning strategies such as reinforcement learning or boosting.

The forthcoming special issue of Pattern Recognition Letters on Philosophical Aspects of Pattern Recognition should provide a timely consolidated view of developments in this area, including connections with epistemology and decision theory.

5.2 Summary of Contributions

The contributions of this work in its three application areas are as follows:

- Improved identification of cell clusters and debris objects in cytology, including segmentation and extraction of features which highlight presence of notches in cluster boundaries.

- Advances in classification of staining patterns in indirect immunofluorescence images, both at cell and at sample level, using texture and shape analysis methods.

- Progress in automated mitosis detection in breast biopsy sections for tumour aggressiveness grading, including stain normalisation and nucleus segmentation methods.

Some of these are based on a shared methodology that is sufficiently flexible to be adapted to multiple domains. For example, the segmentation method

described in Section 2.2.1 involves threshold selection based on a combination of two attributes of the resulting boundary: cross-boundary gradient and contour circularity. The segmentation algorithm used on cell nuclei in Section 4.2.3 is similarly choosing a threshold based on two attributes, but in this case the more suitable combination is one of cross-boundary gradient and internal variance. In both cases the relative weights of the two contributing measurements are determined automatically from the ratio of their respective variances. Contrasting this with other segmentation methods that are based on a cost function with multiple contributing factors, such as the energy function in 'snake' optimisation, whose relative weights have to be tuned to obtain good results, our method is inherently more automatic. As the method is also very fast due to its single search dimension, it is a very versatile approach to threshold selection.

The dictionary construction based on discriminative power in different areas of feature space, described in Section 2.3.1, deserves some extra attention. Although it did not make an enormous difference to the classification performance of the specific application, which is extremely challenging due to high class overlap, it does give a principled basis for drawing quantisation boundaries in feature space when the ultimate goal is one of discrimination, rather than approximation, while being extremely fast in both training and test. A similar principle of class purity increase from a split, measured as information gain, is used in construction of random forests [78], and there may be merit in injecting a degree of randomness and diversity into the discriminative dictionary formulation, perhaps in the order in which dimensions are examined or the position of the putative boundary. This would allow cheap construction of a larger dictionary with multiple trees whose higher dimensionality may lend itself to linear classification methods.

The novel features measuring slope of radial profile around its lowest point, introduced in Section 2.2.2, make a contribution to identification of single nuclei as distinct from clusters or other debris in DAPI images. They are robust to noise and invariant to spatial scaling.

The work on HEp-2 pattern classification highlights the importance of assessing a sample as a whole, and the potential 'cliff-edge' effects that result when individual cells are treated as independent even though they are not. More study remains to be done to determine the best way of utilising the evidence from multiple cells to build up the informational basis for an overall decision in this particular application, but it is essential to remember the patient as primary unit of diagnostic assessment [164]. Interesting parallels can be drawn between the statistical treatment of cells as constituent parts of a patient sample and that of texton patches in a textural bag-of-words model: if an appropriate dictionary could be constructed to represent individual cells and their relevant properties, the distribution of cells in a sample could be viewed in the same way as the distribution of patches in an image, and used to determine the sample class.

Another important contribution arising from the work on HEp-2 pattern is the method for generating a one-dimension texture spectrum from a two-dimensional image, as described under 'DCT based descriptor' in Section 3.2.2. Most DCT processing and analysis is performed in two dimensions, as this retains the full distribution of frequencies within the image. However, for assessment of isotropic textures frequently encountered in pathology imagery the full two-dimensional spectrum spreads the relevant information over too many bins, drowning it in noise. Ideally, the contributions of same spatial frequency at any orientation should be averaged to cancel out some of the noise and boost the desired signal, but this is computationally expensive. Instead, we approximate the same result by sampling the horizontal spectrum at multiple vertical positions and average these to achieve the same improvement in signal-to-noise ratio at a fraction of the computational cost. We also avoid problems associated with the complex shape of the object which can make it difficult to fit a sufficiently large 2-D DCT sampling block without overlapping object edges and significantly distorting the resulting spectrum. As the line sections used to compute the horizontal spectrum can be positioned differently on each line, we can flexibly accommodate complex object shapes and avoid unwanted edge effects without compromising the size of DCT and therefore the resolution of the spectrum.

Our method for stain normalisation of H&E histopathology images, described in Section 4.2.1, also achieves the desired outcome in a very efficient way. By excluding the white areas, whose presence has a strong effect on the colour histograms, but has no bearing on the stain strength and balance, we allow the use of a very fast, robust and simple method of colour balance adjustment, namely histogram matching, for images that were previously treated with complex techniques of logarithmic stain separation.

5.3 Limitations

The chief limitation of any work involving feature extraction is the impossibility of complete proof that the proposed features are the best that could be used. Even the most comprehensive assembly of all known features, followed by the most sophisticated process of feature selection, gives no guarantee that a more suitable feature or set of features, which reflect a more pertinent aspect of the input data or images, will not be dreamt up tomorrow. As the diversity of non-linear functions of the inputs is infinite, we can never try them all, and as we have shown by the experiments with Gaussian Process Latent Variable models (Section 4.3.2), current state-of-the-art techniques for automatic discovery of such functions struggle to produce competitive results in the challenging tasks under consideration.

Limitations of the methods proposed here for each of the tasks of interest largely fall into this category. Some also suffer from a rather limited amount of training and test data: the HEp-2 dataset comes from 28 patient samples, and the AMIDA dataset covers 23 cases. Neither of these are large numbers, but as both are essentially feasibility studies, rather than products ready for clinical introduction, this is understandable.

