UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology* 1175

Automated Tissue Image Analysis Using Pattern Recognition

JIMMY AZAR

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2014

ISSN 1651-6214
ISBN 978-91-554-9028-7
urn:nbn:se:uu:diva-231039

Dissertation presented at Uppsala University to be publicly examined in Häggsalen, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, Monday, 20 October 2014 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Marco Loog (Delft University of Technology, Pattern Recognition & Bioinformatics Group).

Abstract

Azar, J. 2014. Automated Tissue Image Analysis Using Pattern Recognition. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1175. 106 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9028-7.

Automated tissue image analysis aims to develop algorithms for a variety of histological applications. This has important implications in the diagnostic grading of cancer such as in breast and prostate tissue, as well as in the quantification of prognostic and predictive biomarkers that may help assess the risk of recurrence and the responsiveness of tumors to endocrine therapy.

In this thesis, we use pattern recognition and image analysis techniques to solve several problems relating to histopathology and immunohistochemistry applications. In particular, we present a new method for the detection and localization of tissue microarray cores in an automated manner and compare it against conventional approaches.

We also present an unsupervised method for color decomposition based on modeling the image formation process while taking into account acquisition noise. The method is unsupervised and is able to overcome the limitation of specifying absorption spectra for the stains that require separation. This is done by estimating reference colors through fitting a Gaussian mixture model trained using expectation-maximization.

Another important factor in histopathology is the choice of stain, though it often goes unnoticed. Stain color combinations determine the extent of overlap between chromaticity clusters in color space, and this intrinsic overlap sets a main limitation on the performance of classification methods, regardless of their nature or complexity. In this thesis, we present a framework for optimizing the selection of histological stains in a manner that is aligned with the final objective of automation, rather than visual analysis.

Immunohistochemistry can facilitate the quantification of biomarkers such as estrogen, progesterone, and the human epidermal growth factor 2 receptors, in addition to Ki-67 proteins that are associated with cell growth and proliferation. As an application, we propose a method for the identification of paired antibodies based on correlating probability maps of immunostaining patterns across adjacent tissue sections.

Finally, we present a new feature descriptor for characterizing glandular structure and tissue architecture, which form an important component of Gleason and tubule-based Elston grading. The method is based on defining shape-preserving, neighborhood annuli around lumen regions and gathering quantitative and spatial data concerning the various tissue-types.

Keywords: tissue image analysis, pattern recognition, digital histopathology, immunohistochemistry, paired antibodies, histological stain evaluation

Jimmy Azar, Department of Information Technology, Computerized Image Analysis and Human-Computer Interaction, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-9028-7
urn:nbn:se:uu:diva-231039 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-231039)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I J. C. Azar, C. Busch and I. B. Carlbom, "Microarray Core Detection by Geometric Restoration," *Analytical Cellular Pathology*, vol. 35, no. 5–6, pp. 381–393, 2012.

II M. Gavrilovic, J. C. Azar, J. Lindblad, C. Wählby, E. Bengtsson, C. Busch and I. B. Carlbom, "Blind Color Decomposition of Histological Images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 6, pp. 983–994, 2013.

III J. C. Azar, C. Busch, I. B. Carlbom, "Histological Stain Evaluation for Machine Learning Applications," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Nice, France, 2012.

IV J. C. Azar, M. Simonsson, E. Bengtsson and A. Hast, "Image Segmentation and Identification of Paired Antibodies in Breast Tissue," *Journal of Computational and Mathematical Methods in Medicine,* vol. 2014, Article ID 647273, 11 pages, 2014. doi:10.1155/2014/647273.

V J. C. Azar, M. Simonsson, E. Bengtsson and A. Hast, "Automated Classification of Glandular Tissue by Statistical Proximity Sampling," 2014. *Submitted for journal publication.*

The author has significantly contributed to the work in the above papers. In Paper II, the author contributed to the unsupervised aspect of the method. Reprints were made with permission from the respective publishers. The papers in this thesis have been proofread for errors.

Contents

Abbreviations

Unless otherwise specified, the following notation is used:

Vectors are denoted by lower-case boldface letters. Matrices are denoted by upper-case boldface letters. Scalars are denoted by lower-case italic letters.

p	number of features
C^i	label for cluster i
ω_i	label for class i
n	number of training objects
n_i	number of objects in class/cluster i
K or k	number of classes or clusters
$[\cdot]^{\mathrm{T}}$	matrix or vector transpose
$\mathbf{x} = \left[x_1, \dots, x_p\right]^{\mathrm{T}}$	object; column vector
$\lVert\mathbf{x}\rVert$	Euclidean norm of vector **x**
$exp(\cdot)$	exponential function
$ln\ (\cdot)$	natural logarithm
$\mathbf{x} \bullet \mathbf{y}$	scalar product between vectors **x** and **y**
$\mathbf{x} \times \mathbf{y}$	vector product between vectors **x** and **y**
$\hat{\boldsymbol{\mu}}$	sample mean
$\hat{\boldsymbol{\mu}}_i$	sample mean of class or cluster i
$\hat{\boldsymbol{\Sigma}}_i$	sample covariance matrix of class/cluster i (maximum likelihood estimate)
$p(\omega_i)$	prior probability of class i
$p(\omega_i \vert \mathbf{x})$	class posterior probability
$p(\mathbf{x} \vert \omega_i)$	class conditional probability
$\mathbf{S}_{\mathrm{W}} = \sum_{i=1}^{K} \frac{n_i}{n} \hat{\boldsymbol{\Sigma}}_i$	within-class covariance matrix
$\mathbf{S}_{\mathrm{B}} = \sum_{i=1}^{K} \frac{n_i}{n} (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$	between-class covariance matrix
$\mathbf{I}$	identity matrix

1 Introduction

The aim of this thesis is to develop methods for automating tissue image analysis by means of statistical pattern recognition techniques. We begin first by briefly describing the fields of computerized image analysis and pattern recognition, and emphasize the need for developing algorithms for reducing observer variability, increasing the quality of interpretation, and improving the workload efficiency.

Computerized image analysis aims at developing algorithms that operate on digital images, often for the purpose of automating the detection or extraction of objects from images. Algorithms may also be developed for enhancing or modifying the content of images to facilitate visual or automated analysis. The field of digital image processing is very broad and overlaps with many other disciplines such as computer science, optics, signal processing, mathematical topology, color science, and cognitive science, to name a few. The generic aim of image analysis and its input remain the common factors that define the field. Many algorithms and methods from other disciplines have been adapted and applied to images for analyzing various aspects of these.

Pattern recognition, as a branch of Artificial Intelligence (AI), is a slightly younger field. While the classical approach to AI is deductive and model-based (or rule-based), statistical pattern recognition uses an inductive problem-based approach that relies on learning from a (limited) set of examples. Its primary aim is to automatically recognize patterns or predict labels. The learning algorithms employed are often outcome-driven and focus on the ability to generalize to unforeseen cases. The methods commonly used in pattern recognition, whether for supervised classification, clustering, or regression, often rely on rigorous statistical techniques for automatically optimizing and selecting parameters as well as for training and validating these methods. The primary purpose being automation and prediction, in addition to the statistical approach that can be employed throughout, makes pattern recognition extremely useful in digital image analysis, especially for fundamental applications such as image segmentation and object recognition.

Some of the most important applications of image analysis can be found in the medical field. Automated image analysis has successfully found its way into digital radiology and has become a fundamental aspect of various imaging modalities, with its influence ranging from acquisition protocols to diagnostic assessment and treatment planning. However, its integration into histology, which is the study of biological tissue, has been slower despite the importance of histopathology in the grading of cancer and assessment of biomarkers. One reason for this lies in the fact that, as opposed to radiology where images are acquired digitally from the outset, histology has an ‘analog’ component embodied in tissue preparation and staining of biopsy specimens, and an analog-to-digital conversion is necessary before computerized image analysis can be used [1]. Moreover, the reliance on optical microscopy for visual analysis is another obstacle. This latter aspect however has witnessed some improvements with the use of digital microscope-mounted cameras and a major leap with the advent of whole slide imaging scanners, which are capable of high-throughput slide digitization (see Section 2.3). This form of digitization provides hope for eventually having publically available, annotated datasets which can significantly help open up the field and set up proper benchmarking. The lack of accessibility to large, digitized datasets and the continued reliance on local experts and analysts are factors that slow down the exposure of histopathological problems and impede efficient integration with image analysis.

This thesis addresses five main topics relating to tissue image analysis.
Paper I presents a method for the detection and localization of tissue microarray cores using a novel and computationally efficient method that is entirely automated and based on restoring the disc shape of the cores. The approach uses a combination of pattern recognition techniques such as hierarchical clustering and the Davies-Bouldin index for cluster validation, as well as image analysis techniques such as morphological operations and granulometry, in addition to basic analytical geometry.
Paper II presents a method for color decomposition that is based on modeling image formation according to the Beer-Lambert law of light absorption in bright-field microscopy. Contemporary methods require that the absorption spectra for the stains be pre-assigned, and reference colors are often specified *a priori*. The proposed method is novel in that it is unsupervised and is able to overcome this limitation [1] by automatically estimating the reference colors using a Gaussian mixture model that is fitted to the data once this latter is projected onto the Maxwellian plane for decoupling color from intensity. The method also proposes a piece-wise generalization of linear decomposition that proves more accurate in difficult situations where chromaticity clusters are non-linearly separable.

Paper III addresses the choice of histological stain which is often overlooked by pathologists, while image analysts assume it preset. However, the choice of stain color combinations is often the most limiting factor in achieving accurate segmentation and successful automation. The paper advocates the systematic evaluation and comparison of histological stains based on supervised and unsupervised classification criteria as opposed to visual inspection. It presents a framework for carrying out this analysis that is aligned with the objective of automation, and examples are illustrated through the comparison of 13 different stains.

Paper IV addresses immunohistochemistry (see Section 2.4), and particularly the paired antibody problem, in which immunostaining patterns need to be compared across adjacent tissue sections extracted from the Human Protein Atlas project. The paper proposes using a simple and robust, unsupervised image segmentation strategy based on the geometry of the feature space and with the aid of rescaling attributes. Then, the probability maps resulting from soft classification are correlated in pairwise correspondence and combined using the product rule into a single similarity measure. The similarity measure is able to test simultaneously for the adjacency of tissue sections (thereby not placing any assumptions on the grouping of original images), as well as for the similarity in staining patterns across the sections. Lastly, the relative proportions of the individual tissue types are quantified in each section using the derived probability maps.

Paper V presents a novel feature descriptor for characterizing tissue and glandular architecture based on sampling the neighborhoods of lumen regions. Iterative region expansion is used to define shape-preserving, neighborhood strips or rings around each lumen, and within each ring, quantitative and spatial information is collected. The approach does not require the extraction of intricate structural properties, yet it is able to represent tissue architecture in a highly descriptive manner. Furthermore, the method is combined with multiple instance learning to provide an elaborate representation useful for classification.

The publications included in this thesis are compactly summarized in Section 4. We note that throughout that section, it will become evident that the methods used in these publications are based on pattern recognition and image analysis techniques that have been optimized for automation. In particular, parameters relating to the methods have been either kept minimal or automatically selected based on optimization and cross-validation procedures. We also note that the approaches designed often consist of a series of sequential i/o interconnected stages, forming an organized workflow. However, before discussing the different publications, we present a very brief but useful introduction to aspects of histology and tissue preparation in Section 2, in order to highlight the types of problems and datasets that are used throughout this thesis and their origin. This is followed by a brief introduction to some of the pattern recognition and image analysis methods used throughout the publications, summarized in Section 3.

2 Background

Automated tissue image analysis relies upon several preceding processes relating to tissue preparation and staining, image acquisition, and slide digitization. We begin by presenting a brief introduction to some of these topics, while regularly making note of their relevance with regard to the work and publications included in this thesis.

2.1 Tissue Preparation

Once a tissue biopsy or mastectomy section is obtained by surgery, the first step in tissue preparation for histological analysis is to place the specimens in a fixative for preserving the tissue structures and slowing down decomposition. Formaldehyde is often used as a fixative in bright-field microscopy. The second step in tissue preparation is to embed the sections in a paraffin block; the process involves gradually removing the water content by exposing the specimen to a series of high concentration alcohol solutions, which eventually allows for the infiltration of paraffin wax. The third step is to slice the paraffin block into micrometer-thin sections so that they may be viewed under an optical microscope by passing light through the samples. This is carried out using a high-precision cutting tool called a microtome, which is able to provide 3-5 µm thin slices for this purpose. The final step in tissue preparation is to dye the sections with stains that provide proper contrast for visual or computer-based analysis. Higher levels of standardization and quality control for fixation and staining protocols are expected with the proliferation of digital histopathology and slide digitization; consequently, this last step, concerning tissue staining, becomes increasingly important and is discussed in more detail in the following section.

2.2 Histological Staining

In order to visualize structures in the tissue, the tissue section requires staining with a dye combination that provides contrast or highlights specific components. The commonly used H&E stain, which has been the standard stain in histopathology for the past century, consists of two compounds, namely hematoxylin and eosin. Hematoxylin binds to DNA, thus highlighting nuclei in purple-blue, and eosin binds in general to proteins, highlighting other tissue structures such as cytoplasm, epithelium, and stromal regions in pink (see Figure 1(a)).

Additionally, in immunohistochemistry which is a more advanced staining method, an organic compound such as the 3,3’-diaminobenzidine chromogen can be used to highlight specific regions where the antigen-antibody complex is concentrated. Figure 1(b) shows an example of immunohistochemical staining of a tissue section. Histological staining provides the necessary contrast that allows pathologists to visually analyze tissue structures and perform grading; however, despite dedicated training and specialized guidelines, these tasks remain considerably subjective and labor-intensive [2, 3, 4, 5]. The latest recommendations by the American Society of Clinical Oncology as well as the College of American Pathologists advocate the use of quantitative and computer-assisted methods in aim of reducing observer variability among pathologists [1, 6, 7]

The topic of tissue staining is addressed in Paper III, in which we present a methodical way of selecting an optimal stain for a given type of tissue in a manner that focuses on facilitating automation as opposed to visual inspection.

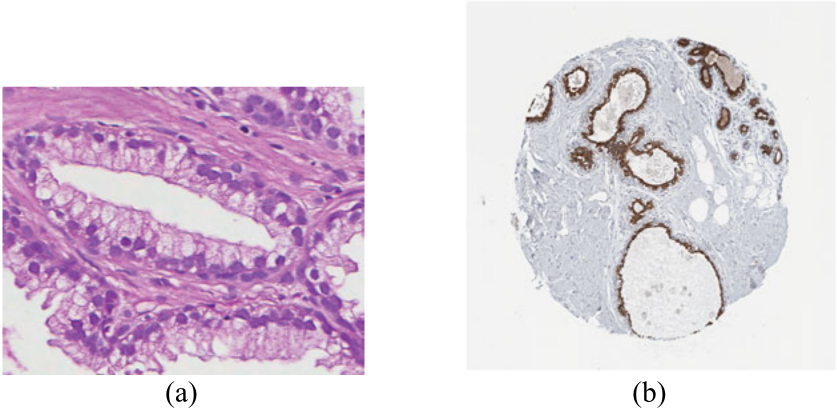


Figure 1. Sample histological stains. (a) Hematoxylin and eosin (H&E) staining. (b) Immunohistochemical staining where the antibody is visualized using DAB (3,3’-diaminobenzidine). Regions stained with the brown DAB are considered as positive areas.

2.3 Tissue Microarrays

Tissue microarrays [8, 9] consist of usually tens to hundreds of tissue cores arranged in an array-like manner within a single slide (see Figure 2). Beginning with a tissue specimen embedded in a paraffin block, a tissue arrayer device punches disc-like cores about 0.6 mm in diameter using a thin needle, and transfers these cores to another recipient paraffin block where the cores are placed in regular array patterns. Finally, the recipient paraffin block can then be sliced into thin sections and stained in preparation for visual analysis under a microscope or, alternatively, for slide digitization. With the advent of whole slide scanners, the entire slide of tissue microarray cores can be digitized using automatic scanning procedures. This allows for high-throughput digitization and histological analysis. Slide scanning is often done at high resolution, thus generating a base image from which several lower resolution images can be derived. The resulting images are stored in a multi-scale pyramidal structure allowing for different magnification views and rapid retrieval at multiple resolutions.

In analyzing tissue microarrays, the individual cores should be detected and localized so that they can be matched with the corresponding specimen from the donor paraffin block [10, 11, 12], and to allow for subsequent segmentation and image analysis within each core. The automation of the detection and localization of tissue microarray cores is the main topic of Paper I of this thesis.

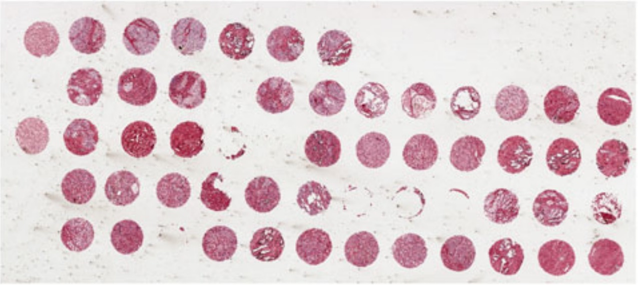


Figure 2. An example of a tissue microarray slide. The disc-shaped tissue cores are arranged in an array-like manner.

2.4 Immunohistochemistry

We have mentioned in Section 2.2 that the DAB chromogen is used in immunohistochemical staining to visualize antibodies which may have been raised against specific antigens or proteins in the tissue (see Figure 1(b)). In general, immunohistochemistry (IHC) is a staining technique that utilizes antibodies that bind to specific antigens in order to highlight their areas of presence and corresponding tissue structures. In order to visualize these areas, the antibody is coupled with a visual marker. In other instances, two antibodies are used: a primary antibody that binds to the specific antigen, followed by a secondary, dye-labeled antibody that binds to the primary antibody. IHC may be used in practice to identify certain receptors such as estrogen (ER), progesterone (PR), human epidermal growth factor 2 (HER2), in addition to the Ki-67 protein, which can be an indicator of cell growth and proliferation [6, 7, 13]. The detection and quantification of such receptors in breast tissue cancer can have significant prognostic value in assessing the risk of recurrence or mortality, as well as predictive value, for example in identifying whether a tumor would respond to endocrine therapy aimed at slowing down the progression of cancer. Several attempts for automating the quantification of these biomarkers and the proportion of positively stained regions can be found in [14, 15]. Numerous publications have indicated that automated image analysis is able to achieve results that are similar to those of trained pathologists [16, 17, 18, 19, 20].

In this context, we address the paired antibody problem in Paper IV, where we present a method for quantifying immunostaining patterns in adjacent tissue sections labeled with IHC staining, and for identifying antibodies that target the same protein, but which may bind to different parts of the protein.

2.5 Cancer Grading

Often the final aim of tissue image analysis is to aid pathologists in the interpretation and cancer grading of biological tissue specimens. Two of the most common types of cancer affecting women and men are, respectively, breast and prostate cancer. In the case of breast cancer, the standard scoring system used at present by pathologists is the Ellis-Elston system (also known as the Nottingham system) [21, 22], which is a modified version of the Bloom-Richardson grading system. For prostate cancer, the standard scoring system in use is the Gleason grading system [23]. We begin by briefly describing these two systems and their focus on glandular structures and tissue architecture, which is also the topic of Paper V.

The Elston grade for breast cancer is based on three different parts: tubule formation, nuclear atypia, and mitotic count. The part concerning tubule formation has similarities with the Gleason pattern description in terms of characterizing glandular/tubular differentiation. Depending on the amount of glands and tubules present in the examined tissue section, a score of 1, 2, or 3 is given, ranging respectively from highly glandular, healthy tissue (score

1) to scarcely glandular tissue as in the case of solid tumors (score 3). The second part of the Elston score is concerned with nuclear size, shape and chromatin texture, i.e., how irregular these appear on a scale from 1 (normal) to 3 (highly atypical). The third and final part is mitotic activity and is concerned with how much the cells are dividing, which corresponds to growth rate. This involves counting dividing cells under high-power microscopic fields (x40 objective), and a score from 1 (low cell count) to 3 (high cell count) is assigned depending on the number of cells counted within a specific area. In the end, the individual scores from the three parts are added up to obtain the final Elston grade. In particular, if the sum is in the range of 3-5 points, it is defined as Elston grade I, 6-7 as grade II, and 8-9 as grade III.

Gleason grading for prostate cancer is based on five different patterns (1-5), and the final score is the sum of the two most occurring patterns, thus ranging from 2 to 10. Note that in case of three visible patterns, the final score is computed as the sum of the primary, most frequent pattern and the higher of the remaining two. The five patterns describe glandular differentiation beginning with Pattern 1, which corresponds to well-differentiated carcinoma that resembles normal tissue, and ending with Pattern 5, which corresponds to poorly-differentiated carcinoma, mostly lacking recognizable glandular units. In general, whether in the case of Elston or Gleason grading, the higher the grade the more potentially aggressive the cancer is, and the higher is the risk of it spreading to healthy tissue, resulting in poor prognosis.

In Paper V, we present a feature descriptor that is based on tissue architecture with the aim of characterizing glandular tissue and tubule formation, which are important components of both Elston and Gleason grading systems.

3 Pattern Recognition & Image Analysis

This section gives a brief introduction to the pattern recognition and image analysis methods that have been used in Papers I-V. It is intended to provide a summary of the main concepts and is written with the aim of facilitating the understanding of these publications. The advanced reader may choose to skip over this section.

3.1 Principal Component Analysis

Principal component analysis (PCA) [24, 25] is a linear, unsupervised method for feature extraction which does not assume any underlying statistical model. It is also known as the Karhunen-Loève transformation in one of its basic forms, and is also equivalent to classical multidimensional scaling [26] operating on a Euclidean dissimilarity matrix. Principal component analysis attempts to find an orthogonal transformation such that the derived variables are uncorrelated. The resulting features are a linear combination of the original ones and are termed the principal components or principal axes. The first principal component is in the direction of maximum variance, followed by the second principal component, etc., while all these remain orthogonal to each other. In the two-dimensional case, the second principal component is automatically the axis perpendicular to the first principal component. However, in three or higher dimensional feature spaces, there is an infinite number of directions that are orthogonal to the first principal axis, and so the second component is found among these in the direction of maximum variance.

Beginning with a p-dimensional dataset consisting of n vectors $\mathbf{x}_i$ and denoted by $\mathcal{L} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, we compute the sample mean and sample covariance matrix over all $\mathbf{x} \in \mathcal{L}$ as follows:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \tag{3.1}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \tag{3.2}$$

The principal components are obtained by performing an eigen-decomposition of the covariance matrix, i.e. by finding the nontrivial solutions to the following set of equations:

$$(\widehat{\mathbf{\Sigma}} - \lambda_i \mathbf{I})\mathbf{e}_i = \mathbf{0}, \qquad i = 1, \dots, p \tag{3.3}$$

where $\mathbf{I}$ is the p x p identity matrix, $\mathbf{e}_1, \dots, \mathbf{e}_p$ are the eigenvectors of the sample covariance matrix, and $\lambda_1, \dots, \lambda_p$ are their corresponding eigenvalues. The variance of the principal component ξ_i is then $var(\xi_i) = \lambda_i$.

Geometrically, principal component analysis represents a rotation of the coordinate axes, where the principal components, i.e. the new axes, are aligned with the directions of maximum variance. This transformation is reversible if all the principal components are retained. However, often the purpose of PCA is to reduce the dimensionality of the dataset by retaining only the first $d < p$ eigenvectors with highest variance. Note that the total variance, as summed up along each axis, does not change with PCA or rotations, and in general this is true as long as the coordinate system remains orthogonal. One common rule for selecting d is to sort the eigenvalues in descending order and retain the first d whose sum corresponds to 90 or 95% of the total variance; that is, the ratio of retained variance would be:

$$\sum_{i=1}^{d} \lambda_i \Bigg/ \sum_{i=1}^{p} \lambda_i \approx 0.9 \tag{3.4}$$

Unlike feature selection techniques, the principal components resulting from a feature extraction method such a PCA are a linear combination of the original features, and therefore the meaning of these variables is often difficult to interpret.

An equivalent way to derive the principal components is by minimizing the least square error to the data. This can be compared to simple regression. In regression, the optimal straight line fit, in the least squares sense, is that which minimizes the square error between the data points and the line model. This error at any given sample point is the vertical distance between the point and the straight line. If the units or scales of the variables are changed, the model (line slope) would change, yet this does not affect the predicted output value. However, unlike regression, the error minimized in PCA is based on the perpendicular distance from a given data point to the straight line. These distances are affected by the rescaling of axes since right angles are generally not preserved by relative changes in scale. Thus, while the predicted values in regression are independent of changes in scale, PCA is sensitive to the rescaling of features. Even when the units of the variables are similar, if one variable has a much larger range than the other, the principal component may favor its direction. It is therefore important to standardize the data prior to applying PCA, such as by normalizing to zero mean and unit variance. As with many feature extraction methods, the criterion used in PCA is not necessarily aligned with the final objective, e.g. classification. There is no guarantee that a subspace spanned by the first d principal components is optimal for classification. Subspace selection by PCA may discard important directions in the data or may include noisy ones; however, PCA remains a very useful technique and a popular tool for data analysis as well as for pattern recognition tasks. In [27], a comparative study was conducted among various feature extraction methods which included PCA, in addition to twelve nonlinear feature extraction techniques such as local linear embedding, Laplacian eigenmaps, Sammon mapping (nonlinear multidimensional scaling), and kernel PCA. Nonlinear techniques for dimensionality reduction have the potential for learning nonlinear manifolds as can be demonstrated using difficult, pre-selected artificial/toy examples; however, despite this advantage, such methods were unable to outperform traditional linear techniques such as PCA when using natural, real-world datasets [27].

