


University of Cagliari

Department of Mathematics and Computer Science
PhD Course in Computer Science
Cycle XXVIII

Ph.D. Thesis

Computer Aided Diagnosis Algorithms for Digital Microscopy

S.S.D. INF/01

Candidate
Lorenzo Putzu

PhD Supervisor
Prof. Cecilia Di Ruberto

PhD Coordinator
Prof. G. Michele Pinna

Final examination academic year 2014/2015

Abstract

Automatic analysis and information extraction from an image is still a highly challenging research problem in the computer vision area, attempting to describe the image content with computational and mathematical techniques. Moreover the information extracted from the image should be meaningful and as most discriminatory as possible, since it will be used to categorize its content according to the analysed problem. In the Medical Imaging domain this issue is even more felt because many important decisions that affect the patient care, depend on the usefulness of the information extracted from the image. Manage medical image is even more complicated not only due to the importance of the problem, but also because it needs a fair amount of prior medical knowledge to be able to represent with data the visual information to which pathologist refer.

Today medical decisions that impact patient care rely on the results of laboratory tests to a greater extent than ever before, due to the marked expansion in the number and complexity of offered tests. These developments promise to improve the care of patients, but the more increase the number and complexity of the tests, the more increases the possibility to misapply and misinterpret the test themselves, leading to inappropriate diagnosis and therapies. Moreover, with the increased number of tests also the amount of data to be analysed increases, forcing pathologists to devote much time to the analysis of the tests themselves rather than to patient care and the prescription of the right therapy, especially considering that most of the tests performed are just check up tests and most of the analysed samples come from healthy patients.

Then, a quantitative evaluation of medical images is really essential to overcome uncertainty and subjectivity, but also to greatly reduce the amount of data and the timing for the analysis. In the last few years, many computer assisted diagnosis systems have been developed, attempting to mimic pathologists by extracting features from the images. Image analysis involves complex algorithms to identify and characterize cells or tissues using image pattern recognition technology. This thesis addresses the main problems associated to the digital microscopy analysis in histology and haematology diagnosis, with the development of algorithms for the extraction of useful information from different digital images, but able to distinguish different biological structures in the images themselves. The proposed methods not only aim to improve the degree of accuracy of the analysis, and reducing time, if

used as the only means of diagnoses, but also they can be used as intermediate tools for skimming the number of samples to be analysed directly from the pathologist, or as double check systems to verify the correct results of the automated facilities used today.

Contents

List of Figures

List of Tables

Acronyms

ALL	Acute Lymphoblastic Leukaemia
ALL-IDB	Acute Lymphoblastic Leukaemia Image Database
ANN	Artificial Neural Networks
ASM	Angular Second Moment
Aut	Autocorrelation
CAD	Computer Aided Diagnosis
CBC	Complete Blood Count
Con	Contrast
Cor	Correlation
CP	Cluster Prominence
CS	Cluster Shade
DAG	Directed Acyclic Graph
DAve	Difference Average
DEnt	Difference Entropy
DVar	Difference Variance
Ene	Energy
Ent	Entropy
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FS	Fuzzy Sets
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Difference Matrix
GLN	Grey Level Non-uniformity
GLRLM	Gray Level Run-Length Matrix
HGLRE	High Grey Level Run Emphasis
Hom	Homogeneity
HSC	Hematopoietic Stem Cells
IDM	Inverse Difference Moment
IFS	Intuitionistic Fuzzy Sets
LBP	Local Binary Pattern
LGLRE	Low Grey Level Run Emphasis
LoG	Laplacian of Gaussian
LRE	Long Run Emphasis
LRHGLE	Long Run High Grey Level Emphasis
LRLGLE	Long Run Low Grey Level Emphasis
MC	Measure of Correlation
MP	Maximum Probability
NB	Naive Bayes

NN	Nearest Neighbors
NNS	Nearest Neighbors Search
PCA	Principal Component Analysis
PM	Product Moment
RBC	Red Blood Cell
RBF	Radial Basis Function
RLN	Run Length Non-uniformity
ROI	Region of Interest
RP	Run Percentage
SAve	Sum Average
SEnt	Sum Entropy
SRE	Short Run Emphasis
SRHGLE	Short Run High Grey Level Emphasis
SRLGLE	Short Run Low Grey Level Emphasis
SVar	Sum Variance
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
Var	Variance
WBC	White Blood Cell
WBCC	White Blood Cell Count

Chapter 1

Introduction

The visual analysis of bodily fluids and tissues focused on diagnosing diseases using a microscope is called microscopic pathology, that is a sub-discipline of pathology. This kind of analysis still constitutes the final step to confirm if and which illness is present. Traditionally, microscopical pathology has been divided into two main branches, namely *cytopathology* and *histopathology*. Cytopathology refers to diagnosis based on the study of cytological images, that are characterized by the presence of single cells and cell clusters, while histopathology refers to diagnosis based on the study of hystologycal images and involves examination of entire human tissues composed of association of cells into structures which deal with a particular function.