Ambiguities of ground-truth annotation plague mitosis detection tasks. In order to establish how many of the 'false positives' produced by automated algorithms were actually mitotic figures missed by the initial panel of experts, a follow-up experiment presented a new panel with false positives from the leading two methods, as well as ground-truth labelled mitotic locations as control. Nearly 30% of the 'false positive' detections from the winning method were labelled as mitoses on re-evaluation, while only 71% of 'ground-truth' locations could pass the same re-inspection [141], a figure that is in line with other studies of inter-observer variation in mitotic labelling [56]. We can only guess on how this level of label noise affects the training and performance evaluation of our own method.

Limitations specific to our approach to mitosis detection include the potentially sub-optimal early choice of cut-off level for candidate location detection: this trade-off between class imbalance and false negative bias may have been improved by a lower value of the threshold, allowing more candidate locations through to the segmentation stage, with a large proportion then rejected based on simple area and contrast limits. More generally, the design and development of a pipe-line algorithm such as the mitosis detection one is very prone to optimisation of one part of the chain while others are not yet in their final form. As all the steps are interconnected, and performance can only be measured for the overall process, this brings a danger that design choice or parameter values of a particular step become sub-optimal when other parts are changed to improve their operation. This is a similar problem to issues of feature selection, but on a somewhat larger scale.

The main limitation of all the examined approaches to HEp-2 pattern classification is the lack of a sufficiently robust method for combining the evidence from individual cells into a final decision for a sample. The majority vote approach loses too much information about each cell by forcing a hard class decision prior to combining cells into a sample. The distribution distance makes strong assumptions of normality, and the cumulative histogram reduces the number of training points too far for successful training. It also fails to account for, and therefore learn from, the patterns of variation present within each sample.

5.4 Future directions

One of the most expensive parts of developing a vision system is acquisition of a sufficiently large labelled dataset, especially in medical domains where the expert's time is an extremely limited resource. To address this, pathology image analysis should make greater use of both semi-supervised algorithms, which leverage additional information from unlabelled images, and active learning paradigms which seek user input on cases that would be of greatest benefit to the system's performance. The use of these techniques in the specific application area of pathology image analysis has not been explored to date, but is a very promising direction based on its contribution in other areas of machine learning [165].

Another very promising avenue, particularly when the exact image features of relevance are not clearly understood, which is frequently the case in complex micro-biological imagery, is deep learning. This approach negates the need for a pre-defined feature extraction stage, instead searching out the most pertinent aspects of the images from a vast space of complex, non-linear functions of the input pixels. Deep learning hierarchies can be constructed from different types of underlying machine learning algorithms [63, 66, 166], and have been shown to be very effective in some histopathology applications already [131]. The training of all such algorithms is extremely computationally intensive, and requires GPU acceleration in order to deliver results in days, rather than years.

In the specific case of mitosis detection, boosting techniques, which can give greater weight to the relatively rare configurations and presentations of the nucleus, could significantly improve the overall recognition rates. They would also be able to place greater emphasis on keeping out the apoptotic nuclei that are similar in appearance to mitoses. Sub-class learning may also be of benefit here, for similar reasons of giving the rarer appearance arrangements a chance to be properly represented and recognised.

For HEp-2 pattern classification, it is essential to develop a proper statistical treatment for combining evidence from the interphase cell appearance, which are the majority of cells within the sample, and the few mitotic cells, which have so far been excluded from most datasets and treatments, despite their massive importance in manual determination of staining pattern. Relatively simple Bayesian inference rules may prove sufficient here.

The future of automated pathology image analysis is about more than replicating the manual diagnostic procedures of pathologists, but also using the image analysis and machine learning methods to discover new properties of prognostic significance [160]. It has already been shown that the search for a more direct correlation between pathology images and patient outcomes *can* discover tissue features that have not been previously identified as being of diagnostic importance [4].

5.5 Conclusions

The original intention of this project was to study the application of computer vision methods to pathology images in order to improve the scope for automation of pathology analysis and consequently increase its reliability, repeatability and, in due course, acceptance in clinical use. Ultimately, we have achieved this original goal, producing new algorithms for cell and tissue analysis, as well as investigating dependencies between cell and sample classification. All stages of a computer vision system have been addressed in one form or another across the three problem domains that we have examined: colour pre-processing for stain normalisation, segmentation in both cytopathology and histopathology, feature extraction in every domain of interest, covering a wide variety of shape and texture analysis algorithms, and finally learning itself, particularly dictionary learning in a supervised setting. The significance of these advances is acknowledged by their publication in peer-reviewed conference proceedings and journals [113, 127, 141].

Across all the research areas, the greatest challenge has consistently been one of separately optimising stages in a processing chain, when the optimal parameters for a later stage actually depend on the method and configuration adopted for the previous stages, and vice versa. For example, the choice of features and their number affects the choice of best classifier, but the features themselves cannot be evaluated and selected without using some classification method. The loop is closed because the results of feature evaluation depend on the exact type and parameters of the classifier used. Design of such systems is notoriously difficult and the result fragile and highly susceptible to over-fitting. None of the algorithms that we have produced can claim to be completely optimal, even within the constraints of available data and processing. Only radically different approaches which optimise the entire process as a single entity can hope to break free of these limitations.

The two most promising trends in current computer vision research, random forests and convolutional neural networks, take different routes to freedom from manual feature extraction design: random forests evaluate the information gain of very large numbers of random features, which is independent of classifier type, while deep CNNs use unsupervised learning in their early stages, simply seeking a more compact representation that suits the input data. Neither approach is optimal in every setting, and both require tuning of parameters in order to achieve their best, as well as large quantities of training data and considerable amounts of computation. However, both of these commodities, data and processing power, are more easily obtainable in the long run than additional supplies of specialist expertise in feature extraction, which do not reliably deliver superior results anyway.