In Paper IV, we regard PCA as a geometric transformation and use it to standardize the representation of the dataset, without performing dimensionality reduction.

3.2 Clustering

In image analysis and many practical applications, it may be that the available dataset is either not labeled or only a small portion of the training objects are labeled. Unsupervised learning methods that try to handle unlabeled data are often referred to as clustering techniques. Cluster analysis aims to find 'natural' groups in the data, thus allowing the data to express itself without imposing training labels. However, defining what a 'natural' grouping constitutes is a diverse matter, and clustering techniques do in fact implicitly pose structure on the data in one way or another. Perhaps, in some contexts, a 'natural' grouping is that which agrees with our human interpretation. In low-dimensional feature spaces where data objects can be visualized, such a validation is to some extent possible, though not always practical. However, with high-dimensional feature spaces, judging clusters by visualization becomes unfeasible. All clustering algorithms can yield a partitioning of the data into clusters or groups, however, the real difficulty lies in validating these groupings. For some applications, this may not be important. For example, in large datasets, clustering may also be used for data reduction where it is computationally efficient to handle representative cluster centers rather than the entire dataset.

A general way to define a cluster is that by a subset of objects such that the resemblance among the objects within the same subset is higher than that to other objects of another subset. Sensibly, objects that are close to each other in feature space should have greater resemblance to each other than to objects that are far away. Inevitably one would have to define this resemblance using some kind of distance measure. It is with the choice of distance measure that the clustering algorithm begins to impose structure on the data. The user has to decide on the type of clustering method to be used and needs to judge the quality of the clustering in some manner. It is therefore very advantageous if the user has some prior knowledge of the type of clusters expected in terms of their shapes and/or sizes or can even visualize the end result. By setup, this is the case in Paper I, where tissue microarray cores can be thought of as spherical, disc-like clusters in 2D.
Often determining the number of clusters, K, in the data is crucial. It is sometimes known *a priori*, however, the choice of this number can be determined by repeating the clustering over a range of values for K and selecting the one for which the cluster validation criterion is optimal or exhibits the greatest improvement. In practice, such an automatic optimization for determining ‘natural’ groupings is most possible in cases when the cluster shapes are known, and thus the clustering method and validation criterion can be chosen accordingly. Fortunately, this happens to be the case in Paper I, where hierarchical clustering with complete linkage was used to account for the disc-shaped microarray cores, and the Davies-Bouldin index was used as a validation criterion well-suited for spherical-shaped clusters.

There are different categories for techniques employed in clustering of which we may distinguish specifically hierarchical methods that operate on a dissimilarity matrix, sum-of-squares methods such as k-means and fuzzy c-means, and mixture models which represent the probability density function as a sum of individual component distributions. An example of the latter is the Gaussian mixture model which is used in Papers II and III; a description of the method can be found in the appendix of Paper III. Fuzzy c-means is used in Paper IV, whereas hierarchical clustering is used in Paper I.

3.2.1 Hierarchical Clustering

Hierarchical clustering [28] is one of the most common methods for representing data structure using a tree diagram, called in this case a dendrogram (see Figure 3). Cutting the tree at a given horizontal level results in a number of disjoint clusters, K, depending on the number of vertical stems the horizontal level cuts through. For example, if the dendrogram in Figure 3 is sectioned at a threshold level of 4, this would partition the data into three clusters as shown by the different colors.

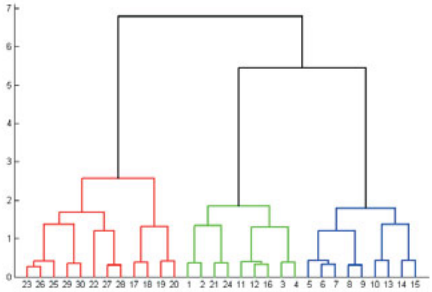


Figure 3. Dendrogram: a nested set of partitions summarized in a tree structure. The numbers on the abscissa axis represent data object labels, and the ordinate axis represents distance values among clusters. The ordering of the data objects is arbitrary but is always selected so that the branches of the tree do not cross.

In this section, we discuss the agglomerative approach for constructing these dendrograms which enable us to partition data into clusters. The idea is to begin with single objects and assume each of these is a cluster; then the closest clusters are merged sequentially forming larger clusters until all the data lies within one cluster. This creates a nested hierarchy of clusters and sub-clusters, which is a very useful way to summarize the structure of the data based on distances.

Suppose a dataset consists of n objects. The algorithm operates on the n x n dissimilarity matrix obtained by computing the distances between these objects. The objects are initially regarded as single clusters, i.e., there are n clusters at the outset. The algorithm then proceeds as follows:

1. The closest pair of clusters is merged into one cluster. In cases of ties, often any one of the pairs can be selected arbitrarily or based on some criterion. This so-called *ties in proximity* problem is discussed in [29] with suggested ways of resolving it in [30].

2. The dissimilarity matrix is updated by computing the distances between the newly formed cluster and all the remaining clusters. The distance between two clusters, i.e. sets of objects, can be defined in many ways.

3. Steps 1-2 are repeated iteratively until all the n objects are in the same, one cluster. Alternatively, a stopping criterion could be a preset number of clusters, K.

However, the choice of distance measure between clusters, referred to as the linkage type, is very important and affects the clustering result. Some of the types of distances used are the single-linkage, complete-linkage, and average-linkage. In single-link clustering, the distance between any two clusters C^i and C^j is defined as the distance between their closest member objects, that is:

$$d_s(C^i, C^j) = min\,\{d(\mathbf{x}, \mathbf{y}) \colon \mathbf{x} \in C^i, \mathbf{y} \in C^j\} \tag{3.5}$$

The inter-object distance $d(\mathbf{x}, \mathbf{y})$ often refers to the Euclidean norm, whereby single-linkage can be rewritten as:

$$d_s(C^i, C^j) = min_{\mathbf{x} \in C^i, \mathbf{y} \in C^j} ||\mathbf{x} - \mathbf{y}||^2 \tag{3.6}$$

In complete-link clustering, this distance becomes:

$$d_c(C^i, C^j) = max_{\mathbf{x} \in C^i, \mathbf{y} \in C^j} ||\mathbf{x} - \mathbf{y}||^2 \tag{3.7}$$

This represents the distance between the farthest two objects, while one of the objects is in the first cluster and the other is in the second cluster.

In average-link clustering, the distance is computed as that between the centers of the two clusters, C^i and C^j, where these may contain n_i and n_j objects respectively, that is:

$$d_a(C^i, C^j) = \frac{1}{n_i n_j} \sum_{\mathbf{x} \in C^i} \sum_{\mathbf{y} \in C^j} ||\mathbf{x} - \mathbf{y}|| \tag{3.8}$$

Single-link clustering tends to find string-like clusters, whereas complete-link clustering results in more compact, spherical-shaped clusters, and finally, average-link clustering falls in between.

In Paper I, we use complete-link clustering to detect microarray cores since the method works well with spherical, compact clusters. To automatically determine the number of clusters, K, in the image, we repeat the clustering over a range of values for K, and for each result, we compute the Davies-Bouldin index [31] to assess the quality of the clustering and compare among the different outcomes. This cluster validation criterion computes the mean and variance measures to assess the location and widths of the clusters, and then uses these to compute a conservative pairwise score per cluster while considering all possible pairs. Finally, these scores are averaged and used as an indicator of how compact and well-separated the clusters are. Using the Euclidean distance, the Davies-Bouldin index is a cluster validation measure that is ideally suited for spherical clusters. An explanation and derivation of this index can be found in Paper I.

3.2.2 Gaussian Mixture Model

A finite mixture model is a distribution that is written as a linear combination of individual distributions, i.e., it has the form:

$$p(\mathbf{x};\psi) = \sum_{j=1}^{K} \alpha_j \, p(\mathbf{x};\theta_j) \tag{3.9}$$

where K is the number of individual distributions, $\psi = \{\alpha_j, \theta_j \,|\, j = 1, \dots, K\}$ is the complete set of parameters for the mixture distribution, and the scalars, α_j, are the mixing coefficients, where $\alpha_j \geq 0$ and $\sum_{j=1}^{K} \alpha_j = 1$. In the case of the Gaussian mixture model [32], the component distributions $p(\mathbf{x};\theta_j), j = 1, \dots, K$, are multivariate normal distributions given by:

$$p(\mathbf{x};\theta_j) = \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}_j|}} exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \mathbf{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right\} \tag{3.10}$$

where $\boldsymbol{\mu}_j$ and $\mathbf{\Sigma}_j$ are the mean vector and covariance matrix of the individual components, and $\theta_j = \{\boldsymbol{\mu}_j, \mathbf{\Sigma}_j\}$ is the set of parameters. The model parameters may be fitted to the data by maximum likelihood estimation. These parameters can be computed efficiently using an expectation-maximization (E-M) procedure. Details of this iterative procedure can be found in Appendix A.2 of Paper III.

While the individual components are Gaussian and can vary in orientation and width across different directions as specified by the covariance matrix, the mixture distribution is rather flexible and can model a variety of realistic distributions. For real-world datasets, the range of shapes Gaussian mixtures are able to model is often sufficient to represent the natural variation in the data. The prevalence of Gaussian distributions in these datasets can be partially explained by the central limit theorem, since real-world measurements are often the sum of a large number of unobserved random events.

The E-M algorithm is a very general procedure and its use in Gaussian mixture models is related to the k-means algorithm which is basically an E-M algorithm. Because the Euclidean norm is often used in k-means to compute distances to cluster centers, k-means imposes a spherical structure on the clusters. The Gaussian mixture model allows for more flexible clusters with Gaussian-like distributions which are described not just by their means but also by their covariance matrices that specify orientation and amount of spread in different directions. Because the number of parameters is higher relative to k-means, more data is usually required to estimate these accurately. The number of parameters is exactly $K\left(p+\frac{1}{2}p(p+1)+1\right)$ with p parameters for the mean, $\boldsymbol{\mu}_j$, and $\{p+(p-1)+\cdots+1\}=p(p+1)/2$ parameters for the symmetric p x p covariance matrix, $\boldsymbol{\Sigma}_j$, in addition to a single parameter for the mixing coefficient, α_j, while these are for each of the K Gaussian components. If the Gaussian components are restricted to a circularly symmetrical shape, i.e., their covariance matrices are diagonal with equal elements along the diagonal, $\boldsymbol{\Sigma}_j=\sigma^2\mathbf{I}$, then as $\sigma\rightarrow 0$, the variances of each component vanish in all directions and the K Gaussian components reduce to K single centroids. In this context, the E-M algorithm used in the Gaussian mixture model reduces to the k-means algorithm, whereby every object is assigned to the closest centroid.

The E-M algorithm requires an initial set of parameters for the Gaussian components before it can converge to a solution. One way is to initialize the parameters randomly and repeat this a few times as in k-means. This also helps avoid selecting a local optimum. A remaining issue is how to select the number of clusters K in an automated way if it is not known or specified *a priori*. For probabilistic models like the Gaussian mixture model, increasing K results in improvements in the log-likelihood estimate, although this improvement is expected to level off gradually. It is possible to select K for which the sequential increase in the log-likelihood curve is largest, i.e. relative to the value when using $K-1$ components. Another approach is to use information criteria such as the Akaike or Bayesian information criterion. These criteria limit the improvement in the log-likelihood estimate by incorporating the complexity of the model, in terms of its parameters or number of components, into the goodness-of-fit measure. By penalizing models with increasingly large values of K, a tradeoff is introduced between maximizing the log-likelihood and minimizing the mixture's complexity. This tradeoff circumvents the problem of overfitting the mixture model to the data as a result of simply increasing the number of components. The information criteria may be computed over a range of values for K, from which an optimal value is selected at the point where the curve attains a minimum. These criteria are described in more detail in Paper V where an example is also presented.

3.3 Supervised Classification

When class labels are available for sample objects, we may train a classifier on this limited set of objects. The aim of the classifier is then to predict the class labels of future objects that it has not been trained on. Supervised classification, which uses learning in the presence of class labels, is therefore very similar to regression in the sense that an output is predicted for a given p-dimensional input object. From this perspective, it can actually be regarded as a special case of regression, where the output is confined to a categorical label (in a multilayer perceptron this distinction is even less strict and rather artificially enforced). The set of labeled objects used to construct or train the classifier is called the *training set*. To test the performance of a classifier, a separate independent set of objects with known labels is used in order to validate whether the predicted labels of these test objects match their true labels. This set of objects is called the *test set* or validation set. The labeled dataset is often split between a training set and a test set. However, apart for simulation experiments, since labeled data may be scarce or difficult to obtain in reality, setting aside labeled objects for testing may weaken the training phase and the classifier's ability to learn and generalize, especially if the training set is not large or descriptive enough to represent the true underlying class distributions in feature space. This results in a pessimistic estimate of the classification rate. However, without validation, constructing a classifier could be meaningless. In the vast majority of cases, k-fold cross-validation is used or similar rotational methods. The labeled dataset is randomly split into k equally-sized, non-overlapping partitions; each of these k subsets is used sequentially as a test set while the classifier is trained on the remaining data consisting of (k-1) subsets. This results in k classification error rates, which are then averaged and reported along with the standard deviation. In the case where k is equal to the number of labeled objects in the dataset, the procedure is called the *leave-one-out* method since only 1 test object is used at every validation round. An undesirable aspect of the k-fold cross-validation procedure is that the training sets used across the rotations are much overlapping (except when k=2). However, the method remains the most widely used standard for testing classification and does work well in practice as long as the size of the dataset and choice of folds remains sensible.

The complexity of the classifier plays an important role in its ability to generalize. The complexity of a classifier can often be controlled by its parameters, and these may be optimized for example using cross-validation so as to avoid overfitting or underfitting the classifier to the data. Classifiers with low complexity such as linear classifiers are susceptible to exhibiting bias even when the size of the training set is large since the classifier is simply not flexible enough to follow the distribution of the data or produce the correct decision boundary, thus resulting in systematic error. This underfitting problem is not encountered in practice as often as the overfitting problem since there is usually an appeal for classifiers with a high complexity range, but when combined with the fact that training sets tend to be very limited and feature spaces high-dimensional, it results in a serious problem that requires careful handling. In this case, although the bias error is reduced, flexible classifiers are prone to exhibit high variance. This can be explained by the bias-variance dilemma. The optimal complexity is that at which the overall bias-variance error in both its components is minimized. In other words, the classifier should be just flexible enough to follow the data distribution while simple enough so as not to overadapt to particular instances. Sample decision boundaries with varying complexity can be seen in Figure 4. Careful testing and cross-validation procedures can be applied to select the optimal complexity of a given classifier. Examples of this can be seen in Papers III-V.

The dimensionality of the feature space, p, in relation to the number of training objects, n, is a very important factor that contributes to the overfitting problem and poor generalization. Particularly, when n is not large with respect to p, the amount of training objects is insufficient to represent the true class conditional distributions in feature space. Classifiers that are based on the Bayesian approach and that need to estimate the class conditional density function tend to suffer the most from this problem. Thus, a classifier trained on these objects may not be able to generalize well over new cases. The sheer volume of the data required to represent a class densely in a high-dimensional feature space becomes problematic, especially that the size of the training set is almost always small or limited. Thus by including too many features, the performance of a classifier begins to deteriorate at some point after it has been initially improving, and hence the classification error curve shows a peak. There is often an optimal subset of features to include for which the classification error rate is minimal. In the context of pattern recognition is this referred to as the *curse of dimensionality* or the *peaking phenomenon*. In statistics, this is known as the *small n, large p* problem. For classifiers that rely on estimating and inverting covariance matrices, the problem is further amplified since it becomes impossible to invert the sample covariance matrix when $n < p$, and the pseudo-inverse or regularization techniques are usually employed to solve the problem. It is often necessary to reduce the dimensionality of the data before the classification task to avoid such problems, and this can be done through either feature selection or feature extraction techniques such as PCA.

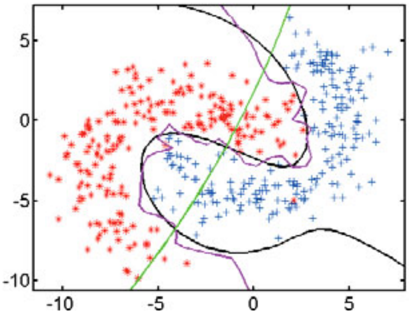


Figure 4. Sample decision boundaries showing varying complexity: the 1-nearest neighbor classifier (purple), support vector classifier with cubic polynomial kernel (black), and the quadratic classifier (green).

3.3.1 Bayes Optimal Classifier

Following the Bayesian approach to classification, if the conditional probability density function is known for the various classes as well as their prior probabilities, then everything is known about the data in order to make decisions. Particularly, Bayes' theorem is used to compute the posterior probabilities upon which decisions are based. These represent the probability of belonging to class ω_i, given the object $\mathbf{x}$. Bayes' theorem states:

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(\mathbf{x})}. \tag{3.11}$$

where $p(\omega_i|\mathbf{x})$ is the class posterior probability, whereas $p(\mathbf{x}|\omega_i)$, is the class conditional probability, and $p(\omega_i)$ is the class prior probability. The term $p(\mathbf{x})$ appearing in the denominator can be written using the law of total probability as:

$$p(\mathbf{x}) = \sum_i p(\mathbf{x}|\omega_i)p(\omega_i) \tag{3.12}$$

This term is often seen as a normalization factor and can be ignored when comparing the class posterior probabilities for a given object in order to make decisions, since $p(\mathbf{x})$ is a common term that will appear across all of the class posterior probability computations. An object can then be assigned to the class with the highest posterior probability.

If the true class conditional density functions and priors are known, then the optimal Bayes classifier is obtained. The error this classifier achieves is the theoretical minimum. However, in practice, the true class conditional density functions are never known and need to be estimated. This can be done using a non-parametric density estimation method such the k-nearest neighbor or

Parzen density estimation, resulting respectively in the k-nearest neighbor classifier or the Parzen classifier, or by means of a parametric method such as using the normal density model, resulting in the quadratic normal-based classifier. Once the class conditional densities are estimated, one can arrive at the posterior probabilities using Bayes' theorem. An exception to this approach is the Logistic classifier which directly models the class posterior probability density function using a sigmoidal function due to the fact that the logarithm of the ratio of class conditional density functions is taken to be linear in terms of $\mathbf{x}$ for any pair of classes, resulting in a linear decision boundary $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$, i.e.,

$$ln\left(\frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)}\right) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \tag{3.13}$$

In summary, since the class conditional densities can only be estimated and the training set available for this estimation is finite, a classifier's error will be higher than the Bayes' error. The larger the number of training data representing the true underlying class distributions and the more accurate the density estimate is, the closer will be the classifier's performance to the Bayes optimal classifier.

3.3.2 Quadratic Classifier

In estimating the class conditional probability density function, one approach is to use a parametric density estimation method, i.e., one that assumes a certain model. The most common parametric density estimation method uses the multivariate normal distribution to estimate each class conditional density function, i.e.,

$$\hat{p}(\mathbf{x}|\omega_i) = \frac{1}{\sqrt{(2\pi)^p|\widehat{\mathbf{\Sigma}}_i|}} exp\left\{-\frac{1}{2}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_i^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)\right\} \tag{3.14}$$

The parameters of each model are estimated from the training objects in the corresponding class. Thus $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\mathbf{\Sigma}}_i$ are respectively the sample mean and covariance matrix of class ω_i as estimated from the available data. If we substitute this expression into the conditional probability density function in Bayes' theorem, we obtain the following expression for the logarithm of the posterior probability:

$$\begin{aligned} ln\big(\hat{p}(\omega_i|\mathbf{x})\big) = \frac{-p}{2} ln(2\pi) - \frac{1}{2} ln\big(|\hat{\boldsymbol{\Sigma}}_i|\big) \\ - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i) + ln\big(\hat{p}(\omega_i)\big) \\ - ln(\hat{p}(\mathbf{x})) \end{aligned} \tag{3.15}$$

When comparing among posterior probabilities for a given object, we may ignore the first and the last terms in this equation since the first term is a constant and the last term is independent of the classes and represents the normalization factor in Bayes' theorem which appears in all these expressions. Equation 3.15 reduces to:

$$\begin{aligned} g_i(\mathbf{x}) = -\frac{1}{2} ln\big(|\hat{\boldsymbol{\Sigma}}_i|\big) - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i) \\ + ln\big(\hat{p}(\omega_i)\big) \end{aligned} \tag{3.16}$$

The discriminant rule becomes: assign object $\mathbf{x}$ to the class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $i \neq j$. This is the same as assigning according to the highest posterior probability. Equivalently, we can construct the discriminant function between two classes as follows:

$$f(\mathbf{x}) = \hat{p}(\omega_1|\mathbf{x}) - \hat{p}(\omega_2|\mathbf{x}) \tag{3.17}$$

The decision rule becomes: assign object $\mathbf{x}$ to ω_1 if $f(\mathbf{x}) > 0$. If we substitute the expressions for $\hat{p}(\omega_1|\mathbf{x})$ and $\hat{p}(\omega_2|\mathbf{x})$ from equation 3.15 and simplify, we can write the discriminant function in the following form:

$$f(\mathbf{x}) = \mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{x} + \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \tag{3.18}$$

where the weights are given by:

$\mathbf{W} = \left(-\frac{1}{2}\hat{\boldsymbol{\Sigma}}_1^{-1} + \frac{1}{2}\hat{\boldsymbol{\Sigma}}_2^{-1}\right)$,
$\mathbf{w} = \left(\hat{\boldsymbol{\Sigma}}_1^{-1}\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\mu}}_2\right)$, and
$w_0 = \Big(-\frac{1}{2} ln\big(|\hat{\boldsymbol{\Sigma}}_1|\big) + \frac{1}{2} ln\big(|\hat{\boldsymbol{\Sigma}}_2|\big) + ln\big(p(\omega_1)\big) - ln\big(p(\omega_2)\big) - \frac{1}{2}\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_1^{-1}\hat{\boldsymbol{\mu}}_1$
$+ \frac{1}{2}\hat{\boldsymbol{\mu}}_2^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\mu}}_2\Big)$

Thus, equation 3.18 shows that the function is quadratic in $\mathbf{x}$ and the decision boundary will in general be a conic section (such as an ellipse or a parabola). Because this classifier uses the normal distribution to model each class conditional density function before applying Bayes' theorem, it is referred to as the *normal-based quadratic classifier*.

Note that the class prior probabilities $\hat{p}(\omega_1)$ and $\hat{p}(\omega_2)$ appear only in the scalar offset term, w_0, of equation 3.18. This means that changing the class priors will cause the decision boundary to shift as if changing the cost of classification among the two classes.

Equations 3.15, 3.16 and 3.18 show that the class sample covariance matrices need to be computed and inverted; however if the number of objects is small, the estimate will be poor. Moreover, if this number is less than the dimensionality of the feature space, i.e., $n < p$, then the sample covariance matrix is rank deficient and therefore cannot be inverted. One possibility may be to use the Moore-Penrose pseudo-inverse $\boldsymbol{\Sigma}_i^{+} = \widehat{\boldsymbol{\Sigma}}_i^{\mathrm{T}}\left(\widehat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{\Sigma}}_i^{\mathrm{T}}\right)^{-1}$. Note that this can be used as well in the Fisher classifier, where the problem of inverting the sample covariance matrix is also present; in this case, the classifier is called the pseudo-Fisher classifier in reference to the pseudo-inverse.

Regularizing the class sample covariance matrix can also prevent this problem by setting: $\widehat{\boldsymbol{\Sigma}}_i' = \widehat{\boldsymbol{\Sigma}}_i + \lambda\mathbf{I}$. The term $\lambda\mathbf{I}$ adds a constant to the diagonal of the covariance matrix which may prevent singularities. As λ increases, the main diagonal of the matrix becomes dominant, and the sample covariance matrix simplifies towards a scaled identity matrix. Simplifying the estimation of the class covariance matrices may lead to the *normal-based linear classifier* and the *nearest mean classifier* as discussed in the following sections.

3.3.3 Normal-based Linear Classifier

To reduce problems with estimating and inverting class sample covariance matrices, one can average these to obtain a more stable estimate and assume all the individual classes have this same average estimate: $\mathbf{S}_{\mathrm{W}} = \sum_{i=1}^{K}(n_i/n)\widehat{\boldsymbol{\Sigma}}_i$. The assumption of equal covariance matrices among the classes simplifies the quadratic classifier to a linear classifier since the first term and the quadratic term in equation 3.16 become invariable across all the classes and can be dropped, simplifying the equation to:

$$g_i(\mathbf{x}) = -\frac{1}{2}\widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}^{-1}\widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{x} + ln\big(\hat{p}(\omega_i)\big) \qquad (3.19)$$

The resulting discriminant rule is linear, and equation 3.18 reduces to:

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \qquad (3.20)$$

since **W** vanishes and the remaining weights simplify to:

$\mathbf{w} = \mathbf{S}_{\mathrm{W}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$, and
$w_0 = \left(ln(\hat{p}(\omega_1)) - ln(\hat{p}(\omega_2)) - \frac{1}{2}\widehat{\boldsymbol{\mu}}_1^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}^{-1}\widehat{\boldsymbol{\mu}}_1 + \frac{1}{2}\widehat{\boldsymbol{\mu}}_2^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}^{-1}\widehat{\boldsymbol{\mu}}_2\right)$.