The measurements and characterization of cells from cytological images can be performed automatically since the late 1950s, when Coulter [Cou56] developed a method for sizing and counting cells, using electrical impedance directly from the blood sample. Nowadays the method proposed by Coulter has been improved in order to analyse different particles. A further improvement of this approach is the Flow Cytometry, a technology that is used to simultaneously measure and analyse multiple physical characteristics, chemical properties and defines the maturation stage of particles, as they flow in a fluid stream rapidly and they pass one-at-a-time through at least one laser. Particle components are fluorescently labelled and then excited by the laser to emit light at varying wavelengths, and then distinguished using an optical-to-electronic coupling system that records the way in which the cell emits fluorescence and scatters incident light from the laser. The properties measured include size, morphology, granularity and internal and external structure of cells in question. This system, due to its complexity, needs many quality controls. Some of these controls are performed internally by the same instrument, but others must be performed externally to check the performance of each component.

1.1 CAD - Computer Aided Diagnosis

For this and many others reasons the microscope is still an essential tool to the pathology laboratory today, since manual observation of samples continues to be performed. In fact it can be used to check the results from an instrument or if a recalibration is needed. The manual microscopic examination involves numerous drawbacks, in particular the results accuracy heavily depends on the operator skills. The operators develop their skills during a complex training periods analysing as many cases as possible of different pathology. Nevertheless many cases require different experts and technical opinions in order to reduce human error. As it can be guessed the process of manual microscopic observation is really slow and time consuming, in particular if it involves different operators for a single diagnosis. Digital microscopes are becoming routine pieces of equipment in laboratories, being a combination of a digital camera and a microscope, are able to scan the samples and store the images for future review. Furthermore digital microscopy adds high-resolution and spatial information that cannot be extracted from flow measurements. Digital slides are also, by nature, easier to share than physical slides thus increasing the possibility of consultations between two or more experts. Digital slides have also the potential to be numerically analysed directly by computer algorithms, useful to automate the manual counting of structures, or for classifying the condition of tissue. The extraction of image-based information by computer technology from digital slide is also known as *Digital Pathology* and can be used both in order to speed up the process of diagnosis and both in order to reduce uncertainty and subjectivity.

For this reasons in the last few years many Computer Aided Diagnosis (CAD) system have appeared in order to automate or aid some stages of the diagnostic process, motivated also by the presence of equipment which allows to automatically obtain slides with a good quality. However, automatic interpretation of microscopy medical images is still an open research question. In particular the main challenge when developing CAD systems is the creation of an effective method to extract meaningful information from the images, such as the number of cells in the film or the position of the different structures in a tissue. This issues becomes more complex, in terms of artificial vision, considering that there is not a colour standardization for the staining and acquisition of digital slides. In fact, there is a huge colour variability between different slides, due to the quality of the biological sample and the sample preparation, such as the quantity of dye used during the staining procedure, or due to different acquisition system and the image capturing parameters, such as the environment illumination. Furthermore, such variability may be present in the same slide, in particular the presence uneven lighting, with a central area very bright and shading areas more marked towards the corners, that can be caused by an excessive use of the light of the microscope.

1.2 Contributions

The purpose of this thesis has been the detailed analysis of the outstanding issues in the CAD from digital microscopy images, studying some possible solutions and proposing strategies and algorithms applied in two specific use cases: histology and haematology. Special efforts have been focused on the extraction of useful information from the digital images, developing general algorithms not dependent on specific dataset. In particular, for what concerns the histology image analysis, the most important contribution has been made with the realisation of a general classification framework, able to manage different medical problems with an high accuracy. In the proposed framework no object detection or segmentation method is needed, since every segmentation algorithm applied to histology images produces an huge number of regions and structure, too difficult to manage singularly. The overall procedure instead is totally based on the analysis of textures, being the most suitable to analyse the tissue structure. Also, great importance has been given to the analysis of colours, considered one of the most interesting contents to be analysed in the histological images, studying not only the internal correlation of various colours, but also by analysing the correlation between different colours. For what concerns the peripheral blood image analysis special effort has been made to address the counting and in particular the segmentation issues, proposing different segmentation algorithms able to isolate the cell of interest from images acquired in different illumination condition and stained with different dye. This effort to correctly segment and count the cells has been made in order to be able to manage each cell singularly, and then to diagnose the presence of leukaemia. Since the importance of this kind of diagnosis different ensembles of descriptors and classifier have been evaluated in order to provide a result as accurate as possible. The scientific results obtained during this PhD work and described in this thesis also appeared in related publications, listed below:

- C. Di Ruberto, L. Putzu, "A Feature Learning Framework for Histology Images Classification", Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology - Systems & Applications, Elsevier/MK book, 2015, (In Press).

- C. Di Ruberto, A. Loddo, L. Putzu, "A Multiple Classifier Learning by Sampling System for White Blood Cells Segmentation", G. Azzopardi, N. Petkov Eds. Computer Analysis of Images and Patterns - 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I. Lecture Notes in Computer Science 9256, Springer 2015, pp. 415-425, ISBN 978-3-319-23191-4.