In terms of the specific application area of pathology image analysis, we must

additionally look forward to machine learning's ability to discover new correlations between visual characteristics of cells and tissues and the biological processes within those tissues that ultimately determine patient outcomes. This could revolutionise not only the delivery of diagnostic pathology services, but also the research into potential treatments and more personalised medicine.

Acknowledgements

I would first of all like to thank my supervisors, Prof. Josef Kittler and Dr. William Christmas, for their unstinting support and encouragement, as well as valuable direction over the three and a half years of my doctoral study. Other fellow members of the Centre for Speech, Vision and Signal Processing have also provided assistance and insight, along with a broader perspective of the research area, that are much appreciated. The Researcher Development Programme of the University of Surrey has provided many courses that I have found immensely useful and helpful, which have made a real contribution to my wider development as an academic researcher, communicator and teacher.

I am grateful to Ikonisys Inc. for supplying the datasets used in Chapter 2, as well as the background information relating to this problem domain.

I am indebted to Dr. Peter Jackson, Consultant Pathologist at the Royal Surrey County Hospital, for his guidance on the realities of pathology practice and the opportunity to see a non-digital version of H&E slides.

I must acknowledge the huge contribution made by the group of Prof. Neil Lawrence at the Sheffield Institute for Translational Neuroscience to the application of Gaussian Process learning to mitosis detection (Chapter 4). Special thanks go to Dr. Teo de Campos and Andreas Damianou for running immensely long simulations, and Dr. James Hensman for tuition in the theoretical foundations.

My own computational experiments would be considerably harder without the use of WEKA machine learning environment [167], OpenCV image processing library [168] and LibSVM [129].

I acknowledge the generous financial support of the UK Engineering and Physical Sciences Research Council (EPSRC) for both tuition fees and stipend during my period of study.

Finally, I must express my appreciation for the much needed moral support from friends and relatives too numerous to mention, and to my children for their patience while mummy spent endless hours staring at various grey, green or pink blobs.

Bibliography

[1] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. A. Zehnbauer, K. Lister, and R. Parwaresch, "Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index," *Mod Pathol*, vol. 18, no. 8, pp. 1067–1078, 2005.

[2] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, L. C. Collins, and T. J. Colgan, "Review of the current state of whole slide imaging in pathology.," *Journal of Pathology Informatics*, vol. 2, no. 36, 2011.

[3] S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

[4] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, p. 108ra113, 2011.

[5] M. Guican, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *Biomedical Engineering, IEEE Reviews in*, vol. 2, pp. 147–171, 2009.

[6] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.

[7] R. L. Cahn, R. S. Poulsen, and G. Toussaint, "Segmentation of cervical cell images.," *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 681–8, 1977.

[8] H. Borst, W. Abmayr, and P. Gais, "A thresholding method for automatic cell image segmentation.," *Journal of Histochemistry & Cytochemistry*, vol. 27, no. 1, pp. 180–7, 1979.

[9] A. Elmoataz, M. Revenu, and C. Porquet, "Segmentation and classification of various types of cells in cytological images," in *Image Processing and its Applications, 1992., International Conference on*, pp. 385–388, Apr. 1992.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-9, pp. 62–6, 01 1979.

[11] Z. Hou, Q. Hu, and W. Nowinski, "On minimum variance thresholding," *Pattern Recognition Letters*, vol. 27, no. 14, pp. 1732–1743, 2006.

[12] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[13] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, pp. 146–168, 2004.

[14] R. Kohler, "A segmentation system based on thresholding," *Computer Graphics and Image Processing*, vol. 15, no. 4, pp. 319–338, 1981.

[15] Q. Liao and Y. Deng, "An accurate segmentation method for white blood cell images," in *Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on*, pp. 245–248, 2002.

[16] D. Cohen, "On active contour models and balloons," in *CVGIP: Image Understanding*, vol. 53, pp. 211–218, Academic Press, March 1991.

[17] M. Hu, X. Ping, and Y. Ding, "Automated cell nucleus segmentation using improved snake," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 4, pp. 2737–2740 Vol. 4, oct. 2004.

[18] Z. Lu, G. Carneiro, and A. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), vol. 8149 of *Lecture Notes in Computer Science*, pp. 452–460, Springer Berlin Heidelberg, 2013.

[19] M. Nosrati and G. Hamarneh, "Segmentation of cells with partial occlusion and part configuration constraint using evolutionary computation,"

in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), vol. 8149 of *Lecture Notes in Computer Science*, pp. 461–468, Springer Berlin Heidelberg, 2013.

[20] E. Meijering, "Cell segmentation: 50 years down the road," *Signal Processing Magazine, IEEE*, vol. 29, pp. 140–145, Sept 2012.

[21] S. Kothari, Q. Chaudry, and M. Wang, "Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, pp. 795–798, july 2009.

[22] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and cell splitting," *Medical Imaging, IEEE Transactions on*, vol. 30, pp. 1661–1677, Sept 2011.

[23] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[24] M. Yang, K. Kpalma, J. Ronsin, *et al.*, "A survey of shape feature extraction techniques," *Pattern Recognition*, pp. 43–90, 2008.

[25] D. A. V. Amaro, *Statistical Shape Analysis for bio-structures: Local Shape Modelling, Techniques and Applications*. PhD thesis, Warwick University, 2009.