For the two-class case, this classifier is more or less equivalent to the Fisher classifier, which attempts to maximize the following criterion with respect to **w**:

$$\mathrm{J}_{\mathrm{F}}(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{w}}\mathbf{w}} \qquad (3.21)$$

where $\mathbf{S}_{\mathrm{W}}$ and $\mathbf{S}_{\mathrm{B}}$ are the within- and between- scatter matrices, respectively, (see Abbreviations list). The solution can be found analytically by setting $\frac{\partial \mathrm{J}_{\mathrm{F}}(\mathbf{w})}{\partial(\mathbf{w})} = 0$ and solving for **w**, which leads to:

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) \qquad (3.22)$$

The resulting classifier has the same direction as the normal-based linear classifier, i.e., the hyperplanes have the same normal vector.

3.3.4 Nearest Mean Classifier

An even greater simplification to the previous analysis is to assume that all the features are uncorrelated and of equal variance, i.e., the class sample covariance matrices become of the form: $\hat{\mathbf{S}} = \sigma^2\mathbf{I}$. Replacing the sample covariance matrix of equation 3.19 by this scaled identity matrix simplifies the discriminant rule into:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left(\widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_i - 2\widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{x}\right) + ln\left(\hat{p}(\omega_i)\right) \qquad (3.23)$$

Also, the decision function in equation 3.20 remains linear but the weights simplify to:

$\mathbf{w} = 1/\sigma^2(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$, and
$w_0 = \left(ln(\hat{p}(\omega_1)) - ln(\hat{p}(\omega_2)) - \frac{1}{2\sigma^2}\widehat{\boldsymbol{\mu}}_1^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_1 + \frac{1}{2\sigma^2}\widehat{\boldsymbol{\mu}}_2^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_2\right)$.
Assuming equal class priors, the offset term reduces further to $w_0 = \left(-\frac{1}{2\sigma^2}\widehat{\boldsymbol{\mu}}_1^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_1 + \frac{1}{2\sigma^2}\widehat{\boldsymbol{\mu}}_2^{\mathrm{T}}\widehat{\boldsymbol{\mu}}_2\right)$. This classifier simply assigns an object **x** to the nearest class mean in the Euclidean sense, and hence its name.
Another generalized estimate of the class sample covariance matrix that is a combination of the above cases is given by:

$$\boldsymbol{\Sigma}_i^{\lambda} = \frac{(1-\lambda)n_i\widehat{\boldsymbol{\Sigma}}_i + \lambda n \mathbf{S}_{\mathrm{W}}}{(1-\lambda)n_i + \lambda n} \tag{3.24}$$

This provides a parameter, $0 \leq \lambda \leq 1$, for controlling the complexity of the classifier ranging from the quadratic classifier, for which $\lambda = 0$, to the linear normal-based classifier, for which $\lambda = 1$. At these extremes, equation 3.24 reduces to:

$$\boldsymbol{\Sigma}_i^{\lambda} = \begin{cases} \widehat{\boldsymbol{\Sigma}}_i\,, & \lambda = 0 \\ \mathbf{S}_{\mathrm{W}}, & \lambda = 1 \end{cases} \tag{3.25}$$

3.3.5 K-Nearest Neighbor Classifier

Returning to the Bayesian approach, we can instead use a non-parametric density estimation method, i.e., one that does not assume a model, for approximating the class conditional densities. One such method is the k-nearest neighbor density estimation; another is Parzen density estimation. In the k-nearest neighbor method, an approximation of the density is obtained locally by fixing the number of neighbors, k, around each object $\mathbf{x}$ and finding the volume of the cell or p-dimensional hypersphere centered at $\mathbf{x}$ and containing those k neighboring objects out of the total number of objects, n. The estimated density is then computed as the fraction (k/n) per volume, i.e.,

$$\hat{p}(\mathbf{x}) = \frac{k/n}{V_{k,\mathbf{x}}} \tag{3.26}$$

When more than one class is available with each having n_i sample objects, i.e. $\sum_i n_i = n$, then for a given object $\mathbf{x}$, the k nearest neighbors are located and from among these, there will be in general k_i objects belonging to class ω_i, where $\sum_i k_i = k$. The class conditional densities can then be estimated as follows:

$$\hat{p}(\mathbf{x}|\omega_i) = \frac{k_i/n_i}{V_{k,\mathbf{x}}} \tag{3.27}$$

whereas, the class prior probabilities are given by $\hat{p}(\omega_i) = n_i/n$. Now, using Bayes' theorem, we may classify an object $\mathbf{x}$ by comparing its class posterior estimates. In particular, the object is assigned to class ω_i if $\hat{p}(\omega_i|\mathbf{x}) > \hat{p}(\omega_j|\mathbf{x})$, for all $j \neq i$; that is, in other terms:

$$\hat{p}(\mathbf{x}|\omega_i)\hat{p}(\omega_i) > \hat{p}(\mathbf{x}|\omega_j)\hat{p}(\omega_j) \tag{3.28}$$

Substitution leads to:

$$\frac{k_i}{n_i V_{k,\mathbf{x}}}\frac{n_i}{n} > \frac{k_j}{n_j V_{k,\mathbf{x}}}\frac{n_j}{n} \tag{3.29}$$

Equation 3.29 simplifies to $k_i > k_j$. Therefore, object $\mathbf{x}$ is assigned to ω_i if $k_i > k_j$ for all $j \neq i$. The classifier thus reduces to an instance-based rule, where an object is labeled by a majority vote among its k nearest neighbors. Ties $(k_i = k_j)$ can be avoided in the two-class case by selecting an odd value for k. In general, ties can be resolved in a multiple of ways such as by either arbitrary assignment or based on the single closest neighbor, the closest $k - 1$ neighbors, or the closest class mean as measured from among the k neighbors. Another method is to weigh the distances of the k neighbors to the test object $\mathbf{x}$, resulting in a weighted vote where closer neighbors carry a larger weight that is inversely proportional to the distance from the test object.

The parameter k of this classifier controls its complexity and its selection is very important. Small values result in a very complex and jagged boundary that may overadapt to single instances (see Figure 4) whereas large values give a smoother boundary. In the limit, as $k \to n$, any test object will be classified the same way, i.e., always to the same class: the one with highest prior probability $\hat{p}(\omega_m) = n_m/n$ from among all the classes. The classification error is then:

$$e_{k\to n} = \sum_{i \neq m} \hat{p}(\omega_i) = 1 - \hat{p}(\omega_m) \tag{3.30}$$

For example, if a dataset consists of 3 classes $\omega_1, \omega_2,$ and ω_3 with 5, 10, and 20 training objects, respectively, then by the 35-nearest neighbor, any test object will always be assigned to class ω_3. The classification error fraction will be (5+10)/35.

As k varies between 1 and n, the classifier's error will attain a minimum for some value of k in this range. One preferred method for selecting k is to use cross-validation, such as the leave-one-out procedure, in order to find this optimal value in a systematic way.

The k-nearest neighbor is a simple classifier but which often performs remarkably well with proper selection of k using cross-validation. Its complexity can be controlled easily using this single parameter. However, with large datasets, the computational weight becomes heavy since distances to all training objects need to be computed. The classifier is also very sensitive to the scaling of features, so normalization to zero mean and unit variance is recommended prior to classification.

3.3.6 Support Vector Classifier

The support vector machine in its most basic form is similar to the perceptron. Both address binary, i.e. two-class, classification and both assume linearly separable classes and result in a linear decision surface. The support vector classifier, however, finds a single unique solution for a given dataset: one that maximizes the margin between the two classes. This margin can be seen in Figure 5 as the distance between the two canonical hyperplanes (dashed lines) where no object lies in between. The separating hyperplane lies midway between these two, and the direction of these hyperplanes as indicated by their normal vector **w**, is found such that this margin is maximized. Mathematically, recall that in 3D space, the distance between a point $A(x_0, y_0, z_0)$ and a plane $(P): ux + vy + wz + r = 0$ with normal vector $\vec{n}(u, v, w)$ is given by:

$$d = \frac{|ux_0 + vy_0 + wz_0 + r|}{\sqrt{u^2 + v^2 + w^2}} \tag{3.31}$$

which follows from the fact that this distance is that from A to its orthogonal projection H on (P), i.e., $d = AH = \frac{|\overrightarrow{AH} \bullet \vec{n}|}{||\vec{n}||}$, while noting that $\overrightarrow{AH}$ and $\vec{n}$ are parallel. The same reasoning applies in higher dimensions, and equation 3.31 becomes:

$$d = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0|}{||\mathbf{w}||} \tag{3.32}$$

Points belonging to the canonical hyperplanes satisfy the equation $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \pm 1$. Thus, the distance between a canonical hyperplane and the separating hyperplane is $d = 1/||\mathbf{w}||$, and the distance between the two canonical hyperplanes is twice as much, i.e., $d = 2/||\mathbf{w}||$. This is the margin to be maximized and is therefore equivalent to minimizing $||\mathbf{w}||$ or more conveniently $\frac{1}{2}||\mathbf{w}||^2 = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$, i.e., in preparation for taking the partial derivative as part of the usual analytic method for finding extrema.

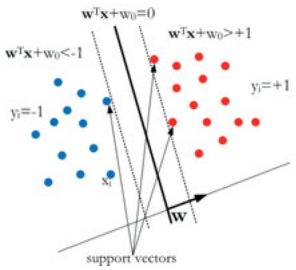

$\mathbf{w}^T\mathbf{x}+w_0=0$
$\mathbf{w}^T\mathbf{x}+w_0>+1$
$\mathbf{w}^T\mathbf{x}+w_0<-1$
$y_i=+1$
$y_i=-1$
x_i
$\mathbf{w}$
support vectors

Figure 5. The support vector classifier in the linearly separable case with two classes of labels $y_i = \pm 1$. The separating hyperplane $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = 0$ and the two canonical hyperplanes $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \pm 1$ (dashed lines) are shown. The support vectors belong to the canonical hyperplanes.

The term, $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$, needs to be minimized but subject to the constraint that the objects fall correctly to the sides of the canonical hyperplanes, i.e., all objects are correctly classified with no object lying between the two hyperplanes. These conditions translate to the following equations:

$$\begin{cases} \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + w_0 \geq +1, & y_i = +1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + w_0 \leq -1, & y_i = -1 \end{cases} \tag{3.33}$$

These constraints can be rewritten in compact form as follows:

$$y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + w_0\right) - 1 \geq 0 \tag{3.34}$$

A standard approach to minimize $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$, subject to the above inequality constraints is to use the method of Lagrange multipliers, where the constraints can be coupled with the main function to be optimized, i.e., $\nabla f - \alpha_i \nabla g_i = 0$, where f is the main function to be optimized and α_i is the Lagrange multiplier associated with each constraint function g_i. Note that the inequality constraints in equation 3.34 are in the form $g_i \geq 0$, and so these are multiplied by positive multipliers, $\alpha_i \geq 0$, and subtracted from the main function. In this case, the formulation would be as follows:

$$min\left\{L_p \equiv \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left(y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{i}} + w_0\right) - 1\right)\right\} \tag{3.35}$$

Note that in equation 3.35 there are as many constraint equations as there are training objects in the dataset, and each is associated with a single Lagrange multiplier, α_i. Taking the partial derivative of the primal form of the Lagrangian, L_p, between braces in equation 3.35 with respect to **w**, setting it to zero, and solving for **w**, gives the result:

$$\frac{\partial L_p}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{3.36}$$

Performing the same with respect to the remaining offset, w_0, gives:

$$\frac{\partial L_p}{\partial w_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.37}$$

Substituting these results back into L_p gives the dual form:

$$L_D \equiv \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \tag{3.38}$$

The inner product term $\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$ becomes very useful in nonlinear support vector machines. Particularly, when the data is transformed through a mapping $\phi(\mathbf{x}_i)$ into an often higher-dimensional, kernel space where the classes are assumed to become linearly separable, the inner product $\phi^{\mathrm{T}}(\mathbf{x}_i)\phi(\mathbf{x}_j)$ can be expressed as a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^{\mathrm{T}}(\mathbf{x}_i)\phi(\mathbf{x}_j)$ without having to define the mapping $\phi(\mathbf{x})$ explicitly or transform the data in practice. In this case, one can even assign the exponential kernel, which is an infinite polynomial. This replacement of the inner product by a general kernel function is known as the *kernel trick*. Apart for the polynomial kernel, the radial basis function (RBF) kernel is commonly used and is defined as:

$$K(\mathbf{x}_l, \mathbf{x}) = exp\left(-\frac{\|\mathbf{x}_l - \mathbf{x}\|^2}{\sigma^2}\right) \tag{3.39}$$

This is very similar to a circularly symmetric Gaussian with covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. The parameter σ controls the complexity, where small values result in highly complex boundaries and large values in smoother boundaries.

The nonlinear support vector classifier is derived in Appendix A.1 of Paper III and is used with a radial basis function kernel in the mentioned paper.

Returning to equation 3.36, note that the normal vector defining the direction of the separating hyperplane, $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$, is expressed in terms of the objects $\mathbf{x}_i$ and Lagrange multipliers α_i. The Lagrange multipliers are obtained by optimizing the dual form L_D in equation 3.38 which is a quadratic function, subject to the constraints $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. This can be done using quadratic programming with several standard software packages available for this purpose. The convex optimization returns a unique global solution, where most of the multipliers are of zero value, whereas the objects associated with $\alpha_i \neq 0$ are called the *support vectors*. These are the landmark objects that define the hyperplanes. Therefore, in fact the expression for $\mathbf{w}$ simplifies to a sparse linear combination of the objects, $\mathbf{w} = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i$, where the summation is only over SV, the set of support vector indices.

The support vector classifier often performs well in high-dimensional feature spaces even with a relatively limited training set since it relies on support vectors and the geometry of the data rather than on estimating the conditional probability density function, which becomes expensive in high-dimensional spaces and requires a large training set. Comparatively, the Bayesian approach appears to acquire much more information than needed before assigning labels to specific objects. The complexity of the nonlinear support vector classifier can be controlled by the kernel parameter (e.g.: the order of the polynomial kernel or the width of the RBF kernel) and the regularization cost parameter C, which appears in the formulation of the nonlinear classifier and is associated with slack variables. A sensible way for selecting these parameters is through k-fold cross-validation or similar testing. With the kernel method, the flexibility of the classifier can be very high and may range from simple to highly complex and detailed boundaries. However, determining these parameters by cross-validation can be expensive. Moreover, the use of quadratic programming for solving the constrained optimization problem is likely to be in most cases computationally expensive, and for large datasets exceeding a few thousand sample objects, often a special-purpose optimizer is employed.

The classifier is formulated based on the binary classification problem. Extension to the multiclass case often involves a one-against-one or a one-against-all classification scheme, both of which increase the computational load. In the first setting, $K(K-1)/2$ binary classifications are performed for a given test object and the results are combined into a final classification. In the second, K binary classifications are performed and in the simplest case, the object $\mathbf{x}$ is assigned to the class for which the distance from the separating hyperplane is the largest, i.e., the most confident case. Another possibility is to couple the K classifications within the constrained optimization problem, however, this strains the quadratic programming routines significantly.

The support vector classifier’s formulation leading to an optimization problem is mathematically elegant and yields a unique global solution for a given data, kernel parameter, and regularization parameter. In some contexts, this has been compared against feed-forward artificial neural network classifiers where the solution in these latter can vary with each run depending on the random initialization, while noting that the neural network classifier has the disadvantage that it often converges to a local minimum in its error minimization. However, over time it has been acknowledged that it is this seeming ‘disadvantage’ which makes the classifier resilient and highly adaptive and tunable in relation to the bias-variance tradeoff. The burden remains in selecting a suitable network architecture and avoiding overfitting in the presence of a large number of network weight parameters. A common method is to use cross-validation in the training of the network, whereby after every fixed number of iterations the classifier is tested using the validation set. Although the apparent error might always decrease, as soon as the test error begins to rise, the training is stopped so as to prevent overfitting and avoid large network weights.

3.3.7 Multiple Instance Learning

Multiple instance learning (MIL) can be seen as a generalization of supervised learning in which the labels of single objects are not known, but rather only the label of a bag (set of instances) is known. The bag can be regarded as a compound, non-simple object. When the bag is reduced to a single instance, multiple instance learning simplifies to standard supervised learning. The MIL problem is a naturally occurring one, but it mostly came into light with the drug activity prediction problem that was addressed by Dietterich et al. in [33]. In particular, a molecule has a potential for qualifying or developing into a drug if it is able to bind to its target area effectively. Dietterich et al. sought to develop a learning algorithm that may predict such a potential. The problem was that a molecule may have numerous alternative conformations or shapes, most of which will not bind to the target site. Only the end result of whether the molecule is effective or not could be recognized by the biochemist, but it was not known which of the molecule’s many configurations was responsible for locking onto the binding site. In other words, the multiple configurations of a molecule, i.e. the single instances, are unlabeled whereas the bag, i.e. the type of molecule itself, carries a label that can either be positive or negative. The bag has a negative label if none of its instances are able to bind to the target site, whereas it carries a positive label if at least one of its instances is able to bind to the target site. This asymmetry in assigning labels complicates the classification task. This type of problem can also be seen in image analysis, where one can think of an image as a bag consisting of a complex assortment of instances derived from local patches, sub-regions, or clustered areas spread over different parts of the image. In medical image classification, this approach can be used for example when a subject is known to have a certain illness or not, while it is unclear which part of the image signals the pathology or which feature vectors contribute to this labeling. In natural scene classification, an image may be transformed or clustered into a matrix of blobs, and the feature vectors derived from each of these sub-regions become the unlabeled instances, whereas only the image as a whole has a known label.

The MIL problem can be summarized using our usual machine learning notation as follows. Given a training set of objects, $\mathcal{L} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, the objects themselves are not labeled but are grouped within bags, $B_1, \dots, B_k$, where $B_G = \{\mathbf{x}_i : i \in G\}$, and G is the set of object indices belonging to bag B_G. The number of instances contained in each bag can vary from one bag to another. Each bag B_G carries a label $Y_G = +1$ or $Y_G = -1$. In the case when the bag label is negative, $Y_G = -1$, then all the included instances can be assigned a negative label, $y_i = -1$, for all $i \in G$. Otherwise, if the bag label is positive, $Y_G = +1$, then we are certain that there exists at least one object $\mathbf{x}_i$ in that bag which can be assigned a positive label, $y_i = +1$. Such an object is an example of an instance that would belong to the so-called *concept*, i.e., it is a key object that contributes to the positive bag label. The MIL problem reduces to standard supervised learning when each instance in the data is its own bag, i.e., $B_i \equiv \mathbf{x}_i$.

Several classifiers have been formulated for the multiple instance learning problem (e.g. [34, 35]), however a recent approach in [36] is particularly attractive as it uses the dissimilarity representation by computing the distances among the bags in order to transform the MIL problem back into a standard supervised learning setting, where any typical classifier can be utilized. This approach has been adopted in Paper V and was used for the classification of glandular tissue.

In the following sections, we turn our attention to some of the digital image analysis methods used in Papers I and V, namely, morphological operations and granulometry.

3.4 Digital Image Processing

Most commonly, an image can be defined as a matrix of numbers, $\mathbf{M}[i,j]$. This matrix can be viewed on a screen using grayscale shades in the form of an image display (note that a color image is a stack of 3 such matrices, one for each of the red, blue, and green components superimposed). Every element in this matrix is called a pixel. For a 1-bit image, the only value a pixel can take is either 0 or 1, resulting in what is called a binary or black-and-white image. For an 8-bit image, the possible range of values are: $[0,\ldots,255]$, and this is the most standard representation. In general, for a b-bit image, the number of possible grayscale values is 2^b, ranging from 0 to $2^b - 1$.

3.4.1 Neighborhood Connectivity

Many distance measures can be defined between pixel locations. A distance measure between two pixel locations $\mathbf{p}(p_x, p_y)$ and $\mathbf{q}(q_x, q_y)$ is a *metric* if it satisfies the following 4 conditions:

1) $D(\mathbf{p},\mathbf{q}) \geq 0$.
2) $D(\mathbf{p},\mathbf{q}) = 0$, if and only if $\mathbf{p} = \mathbf{q}$.
3) $D(\mathbf{p},\mathbf{q}) = D(\mathbf{q},\mathbf{p})$.
4) $D(\mathbf{p},\mathbf{z}) \leq D(\mathbf{p},\mathbf{q}) + D(\mathbf{q},\mathbf{z})$.

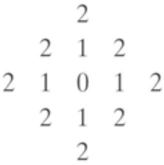
The last condition is known as the triangle inequality. The most common distance metric is the Euclidean distance defined as:

$$D_e(\mathbf{p},\mathbf{q}) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \qquad (3.40)$$

Another distance function is the city-block distance, D_4, defined as follows:

$$D_4 = |p_x - q_x| + |p_y - q_y| \qquad (3.41)$$

The city-block distances around a central pixel are shown in Figure 6. The set of pixels at which the D_4 distance from the central pixel $\mathbf{p}$ equals 1 form the 4-connected neighbors or simply the *4-neighbors* of $\mathbf{p}$, denoted by $N_4(\mathbf{p})$. These are basically the immediate horizontal and vertical neighbors.



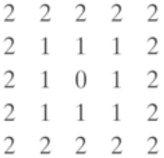
		2		
	2	1	2	
2	1	0	1	2
	2	1	2	
		2		

Figure 6. City-block distances around a central pixel.

Another distance metric is the chessboard distance or D_8 distance, defined as follows:

$$D_8 = max(|p_x - q_x|, |p_y - q_y|) \tag{3.42}$$

The chessboard distances around a central pixel are shown in Figure 7. The set of pixels at which the D_8 distance from the central pixel $\mathbf{p}$ equals 1 form the 8-connected neighbors of $\mathbf{p}$, denoted by $N_8(\mathbf{p})$. These are the immediate horizontal and vertical neighbors, in addition to the diagonal neighbors.



2	2	2	2	2
2	1	1	1	2
2	1	0	1	2
2	1	1	1	2
2	2	2	2	2

Figure 7. The chessboard distances around a central pixel.

Distance transforms can transform a binary image into a distance map by labeling each pixel with the distance to its closest foreground or object pixel. For computational considerations, these transforms are often computed by propagating local distances using small neighborhoods. Such local distance functions can for example be defined based on the city-block or chessboard distances. Optimal local distance functions that minimize the maximum difference of the distance transform to the Euclidean distance were proposed in [37].

3.4.2 Morphological Operations

A pixel location in an image coordinate system can be regarded as a vector $\mathbf{a}$ in $\mathbb{Z}^2$. An object or region of interest in an image can then be defined by a set of vectors, A. The *translation* of A by some vector $\mathbf{z}$ is denoted by: $(A)_{\mathbf{z}} = \{\mathbf{c}|\mathbf{c} = \mathbf{a} + \mathbf{z},\ \text{for}\ \mathbf{a} \in A\}$. A *reflection* of a set A is defined as: $\hat{A} = \{\mathbf{w}|\mathbf{w} = -\mathbf{a},\ \text{for}\ \mathbf{a} \in A\}$. Following this, we may define a *dilation* of a set A by a set B as follows: $A \oplus B = \{\mathbf{z}|(\hat{B})_{\mathbf{z}} \cap A \neq \emptyset\}$. The set B is then called a *structuring element*. One can think of a dilation as an operation that tries to expand the object A by means of the structuring element B. This latter is first flipped around its origin, and is imagined to be sliding over A, trying to escape from it from all directions but without ever being allowed to completely detach from A. The extent the origin of $\hat{B}$ can reach outside of A in any direction then becomes part of the dilated version of A.

An *erosion* of a set A by a structuring element B is given by: $A \ominus B = \{\mathbf{z} | (B)_{\mathbf{z}} \subseteq A\}$. Erosion can be thought of as shrinking the size of object A by means of the structuring element. More precisely, it is the possible positions that the origin of the structuring element B can assume in the image while B remains completely entrapped within A.

An *opening* of a set A by a structuring element B is defined as: $A \circ B = (A \ominus B) \oplus B$. By first eroding, then dilating by a structuring element B, one can for example eliminate narrow bridges or protrusions between connected subregions of the object A since these areas may not survive the erosion operation and may become completely disconnected before dilation is applied. Any small isolated regions can also be removed permanently by this operation if they happen to be smaller in size than the structuring element. This latter property is specifically of interest since it is used in morphological granulometry.

A *closing* of a set A by a structuring element B is defined as: $A \bullet B = (A \oplus B) \ominus B$. By first dilating, one can close small holes or gaps within the object A, which once filled are not affected by the subsequent erosion operation, and thus the closure of these regions would be irreversible.

In the above discussions, we have assumed that the image is binary, that is with only two possible pixel values: zero (black) or one (white). However, these operations also extend to grayscale images, where they affect dark and bright details in the image. In the context of this thesis, we limit the discussion to the binary case in accordance with Papers I and V. Moreover, morphological opening is of particular interest in this setting since it is applied in granulometry, which has been used in Paper I.

3.4.3 Morphological Granulometry

In image analysis, granulometry is a method for determining the size distribution of particles in an image without having to segment the image or separate the particles. It is a robust method that uses a series of morphological opening operations applied over the image. The process is analogous to applying a sequence of sieves with different hole sizes for separating the different types of grains based on their sizes. For our purpose, we deal with binary images consisting of disc-like microarray cores in Paper I. By applying a series of openings with a disc structuring element of increasing size (controlled by its radius), the cores that are smaller in size than the structuring element will disappear sequentially from the image. The total surface area of the foreground is computed after each opening, and this area will be monotonically decreasing as a function of the structuring element's radius since the particles will vanish from the image gradually from the smallest to the largest if there is more than one group size. The finite difference of this monotonically decreasing curve can reveal the size distribution of the particles. This is simply indicated by the peaks in the finite difference curve, which signal large drops in the surface area, at which an entire group of similarly-sized particles disappear from an image.