[26] D. Zhang and G. Lu, "A Comparative Study on Shape Retrieval Using Fourier Descriptors with Different Shape Signatures," in *International Conference on Intelligent Multimedia and Distance Learning*, (Fargo, ND, USA), pp. 1–9, June 2001.

[27] M. Mirmehdi, X. Xie, and J. Suri, *Handbook of Texture Analysis*. London, UK, UK: Imperial College Press, 2009.

[28] N. Pressman, R. Haralick, H. Tyrer, and J. Frost, "Texture analysis for biomedical imagery," tech. rep., Dahlem Workshop on Biomedical Pattern Recognition and Image Processing, May 1979.

[29] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-6, no. 4, pp. 269–285, 1976.

[30] B. V. Levienaise-Obadia, J. Kittler, and W. J. Christmas, "Comparative study of strategies for illumination-invariant texture representations," vol. 3656, pp. 653–664, 1998.

[31] C. Christodoulou, S. Michaelides, and C. Pattichis, "Multifeature texture analysis for the classification of clouds in satellite imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 41, no. 11, pp. 2662–2668, 2003.

[32] P. Linares, P. McCullagh, N. Black, and J. Dornan, "Feature selection for the characterization of ultrasonic images of the placenta using texture classification," in *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, pp. 1147–1150 Vol. 2, 2004.

[33] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[34] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[35] S. Heynen, E. Hunter, and J. Price, "Review of cell nuclear features for classification from fluorescence images," *Proceedings of the SPIE - The International Society for Optical Engineering*, vol. 3921, pp. 54–65, 2000.

[36] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975.

[37] T. Kurita and N. Otsu, "Texture Classification by Higher Order Local Autocorrelation," *Proc. ACCV*, vol. 93, pp. 175–178, 1993.

[38] T. Ojala, M. Pietikänen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Computer Vision - ECCV 2000*, vol. 1842 of *Lecture Notes in Computer Science*, pp. 404–420, Springer Berlin Heidelberg, 2000.

[39] T. Ojala, M. Pietikäinen, and T. Mäenpää, "A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification," in *Advances in Pattern Recognition ICAPR 2001* (S. Singh, N. Murshed, and W. Kropatsch, eds.), vol. 2013 of *Lecture Notes in Computer Science*, pp. 399–408, Springer Berlin Heidelberg, 2001.

[40] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[41] W.-H. Liao, "Region description using extended local ternary patterns," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 1003–1006, 2010.

[42] G. Zhao, T. Ahonen, J. Matas, and M. Pietikäinen, "Rotation-invariant image and video description with local binary pattern features," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1465–1477, 2012.

[43] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[44] M. Varma and A. Zisserman, "A Statistical Approach to Material Classification Using Image Patch Exemplars," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, pp. 2032–2047, 2009.

[45] G. V. D. Wouwer, B. Weyn, P. Scheunders, W. Jacob, E. V. Marck, and D. V. Dyck, "Wavelets as Chromatin Texture Descriptors for the Automated Identification of Neoplastic Nuclei," *Journal of Microscopy*, vol. 197(pt 1), pp. 25–35, 2000.

[46] S. Niwas, P. Palanisamy, and K. Sujathan, "Complex wavelet based texture features of cancer cytology images," in *Industrial and Information Systems (ICIIS), 2010 International Conference on*, pp. 348–353, Aug. 2010.

[47] C. Kocur, S. Rogers, L. Myers, T. Burns, M. Kabrisky, J. Hoffmeister, K. Bauer, and J. Steppe, "Using neural networks to select wavelet features for breast cancer diagnosis," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 15, no. 3, pp. 95–102, 108, 1996.

[48] R. Lopes and N. Betrouni, "Fractal and multifractal analysis: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 634–649, 2009.

[49] D. Sabino, E. Nakamura, L. Costa, R. Calado, and M. Zago, "Chromatin texture characterization using multiscale fractal dimension," in *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, vol. 2, pp. 529–533 vol.2, 2002.

[50] P. Soille and J.-F. Rivest, "On the Validity of Fractal Dimension Measurements in Image Analysis," *Journal of Visual Communication and Image Representation*, vol. 7, no. 3, pp. 217–229, 1996.

[51] N. Theera-Umpon and S. Dhompongsa, "Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 11, pp. 353–359, may 2007.

[52] Y. Q. Chen, M. S. Nixon, and D. W. Thomas, "Statistical geometrical features for texture classification," *Pattern Recognition*, vol. 28, no. 4, pp. 537–552, 1995.

[53] R. Walker and P. Jackway, "Statistical geometric features - extensions for cytological texture analysis," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 2, pp. 790–794 vol.2, 1996.

[54] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing* (A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, eds.), vol. 5099 of *Lecture Notes in Computer Science*, pp. 236–243, Springer Berlin Heidelberg, 2008.

[55] V. Arvis, C. Debain, M. Berducat, and A. Benassi, "Generalization of the cooccurrence matrix for colour images: Application to colour texture classification," *Image Analysis & Stereology*, vol. 23, no. 1, 2011.

[56] C. Malon, E. Brachtel, E. Cosatto, H. P. Graf, A. Kurata, M. Kuroda, J. S. Meyer, A. Saito, S. Wu, and Y. Yagi, "Mitotic figure recognition: Agreement among pathologists and computerized detector," *Analytical Cellular Pathology*, vol. 35, no. 2, pp. 97–100, 2012.

[57] S. Fefilatyev, M. Shreve, K. Kramer, L. Hall, D. Goldgof, R. Kasturi, K. Daly, A. Remsen, and H. Bunke, "Label-noise reduction with support vector machines," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3504–3508, 2012.

[58] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *SIGKDD Explor. Newsl.*, vol. 6, pp. 40–49, June 2004.

[59] W. Liu and S. Chawla, "Class confidence weighted kNN algorithms for imbalanced data sets," in *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD'11, (Berlin, Heidelberg), pp. 345–356, Springer-Verlag, 2011.

[60] D. Yin, C. An, and H. Baird, "Imbalance and concentration in k-NN classification," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 2170–2173, 2010.

[61] N. Thacker, "Defining probability for science," tech. rep., TINA Memo 2007-008, 2007.

[62] R. M. Haralick, "Performance characterization in computer vision," in *Computer Analysis of Images and Patterns* (D. Chetverikov and W. Kropatsch, eds.), vol. 719 of *Lecture Notes in Computer Science*, pp. 1–9, Springer Berlin Heidelberg, 1993.

[63] Y. Bengio and A. Courville, "Deep Learning of Representations," in *Handbook on Neural Information Processing* (M. Bianchini, M. Maggini, and L. C. Jain, eds.), vol. 49 of *Intelligent Systems Reference Library*, pp. 1–28, Springer Berlin Heidelberg, 2013.

[64] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[65] R. Ludovic, R. Daniel, L. Nicolas, K. Maria, I. Humayun, K. Jacques, C. Frédérique, G. Catherine, L. Gilles, N. Metin, *et al.*, "Mitosis detection in breast cancer histological images An ICPR 2012 contest," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 8, 2013.

[66] A. C. Damianou and N. D. Lawrence, "Deep Gaussian Processes," *JMLR*, 2013.

[67] L. Van der Maaten, E. Postma, and H. Van den Herik, "Dimensionality reduction: A comparative review," *Technical Report TiCC TR 2009-005*, 2009.

[68] A. J. Izenman, "Introduction to manifold learning," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 5, pp. 439–446, 2012.

[69] M. Titsias and N. Lawrence, "Bayesian Gaussian process latent variable model," *JMLR*, 2010.

[70] J. Snoek, R. P. Adams, and H. Larochelle, "On nonparametric guidance for learning autoencoder representations," *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.

[71] R. Urtasun and T. Darrell, "Discriminative gaussian process latent variable model for classification," in *Proceedings of the 24th international conference on Machine learning*, pp. 927–934, ACM, 2007.

[72] X. Gao, X. Wang, D. Tao, and X. Li, "Supervised gaussian process latent variable model for dimensionality reduction," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 2, pp. 425–434, 2011.

[73] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI2013)*, 2013.

[74] B. Nielsen, F. Albregtsen, and H. Danielsen, "Low dimensional adaptive texture feature vectors from class distance and class difference matrices," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 1, pp. 73–84, 2004.

[75] S. Lloyd, "Least squares quantization in PCM," *Information Theory, IEEE Transactions on*, vol. 28, no. 2, pp. 129–137, 1982.

[76] S. Lazebnik and M. Raginsky, "Supervised Learning of Quantizer Codebooks by Information Loss Minimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, pp. 1294–1309, july 2009.

[77] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 1–8, 2008.

[78] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*. Advances in Computer Vision and Pattern Recognition, Springer London, 2013.

[79] M. Yaqub, M. Javaid, C. Cooper, and A. Noble, "Investigation of the role of feature selection and weighted voting in random forests for 3D volumetric segmentation," *Medical Imaging, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[80] E. Cosatto, M. Miller, H. Graf, and J. Meyer, "Grading nuclear pleomorphism on histological micrographs," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, dec. 2008.

[81] M. Horn and M. Berthold, "Towards active segmentation of cell images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 177–181, 30 2011-april 2 2011.

[82] O. Oechsle, *Towards the Automatic Construction of Machine Vision Systems using Genetic Programming*. PhD thesis, University of Essex, 2009.

[83] C.-H. Chan, J. Kittler, and M. Tahir, "Kernel fusion of multiple histogram descriptors for robust face recognition," in *Structural, Syntactic, and Statistical Pattern Recognition* (E. Hancock, R. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, eds.), vol. 6218 of *Lecture Notes in Computer Science*, pp. 718–727, Springer Berlin Heidelberg, 2010.

[84] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, pp. 78–87, Oct. 2012.

[85] D. Zink, A. H. Fischer, and J. A.Nickerson, "Nuclear structure in cancer cells," *Nature Reviews*, vol. 4, pp. 677–687, Sept 2004.

[86] S. Mohapatra, D. Patra, and S. Satpathi, "Image analysis of blood microscopic images for acute leukemia detection," in *Industrial Electronics, Control Robotics (IECR), 2010 International Conference on*, pp. 215–219, 2010.

[87] T. Kiyuna, A. Saito, E. Kerr, and W. Bickmore, "Characterization of chromatin texture by contour complexity for cancer cell classification," in *BIBE 2008 (8th IEEE International Conference on BioInformatics and BioEngineering, 2008)*, pp. 1–6, Oct. 2008.

[88] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *Communications, IEEE Transactions on*, vol. 28, pp. 84–95, Jan 1980.

[89] J. Krapac, J. Verbeek, and F. Jurie, "Learning tree-structured descriptor quantizers for image categorization," in *British Machine Vision Conference*, 2011.

[90] A. S. Wiik, M. Høier-Madsen, J. Forslid, P. Charles, and J. Meyrowitsch, "Antinuclear antibodies: a contemporary nomenclature using HEp-2 cells," *Journal of autoimmunity*, vol. 35, no. 3, pp. 276–290, 2010.