Morphological granulometry as employed in Paper I can be summarized using the following pseudo-code:

```
for $r = 0$ to $r_{max}$
    $\mathbf{se}(r)$ ← disc structuring element with radius $r$
        $\mathbf{I}_o$ ← opening of image $\mathbf{I}$ by $\mathbf{se}(r)$
      $S(r)$ ← compute foreground surface area in $\mathbf{I}_o$
end
```

The curve of $S(r)$ is monotonically decreasing, and the finite difference curve, $|S(r) - S(r-1)|$, over the proper domain reveals the particle size distribution.

4 Summary of Publications

Through Papers I to V of this thesis, we move gradually from first trying to detect tissue cores arranged in large microarrays, then to segmentation and color decomposition methods for analyzing these tissue cores once detected, and ending with two subsequent applications relating to the identification of immunostaining patterns and classification of glandular tissue. Particularly, in Paper I, we discuss how to automate the localization of single tissue cores in large microarrays. The remaining papers focus on methods relating to these single tissue cores and biopsies. In Paper II, we propose a new method for segmenting stained tissue sections based on a blind color decomposition scheme. In Paper III, we discuss how to optimize the use of staining and color combinations in tissue sections for automation purposes and present a framework for doing so. Finally, Papers IV and V present two applications that are based on the preceding papers. In Paper IV, we use color segmentation as a primary step and propose a combined, correlation-based analysis to identify paired antibodies across adjacent tissue core sections. In Paper V, we start with color decomposition and present a new and practical feature descriptor for classifying glandular tissue as a basis for Gleason grading and tubule-based Elston grading in prostate and breast tissue respectively.

4.1 Paper I: Microarray Core Detection by Geometric Restoration

4.1.1 Problem Description

The aim of Paper I is to accurately detect and localize tissue microarray cores in an automated manner that allows for high-throughput processing of images from whole slide scanners. Tissue microarrays often consist of tens to hundreds of tissue cores or discs that are arranged in arrays [8]. Each core represents a tissue cross-section that has been cut out from a gland and stained for histology. Whole slide imaging scanners are able to scan the entire arrangement from one slide in a timely manner and provide images of these tissue cores collectively as in Figure 8. Automating the analysis of these images would thus allow for high-throughput processing of tissue cores, which is needed for a variety of purposes such as paired antibody identification (see Paper IV) or glandular tissue classification and cancer grading (see Paper V). The task however is not made easy due to the following problems which are often present in images of tissue microarrays.

1. The preparation procedure of microarrays often results in many irregularities in core alignment and grid geometry due to mechanical strain, human error, and acquisition factors. Examples of these irregularities include geometric distortions, misalignments, irregular distributions, and variable inter-core spacing (see Figure 8).

2. Often many cores in the microarray will be either missing entirely or missing parts of their disc, and this fact along with variable array-spacing can affect several methods found in the literature. Inhomogeneity in the image can also affect methods that rely upon profiling gray-level intensities and which are sensitive to such variations as in [38].

Common existing methods for detecting microarray cores often rely on template matching and the Hough transform for circle detection. However, these methods bear several drawbacks. Template matching often attempts to fit a rectangular grid over the microarrays. Fitting a grid over the tissue cores is not always justifiable due to the possibility of having an irregular grid geometry. In addition, the method requires prior knowledge of grid parameters such as the size of the cells or distance between grid points, which may vary from case to case and thus hinder high-throughput analysis. The result is similar to a hit or miss scenario: if the grid geometry is regular, the assumptions are fulfilled, and hence high detection rates may be achieved; however, if the grid assumptions are violated, very low detection rates would incur. Template matching is often computationally expensive due to the range of rotations and translations it must account for. Another common approach for core detection is the Hough transform [39, 40]. In this context, the method operates in a three-dimensional parameter space and is also computationally expensive, requiring careful optimization for managing the 3D accumulator. Reliance on maxima detection in parameter space is a difficult task and makes the method very sensitive to noise and variations, especially with the fact that there is less evidence in parameter space for defective cores with missing parts than for cores that are complete. While, generally, the Hough transform for circle detection is a valuable technique for exclusively detecting circles from among several different geometric shapes in an image, it is unnecessary in this particular application, since otherwise we would be tackling a different and more general problem. In principle, tissue microarrays contain only disc shapes, and any deviations from these (including irregular shapes arising from missing core portions) should not be discarded as being non-circles, but should be detected as if the cores were complete.

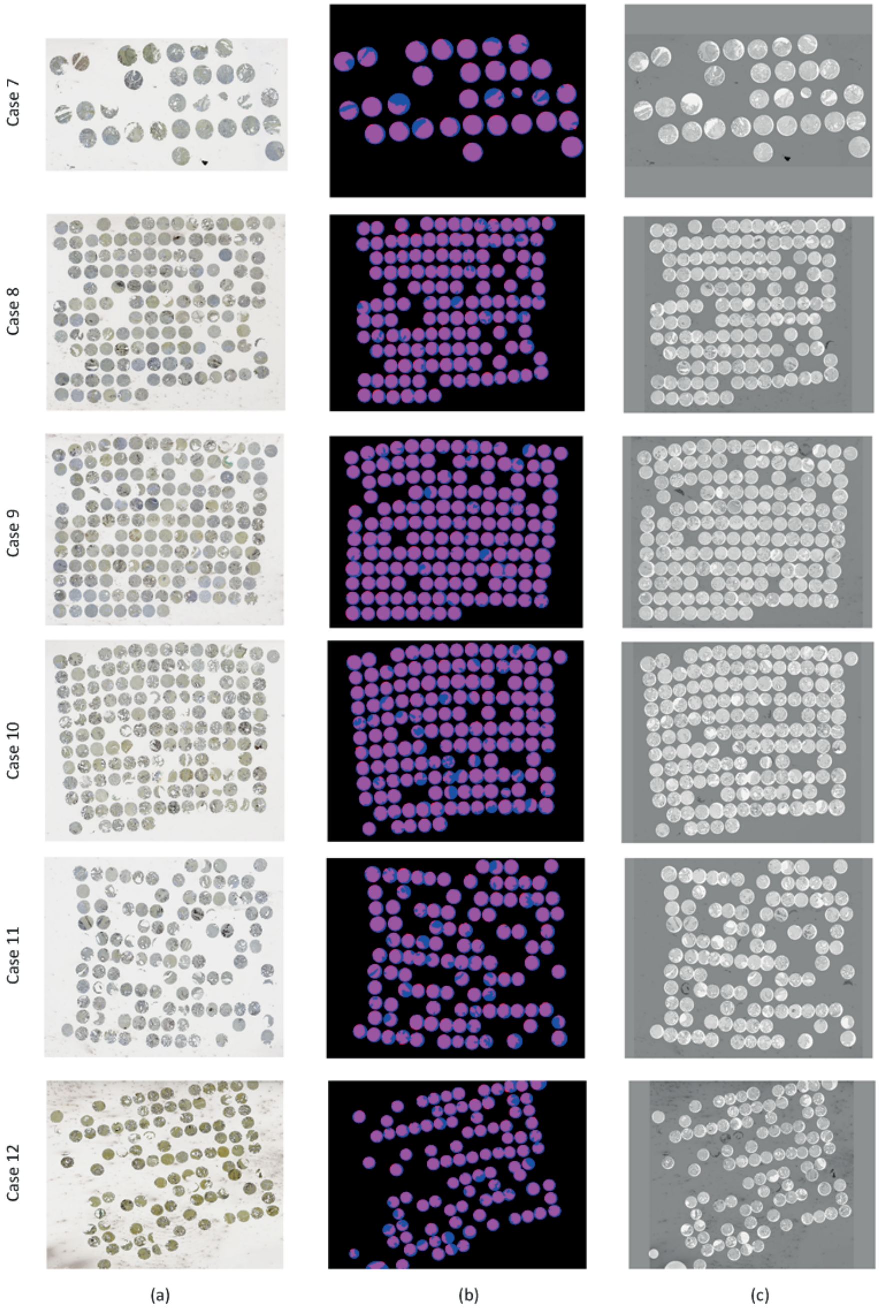


Figure 8. (a-b-c). The application of the proposed method to tissue microarray images. (a) Original TMA image. (b) Objects following morphology overlaid against reconstructed circles. (c) Original image overlaid against reconstructed circles.

4.1.2 Proposed Method

In this section, we explain the proposed method and summarize the results of Paper I for solving the microarray core localization problem; however, we have chosen to present the different stages of the method in reverse order as compared to the published version of the paper for the purpose of explaining the motivation behind the approach.

The main idea of the method proposed in Paper I for detecting microarray cores is to apply very basic, rapid morphological operations to detect bits and pieces of the cores (see Figure 9), followed by a randomized three-point method for reconstructing ideal discs around the detected core portions such that each disc covers the entire core beneath (see Figure 10). Morphological operations alone are often unreliable in detecting cores in a way that preserves their shape and contours because one would have to change the parameters of the structuring element from case to case, sometimes even within the same image. Therefore, by placing minimal expectations on these operations, we have in designing our method, shifted the burden unto circle reconstruction, starting from core portions as detected by morphology. A circle can be reconstructed using any three non-collinear points of the boundary. Thus, the method first computes the convex hull of the detected core region to discard concave sections of the boundary, and three points are randomly selected from the set of points of the convex hull such that the points are maximally distant from each other, i.e., apart by about one third the cardinality of the set. We repeat this randomization ten times, hence generating ten candidate circles around each core portion, and then select the largest of these circles as the winner, in order to ensure that the entire core portion (and its idealization) is covered as shown in Figure 10. In summary, the strategy of the method is to detect less-than-ideally, and thereafter idealize. The detection phase using morphology is less than ideal but rapid and efficient, and the idealization phase using circle reconstruction is analytic and thus highly efficient and computationally inexpensive.

However, in order to completely automate the process described above, we require some further provisions, which necessitated adding Stages 1-3 to the overall algorithm (see Figure 11). Circle reconstruction is the last stage of the method (Stage 5) and operates on the convex hull boundary resulting from the previous stage consisting of morphological operations (Stage 4); the reconstruction requires no active parameters of its own. However the basic morphological operations require as input: (1) the type of structuring element to be used for these operations and (2) its corresponding size. Because the tissue cores are disc-like, the type of structuring element used was also a disc. The disc has only a radius parameter which is to be set in direct proportion to the radius of the microarray cores present in the image.

To determine the radius of these cores in an automated manner, we convert original images into binary form (Stage 1) and use two levels of estimation for this radius, namely, a first rough estimate using cluster validation (Stage 2), which we then use to obtain a second refined estimate of the radius using morphological granulometry (Stage 3). In Stage 2, we use hierarchical clustering [41] with complete linkage since it is ideal for spherical clusters, while setting the number of clusters to vary over a large range of values. At each of these values, we assess the clustering result using the Davies-Bouldin index [31] for cluster validation since it is suitable for validating spherical clusters. We hence obtain a curve as in Figure 12(c); the minimum of the curve indicates the optimal number of clusters to select for the image. Once this number is determined, the overall foreground area of the tissue cores is divided by this number to obtain a coarse estimate of the area of a single core or disc, from which we directly deduce the radius of the disc. This first estimate of the radius is then used to define a suitable range of radii for the morphological granulometry of Stage 4 for re-determining a better estimate of the radius itself. Morphological granulometery is a very systematic, robust, and well-tested method for determining the size distribution of particles in an image. Using a disc structuring element, morphological granulometry requires as input only a range of radii to be tested. It proceeds by applying morphological opening while sequentially increasing the disc radius; all tissue cores smaller than the size of the disc structuring element disappear by the opening operations. Once a distribution profile is obtained as in Figure 12(d), we compute its first-order finite difference since we are interested in areas of the profile with large jumps at which entire groups of cores disappear from the image (see Figure 12(e)). In summary, this allows us to detect a refined estimate of the radius of tissue cores so that the entire process becomes automated from beginning to end.

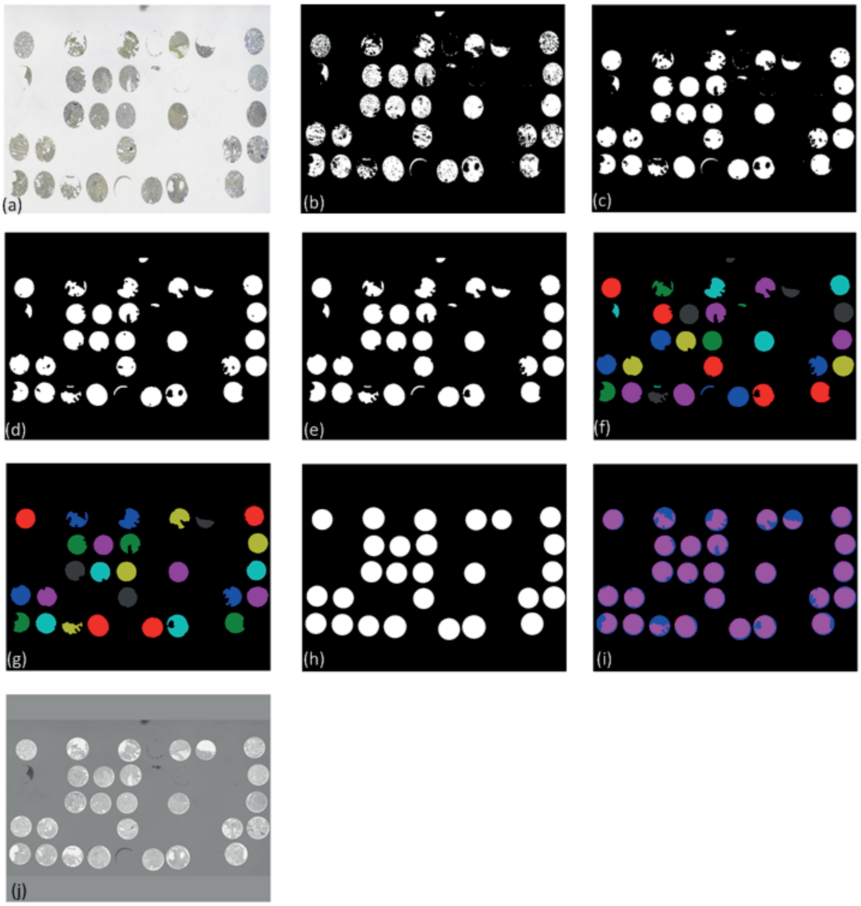


Figure 9. An example showing the morphological operations (Stage 4) in sequential order and the result of circle reconstruction (Stage 5). (a) Original TMA image. (b) Global thresholding by Otsu's method. (c) Morphological closing. (d) Median filtering to remove speckles. (e) Region filling of holes. (f) Object labeling by 8-connectivity. (g) Filtering out any small objects by their respective areas using the size condition mentioned in Stage 4 of the paper. (h) Reconstruction of circles from the objects in (g). (i) Superimposing the objects with their reconstructions, that is (i)=(g)+(h). (j) An overlay of the original TMA image with the reconstructed circles, that is (j) = (a) + (h).

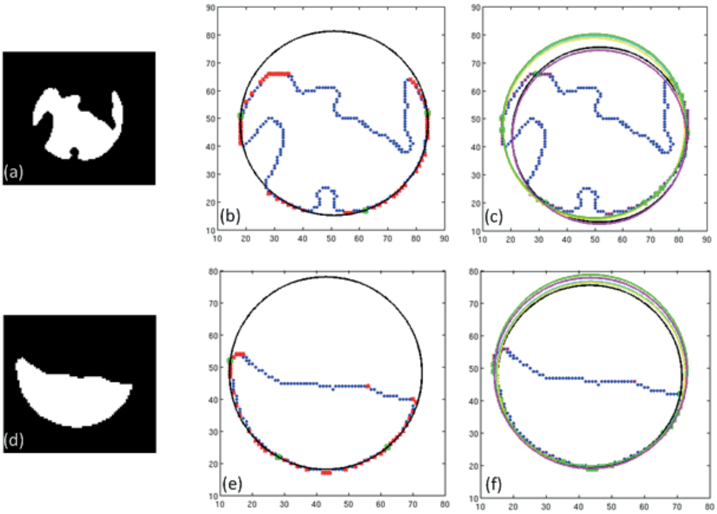


Figure 10. The circle restoration algorithm (Stage 5) applied over two defective cores. (a)/(d) Object to be reconstructed into a perfect circle. (b)/(e) A single circle reconstructed from three vertices (marked by green squares) chosen randomly from the convex hull which consists of pixels marked in red. (c)/(f) Multiple circles are generated (shown in different colors) by randomly varying the initial three vertices; only the largest of the circles is retained.

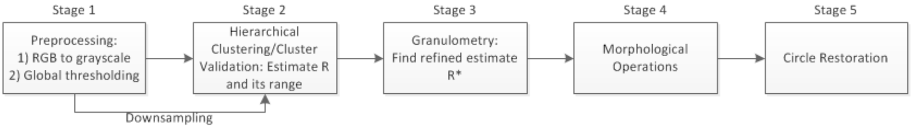


Figure 11. Flowchart of the proposed method. '*R*' represents the approximate radius of the cores. Stages 1-3 are designed for obtaining a radius estimate to use for the disc structuring element in Stage 4. The randomized 3-point circle reconstruction method is applied in Stage 5.

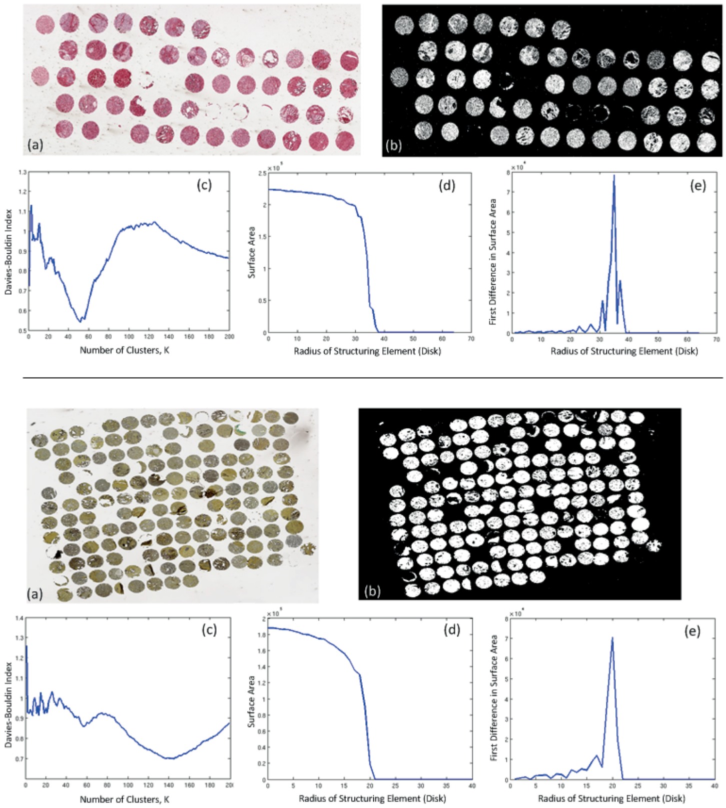


Figure 12. An illustration of morphological granulometry. (a) Original TMA image. (b) Global thresholding by Otsu's method. (c) Davies-Bouldin index versus number of clusters. (d) Granulometry function. (e) Object size distribution. The value at which the Davies-Bouldin index attains a minimum in (c) is considered to be the optimal number of clusters to use. The curve in (e) is the finite difference of that in (d).

4.1.3 Contributions

In Paper I, we have addressed the automation of tissue microarray core detection and localization as a precursor for high-throughput image processing in histopathology. The novelties of the proposed approach can be summarized by the following points.

1. The problem was transformed into that of shape restoration by decoupling detection from localization and by de-emphasizing reliance on the detection phase, while shifting the weight unto the restoration phase. The locations of tissue cores are restorable from remnants of the original cores, and the reconstructed discs cover the entire core regions.

2. An unsupervised method with cluster validation was used in order to provide morphological granulometry with the required input, thereby making the entire process completely automated.

3. The randomized, three-point reconstruction method used in the paper is made possible through augmenting the image coordinate system to perform cross product calculations. The reconstruction is analytic and simple; thereby it reduces computational time significantly, especially in comparison to template matching and Hough transform methods.

4. The proposed method does not place assumptions on core arrangement or grid geometry, and thus the usual problems that other methods experience such as with misalignments and partial cores become irrelevant in this context.

5. The method is capable of detecting defective cores despite their irregular shapes, just as simply as complete cores, while the pathologist retains the choice of accepting or rejecting these cores for subsequent analysis.

In summary, the methods we have found in the literature regarding microarray detection and localization, proceed by determining linear or curvilinear boundaries or attempt to separate the cores through various ways. While the approaches are diverse, they do not alter these premises but rather attempt to improve the detection performance in one way or another. In the proposed approach we have succeeded in altering these premises altogether by decoupling detection from localization. We have used instead simple morphology for a rough detection of core portions in the image, followed by an analytic restoration of the disc shapes around these regions. This in turn allowed us to circumvent common obstacles that other approaches tend to face. The accuracy, simplicity, and computational aspect of the method make it specifically designed for high-throughput analysis of microarray images.

4.2 Paper II: Blind Color Decomposition of Histological Images

4.2.1 Problem Description

An important aspect in automating tissue image analysis is the segmentation of an image into its various tissue types. A decisive feature for performing this segmentation is often color, which is a raw representation of a 3-channel image from which various other feature measures may be derived through different transformations. Paper II addresses the topic of color decomposition of stained tissue images and introduces a novel approach that is unsupervised and outperforms existing methods in the literature. Obstacles that may hinder color decomposition arise from factors relating to tissue preparation, natural variation in spectral signatures of identically stained tissue (biochemical noise [42]), overlap between spectral signatures of different tissue components, and acquisition noise. The automation of the process for high-throughput analysis is often hampered by the need for user input in the form of pre-assigning or interactively pinpointing reference colors in color space. In the proposed approach, we formulate the color decomposition problem using a light-absorption model, while also modeling sensor noise to improve the accuracy of the method. We adopt an unsupervised strategy for determining reference colors that does not require training or feedback from the pathologist.

4.2.2 Proposed Method

The proposed method, referred to as Blind Color Decomposition (BCD), consists of a series of sequential steps. Figure 13 shows the different stages of the method in a flowchart. In summary, the algorithm proceeds as follows:

1. Any light-scattering stains [43, 44], such as the dark brown dye 3,3’-diaminobenzidine (DAB), which might be present in the image are detected *a priori* and the regions are masked out from subsequent analysis since these stains do not obey the Beer-Lambert law of absorption, and the corresponding density cannot be modeled properly.

2. Charge-coupled device (CCD) sensor noise is modeled to improve the accuracy of locating reference colors at a subsequent stage. This type of photon noise is Poisson-distributed [45, 46], and is most visible in areas of the image where the optical density is low. To model this noise, we use either a blank image or ‘white’ areas of the image, and the corresponding variances are estimated from the histograms obtained for

each color channel. These values are used to smooth the image data by generating a 3D cloud of N points around each data point and then reweighing the data points to give emphasis to those with high optical density as in [47]. However, in the end only $1/N$ of the overall data points with highest scores are retained so that the number of points is reduced back to its original value.

3. The data is linearized using the Beer-Lambert law of absorption [44, 48] and projected onto the Maxwellian chromaticity plane [49, 50] using a perspective transformation [51] with the center lying at the origin of the 3D color space and the projection plane lying at a distance of $1/\sqrt{3}$ from the origin. This transformation decouples color from intensity, whereby distances between points in the Maxwellian plane represent chromaticity differences between the corresponding colors. Geometrically, the plane is defined by an equilateral triangle within the 3D color cube, where pure RGB colors project onto the vertices of this triangle and the achromatic axis (one of the cube's diagonals) projects onto the circumcenter of the triangle [49, 50].

4. We use a Gaussian mixture model [52, 53] trained using expectation-maximization (E-M) as an unsupervised method for locating chromaticity clusters in the Maxwellian plane. We then extract the cluster centers of the Gaussian distributions resulting from the E-M fit in order to obtain the reference colors (see Figures 14-15).

5. The reference colors form the columns of the mixing matrix. Linear decomposition (or piece-wise linear decomposition, PW-LD) is then carried out by inverting the mixing matrix (or submatrices in case of PW-LD) to determine the density maps, one for each tissue type. To account for the cases where chromaticity clusters are not completely separable from each other, we have extended the commonly used linear decomposition method to piece-wise linear decomposition as a useful generalization. Piece-wise linear decomposition was implemented by considering pairs of chromaticity clusters, and thus only selecting the corresponding submatrices of the mixing matrix before computing the pseudo-inverse. Figures 16-17 show that PW-LD is able to outperform linear decomposition in difficult cases.

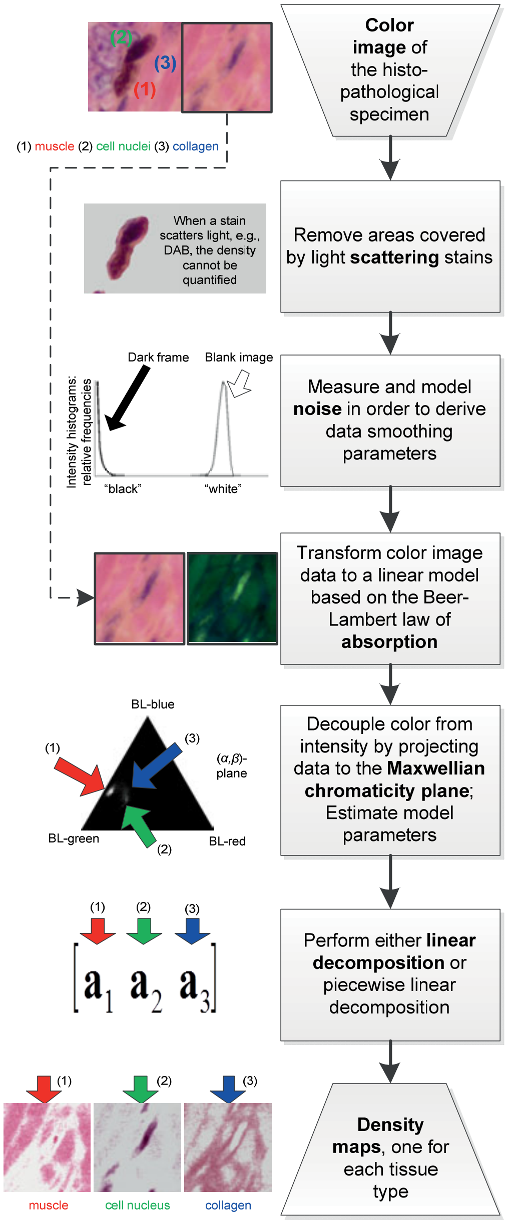

Color image of the histo-pathological specimen
(2)
(3)
(1)
(1) muscle (2) cell nuclei (3) collagen
Remove areas covered by light scattering stains
When a stain scatters light, e.g., DAB, the density cannot be quantified
Measure and model noise in order to derive data smoothing parameters
Dark frame
Blank image
Intensity histograms: relative frequencies
"black"
"white"
Transform color image data to a linear model based on the Beer-Lambert law of absorption
Decouple color from intensity by projecting data to the Maxwellian chromaticity plane; Estimate model parameters
BL-blue
(1)
(3)
(α,β)-plane
BL-green
(2)
BL-red
Perform either linear decomposition or piecewise linear decomposition
(1)
(2)
(3)
$[a_1\ a_2\ a_3]$
Density maps, one for each tissue type
(1)
(2)
(3)
muscle
cell nucleus
collagen

Figure 13. Flowchart of the BCD method (© 2013 IEEE).

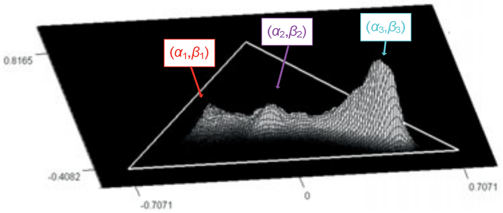


Figure 14. The initial transformation of input RGB data is based on the Beer−Lambert law, followed by a perspective projection to the Maxwellian chromaticity plane. The three vertices are associated with the transformed pure colors, which we refer to as Beer−Lambert red, green, and blue, respectively; (these colors correspond to cyan, magenta, and yellow in the original red−green−blue space). The coordinates (α_1,β_1), (α_2,β_2), and (α_3,β_3) in the plane determine the three reference colors (© 2013 IEEE).

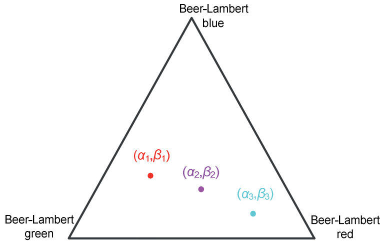


Figure 15. Position of the three reference colors (α_j,β_j) in the Maxwellian chromaticity plane (© 2013 IEEE).

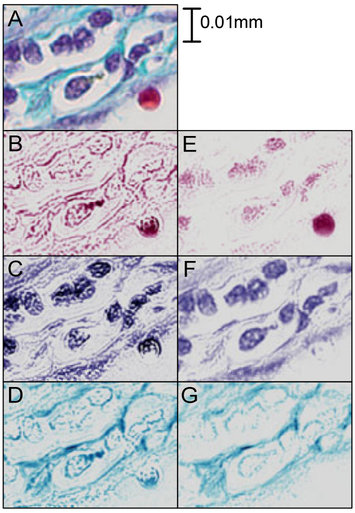


Figure 16. Pseudo colored result of decomposition of Gomori trichrome stain (A) to density maps using: in B, C, and D linear decomposition, and in E, F, and G piece-wise linear decomposition rules. In B and E erythrocytes appear red, and in C and F cell nuclei of fibroblasts, lymphocytes and smooth muscle are purple, and in D and G collagen is turquoise (© 2013 IEEE).

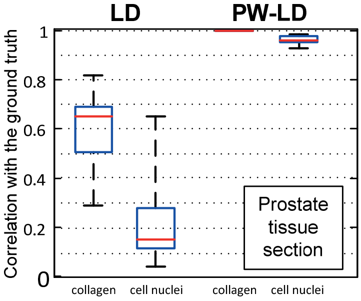


Figure 17. Box plot with the results of the comparisons of the linear decomposition (LD) and piece-wise linear decomposition (PW-LD) to ground truth for 16 fields of view of stomach tissue sections stained with Gomori trichrome (© 2013 IEEE).

To assess the BCD method, ground truth was collected by a pathologist through marking regions under many fields of view (FOV) in images of stained tissue types. The total ground truth was computed as the median of the FOV ground truths. For comparison, the same linear decomposition method and conditions were applied to BCD as to several other methods for color decomposition such as non-negative matrix factorization (NMF) [54, 55], independent component analysis (ICA) [54], principal component analysis (PCA) [56, 57], resulting in density maps for each method. By means of the Pearson correlation measure, the resulting density maps were compared against those obtained using total ground truth, and the results are shown in Figure 18.

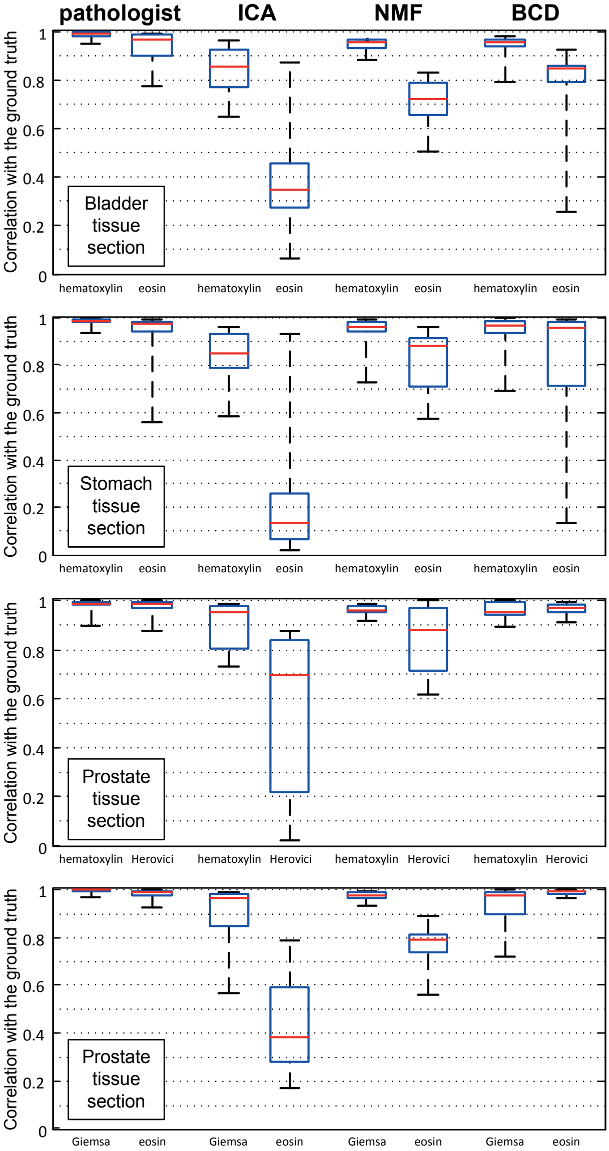


Figure 18. Box plots with the results of comparisons of the BCD method and other blind methods implemented with identical preprocessing and linear decomposition. The density maps are derived by manual selection by a pathologist (pathologist), ICA, NMF, and color decomposition based on reference colors extracted from the Maxwellian chromaticity plane. The figure shows correlations for bladder neck tissue stained with H&E, stomach tissue stained with H&E (hematoxylin-and-eosin), and prostate tissue stained with H&H (hematoxylin-and-Herovici), and G&E (Giemsa-and-eosin), respectively (© 2013 IEEE).

4.2.3 Contributions

In Paper II, we presented a novel, unsupervised method for color decomposition and have demonstrated its application in histological images in comparison with contemporary methods.

The contributions of the proposed method may be summarized as follows:

1. Color decomposition was formulated and dealt with not exclusively through data analysis but also using a physical model that takes into account light absorption characteristics and the presence of acquisition noise.

2. The method reduces sensitivity to stain intensity variations by projecting the data onto the Maxwellian chromaticity plane, thereby decoupling intensity from color information.

3. The method used to determine reference colors in the Maxwellian plane is unsupervised and uses expectation-maximization to fit a Gaussian mixture model. Thus, the pathologist does not need to provide training samples or pre-assign reference colors manually in color space.

4. Through modeling sensor noise, the method is able to improve the accuracy of locating reference colors by re-weighing the contribution of data points using the noise model.

5. Linear decomposition, which is common to most methods in this context, was extended and generalized into piece-wise linear decomposition in order to solve difficult cases where a spectral signature of a stained tissue component is not linearly separable from the spectral signatures of the remaining tissue components.

6. The method performs soft, rather than binary, classification and provides density maps, one for every stained tissue type. Due to piece-wise linear decomposition, it is also able to account for both separable and partially separable chromaticity clusters in the Maxwellian plane.

   In summary, the proposed approach uses an accurate model for light absorption and sensor noise, determines reference colors in an unsupervised manner, and performs piece-wise linear decomposition for obtaining density maps. Results have shown that it outperforms contemporary methods for color decomposition, including non-negative matrix factorization.

4.3 Paper III: Histological Stain Evaluation for Machine Learning Applications

4.3.1 Problem Description

Paper III presents a framework for selecting histological stains through objective testing and comparison such that the stain is optimal with regard to automation. When designing an automated process for a certain application in tissue image analysis, it is not uncommon that the dataset is taken for granted, without questioning possible alternatives. Yet it is usually the dataset that places a severe limit on the performance of the entire system, regardless of how advanced the algorithms employed might be. For example, when it comes to the color combinations of a stain, the intrinsic overlap between the classes corresponding to the different tissue types in color space sets a strict limitation on the classification rate, irrespective of what method is used for classification. For almost a century, the hematoxylin & eosin stain has been the standard choice for use in histopathology. With the advancement of digital histopathology and experimentation with staining protocols, it becomes necessary to question the optimality of a given stain for diagnosis and automation. When visually inspecting images, pathologists do not perceive color information out of context. Their processing is contextual and simultaneously takes into account factors such as texture, morphology, prior expectations, and high-level image understanding of anatomy and cell structure. In automating image segmentation based on color, one could evaluate different stains among possible candidates in order to choose an optimal stain for a given type of tissue. This is what Paper III advocates in principle through presenting a systematic way of carrying out this evaluation in a manner that is aligned with the final objective of automating the analysis, rather than relying on visual examination under a microscope. We present standard testing procedures for either case where the required automation is supervised or unsupervised, and we explain the motivation behind this evaluation and the ideas relating to chromaticity cluster overlap in feature space, which we believe is how the problem should be addressed if automation is desired.

4.3.2 Proposed Method

An ideal stain-tissue combination can be seen as one that results in chromaticity clusters that are compact and distant from each other in color space. The proposed method for selecting an optimal stain for a given application proceeds by performing comparative analysis among candidate stains using objective classification criteria. Since these criteria depend on the kind of automation desired, i.e., whether the automation is supervised or unsupervised, we present three different sets of evaluations: one for supervised methods, another for unsupervised methods, and a third concerning general class separability. In what follows, we summarize these three evaluation procedures and give examples for each using prostate tissue samples. However, we begin first by briefly describing the dataset used and the ground truth acquired for assessing these criteria.

Ground Truth

Thirteen different stains, listed in Table 1, were used for comparison based on classification performance. Some of these stains are shown in Figure 19. For acquiring ground truth, a pathologist labeled pixels by delineating sample regions of nuclei, stroma, and cytoplasm for each of the thirteen stains. The labeling was carried out in a conservative manner, and the dataset was balanced with regard to the number of pixels per class. Pixels selected for each of the three classes were assigned the label of the corresponding class. These labels were used to validate the classification performance of supervised and unsupervised methods against that of ground truth. All types of classifications were carried out in the Maxwellian plane [49, 50], that is, after transforming the 3D color data using the Beer-Lambert law of absorption [44, 48], and then projecting the data onto the mentioned plane (refer to Paper II). This effectively decouples intensity from color and allows us to assess the separability among the chromaticity clusters that are formed within the Maxwell triangle. Figure 20 shows an example of this data transformation for two different stains.

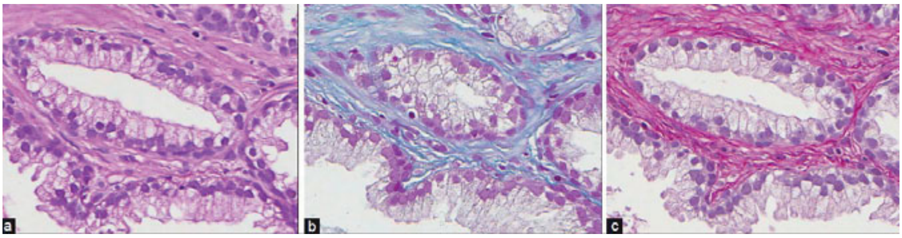


Figure 19. (a) Hematoxylin-eosin stain. (b) Mallory's trichrome stain. (c) Sirius-hematoxylin stain.

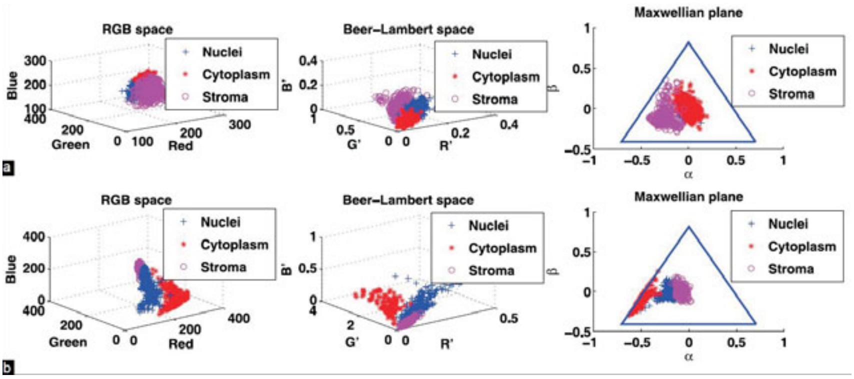


Figure 20. (a-b) Scatter plots in RGB space (left), Beer-Lambert space (center), and Maxwellian plane (right). (a) MILLERS-E stain. (b) CYTK stain. The amount of class overlap and separability based on color differs for each stain.

Supervised Classification

To evaluate a stain in terms of color decomposition in the supervised case, we used a support vector classifier with a radial basis kernel. The classifier has a wide range of complexity and two parameters that control it: the width of the RBF kernel, γ, and the regularization parameter C. These were optimized using a grid search method over a range of values for each parameter [58]. A random 25% portion of the dataset was used to perform this optimization in order to determine the best possible classifier parameters. At each point of the grid, i.e., for every possible pair of parameter values (C, γ), 10-fold cross-validation was carried out to yield a classification error. Figure 21 shows the resulting plot of classification errors at every grid point. The point corresponding to the lowest value in this plot corresponds to the optimal parameter values to be selected. Once the parameters of the classifier have been determined, we consider the remaining 75% portion of the dataset that has not been used for parameter optimization, as an independent validation set. That is, using this set, the classifier with the chosen parameters is trained and tested using 10-fold cross-validation. The 1:3 split between a dataset for parameter optimization and that for validation is random, and so we repeat the entire process 10 times and report the final average error rate and standard deviation as shown in Figure 22.

Other types of classifiers can be used, however the evaluation procedure would be similar nonetheless. The testing we have adopted is intensive, but it is possible to use less stringent validation which is not as computationally expensive, for example by reducing the number of folds used for cross-validation or the number of repetitions for splitting the dataset between parameter optimization and testing.

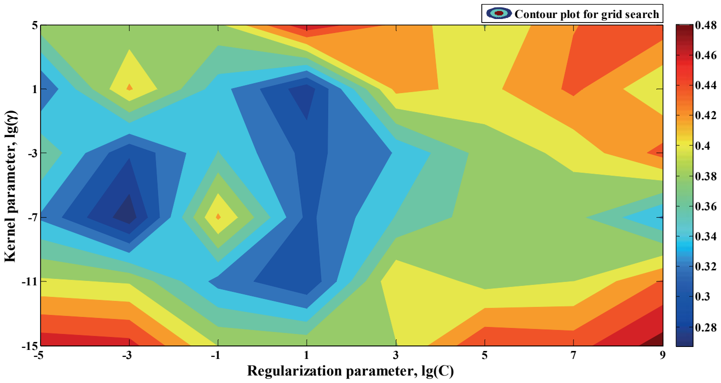


Figure 21. Optimization of SVM radial basis function kernel parameter γ and regularization parameter C by grid search. Plotted values represent 10-fold cross-validation error. Note that lg(.) is the base 2 logarithm. The optimal values for C and γ are those at which the error is minimum.

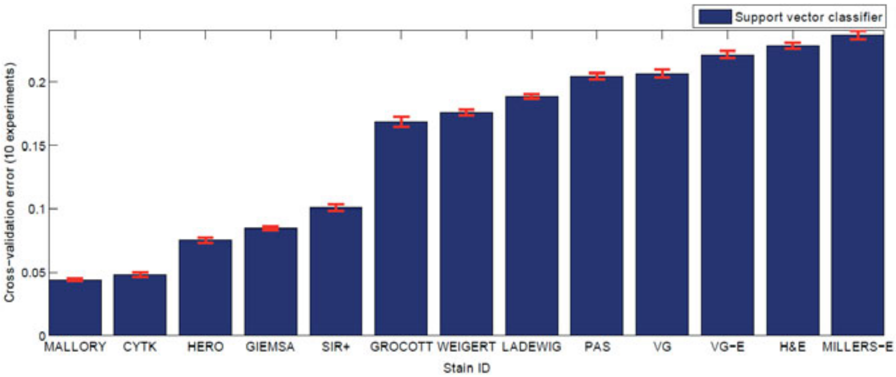


Figure 22. Ten-fold cross-validation error using an optimized SVM classifier for each stain.

Unsupervised Classification

In many situations, providing a labeled training set is seen as a laborious manual task, and therefore unsupervised learning may be desired instead. In this paper, we have used the Gaussian mixture model [52, 53] for determining the chromaticity clusters in the Maxwellian plane using the expectation-maximization algorithm, which tries to fit the mixture model over the data points while maximizing the likelihood of this fit. The model parameters are initialized randomly over several repetitions and the model with the highest log-likelihood fit is selected. The mixture of Gaussians is a flexible model regarding the final distributions it can adapt to. Figure 23 shows the final result of this fit in the Maxwellian plane for a number of stains. The contour plots in the figure are at 25% below the peak of each Gaussian distribution.

Other evaluation criteria were also used to assess clustering performance quantitatively since data labels are available. These criteria were the Rand index [59] in addition to the F_1-measure [60], which makes use of precision and recall. Computing these two criteria is not straightforward in the case of clustering since the cluster labels do not have to match the available ground truth labels but may permute and still be correct; thus the process is done in a pairwise manner over all the data points. Given any pair of points, if they happen to belong to the same cluster (C^1) while also having similar ground truth labels $(\omega_1; \omega_1)$, the pair counts towards a true positive whereas if their ground truth labels are different $(\omega_1; \omega_2)$, the pair counts towards a false positive, etc. Results of this analysis are shown in Figure 24, and the optimal stains in this case were Mallory's trichrome and sirius-hematoxylin.

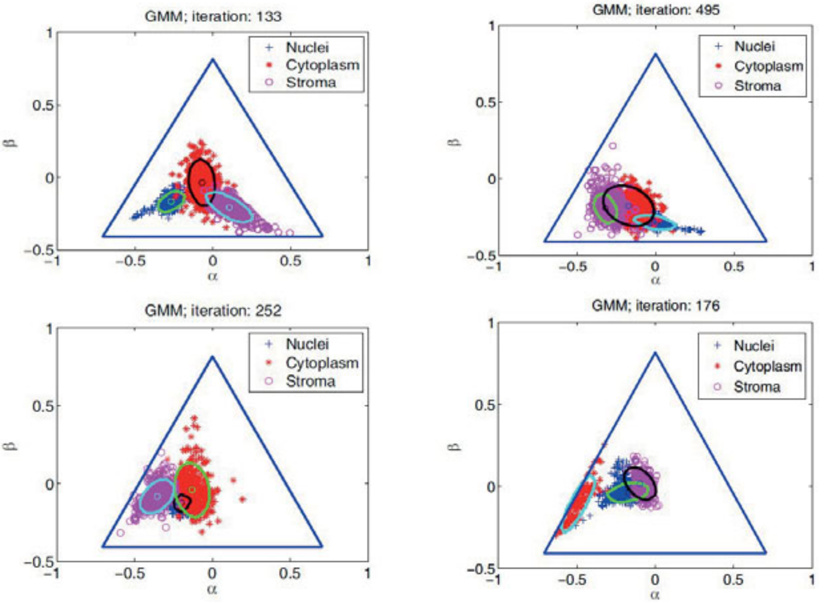


Figure 23. Gaussian mixture model clustering for different stains: Mallory (upper left), Giemsa (upper right), SIR+ (lower left), CYTK (lower right).

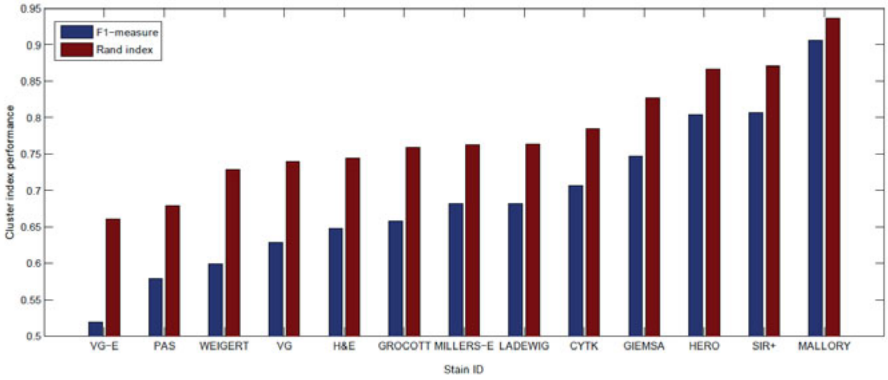
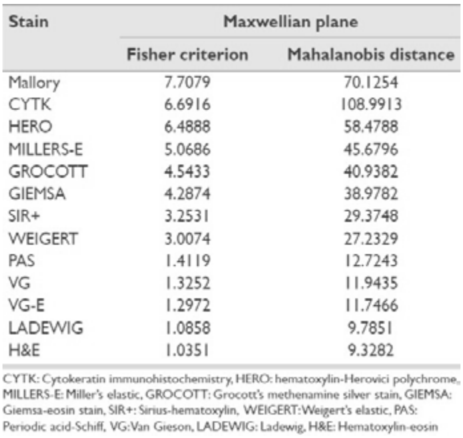


Figure 24. Cluster evaluation using the F_1-measure and Rand index.

Class Separability Measures

Regardless of whether a supervised or unsupervised approach is used for classification and automation, class separability may be assessed independently using scatter-based criteria. We used two types of measures for assessing separability in the Maxwellian plane, namely a Fisher-based criterion [61, p. 311] and the sum-of-squared Mahalanobis distance [61, p. 314], which is computed by summing up the squared Mahalanobis distances as measured between every pair of classes in the Maxwellian plane. These criteria can indicate how compact and distant class distributions appear in color space. The more compact and well-separated the classes are, the higher these measures should be and the better is the prospect for obtaining successful classification in principle. Table 1 summarizes the results as computed for all thirteen stains. Although scatter-based criteria provide useful insight into class overlap, one must keep in mind that they are only predictive of the performance of classification and may not be as accurate an indicator as the actual classification error rates. For example, it might happen as in the case of Miller's elastic, that two of the classes are almost completely overlapping which would preclude accurate classification; however, the third class happens to lie very far from the other two, hence influencing the Mahalanobis and Fisher criteria positively, yet without much consequence regarding the classification rate. In conclusion, the evaluations conducted above for supervised and unsupervised classification over prostate tissue indicate that some stains such as cytokeratin, Mallory's trichrome, and sirius-hematoxylin consistently outperform other stains.

Table 1. Values for the Fisher criterion and sum-of-squared Mahalanobis distance for the different stains.



Stain	Maxwellian plane	
	Fisher criterion	Mahalanobis distance
Mallory	7.7079	70.1254
CYTK	6.6916	108.9913
HERO	6.4888	58.4788
MILLERS-E	5.0686	45.6796
GROCOTT	4.5433	40.9382
GIEMSA	4.2874	38.9782
SIR+	3.2531	29.3748
WEIGERT	3.0074	27.2329
PAS	1.4119	12.7243
VG	1.3252	11.9435
VG-E	1.2972	11.7466
LADEWIG	1.0858	9.7851
H&E	1.0351	9.3282

CYTK: Cytokeratin immunohistochemistry, HERO: hematoxylin-Herovici polychrome, MILLERS-E: Miller's elastic, GROCOTT: Grocott's methenamine silver stain, GIEMSA: Giemsa-eosin stain, SIR+: Sirius-hematoxylin, WEIGERT: Weigert's elastic, PAS: Periodic acid-Schiff, VG: Van Gieson, LADEWIG: Ladewig, H&E: Hematoxylin-eosin

4.3.3 Contributions

In Paper III, we presented a method for quantitatively comparing histological stains based on their classification rates and clustering performance, which is well-aligned with the purpose of automation.