[91] P. L. Meroni, M. Biggioggero, S. S. Pierangeli, J. Sheldon, I. Zegers, and M. O. Borghi, "Standardization of autoantibody testing: a paradigm for serology in rheumatic diseases," *Nature Reviews Rheumatology*, 2013.

[92] N. Agmon-Levin, J. Damoiseaux, C. Kallenberg, U. Sack, T. Witte, M. Herold, X. Bossuyt, L. Musset, R. Cervera, A. Plaza-Lopez, C. Dias, M. José Sousa, A. Radice, C. Eriksson, O. Hultgren, M. Viander,

M. Khamashta, S. Regenass, L. E. Coelho Andrade, A. Wiik, A. Tincani, J. Rönnelid, D. B. Bloch, M. J. Fritzler, E. K. L. Chan, I. Garcia-De La Torre, K. N. Konstantinov, R. Lahita, M. Wilson, O. Vainio, N. Fabien, R. A. Sinico, P. Meroni, and Y. Shoenfeld, "International recommendations for the assessment of autoantibodies to cellular antigens referred to as anti-nuclear antibodies," *Annals of the Rheumatic Diseases*, 2013.

[93] N. Bizzaro, R. Tozzoli, E. Tonutti, A. Piazza, F. Manoni, A. Ghirardello, D. Bassetti, D. Villalta, M. Pradella, and P. Rizzotti, "Variability between methods to determine ana, anti-dsdna and anti-ena autoantibodies: a collaborative study with the biomedical industry," *Journal of Immunological Methods*, vol. 219, no. 1–2, pp. 99–107, 1998.

[94] P. Perner, "Image analysis and classification of hep-2 cells in fluorescent images," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 2, pp. 1677–1679 vol.2, 1998.

[95] Y.-L. Huang, Y.-L. Jao, T.-Y. Hsieh, and C.-W. Chung, "Adaptive automatic segmentation of hep-2 cells in indirect immunofluorescence images," in *Sensor Networks, Ubiquitous and Trustworthy Computing, 2008. SUTC '08. IEEE International Conference on*, pp. 418–422, 2008.

[96] U. Sack, S. Knoechner, H. Warschkau, U. Pigla, F. Emmrich, and M. Kamprad, "Computer-assisted classification of hep-2 immunofluorescence patterns in autoimmune diagnostics," *Autoimmunity Reviews*, vol. 2, no. 5, pp. 298–304, 2003.

[97] C. Plata, H. Perner, S. Speth, K. J. Lackner, and P. von Landenberg, "Automated classification of immunofluorescence staining of hep-2 cells in clinical routine diagnostics," *Transactions on Mass-Data Analysis of Images and Signals*, vol. 1, no. 2, pp. 147–159, 2009.

[98] K. Egerer, D. Roggenbuck, R. Hiemann, M.-G. Weyer, T. Buttner, B. Radau, R. Krause, B. Lehmann, E. Feist, and G.-R. Burmester, "Automated evaluation of autoantibodies on human epithelial-2 cells as an approach to standardize cell-based immunofluorescence tests," *Arthritis Research & Therapy*, vol. 12, no. 2, p. R40, 2010.

[99] P. Perner, H. Perner, and B. Müller, "Mining knowledge for hep-2 cell image classification," *Artificial Intelligence in Medicine*, vol. 26, no. 1, pp. 161–173, 2002.

[100] R. Hiemann, N. Hilger, J. Michel, J. Nitschke, A. Böhm, U. Anderer, M. Weigert, and U. Sack, "Automatic analysis of immunofluorescence

patterns of HEp-2 cells," *Annals of the New York Academy of Sciences*, vol. 1109, no. 1, pp. 358–371, 2007.

[101] P. Soda and G. Iannello, "A Hybrid Multi-Expert Systems for HEp-2 Staining Pattern Classification," in *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pp. 685–690, sept. 2007.

[102] T.-Y. Hsieh, Y.-C. Huang, C.-W. Chung, and Y.-L. Huang, "HEp-2 cell classification in indirect immunofluorescence images," in *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*, pp. 1–4, dec. 2009.

[103] N. Bizzaro, A. Antico, S. Platzgummer, E. Tonutti, D. Bassetti, F. Pesente, R. Tozzoli, M. Tampoia, and D. Villalta, "Automated antinuclear immunofluorescence antibody screening: A comparative study of six computer-aided diagnostic systems," *Autoimmunity Reviews*, 2013.

[104] "HEp-2 cells classification contest," 2012. http://nerone.diiie.unisa.it/hep2contest/.

[105] "Competition on cells classification by fluorescent image analysis." Hosted at ICIP, 2013. http://nerone.diiie.unisa.it/contest-icip-2013.

[106] P. Agrawal, M. Vatsa, and R. Singh, "HEp-2 cell image classification: A comparative analysis," in *Machine Learning in Medical Imaging* (G. Wu, D. Zhang, D. Shen, P. Yan, K. Suzuki, and F. Wang, eds.), vol. 8184 of *Lecture Notes in Computer Science*, pp. 195–202, Springer International Publishing, 2013.

[107] "SNP HEp-2 dataset," Sept 2013. http://itee.uq.edu.au/~lovell/snphep2/.

[108] P. Foggia, G. Percannella, P. Soda, and M. Vento, "Benchmarking HEp-2 cells classification methods," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 10, pp. 1878–1889, 2013.

[109] S. Di Cataldo, A. Bottino, E. Ficarra, and E. Macii, "Applying textural features to the classification of HEp-2 cell patterns in IIF images," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3349–3352, 2012.