In conclusion, we summarize the contributions and key points of the paper as follows:

1. Prior to the designing of an automated system for tissue image analysis, the choice of stain should not go unquestioned, but may rather be treated as a variable when alternative stains are available.

2. For the purpose of automation, the choice of stain should be motivated by objective criteria, such as classification error and clustering performance, which are aligned with the final aim of the design process, rather than be based on visual inspection or habitude.

3. Some histological stains tend to consistently rank better than others for a given application, and the optimal stain is expected to vary with different applications and types of tissue. For every intended application, rigorous validation should be performed.

4. In the case of supervised classification, one may choose a classifier with a wide range of complexity. The complexity can then be determined through an optimization of the classifier parameters by means of cross-validation, and the classifier's performance is re-evaluated on an independent test set.

5. For unsupervised automation, there is a large number of measures that can be used for assessing clustering performance. Two effective criteria are the Rand index and the F-measure, computed from a pairwise comparison among the data point labels with ground truth.

6. Scatter-based criteria may also be used to gain insight concerning class separability and predict the performance of classification. These criteria are generally simple and fast to compute; however, they can be misleading in certain situations and should be treated with caution.

4.4 Paper IV: Image Segmentation and Identification of Paired Antibodies in Breast Tissue

4.4.1 Problem Description

In Paper IV, we address the problem of identifying paired primary antibodies in adjacently cut tissue sections, which therefore contain similar cell structures and anatomy. These types of antibodies may bind to the same target protein, yet to different locations of the protein. If the protein is present in particular cell structures of the tissue, then a stain such as the dark brown 3,3'-diaminobenzidine can be used, through a certain process, to stain those regions where the antibody (and thus protein) is present. Paired antibodies designed to recognize all isoforms of a protein are expected to produce similar staining patterns across adjacent tissue sections. However, in general, since paired antibodies may bind to different parts of the target protein, it is possible that a second protein contains a similar critical component as the target protein, thus resulting in unspecific binding and a different staining pattern. Therefore, comparing the staining patterns that paired antibodies produce across images of adjacent tissue sections for a specific protein provides quality control over the binding and the ability of the antibodies to identify correctly and exclusively the target antigen [62]. This may also help identify outlier antibodies that result in unspecific binding or that may have weak affinity. The study of paired antibodies in pathology can lead to the development of biomarkers for diseases and can improve prognostics. In Paper IV, we present a method for accurately segmenting and quantifying antibody immunostaining patterns in images of the Human Protein Atlas [63], as well as for automatically identifying paired antibodies through a combined, normalized cross-correlation analysis involving posterior probability maps resulting from soft segmentation of the images. The automation of this process also facilitates high-throughput protein expression analysis, which is becoming an increasingly important application. The paired-antibody problem can be studied generally for any type of protein or tissue. Some of the problems obstructing the identification of paired antibodies include variations in antigen affinity and weak absorption of stain, imperfect alignment among adjacent sections, the need for a simple and robust method for segmenting tissue types reliably, and a difficulty in how to generally assess similar/dissimilar patterns.

4.4.2 Proposed Method

The dataset used in Paper IV consisted of adjacent tissue sections obtained from the Human Protein Atlas project. These are microarray cores that have been cut out from a gland consecutively, one thin slice after the other. Adjacent sections therefore contain almost the same tissue and cell structures and are suitable for assessing the paired-antibody problem through identifying similar or dissimilar staining patterns. Each antibody is exposed to a different tissue section, and the brown DAB chromogen is used to visualize the antibody, whereas the blue hematoxylin dye is used as a (non-specific) counterstain that basically stains the tissue regardless of where the antibody is concentrated. The total number of images was 49, and these are listed in Table 2 according to their antigen groups. For a depiction of how these tissue sections appear, please refer to Figure 25.

Table 2. HPA-based image dataset.

All cell types negative	Gene AASS (6 cases)
	Gene ACOT7 (6 cases)
	Gene ANKRD2 (4 cases)
	Gene APEH (4 cases)
Glandular cells positive; Adipocytes negative	Gene ALDH6A1 (8 cases)
	Gene CTNNB1 (2 cases)
	Gene ZWINT (3 cases)
Adipocytes positive; other cell types negative	Gene PLIN1 (3 cases)
Group R (miscellaneous cases C1-C13)	C1,C2: Paired antibodies for collagen protein from gene COL15A1.
	C3,C5,C4,C6: Paired antibodies for protein product from gene FAM54B
	C7,C8,C9: Paired antibodies for the cingulin protein from gene CGN
	C10,C11: Paired antibodies for the protein product of the gene AC008073.5
	C12,C13: Paired antibodies for the protein from gene C16orf70

We began by investigating chromaticity clusters in RGB color space through sampling regions and plotting the resulting pixels in a scatter plot as shown in Figure 26(A). With the aim of performing unsupervised classification, the proximity and relative position of clusters is important, and the scaling of features plays a considerable role in modifying these factors as illustrated in Figure 27. In order to control this scaling, we use principal component analysis (PCA) as in Figure 26(B) to rotate the feature space so that the first feature is always aligned with the principal component or direction of maximum variance in the data. This standard alignment allows us to scale the features in a consistent way from case to case. Figure 26 shows two main branches, one for the brown DAB and another corresponding to the blue hematoxylin stain. Initial experimentation with clustering revealed that it is difficult to cluster the two branches while simultaneously detect clusters along each branch extension. The relative scaling of features 1 and 3 in Figure 26(B) proved critically sensitive and would need to be adapted from case to case. Therefore, we decided to separate the two branches using supervised classification, while the remaining extensions which spread naturally along the hematoxylin branch would be clustered following a rescaling of features. Thus, in summary, by removing the DAB branch from the analysis and emphasizing feature 1 over feature 3, the detection of the 3 clusters along the hematoxylin branch could be done in a consistent and reliable manner, whereas a direct attempt at clustering the space into four groups proved overly sensitive to the scaling of features.

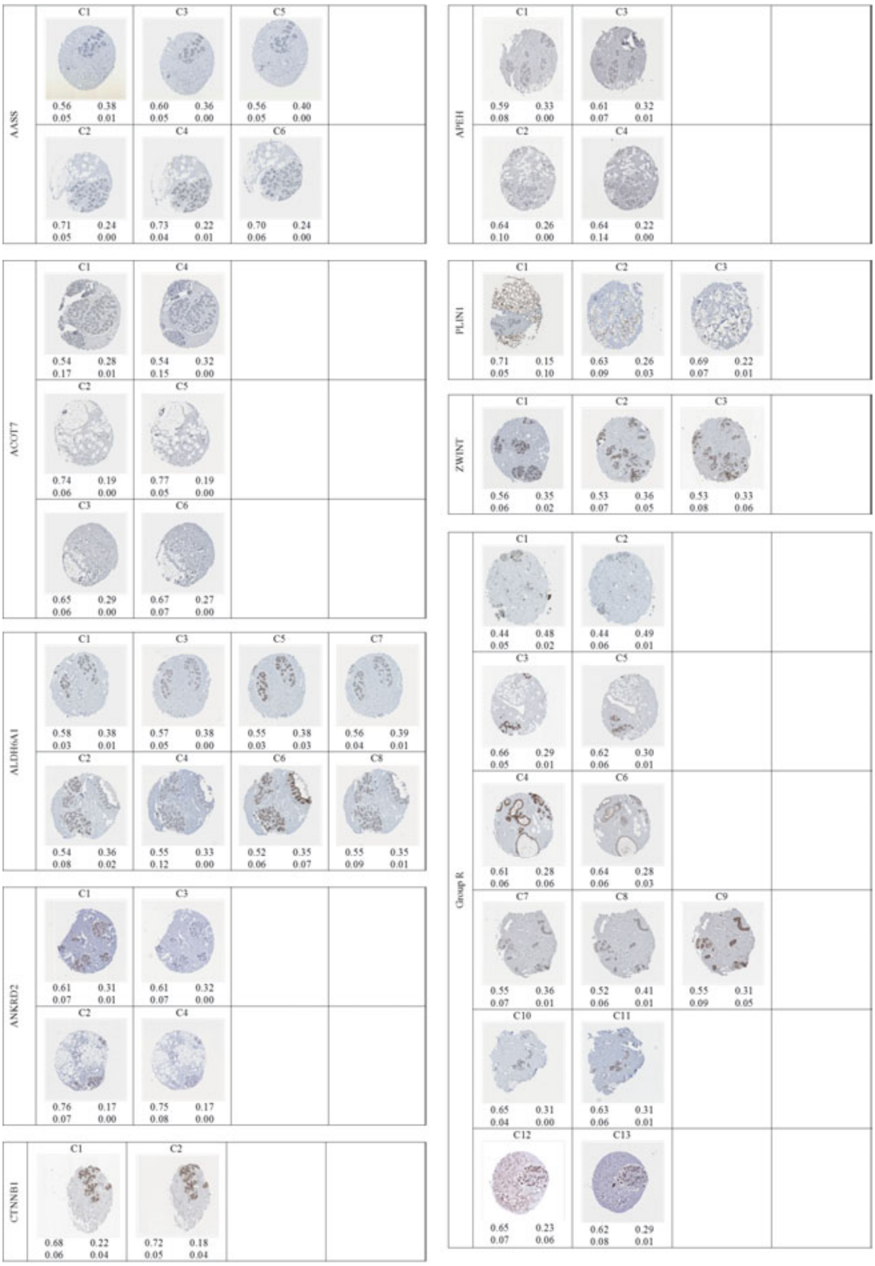


Figure 25. Sample dataset of images of breast tissue microarray sections corresponding to different antigen groups. The numeric values in each case represent the fractional ratio of the four classes in the following order $\begin{bmatrix} Lumen & Stroma \\ Nuclei & DAB \end{bmatrix}$ as quantified by segmentation.

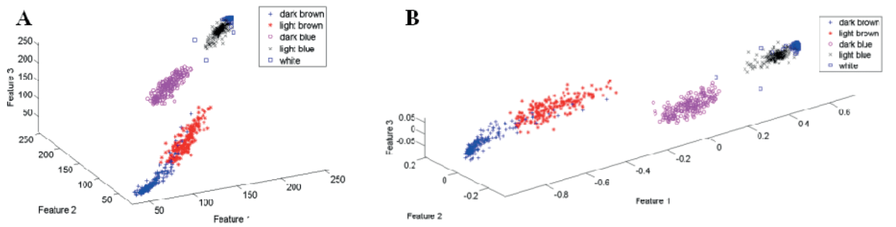


Figure 26. (A) RGB feature space for a balanced dataset; the principal direction of variance extends diagonally. (B) Feature space after PCA; the principal component is aligned to Feature 1, which may now be scaled directly prior to clustering.

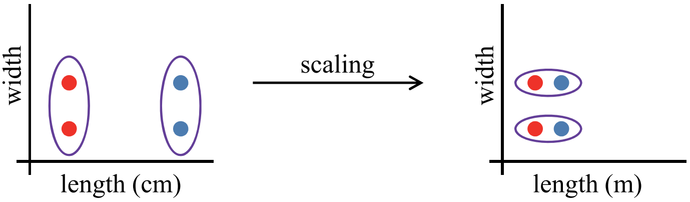


Figure 27. A schematic showing that the effect of scaling a feature with respect to another may influence the clustering result.

Thus, concerning image segmentation based on color, the proposed method consists of the following steps:

1. Perform PCA without feature reduction, and rescale the features.
2. Detect DAB regions (supervised)
3. Detect clusters along the hematoxylin branch (unsupervised).

In what follows, we summarize the aspects of the classifications done in steps 2 and 3 above, before moving on to the analysis concerning paired-antibody identification and matching, which requires the above segmentation as prerequisite.

Detection of DAB (by supervised classification)

As mentioned in Papers II and III, DAB is a light-scattering stain which does not obey the Beer-Lambert law of absorption [44, 48]. Moreover in Paper II, it was recommended that DAB-stained regions be removed *a priori* and excluded from color decomposition. We have seen in Figure 26 that DAB also impedes the use of clustering in a direct manner for this application due to the curved elongation of its branch towering over the hematoxylin branch. Therefore in order to exclude DAB-stained regions from the analysis, we identified those regions against blue hematoxylin regions and white lumen regions using a trained classifier. Our dataset consisted of sample regions from Group R, which contains 13 images from the overall 49 images of the dataset. In the end, this training is performed only once using the mentioned subset, but we also evaluate the classification using cross-validation as shown in Figure 28. The number of pixels for each of these three classes was 27000, and the classifier used was the quadratic (normal-Bayes) classifier which was trained in a dissimilarity feature space [64] derived from the RGB space by randomly selecting a representation set consisting of 50 prototypes per class. The 10-fold cross-validation error rate for this classifier, labeled 'FeatDisSpace', is shown in Figure 28(B), along with other classifiers (Naïve Bayes, linear discriminant, and Fisher) shown for basic comparison. Figure 28(A) shows the learning curves for these classifiers, from which one can observe that the error rate for the chosen classifier drops rapidly as the training set size is increased, and it levels off below that of the other classifiers. The quadratic classifier is also computationally efficient as compared to classifiers that have a higher complexity range and which may require serious optimization. Figure 29 shows the result for the detection of DAB-stained regions using the proposed classification; note that the cases are images from Group ALDH6A1, which were not part of the training/validation procedure.

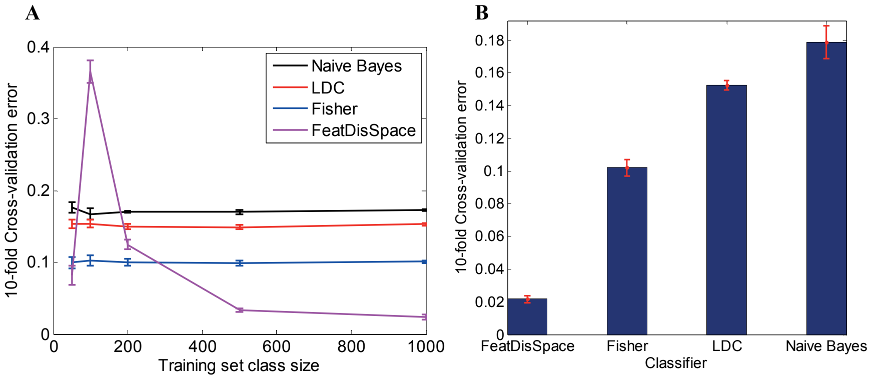


Figure 28. (A) Classifier learning curves. (B) Overall classification error using 10-fold cross-validation.

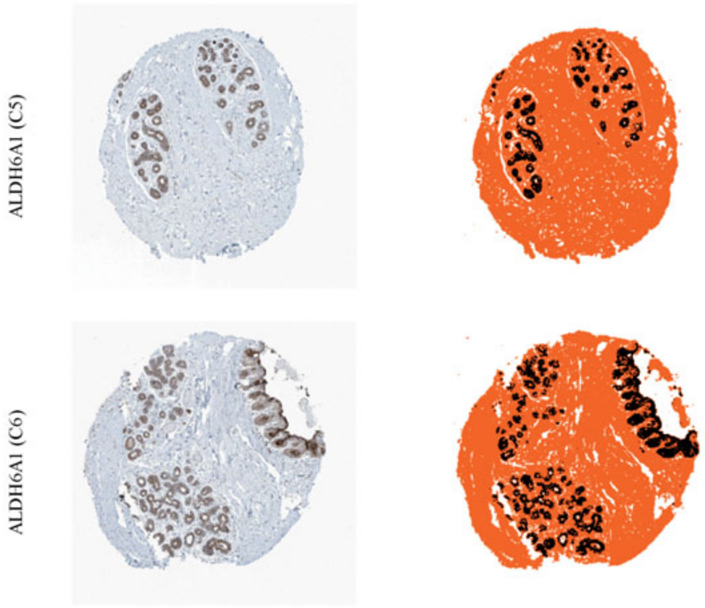


Figure 29. Detection of DAB-stained regions. Pseudo-colors were assigned as follows: DAB appears as black, lumen/background as white, and the remaining structures as orange.

Unsupervised Classification

With the removal of DAB-stained areas, the remaining regions can be clustered in a consistent and accurate manner. We have used soft classification using fuzzy c-means. The regions were clustered into three groups yielding lumen (white), nuclei (dark blue), and stroma (light blue), and the membership function exponent used in fuzzy c-means was m = 2. Recall that in Paper II, we used the Gaussian mixture model to cluster data using the expectation-maximization (E-M) algorithm, which makes use of Bayes' theorem in its E-step to compute posterior probabilities. To remain consistent with regard to Paper II and for convenience, we refer to the resulting three membership images in this context as (posterior) probability maps, assimilating the probability of belonging to the different classes given a particular pixel (note also that the quadratic, normal-Bayes classifier discussed earlier under supervised classification returns as raw output, posterior probability maps). These maps which are in the range [0,1], are sorted using average grayscale intensity so as to standardize the order of the corresponding tissue types. This becomes useful later when we try to correlate probability maps across adjacent tissue sections to identify similar/dissimilar staining patterns.

In displaying the final segmentation result, we have used pseudo-coloring by multiplying each probability map with a 3-channel, uniform color image element-wise, and summing up across the maps. Finally, the image parts corresponding to DAB which were detected separately using supervised classification are re-introduced into the final display. Figure 30 shows some examples of the final segmentation.

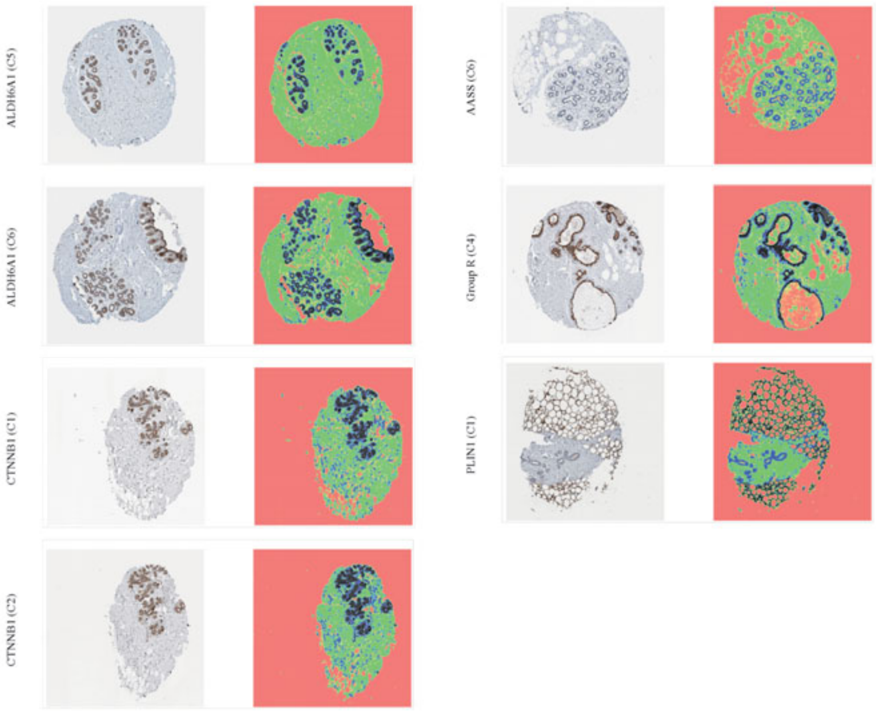


Figure 30. Segmentation of sample cases into four classes: DAB, lumen, stroma, and nuclei. Pseudo-colors were assigned as follows: DAB appears as dark-brown, lumen/background as orange, stroma as green, and nuclei as blue.

Following the segmentation of tissue sections into probability maps, we now address how this can be used to identify paired antibodies across tissue sections. Comparing immunostaining patterns across two adjacent sections can be problematic since the sections may not be perfectly aligned due to the cutting and physical handling of microarrays. Figure 31(A) shows a case where two adjacent sections are not completely aligned. While rotational differences are not discernible, the sections typically have a translational offset with respect to each other. We correct for the translational shift among images using the normalized cross-correlation [65, 66]. The cross-correlation is not carried out over the original RGB images such as those shown in Figure 31, but rather over the corresponding pairs of probability maps which have already been ordered according to average grayscale intensity as mentioned previously. Thus, each probability map of the first image is cross-correlated with its corresponding probability map from the second image, resulting in a matrix of correlation coefficients. We consider the absolute value of the coefficients in the matrix and select the maximum. To obtain a combined measure for the overall match between staining patterns, we multiply the selected coefficients, thus arriving at a single overall measure. Using this product rule gives individual coefficients influence over the final combined measure especially if a single coefficient is near zero. Therefore, a mismatch among any pair of probability maps can significantly reduce the value of the final similarity measure. The number of map pairs used was four, corresponding to nuclei, stroma, lumen, and DAB-stained regions. Although only DAB regions are indicative of positive areas, the inclusion of all four components in the analysis ensures that the tissue sections are also simultaneously tested for adjacency and for having similar structures. In other words, the procedure accounts for the hypothetical, though unlikely, case where two non-adjacent, unrelated tissue sections happen to have similar DAB patterns, coincidentally. By performing pairwise comparisons among corresponding probability maps, we avoid relying on the assumption of adjacency of tissue sections or image prearrangement. Tables 3-4 show results of the cross-correlation analysis over sample groups from the dataset. The highlighted cases can be compared with corresponding tissue sections in Figure 25.

Finally, we also quantify the amount of each of the four classes by using the maximum posterior probability rule across the probability maps to obtain crisp labels and then weigh the presence of each class with respect to the image. This gives a rough indication of the relative area covered by the tissue types.

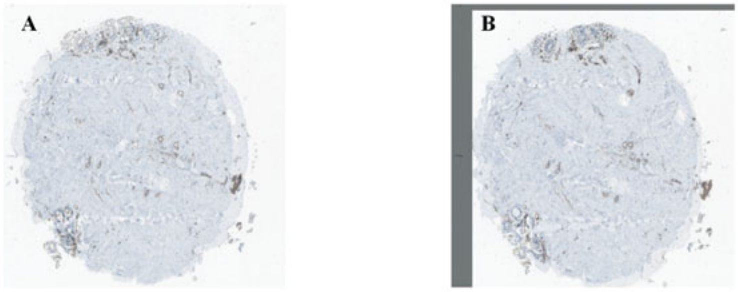


Figure 31. (A) Images of two adjacent sections (cases C1 and C2 of Group R) superimposed to show that they are not entirely overlapping but require registration or the use of a cross-correlation measure. (B) The images are registered using normalized cross-correlation.

Table 3. Similarity measure between segmented images for sample groups: CTNNB1, PLIN1, and ZWINT. The number of probability maps is 4. The largest significant values are highlighted.

		C1	C2	
CTNNB1	C1	1.0000	**0.0387**	
	C2		1.0000	
		C1	C2	C3
PLIN1	C1	1.0000	0.0019	0.0007
	C2		1.0000	0.0026
	C3			1.0000
		C1	C2	C3
ZWINT	C1	1.0000	0.0094	0.0107
	C2		1.0000	**0.0364**
	C3			1.0000

Table 4. Similarity measure between segmented images for Group R. The number of probability maps is 4. The largest significant values are highlighted.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**
C1	1.0000	**0.0132**	0.0082	0.0037	0.0064	0.0034	0.0049	0.0051	0.0050	0.0038	0.0062	0.0021	0.0068
C2		1.0000	0.0058	0.0035	0.0056	0.0033	0.0048	0.0051	0.0054	0.0048	0.0039	0.0017	0.0045
C3			1.0000	0.0063	**0.0167**	0.0072	0.0058	0.0084	0.0083	0.0062	0.0107	0.0087	0.0114
C4				1.0000	0.0071	**0.0181**	0.0050	0.0068	0.0066	0.0025	0.0044	0.0056	0.0052
C5					1.0000	0.0059	0.0070	0.0074	0.0083	0.0048	0.0071	0.0042	0.0061
C6						1.0000	0.0053	0.0062	0.0071	0.0036	0.0045	0.0055	0.0060
C7							1.0000	**0.0250**	**0.0530**	0.0040	0.0064	0.0036	0.0054
C8								1.0000	**0.0338**	0.0039	0.0056	0.0038	0.0055
C9									1.0000	0.0035	0.0084	0.0046	0.0057
C10										1.0000	**0.0361**	0.0024	0.0073
C11											1.0000	0.0039	0.0075
C12												1.0000	**0.0215**
C13													1.0000

4.4.3 Contributions

We summarize the contributions of Paper IV as follows:

1. Variations in the affinity of an antibody, in addition to factors regarding tissue preparation, absorption, and staining give rise to intensity variations in images of tissue sections. This makes automation sensitive to such variations. However, in Paper IV we have overcome this problem by introducing a rescaling of features based on a study of the feature space for this application.

2. The results of the proposed method are comparable to those of commercial software such as Genie [67, 68, 69] or ilastik [70], which are supervised methods that we have experimented with during our work. The supervised aspect of the proposed method concerns only the detection of DAB-stained regions, and the training phase is arranged *a priori*. The remaining bulk of the algorithm performs clustering, and no region delineation or feedback is required from the user.

3. Clustering is carried out in a low-dimensional feature space using fuzzy c-means and converges rapidly, which is suitable for high-throughput processing of immunostaining patterns across tissue microarrays.

4. We used a similarity measure to compare staining patterns and test for adjacency of tissue sections, simultaneously. In doing so, we have applied the normalized cross-correlation over the probability maps generated from image segmentation in a pairwise manner, and combined the correlation coefficients using the product rule. By considering the various probability maps, we have taken into account both the antibody-specific patterns visualized by DAB as well as the remaining tissue structures in the image. The method is consistent and capable of generalizing well over cases.

5. The approach is based on general concepts which are applicable to various types of tissues and proteins, and are not limited to the Human Protein Atlas database.

4.5 Paper V: Automated Classification of Glandular Tissue by Statistical Proximity Sampling

4.5.1 Problem Description

An important component of Elston [21] and Gleason [23] grading of cancer in breast and prostate tissue, respectively, is based on glandular structure and tissue architecture. One of the three subcomponents of Elston grading for example is concerned with tubule formation, where the presence of glandular tissue is given a 1-3 score, ranging from healthy tissue (prevalently glandular) to solid tumors (scarcely glandular). Gleason grading is based on five patterns that closely describe changes in tissue and glandular architecture throughout the different malignancy grades. In assisting pathologists with the grading process, it is therefore very important if one is able to automatically derive a set of features that can accurately describe and quantify tissue characteristics in the image. In Paper V, we present a new feature descriptor for describing tissue architecture based on collecting statistics concerning the quantity and the spatial order of different tissue types around tubule regions. The method requires only a basic decomposition of the image into various tissue types such as by methods presented in Papers II-IV.

The transformation of glandular structures across various cancer grades is a complex process. Some of the problems that hinder the automation of tissue characterization and grading include variations in staining and tissue absorption; however, a main difficulty lies in explicitly extracting properties of glandular structures in a reliable manner, given the complexity and diversity of tissue architecture. Pathologists often rely on a combination of observations relating to color, texture, morphology, density, and prior knowledge of anatomy, in order to characterize tissue. The grading process remains subjective, especially across intermediate grades, and often requires consensus among at least two pathologists before a final grade can be assigned. Thus, there is in general a serious need for reducing the workload in carrying out these tasks by means of automation and computer-aided methods. Translating the knowledge of the human expert into clearly defined features that can be extracted from images is fraught with problems, partly because experts rely on a large combination of features perceived simultaneously, and partly because extracting high-level image features often runs into generalization issues. There is typically a sensitive tradeoff between designing a method that over-adapts to the complexity of the problem and thereby loses its ability to generalize, and between one that weakly accounts for the complexity and is thereby unable to provide useful discrimination, such as when considering global texture features solely without regard to tissue architecture. The method proposed in Paper V relies on the decomposition of images into tissue types based on Papers II-IV, but avoids computing structural complexities such as by explicitly extracting properties of individual tissue components, which can widely vary. However, the method focuses on describing the architecture of glandular tissue starting from lumen areas by applying sequential dilation-subtraction operations to form concentric rings or neighborhood strips from which the relative proportions of various tissue types are recorded. The procedure preserves the shape of the tubule regions (as long as the number of rings is not too large), and due to the sequential rings, it accounts for the spatial order of the measured quantities. In the following section, we discuss the method in greater detail, and explain how it may be used with multiple instance learning as a general feature descriptor for representing images with complex objects, such as biological tissue structures.

4.5.2 Proposed Method

In Paper V, we present a method for deriving a new type of feature that is based on statistically collecting information from the neighborhood of lumen regions as starting points for describing glands and tissue architecture. The proposed method, denoted as *statistical proximity sampling*, can be summarized by the following steps:

1. The tissue image is segmented by soft classification into a set of K probability maps, one for each tissue type. As an example, refer to Figure 32, which was chosen such that the number of lumen regions is small in order to keep track of the resulting profile curves and be able to analyze them visually for illustrating the method.

2. Lumen regions in the image are expanded using dilation operations. This is done in sequential steps by dilating the lumen region and then removing the preceding area from the dilated version, resulting in an annulus or ring at each step. The rings gradually progress away from the central lumen region, while the shape of that region is mostly preserved in the process (see Figure 33). The resulting set of concentric rings are then regarded as neighborhood strips from which we gather statistics concerning the relative quantities of the different tissue types as computed from the probability maps. This feature description also naturally contains spatial information due to the progressing position (order) of the rings.

3. For every ring, we compute the fraction of each tissue type using their corresponding probability maps, which were derived in Step-1. The result is a vector consisting of K scalar elements, as for example [0.1 0.5

0.3 0.1], indicating that the relative proportions of the 4 tissue types in this case are 10%, 50%, 30%, and 10%. Each of these ratios is computed from the respective probability map by summing up the values which happen to lie within the ring and dividing by the total number of pixels in the ring.

4. The K-element vectors are concatenated one behind the other starting with the one corresponding to the first ring (closest to the central lumen). Thus with R rings, the resulting feature vector will be of length R x K. Each lumen region is hence described by one such feature vector. Figure 35 shows an example where there are 4 feature vectors plotted using different colors corresponding to the 4 extracted regions shown in Figure 34 (note that the colors of the curves were set to match those of the lumen regions). The number of rings used in this example was $R = 30$.

5. Under the framework of multiple instance learning (MIL), we treat the tissue image as a bag or compound object consisting of a collection of instances; the instances consist of the derived feature vectors representing different regions of the image. Thus, each image can then be classified based on its contents and assigned a grade. For constructing a classifier, we have used the bag dissimilarity approach proposed in [36], which allows us to decouple the classification task from the multiple instance formulation, thus permitting the use of various types of classifiers in a standard manner.

The feature plots in Figure 35 describe how class quantities vary as one progresses spatially away from lumen regions. This is an important attribute, since for instance in Gleason grading, as the cancer grade advances, the tissue architecture around glandular units begins to change markedly as cells invade surrounding tissue, and as glands take on cribriform shape or split into multiple lumen areas. For such changes, the profile curves would exhibit different patterns. In interpreting these profiles for the case in Figure 35, one may notice that the cyan curve for instance shows a gradually increasing fraction of lumen (first part of the curve) due to the presence of an adjoining lumen region marked in red in Figure 34; it also reveals an absence of nearby stromal components since the fourth part of the curve is flat, and this may be validated from Figure 34.

Figure 36 shows an alternative way of showing the feature curves for illustration purposes. In this case, the four sequential parts of each feature vector shown in Figure 35 are plotted using a relative frequency pie chart as shown by the subplots of Figure 36. Each subplot represents a luminal region and describes how the four tissue types vary around it across the rings.

Therefore, Figure 36 reveals how the different luminal regions compare to one another in terms of how each one’s neighborhood is characterized and tends to vary.

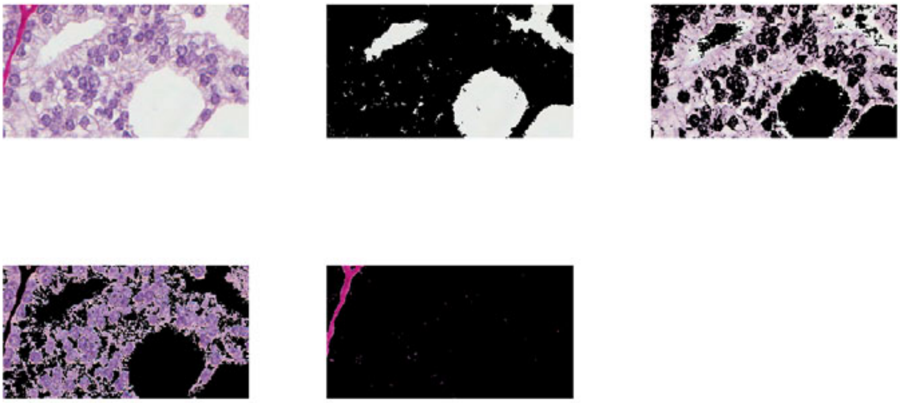


Figure 32. An image of prostate tissue (upper left) is decomposed into four classes: lumen, epithelium, nuclei, and stromal regions. The probability maps in this example were thresholded at a level of 10% for enhancing visibility.

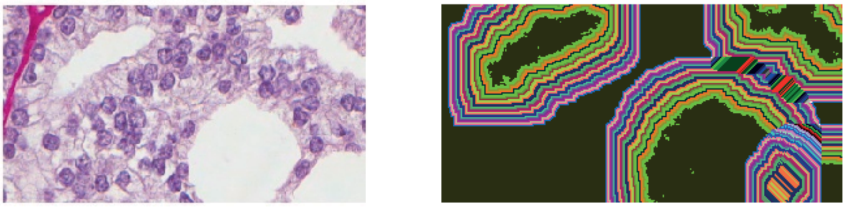


Figure 33. The sampling rings growing away from each lumen region for the example in Figure 32. Each ring is obtained by first dilating the lumen region and then subtracting it from the dilated version. This is done sequentially and the number of rings in this example is 30.

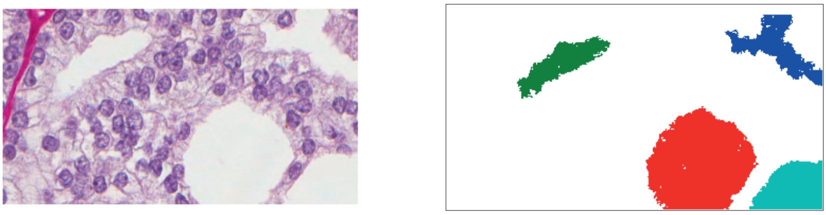


Figure 34. The lumen class from the original image shown in Figure 32 is extracted. The different regions are labeled according to their 4-connectivity. These regions form the basis and starting point of our algorithm for deriving features.

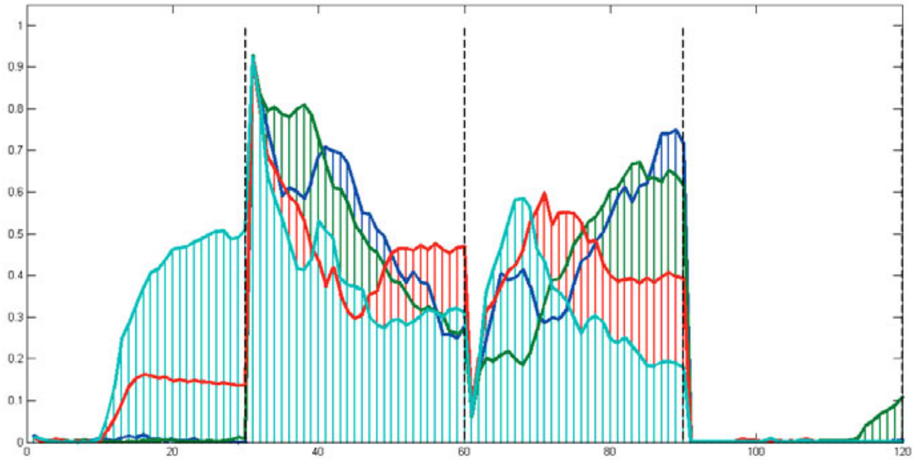


Figure 35. Feature vectors are shown, one for each of the four lumen regions illustrated in Figure 34. The vectors are divided into 4 parts delineated by black vertical lines: the first part depicts the first 30 elements representing the fraction of lumen within the 30 sampling rings, the second part depicts that for the epithelium component, the third for the nuclei component, and the fourth for the stromal component.

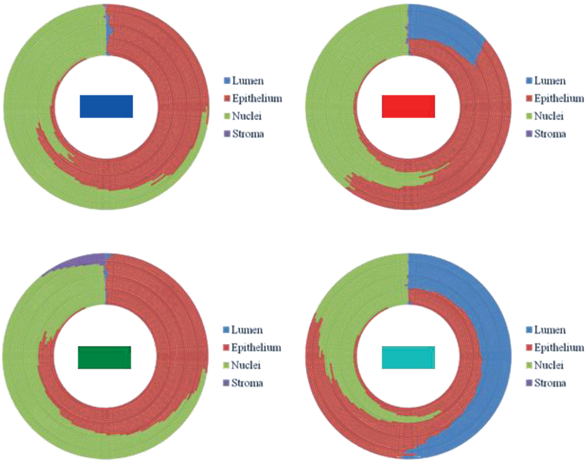


Figure 36. The four parts of each feature vector illustrated in Figure 35 are depicted in a relative frequency graph for each luminal region in this example. The color of the central rectangle in each subplot indicates the corresponding lumen region shown in Figure 34.

With the aim of predicting the label of a given image, we derive a set of feature vectors from lumen regions using the statistical proximity method, and we regard the image as a bag or collection of these features vectors or instances. Only the bag carries a label (which is the grade of the image as determined by the pathologist), whereas instances remain unlabeled. This type of formulation falls under the framework of multiple instance learning (MIL) [33], through which we are able to represent an image using a multitude of features vectors, instead of a single feature vector. Tissue images often contain many critical subregions with varying characteristics, and may be too complex to represent using a single feature vector. Using this multiple instance representation, we adopt a dissimilarity-based approach [36] for constructing a classifier. Images or bags are viewed as a set of points in an (R x K)-dimensional feature space, and the distances between all bags are then computed using the linear assignment distance as defined between sets [36, 71]. Thus, the original feature space is transformed by this manner to a dissimilarity space where the new features represent the distances among the bags. The dissimilarity space is then treated as a standard representation, where typically any classifier can be constructed, hence decoupling the classification task from the complex MIL representation. In general, this type of transformation also allows for a natural extension to the case of multi-class situations, thereby relaxing the original MIL formulation which is based on binary classification, where a bag is assigned either a positive or negative class label.

To validate the proposed method, we used a dataset consisting of breast tissue images obtained from the Human Protein Atlas [63], where some sample tissue sections are shown Figure 37. The images were graded by a pathologist for the presence (class $C1$) or absence (class $C0$) of tubules, which is associated with tubule-based Elston scoring. The microscopy images were acquired at a magnification level of 40X. The characteristics of the dataset are summarized in Table 5. A total of 104 images were graded (49 for $C0$ and 55 for $C1$). Ten sampling rings were used in all cases resulting in a 40-element feature vector to account for the following classes: nuclei, stroma, lumen, and DAB. Note that 4 additional features were later appended to the main proximity feature vector in order to assess any changes in classification following this addition. These additional features were 4 classical attributes relating to the area and shape of the central lumen region, namely: its size, bending energy [72], area-to-perimeter ratio, and convexity ratio as computed from the convex hull. The features were normalized to the same range (i.e., [0,1]) as the elements of the proximity feature vector, and classification was carried out using the MIL toolbox [73, 74].

Figure 38 shows the 10-fold cross-validation error rates for a number of classifiers using the features derived from the proximity sampling method, whereas Figure 39 shows the same error rates, yet after appending the 4 classical lumen-based features. Note that a 25% random fraction of the dataset was used to optimize the parameters of the support vector classifier and k-nearest neighbor classifier using leave-one-out cross-validation, whereas the remaining 75% fraction of the dataset was used to obtain the 10-fold cross-validation error. To reduce sensitivity to the random split of the dataset, the entire procedure was repeated five times, and the average results are reported in the figures. We note that the inclusion of the 4 classical lumen-based features did not have a significant effect on the classification rates (see Table 6).

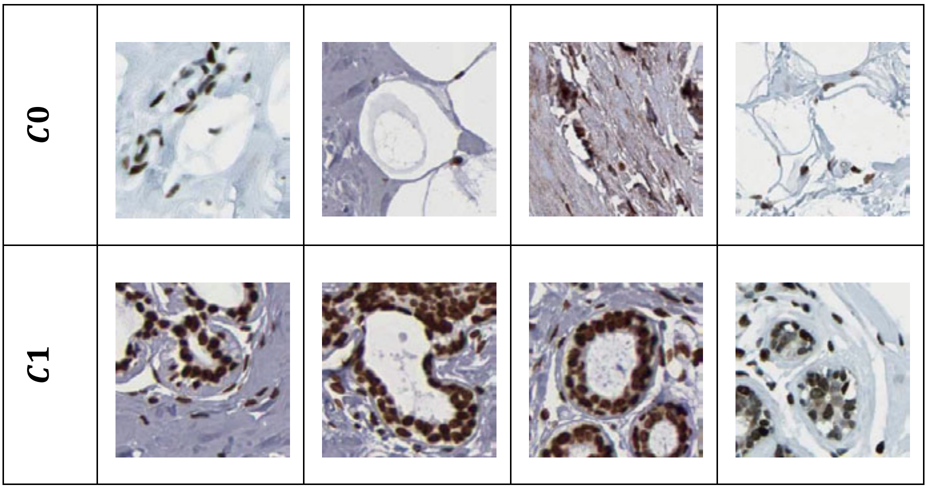
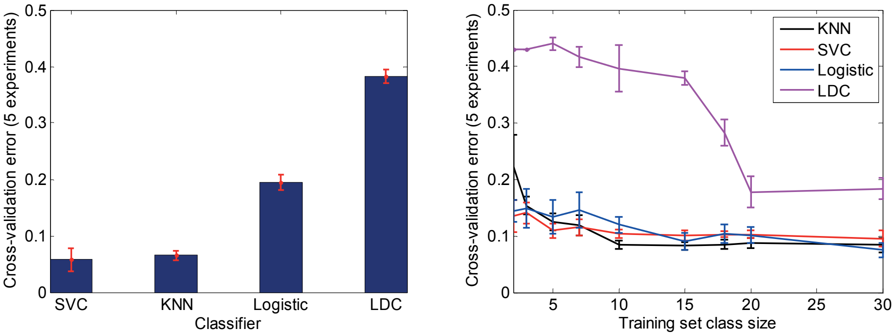


Figure 37. A few sample cases of the breast dataset consisting of images of breast tissue sections labeled as $C1$ and $C0$ based on the presence or absence of milk ducts, respectively.

Table 5. Summary statistics concerning the dataset used in this paper.

Dataset	**Number of instances**	**Dimensionality**	**Number of bags**	**Number of instances per bag**		
				Minimum	Median	Maximum
Breast	1309	40/44	104	1	10	37



(a) (b)

Figure 38. Classification results for the breast dataset using only the features derived by *statistical proximity sampling*. Classifiers used are the linear support vector classifier (SVC), k-nearest neighbor classifier (KNN), the logistic classifier (Logistic), and normal-based linear discriminant classifier (LDC). (a) 10-fold cross-validation error. (b) Classifier learning curves. Error bars represent one standard deviation.

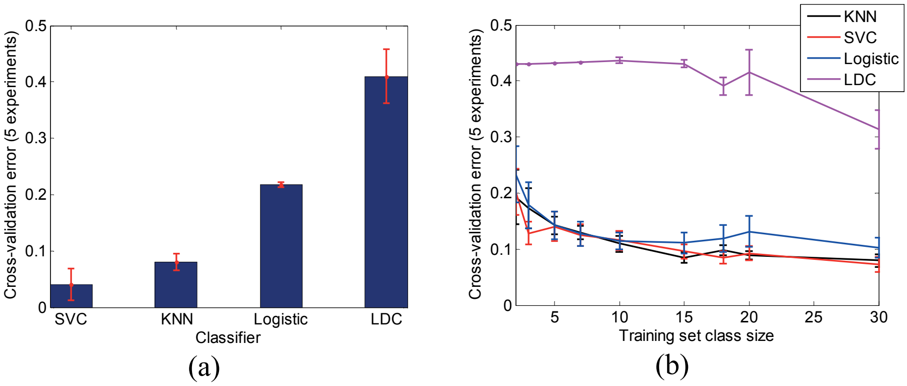


Figure 39. Classification results for the breast dataset using the proximity feature in addition to classical lumen shape features. Classifiers used are the linear support vector classifier (SVC), k-nearest neighbor classifier (KNN), the logistic classifier (Logistic), and normal-based linear discriminant classifier (LDC). (a) 10-fold cross-validation error. (b) Classifier learning curves. Error bars represent one standard deviation.

Table 6. Classification rates for the breast dataset using different classifiers. Classification was done using 10-fold cross-validation and results are reported as: percentage of correct classification ± standard deviation.

Classifier	**Classification rate**	
	Proximity features (Figure 38)	Proximity + classical features (Figure 39)
SVC	94.2 ± 2.0	95.9 ± 2.7
KNN	93.4 ± 0.8	91.9 ± 1.4
Logistic	80.5 ± 1.3	78.2 ± 0.4
LDC	61.7 ± 1.2	59.0 ± 4.7

Unsupervised Approach

We end the analysis by noting the possibility of using clustering, coupled with an information criterion, instead of supervised classification in order to classify the images. Clustering can be applied over the proximity feature vectors, i.e., unlabeled instances, without regard to bag or image labels. In doing so, we used the Gaussian mixture model (GMM) along with the Bayesian or Akaike information criteria. In particular the Bayesian information criterion is defined as: $BIC = -2\,ln(L) + k\,ln(n)$ where L represents the likelihood measure of the mixture fit, k represents the number of clusters, and n represents the number of instances in the dataset. We varied the number of clusters from 1 to 10, and at each value we allowed the mixture model to converge to a solution using the expectation-maximization algorithm over 10 different runs with random initializations. In each of these runs, we computed the Bayesian information criterion and reported the average value and standard deviation in Figure 40. Increasing the number of clusters would normally increase the log-likelihood value, however by factoring in model parameters such as k and n into the tradeoff, the BIC curve can attain a minimum value at which the log-likelihood is highest, yet at the lowest possible model complexity, k. Figure 40 shows that the BIC curve attains its minimum at $k = 2$ clusters. This may give an idea about the possible range or number of groups that data instances naturally fall into. The choice of information criterion however does affect the outcome of this selection as the same figure shows. Once cluster labels are determined, the bag (image) can be labeled according to a voting scheme based on the cluster labels of the instances belonging to the bag.

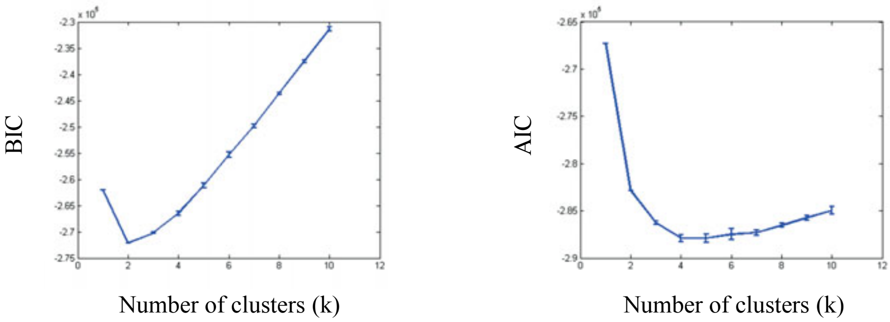


Figure 40. Optimal number of clusters using the Bayesian information criterion (BIC) and Akaike information criterion (AIC) over the breast dataset. The error bars represent the standard deviation at each value of ‘k’.

4.5.3 Contributions

Paper V presented a novel feature descriptor called *statistical proximity sampling* for encoding quantitative and spatial characteristics of tissue architecture. The contributions and attributes of the method can be summarized as follows:

1. The new approach is able to represent complex patterns in biological tissue such as breast or prostate, while avoiding the extraction of explicit properties regarding tissue structures. It relies on region expansion starting from lumen areas to define neighborhoods as concentric annuli, while preserving the shape of the lumen boundary.

2. The method is able to quantify the relative proportions of various tissue types within the neighborhood sampling rings. The positions of the rings as they progress away from the central lumen region encode spatial information as well, since they indicate how these relative proportions vary with distance.

3. The proximity sampling method is an independent descriptor. It requires as input only a set of probability or binary maps resulting from soft or crisp classification, regardless of what supervised or unsupervised method is employed to generate these maps. The method's reliance on the quality of the resulting maps is minimal compared to approaches that require accurate derivations of glandular or cellular properties based on these maps.

4. The neighborhood-based approach can be combined with multiple instance learning to provide a highly descriptive representation for complex images. Therefore, we circumvent the use of a single feature vector representation, which may not be suitable due to the natural complexity of tissue and the grading process.

5. In principle, it is possible to use the method as a general feature descriptor in favor of images that are composed of a combination of complex structures, as well as for a variety of purposes, not necessarily limited to biological tissue. Such use could be a basis for future exploration.

5 Conclusions

We have addressed several topics related to tissue image analysis in this thesis. We end our discussion with a review of the most important conclusions summarized as key messages, one per publication, and numbered according to the included papers.

I In detecting microarray cores, it is unnecessary to use conventional methods such as template matching or the Hough transform. The specificity of the problem allows the use of fast morphology for a basic detection of cores, followed by an analytical, randomized 3-point circle method for restoring the disc-shapes around each core. It is possible to design the entire workflow from input to output in an automated manner, such as with the aid of cluster validation and morphological granulometry.

When possible, one should avoid solving a more general problem (such as circle detection from among different geometric shapes) before arriving at a solution for a more specific problem. This broad viewpoint falls under Vapnik's approach to statistical learning, which is highly outcome-driven and is most apparent in his description of transductive inference [75].

II A common approach for decomposing histological images is by performing supervised classification or clustering to arrive at either probability or binary maps. However, another approach is to first model optical densities using the Beer-Lambert law of light absorption, while accounting for sensor noise and decoupling intensity, and then use linear decomposition to determine the stain concentrations at each pixel. We have shown that this latter approach can be made unsupervised through the estimation of reference colors using a Gaussian mixture model trained by expectation-maximization, thus overcoming previous limitations on having to explicitly specify absorption spectra for the individual stain components.

III Machine learning techniques for automated tissue image analysis often employ blind or supervised classification, or a combination of these. However, regardless of the method used, the intrinsic class overlap in feature space sets a main limitation on performance. In reference to tissue image analysis, the overlap between chromaticity clusters in color space becomes the defining factor for color decomposition. These clusters are defined by the color combinations used in a given stain. Thus it is essential that the choice of stain is not overlooked, but rather selected in a way that optimizes classification performance based on error criteria to ensure the compactness and separability of clusters in feature space. For automation purposes, selecting an optimal stain should be incorporated into the image analysis task rather than be preset based on visual inspection.