[110] I. Ersoy, F. Bunyak, J. Peng, and K. Palaniappan, "HEp-2 cell classification in IIF images using Shareboost," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3362–3365, 2012.

[111] S. Ghosh and V. Chaudhary, "Feature analysis for automatic classification of HEp-2 florescence patterns : Computer-aided diagnosis of auto-immune diseases," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 174–177, 2012.

[112] K. Li, J. Yin, Z. Lu, X. Kong, R. Zhang, and W. Liu, "Multiclass boosting SVM using different texture features in HEp-2 cell staining pattern classification," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 170–173, 2012.

[113] V. Snell, W. Christmas, and J. Kittler, "Texture and shape in fluorescence pattern identification for auto-immune disease diagnosis," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3750–3753, 2012.

[114] P. Strandmark, J. Ulén, and F. Kahl, "HEp-2 Staining Pattern Classification," in *ICPR*, 2012.

[115] G. Thibault and J. Angulo, "Efficient statistical/morphological cell texture characterization and classification," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 2440–2443, 2012.

[116] W. Bel Haj Ali, D. Giampaglia, M. Barlaud, P. Piro, R. Nock, and T. Pourcher, "Classification of biological cells using bio-inspired descriptors," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3353–3357, Nov 2012.

[117] "MIVIA HEp-2 Images Dataset," 2012. http://mivia.unisa.it/datasets/biomedical-image-datasets/hep2-image-dataset/.

[118] S. D. Cataldo, A. Bottino, I. U. Islam, T. F. Vieira, and E. Ficarra, "Subclass discriminant analysis of morphological and textural features for HEp-2 staining pattern classification," *Pattern Recognition*, no. 0, pp. –, 2013.

[119] M. Zhu and A. Martinez, "Subclass discriminant analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 8, pp. 1274–1286, 2006.

[120] R. Nosaka and K. Fukui, "HEp-2 cell classification using rotation invariant co-occurrence among local binary patterns," *Pattern Recognition*, no. 0, pp. –, 2013.

[121] X. Kong, K. Li, J. Cao, Q. Yang, and L. Wenyin, "HEp-2 cell pattern classification with discriminative dictionary learning," *Pattern Recognition*, no. 0, pp. –, 2013.

[122] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "HEp-2 cells classification via sparse representation of textural features fused into dissimilarity space," *Pattern Recognition*, no. 0, pp. –, 2013.

[123] A. Wiliem, Y. Wong, C. Sanderson, P. Hobson, S. Chen, and B. Lovell, "Classification of human epithelial type 2 cell indirect immunofluoresence images via codebook based descriptors," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pp. 95–102, 2013.

[124] M. Faraki, M. T. Harandi, A. Wiliem, and B. C. Lovell, "Fisher tensors for classifying human epithelial cells," *Pattern Recognition*, no. 0, pp. –, 2013.

[125] A. Wiliem, C. Sanderson, Y. Wong, P. Hobson, R. F. Minchin, and B. C. Lovell, "Automatic classification of human epithelial type 2 cell indirect immunofluorescence images using cell pyramid matching," *Pattern Recognition*, no. 0, pp. –, 2013.

[126] Y. Yang, A. Wiliem, A. Alavi, B. C. Lovell, and P. Hobson, "Visual learning and classification of human epithelial type 2 cell images through spontaneous activity patterns," *Pattern Recognition*, no. 0, pp. –, 2013.

[127] V. Snell, W. Christmas, and J. Kittler, "HEp-2 fluorescence pattern classification," *Pattern Recognition*, vol. 47, no. 7, pp. 2338–2347, 2014.

[128] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "HEp-2 Cells classification via fusion of morphological and textural features," in *Bioinformatics Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, pp. 689–694, nov. 2012.

[129] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[130] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1005–1018, 2007.

[131] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pp. 411–418, Springer, 2013.

[132] A. Khan, H. El-Daly, and N. Rajpoot, "A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 149–152, 2012.

[133] A. Khan, H. El-Daly, E. Simmons, and R. NM., "HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images," *Journal of Pathology Informatics*, vol. 4, no. 1, 2013.

[134] H. Irshad, S. Jalali, L. Roux, D. Racoceanu, L. J. Hwee, G. Le Naour, and F. Capron, "Automated mitosis detection using texture, sift features and hmax biologically inspired approach," *Journal of pathology informatics*, vol. 4, no. Suppl, 2013.

[135] H. Irshad, "Automated mitosis detection in histopathology using morphological and multi-channel statistics features," *Journal of pathology informatics*, vol. 4, 2013.

[136] C. D. Malon and E. Cosatto, "Classification of mitotic figures with convolutional neural networks and seeded blob features," *Journal of pathology informatics*, vol. 4, 2013.

[137] F. B. Tek, "Mitosis detection using generic features and an ensemble of cascade adaboosts," *Journal of pathology informatics*, vol. 4, 2013.

[138] A. Albayrak and G. Bilgin, "Detection of mitotic cells in histopathological images using textural features," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, pp. 1–4, 2013.

[139] E. Aptoula, N. Courty, and S. Lefevre, "Mitosis detection in breast cancer histological images with mathematical morphology," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*, pp. 1–4, 2013.

[140] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. Woosley, X. Guan, C. Schmitt, and N. Thomas, "A method for normalizing histology slides for quantitative analysis," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, pp. 1107–1110, 2009.