IV Immunohistochemistry is an advanced form of staining that allows using antibodies for localizing antigens in tissue. Detecting paired antibodies based on the immunostaining patterns across different adjacent tissue sections is not an easy task due to the variability in antigen affinity and tissue structures across the sections. However, by adopting a simple and robust, soft classification method for decomposing the sections into probability maps, it is possible to use correlation analysis reliably based on these maps and to obtain a combined similarity measure for comparing the staining patterns and testing the adjacency assumption.

V Histological image analysis often relies on the extraction of properties from cellular structures and on the measurement of attributes such as nuclear shape and inter-glandular distances. Such tasks can be difficult due to the complexity and diversity of tissue architecture and variations in staining. These methods also often place strong assumptions on image understanding, thus limiting generalizability. Applications requiring high-level image understanding are not easily solvable using ‘standard’ techniques. In working around this problem, a possible approach is to use an implicit representation that is expressive enough to describe tissue architecture. In sampling shape-preserving, neighborhood rings around lumen regions throughout the image, quantitative information can be obtained by computing the relative tissue-type proportions within each ring. Spatial information is encoded by the location of the neighborhood rings as they expand away from the lumen region. Representing complex tissue images in this manner can be naturally combined with multiple instance learning techniques. The feature descriptor places very little dependence on the quality of the probability maps that result from image segmentation as compared to standard methods where this quality affects the accuracy of the derived measurements.

A final note concerning developing methods for automated tissue image analysis: the aim of the analysis should be aligned with its methods, and should support the validation of the final design, its outcome, and its ability to generalize. When using statistical learning methods such as pattern recognition, there is a basic tradeoff between over-adapting to complexity and over-simplification, both of which can lead to poor generalization, as described by the bias-variance dilemma. To succeed under this paradigm, a good understanding of the problem is necessary before moving on to investigate solutions. This helps prevent one from drifting too far from Occam's razor in relation to simplicity and efficiency. The optimal solution in the end is one that results in the best performance, while taking into account complexity and generalizability. The number and type of parameters in a design can be a useful indicator of its complexity. The effect of non-idle parameters and their sensitivities, which can influence the outcome should be kept minimal or adjoined into a more stable stage of the algorithm. The selection of such parameters if present can be optimized using standard statistical methods such as cross-validation and grid search techniques. The final aim is a knowledgeable design that could work in practice – one whose components are often simple enough to be able to generalize and have predictive value.

Bibliography

[1] M. Veta, J. P. W. Pluim, P. J. van Diest and M. A. Viergever, "Breast Cancer Histopathology Image Analysis: A Review," *IEEE Transactions on Biomedical Engineering,* vol. 61, no. 5, pp. 1400-1411, 2014.

[2] Y. Al-Kofahi, W. Lassoued, K. Grama, S. K. Nath, J. Zhu, R. Oueslati, M. Feldman, W. M. F. Lee and B. Roysam, "Cell-based quantification of molecular biomarkers in histopathology specimens," *Histopathology,* vol. 59, no. 1, pp. 40-54, 2011.

[3] A. E. Rizzardi, A. T. Johnson, R. Vogel, S. E. Pambuccian, J. Henriksen, A. P. N. Skubitz, G. J. Metzger and S. C. Schmechel, "Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring," *Diagnostic Pathology,* vol. 7, no. 42, 2012.

[4] 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," *Modern Pathology,* vol. 18, no. 8, pp. 1067-1078, 2005.

[5] E. A. Perez, V. J. Suman, N. E. Davidson, S. Martino, P. A. Kaufman, W. L. Lingle, P. J. Flynn, J. N. Ingle, D. Visscher and R. B. Jenkins, "HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial," *Journal of Clinical Oncology,* vol. 24, no. 19, pp. 3032-3038, 2006.

[6] A. C. Wolff, M. E. H. Hammond, J. N. Schwartz, K. L. Hagerty, D. C. Allred, R. J. Cote, M. Dowsett, P. L. Fitzgibbons, W. M. Hanna, A. Langer, L. McShane, S. Paik, M. D. Pegram, E. A. Perez, M. F. Press, A. Rhodes, C. Sturgeon, S. E. Taube and et al., "American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer," *Archives of Pathology & Laboratory Medicine,* vol. 131, no. 1, pp. 18-43, 2007.

[7] M. E. H. Hammond, D. F. Hayes, A. C. Wolff, P. B. Mangu and S. Temin, "American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer," *Journal of Oncology Practice,* vol. 6, no. 4, pp. 195-197, 2010.

[8] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter and O. P. Kallioniemi, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine,* vol. 4, no. 7, pp. 844-847, 1998.

[9] M. Skacel, B. Skilton, J. D. Pettay and R. R. Tubbs, "Tissue microarrays: A powerful tool for high-throughput analysis of clinical specimens: A review of the method with validation data," *Applied Immunohistochemistry & Molecular Morphology,* vol. 10, no. 1, pp. 1-6, 2002.

[10] Y. Wang, K. Savage, C. Grills, A. McCavigan, J. A. James, D. A. Fennell and P. W. Hamilton, "A TMA de-arraying method for high throughput biomarker discovery in tissue research," *PLOS ONE,* vol. 6, no. 10, p. e26007, 2011.

[11] B. Lahrmann, N. Halama, K. Westphal, C. Ernst, Z. Elsawaf, P. Sinn, F. X. Bosch, H. Dickhaus, D. Jäger, P. Schirmacher and N. Grabe, "Robust gridding of TMAs after whole-slide imaging using template matching," *Cytometry Part A,* vol. 77A, no. 12, pp. 1169-1176, 2010.

[12] D. J. Foran, L. Yang, W. Chen, J. Hu, L. A. Goodell, M. Reiss, F. Wang, T. Kurc, T. Pan, A. Sharma and J. H. Saltz, "ImageMiner:A software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology," *Journal of the American Medical Informatics Association,* vol. 18, no. 4, pp. 403-415, 2011.

[13] P. L. Fitzgibbons, D. L. Page, D. Weaver, A. D. Thor, D. C. Allred, G. M. Clark, S. G. Ruby, F. O'Malley, J. F. Simpson, J. L. Connolly, D. F. Hayes, S. B. Edge, A. Lichter and S. J. Schnitt, "Prognostic factors in breast cancer," *Archives of Pathology & Laboratory Medicine,* vol. 124, no. 7, pp. 966-978, 2000.

[14] V. J. Tuominen, S. Ruotoistenmäki, A. Viitanen, M. Jumppanen and J. Isola, "ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67," *Breast Cancer Research,* vol. 12, no. 4, p. R56, 2010.

[15] V. J. Tuominen, T. T. Tolonen and J. Isola, "ImmunoMembrane: a publicly available web application for digital image analysis of HER2 immunohistochemistry," *Histopathology,* vol. 60, no. 5, pp. 758-767, 2012.

[16] M. C. Lloyd, P. Allam-Nandyala, C. N. Purohit, N. Burke, D. Coppola and M. M. Bui, "Using image analysis as a tool for assessment of prognostic and predictive biomarkers for breast cancer: How reliable is it?," *Journal of Pathology Informatics,* vol. 1, 2010.

[17] A. Brügmann, M. Eld, G. Lelkaitis, S. Nielsen, M. Grunkin, J. D. Hansen, N. T. Foged and M. Vyberg, "Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains," *Breast Cancer Research and Treatment,* vol. 132, no. 1, pp. 41-49, 2012.

[18] M. A. Gavrielides, B. D. Gallas, P. Lenz, A. Badano and S. M. Hewitt, "Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy," *Archives of Pathology & Laboratory Medicine,* vol. 135, no. 2, pp. 233-242, 2011.

[19] Z. M. A. Mohammed, J. J. Going, D. C. McMillan, C. Orange, E. Mallon, J. C. Doughty and J. Edwards, "Comparison of visual and automated assessment of HER2 status and their impact on outcome in primary operable invasive ductal breast cancer," *Histopathology,* vol. 61, no. 4, pp. 675-684, 2012.

[20] A. Nassar, C. Cohen, S. S. Agersborg, W. Zhou, K. A. Lynch, M. Albitar, E. A. Barker, B. L. Vanderbilt, J. Thompson, E. R. Heyman, H. Lange, A. Olson and M. T. Siddiqui, "Trainable immunohistochemical HER2/neu image analysis: A multisite performance study using 260 breast tissue specimens," *Archives of Pathology & Laboratory Medicine,* vol. 135, no. 7, pp. 896-902, 2011.

[21] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology,* vol. 19, no. 5, pp. 403-410, 1991.

[22] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani, J. Palacios, A. L. Richardson, S. J. Schnitt, F. C. Schmitt, P.-H. Tan, G. M. Tse, S. Badve and I. O. Ellis, "Breast cancer prognostic classification in the molecular era: The role of histological grade," *Breast Cancer Research,* vol. 12, no. 4, p. 207, 2010.

[23] J. I. Epstein, W. C. Allsbrook, M. B. Amin, L. L. Egevad and the ISUP Grading Committee, "The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," *American Journal of Surgical Pathology,* vol. 29, no. 9, pp. 1228-1242, 2005.

[24] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine,* vol. 2, pp. 559-572, 1901.

[25] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology,* vol. 24, no. 6, pp. 417-441, 1933.

[26] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika,* vol. 17, no. 4, pp. 401-419, 1952.

[27] L. van der Maaten, E. Postma and J. van den Herik, "Dimensionality reduction: A comparative review," Technical Report TiCC TR 2009–005, 2009.

[28] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika,* vol. 32, no. 3, pp. 241-254, 1967.

[29] B. J. T. Morgan and A. P. G. Ray, "Non-uniqueness and inversions in cluster analysis," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 44, no. 1, pp. 117-134, 1995.

[30] A. Fernández and S. Gómez, "Solving non-uniqueness in agglomerative hierarchical," *Journal of Classification,* vol. 25, no. 1, pp. 43-65, 2008.

[31] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 1, no. 2, pp. 224-227 , 1979.

[32] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 39, no. 1, pp. 1-38, 1977.

[33] T. G. Dietterich, R. H. Lathrop and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence,* vol. 89, no. 1-2, pp. 31-71, 1997.

[34] Y. Chen, J. Bi and J. Z. Wang, "MILES: multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 1931-1947, 2006.

[35] S. Andrews, I. Tsochantaridis and T. Hofmann, "Support vector machines for multiple-instance learning," *Advances in Neural Information Processing Systems,* vol. 15, pp. 561-568, 2003.

[36] D. M. J. Tax, M. Loog, R. P. W. Duin, V. Cheplygina and W.-J. Lee, "Bag dissimilarities for multiple instance learning," *Similarity-Based Pattern Recognition,* vol. 7005 of Lecture Notes in Computer Science, pp. 222-234, 2011.

[37] G. Borgefors, "Distance transformations in digital images," *Computer Vision, Graphics, and Image Processing,* vol. 34, no. 3, pp. 344-371, 1986.

[38] Q. Li, C. Fraley, R. E. Bumgarner, K. Y. Yeung and A. E. Raftery, "Donuts, scratches and blanks: robust model-based segmentation of microarray images," *Bioinformatics,* vol. 21, no. 12, pp. 2875-2882, 2005.

[39] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition,* vol. 13, no. 2, pp. 111-122, 1981.

[40] R. K. Yip, P. K. Tam and D. N. Leung, "Modification of Hough transform for circles and ellipses detection using a 2-dimensional array," *Pattern Recognition,* vol. 25, no. 9, pp. 1007-1022, 1992.

[41] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., New York: Springer, 2009, pp. 520-528.

[42] K. R. Castleman, R. Eils, L. Morrison, J. Piper, K. Saracoglu, M. A. Schulze and M. R. Speiche, "Classification accuracy in multiple color fluorescence imaging microscopy," *Cytometry Part A,* vol. 41, no. 2, pp. 139-147, 2000.

[43] Y. Garini, I. T. Young and G. McNamara, "Spectral imaging: principles and applications," *Cytometry Part A,* vol. 69A, no. 8, pp. 735-747, 2006.

[44] C. M. van der Loos, "Multiple immunoenzyme staining: methods and visualizations for the observation with spectral imaging," *Journal of Histochemistry & Cytochemistry,* vol. 56, no. 4, pp. 313-328, 2008.

[45] J. C. Mullikin, L. J. van Vliet, H. Netten, F. R. Boddeke, G. van der Feltz and I. T. Young, "Methods for CCD camera characterization," in *SPIE: Image Acquisition and Scientific Imaging Systems*, 1994.

[46] G. Polder and G. W. van der Heijden, "Calibration and characterization of spectral imaging systems," in *SPIE: Multispectral and Hyperspectral Image Acquisition and Processing*, 2001.

[47] M. Gavrilovic and C. Wählby, "Quantification of colocalization and cross-talk based on spectral angles," *Journal of Microscopy,* vol. 234, no. 3, pp. 311-324, 2009.

[48] W. W. Parson, Modern Optical Spectroscopy, Berlin-Heidelberg: Springer, 2009, pp. 3-26.

[49] J. C. Maxwell, "On the theory of compound colours, and the relations of the colours of the spectrum," *Philosophical Transactions of the Royal Society of London,* vol. 150, pp. 57-84, 1860.

[50] D. B. Judd, "A Maxwell triangle yielding uniform chromaticity scales," *Journal of the Optical Society of America,* vol. 25, no. 1, pp. 24-35, 1935.

[51] J. D. Foley, A. van Dam, S. K. Feiner and J. F. Hughes, Computer Graphics: Principles and Practice in C, Reading, Massachusetts: Addison-Wesley, 1996, pp. 213-281.

[52] S. Theodoridis and K. Koutroumbas, Pattern Recognition, 4th ed., Massachusetts: Academic Press, 2009, pp. 44-48.

[53] C. M. Bishop, Pattern Recognition and Machine Learning, New York: Springe, 2006, pp. 430-450.

[54] A. Rabinovich, S. Agarwal, C. Laris, J. H. Price and S. J. Belongie, "Unsupervised color decomposition of histologically stained tissue samples," in *Advances in Neural Information Processing Systems*, 2003.

[55] J. Newberg and R. F. Murphy, "A Framework for the automated analysis of subcellular patterns in Human Protein Atlas images," *Journal of Proteome Research,* vol. 7, no. 6, pp. 2300-2308, 2008.

[56] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. Kumar, D. Verbel, A. Kotsianti and O. Saidi, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Transactions on Medical Imaging,* vol. 26, no. 10, pp. 1366-1378, 2007.

[57] A. Tabesh and M. Teverovskiy, "Tumor classification in histological images of prostate using color texture," in *40th Asilomar Conference on Signals, Systems and Computers*, 2006.

[58] C. W. Hsu, C. C. Chang and C. J. Lin, "A practical guide to support vector classification," National Taiwan University, Taipei, 2010.

[59] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association,* vol. 66, no. 336, pp. 846-850, 1971.

[60] C. J. Van Rijsbergen, Information retrieval, 2nd ed., London: Butterworths, 1979.

[61] A. R. Webb, Statistical pattern recognition, 2nd ed., New Jersey: Wiley, 2002.

[62] I. N. Swamidoss, A. Kårsnäs, V. Uhlmann, P. Ponnusamy, C. Kampf, M. Simonsson, C. Wählby and R. Strand, "Automated classification of immunostaining patterns in breast tissue from the human protein atlas," *Journal of Pathology Informatics,* vol. 4, no. 14, 2013.

[63] F. Pontén, J. M. Schwenk, A. Asplund and P.-H. Edqvist, "The Human Protein Atlas as a proteomic resource for biomarker discovery," *Journal of Internal Medicine,* vol. 270, no. 5, pp. 428-446, 2011.

[64] R. P. W. Duin and E. Pękalska, "The dissimilarity representation for structural pattern recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Pucón, 2011.

[65] J. P. Lewis, "Fast template matching," *Vision Interface,* vol. 95, pp. 120-123, 1995.

[66] R. M. Haralick and L. G. Shapiro, Computer and Robot Vision (Volume II), Reading, Massachusetts: Addison-Wesley, 1992, pp. 316-317.

[67] "GENIE, GENetic Imagery Exploitation," Los Alamos National Security, [Online]. Available: http://genie.lanl.gov/.

[68] S. Perkins, J. Theiler, S. P. Brumby, N. R. Harvey, R. Porter, J. J. Szymanski and J. J. Bloch, "GENIE: a hybrid genetic algorithm for feature classification in multispectral images," in *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation III*, 2000.

[69] N. R. Harvey, J. Theiler, S. P. Brumby, S. Perkins, J. J. Szymanski, J. J. Bloch, R. B. Porter, M. Galassi and A. C. Young, "Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 40, no. 2, pp. 393-404, 2002.

[70] C. Sommer, C. Straehle, U. Köthe and F. A. Hamprecht, "Ilastik: Interactive Learning and Segmentation Toolkit, version 0.5," in *8th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Chicago, 2011.

[71] H. W. Kuhn, "The Hungarian method for the assignment problem;," *Naval Research Logistics Quarterly,* vol. 2, no. 1-2, pp. 83-97, 1955.

[72] L. J. Van Vliet and P. W. Verbeek, "Curvature and bending energy in digitized 2D and 3D images," *8th Scandinavian Conference on Image Analysis,* vol. 2, pp. 1403-1410, 1993.

[73] D. M. J. Tax, "MIL, A Matlab Toolbox for Multiple Instance Learning, version 0.8.1," March 2013. [Online]. Available: http://prlab.tudelft.nl/david-tax/mil.html.

[74] R. P. W. Duin, P. Juszczak, P. Paclík, E. Pękalska, D. DeRidder and D. M. J. Tax, "A Matlab Toolbox for Pattern Recognition, PRTools4 version 4.2.5," July 2013. [Online]. Available: http://www.37steps.com/.

[75] V. N. Vapnik, Statistical learning theory, New York: Wiley, 1998, pp. 339-371.

Acknowledgements

I wish to thank Ewert Bengtsson for his guidance and contribution regarding all aspects of this thesis, related publications, and research project; this thesis would not have been possible without him. I am grateful to my main supervisor Anders Hast for his insight and support and for helping me complete this work. I also wish to acknowledge Martin Simonsson for his contribution to the project and datasets. I also thank Ingrid Carlbom and Christer Busch for their supervision and early contribution to this work.

Sammanfattning på svenska

Den här avhandlingen beskriver metoder för automatisk bildanalys av mänskliga vävnader. Målet har varit att utveckla algoritmer som kan minska de skillnader i bedömningar som uppstår mellan mänskliga granskare, öka kvaliteten i tolkningen av proverna och minska arbetsbelastningen inom patologin. Detta är framför allt viktigt för diagnostisk gradering av bröst- och prostatacancer, såväl som kvantifiering av biomarkörer för prognos och tidig upptäckt av cancer. Sådana markörer kan bidra till att bedöma risken för återfall av tumörer och hur väl exempelvis endokrin terapi fungerar.

Några av de viktigaste användningsområdena för datoriserad bildanalys finns inom medicinen. Datoriserad bildanalys har med framgång kommit till användning inom radiologin och har blivit en grundläggande komponent av flera olika diagnostiska metoder där det påverkar såväl själva bildfångsten som den diagnostiska bedömningen och behandlingsplaneringen. Introduktionen av datoriserad bildanalys inom histologin och histopatologin, d.v.s. studien av frisk respektive sjuk mänsklig vävnad, har dock gått betydligt långsammare trots att histopatologin har en central roll vid diagnostik och gradering av cancer liksom vid bedömningen av olika biomarkörer. En orsak till detta är att i motsats till den moderna radiologin, som är alltigenom digital, så har histologin en optisk analog komponent när proverna bereds och avbildas i ett mikroskop. Det är därför nödvändigt med en särskild analog-till-digital konvertering av bilderna innan det går att använda datoriserad bildanalys på dessa. Patologer använder också mycket olika optiska upplösningar när de granskar de histologiska proverna, alltifrån översikter över ett helt vävnadsutsnitt till detaljerad granskning av enskilda celler och cellulära beståndsdelar. Att digitalisera ett histologiskt prov leder därför till väldigt mycket data, ett flertal gigabyte, om det skall ske så att högupplöst analys är möjlig överallt. Introduktionen av högupplösande digitala kameror som kan monteras på mikroskop har gjort det möjligt att digitalisera bilder från histologiska prover. Det är dock främst genom de senaste årens utveckling av mikroskopskanners som snabbt kan digitalisera ett helt prov som datoriserade bildanalysmetoder nu på allvar börjar introduceras inom histologin. Tillgängligheten av dataset över histologiska prover med känd, expertverifierad diagnos kan förväntas komma att få stor betydelse för områdets utveckling eftersom det möjliggör jämförelser mellan olika analysmetoder med statistiskt hållbara metoder. Bristen på sådana digitala dataset har hittills medfört beroenden av lokala experter och väsentligt försvårat jämförelser mellan olika nyutecklade analysmetoder vilket begränsat områdets utveckling.

Datoriserad bildanalys vinner alltså idag insteg inom histopatologin efterhand som mikroskopskanners som möjliggör snabb digitalisering av hela preparat införs. Tidigare användning av vanliga mikroskop med en digital kamera monterad gjorde det alltför tidsödande och besvärligt att rutinmässigt använda digitala analysmetoder. Många av metoderna för datoriserad bildanalys av vävnad, bygger på mönsterigenkänning. Medan den klassiska ansatsen till artificiell intelligens är deduktiv och modell- eller regelbaserad, så använder sig statistisk mönsterigenkänning av en induktiv, problembaserad ansats, som bygger på att lära av ett begränsat antal exempel. Dess förnämsta mål är att automatiskt känna igen mönster eller förutspå etiketter. De algoritmer för lärande som används är ofta drivna av resultatet och fokuserar på förmågan att kunna generaliseras också för oförutsägbara fall. De metoder som är vanliga inom mönsterigenkänning, vare sig det är fråga om övervakad klassificering, klustring eller regression, bygger ofta på rigorösa statistiska tekniker för automatisk optimering och urval av parametrar, såväl som för att träna och validera dessa metoder. Det primära målet är automatisering av bedömning och förutsägelse av diagnos eller gradering. Den statistiska ansatsen gör mönsterigenkänning mycket användbart, speciellt för grundläggande tillämpningar som bild-segmentering och igenkänning av objekt.

I den här avhandlingen använder vi mönsterigenkänning och bildanalys för att lösa flera olika tillämpade problem inom histopatologi och immunohistokemi.

Vi presenterar en ny metod för att detektera och lokalisera vävnadskärnor i prover med många sådana kärnor utspridda som en matris på ett mikroskopglas, så kallade "*tissue microarrays*". Vår metod är helt automatisk och robust och vi jämför den med tidigare beskrivna sätt att göra detta.

Vi presenterar också en automatisk metod för att dela upp en färgbild i olika vävnadskomponenter som bygger på en fysisk modell över hur bilden skapas och som tar hänsyn till det brus som då uppstår. Det nya med metoden är att den inte kräver någon styrning, den är "*unsupervised*". Därmed kräver den inte att man i förväg specificerar spektra för de infärgningar som man önskar skilja från varandra. Detta åstadkoms genom att referensfärger uppskattas genom att en modell bestående av blandade normalfördelningar, "*Gaussian mixture models*" tränas med hjälp av maximering av förväntningsvärdet.

En annan ofta förbisedd faktor i histopatologi är valet av infärgning. De olika färgerna som åstadkoms genom de valda kombinationerna av infärgningar avgör i vilken utsträckning de resulterande färgmolnen kommer att överlappa varandra i färgrymden. Denna fundamentala överlappning sätter gränser för hur väl det går att separera de olika vävnadstyperna och därmed för hur bra olika klassificeringsmetoder kan fungera, oavsett hur de är utformade eller hur komplexa de är. I den här avhandlingen presenterar vi ett ramverk för hur man kan optimera valet av histologiska infärgningstekniker genom en systematisk utvärdering som genomförs utifrån att det slutliga målet är automatisk analys och inte den konventionella visuella analysen.

Immunohistokemi är en relativt avancerad infärgningsteknik som använder antikroppar för att detektera antigener och som sedan markerar dessas närvaro genom kromogener som exempelvis 3,3'-diaminobenzidine. Dessa metoder kan underlätta kvantifieringen av biomarkörer som östrogen, progesteron och mänsklig epidermisk tillväxtfaktor 2. Dessutom kan också mängden specifika proteiner exempelvis Ki-67 proteiner som är kopplade till celltillväxt och spridning kvantifieras. Som en tillämpning har vi utvecklat en metod som identifierar par av relaterade antikroppar, så kallade syskon-antikroppar, genom att korrelera sannolikhetskartor över immunohistokemiskt infärgade intilliggande vävnadssnitt.

Till sist så presenterar vi ett nytt sätt att beskriva körtelstrukturer och den vävnadsarkitektur som omger körtlarna. Detta är en viktig komponent i Gleason gradering av prostatacancer och även en komponent i tub-baserad Elston gradering av bröstcancer. Metoden är grundad på att vi definierar formbevarande ringar runt körtelgångarna och sedan ur dessa ringar samlar in kvantitativ information om den rumsliga fördelningen av de olika vävnadstyperna runt körtlarna.

Sammantaget så bidrar avhandlingen med ett antal viktiga komponenter till framtidens system för datorstödd och automatiserad histopatologisk diagnostik och vävnadsgradering.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1175

Editor: The Dean of the Faculty of Science and Technology

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2014

Distribution: publications.uu.se
urn:nbn:se:uu:diva-231039