[141] M. Veta, P. J. van Diest, M. A. Viergever, H. Wang, A. Madabhushi, F. Gonzalez, A. A. C. Roa, A. B. L. Larsen, J. S. Vestergaard, A. B. Dahl, D. C. Cireșan, J. Schmidhuber, A. Giusti, L. M. Gambardella, F. B. Tek, T. Walter, C.-W. Wang, S. Kondo, B. J. Matuszewski, F. Precioso, V. Snell, J. Kittler, T. E. de Campos, A. M. Khan, N. M. Rajpoot, E. Arkoumani, M. M. Lacle, S. Willems, and J. P. W. Pluim, "Assessment of mitosis detection algorithms in breast cancer histopathology images." Submitted to 'Medical Image Analysis', 2014.

[142] H. Chang, L. A. Loss, and B. Parvin, "Nuclear segmentation in H and E sections via multi-reference graph-cut (MRGC)," in *International Symposium Biomedical Imaging (ISBI)*, 2012.

[143] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *J. Mach. Learn. Res.*, vol. 6, pp. 1783–1816, Dec. 2005.

[144] A. Khan, H. El-Daly, and N. Rajpoot, "A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 11, 2013.

[145] D. Magee, D. Treanor, P. Chomphuwiset, and P. Quirke, "Context aware colour classification in digital microscopy.," in *Proceedings Medical Image Understanding and Analysis*, pp. 1–5, BMVA, 2010.

[146] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3738–3750, 2012.

[147] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, pp. 2189–2202, Nov 2012.

[148] N. Stathonikos, M. Veta, A. Huisman, and P. van Diest, "Going fully digital: Perspective of a Dutch academic pathology lab," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 15, 2013.

[149] K. Kayser, J. Görtler, M. Bogovac, A. Bogovac, T. Goldmann, E. Vollmer, and G. Kayser, "Ai (artificial intelligence) in histopathology-from image analysis to automated diagnosis.," *Folia Histochemica et Cytobiologica*, vol. 47, no. 3, 2009.

[150] S. Park, A. V. Parwani, R. D. Aller, L. Banach, M. J. Becich, S. Borkenfeld, A. B. Carter, B. A. Friedman, M. G. Rojo, A. Georgiou, *et al.*, "The history of pathology informatics: A global perspective," *Journal of pathology informatics*, vol. 4, 2013.

[151] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, 2013.

[152] A. Willitzki, R. Hiemann, V. Peters, U. Sack, P. Schierack, S. Rödiger, U. Anderer, K. Conrad, D. P. Bogdanos, D. Reinhold, *et al.*, "New platform technology for comprehensive serological diagnostics of autoimmune diseases," *Clinical and Developmental Immunology*, vol. 2012, 2012.

[153] C.-H. Veillard, A. Roux, L. Loménie, N. Racoceanu, and D. Huang, "Time-efficient sparse analysis of histopathological whole slide images," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 579–591, 2011.

[154] T. Amaral, S. McKenna, K. Robertson, and A. Thompson, "Classification and immunohistochemical scoring of breast tissue microarray spots," *Biomedical Engineering, IEEE Transactions on*, vol. 60, pp. 2806–2814, Oct 2013.

[155] S. Kothari, Q. Chaudry, and M. Wang, "Extraction of informative cell features by segmentation of densely clustered tissue images," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 6706–6709, sept. 2009.

[156] K. Nandy, P. Gudla, K. Meaburn, T. Misteli, and S. Lockett, "Automatic nuclei segmentation and spatial FISH analysis for cancer detection," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 6718–6721, sept. 2009.

[157] A. Chekkoury, P. Khurd, J. Ni, C. Bahlmann, A. Kamen, A. Patel, L. Grady, M. Singh, M. Groher, N. Navab, *et al.*, "Automated Malignancy Detection in Breast Histopathological Images ," *SPIE Medical Imaging*, vol. 8315, 2012.

[158] N. V. Orlov, A. T. Weeraratna, S. M. Hewitt, C. E. Coletta, J. D. Delaney, D. Mark Eckley, L. Shamir, and I. G. Goldberg, "Automatic detection of melanoma progression by histological analysis of secondary sites," *Cytometry Part A*, vol. 81A, no. 5, pp. 364–373, 2012.

[159] R. Riber-Hansen, B. Vainer, and T. Steiniche, "Digital image analysis: a review of reproducibility, stability and basic requirements for optimal results," *APMIS*, vol. 120, no. 4, pp. 276–289, 2012.

[160] L. Cooper, A. Carter, A. Farris, F. Wang, J. Kong, D. Gutman, P. Widener, T. Pan, S. Cholleti, A. Sharma, T. Kurc, D. Brat, and J. Saltz, "Digital pathology: Data-intensive frontier in medical imaging," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 991–1003, 2012.

[161] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, pp. 1341–1390, October 1996.

[162] T. Lattimore and M. Hutter, "No Free Lunch versus Occam's Razor in Supervised Learning," in *Proc. Solomonoff 85th Memorial Conference*, (Melbourne, Australia), 2011.

[163] J. L. Carroll, *A Bayesian Decision Theoretical Approach to Supervised Learning, Selective Sampling, and Empirical Function Optimization*. PhD thesis, Brigham Young University, 2010.

[164] O. Tsybrovskyy and A. Berghold, "Primary unit for statistical analysis in morphometry: patient or cell?," *Analytical Cellular Pathology*, vol. 18, no. 4, pp. 191–202, 1999.

[165] F. Schwenker and E. Trentin, "Pattern classification and clustering: A review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, no. 0, pp. 4–14, 2014.

[166] Q. V. Le, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, "Building high-level features using large scale unsupervised learning," in *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[167] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[168] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.