- C. Di Ruberto, G. Fodde, L. Putzu, "On Different Colour Spaces for Medical Colour Image Classification", G. Azzopardi, N. Petkov Eds , Computer Analysis of Images and Patterns - 16th International Conference, CAIP 2015,

Valletta, Malta, September 2-4, 2015 Proceedings, Part I. Lecture Notes in Computer Science 9256, Springer 2015, pp. 477-488, ISBN 978-3-319-23191-4.

- C. Di Ruberto, A. Loddo, L. Putzu, " Learning by Sampling for White Blood Cells Segmentation", V. Murino, E. Puppo Eds.: Image Analysis and Processing - ICIAP 2015 - 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part I. Lecture Notes in Computer Science 9279, Springer 2015, pp. 557-567, ISBN 978-3-319-23230-0.

- C. Di Ruberto, G. Fodde, L. Putzu, "Comparison of Statistical Features for Medical Colour Image Classification", L. Nalpantidis, V. Krger, J.O. Eklundh, A. Gasteratos Eds., Computer Vision Systems - 10th International Conference, ICVS 2015, Copenhagen, Denmark, July 6-9, 2015, Proceedings. Lecture Notes in Computer Science 9163, Springer 2015, pp. 3-13. ISBN 978-3-319-20903-6.

- C. Di Ruberto, L. Putzu, "Accurate blood cells segmentation through intuitionistic fuzzy set threshold", Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014, 23-27, November 2014 - Marrakech, Morocco. 978-1-4799-7978-3/14, DOI 10.1109/SITIS.2014.43, pp. 57-64.

- L. Putzu, G. Caocci, C. Di Ruberto, "Leucocyte classification for leukaemia detection using image processing techniques", Artificial Intelligence in Medicine, published online September 14 2014, ISSN 0933-3657, vol. 62, Issue 3, November 2014, pp. 179191, http://dx.doi.org/10.1016/j.artmed.2014.09.002.

- L. Putzu, C. Di Ruberto, "Investigation of different classification models to determine the presence of leukemia in peripheral blood image", 17th International Conference on Image Analysis and Processing, ICIAP 2013, Naples, Italy, 9-13 September 2013, vol. 8156 LNCS, Issue PART 1, 2013, Pages 612-621, ISSN: 0302-9743 ISBN: 978-364241180-9 DOI: 10.1007/978-3-642-41181-6_62.

- L. Putzu, C. Di Ruberto, "White blood cells identification and classification from leukemic blood image", Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2013, 18-20 March 2013, Granada, Spain, pp. 99 106, Copicentro Editorial, ISBN: 978-84-15814-13-9.

- L. Putzu, C. Di Ruberto, "White blood cells identification and counting from microscopic blood images", Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Engineering, ICBCBBE 2013, 14-15 January 2013, Zurich, Switzerland, World Academy

of Science, Engineering and Technology, no. 73, January 2013, pp. 268-275, pISSN: 2010-376X, eISSN: 2010-3778.

1.3 Dissertation structure

The remainder of this dissertation describes in more detail what has been anticipated in this introduction and it is organised as follows:

- Part I illustrates a typical schema of CAD system for digital microscope images. From this schema four main phases have been highlighted and then analysed in detail in chapters 2, 3, 4 and 5, giving an idea of the most used techniques and illustrating the basic concepts useful to the comprehension of the proposed CAD systems.

- Part II describes the proposed CAD system for histology image analysis, giving firstly a more detailed idea on what this system can be used for and then illustrating in detail the chosen solutions.

- Part III describes the proposed CAD system for peripheral blood image analysis, clarifying the issues related to images acquired from the blood smear and then showing different solutions for each phase. In particular, in chapter 7 different solutions for segmentation, in chapter 8 a method for leukocyte count and in chapter 9 a successful approach for classification of cells affected from leukaemia are presented.

- Part IV concludes the dissertation giving some final comments on the proposed approaches, discussing the choices made with the obtained results. The experimental results obtained have brought to further ideas for the future, not just to improve the overall procedures but also to extend the proposed CAD systems to further issues and to different medical problems.

Although the ultimate goal of this work is to recognize and diagnose the presence of diseases through automated methods, to relieve the reader from excessively onerous medical definitions, some details have been moved to the Appendix A that outlines the types of diseases that are directly connected with an abnormal number of cells in the peripheral blood stream and shows how the presence of diseases or parasites may affect the morphology of the cells themselves.

Part I

Digital Microscopy CAD

CAD systems could differ a lot between them, not only because they deal with different medical problems, but also because they manage different kind of images and data. Many other differences can be observed among CAD systems that deal with images acquired with a digital microscope. Despite these major differences CAD systems for digital microscope based on image processing techniques generally consist of the same phases:

- Image pre-processing: noise suppression added to the image during the phase of acquisition and improvement of the luminance and contrast of the images.
- Segmentation: partitioning the image in order to isolate the objects of interest in the image. This phase is the most important in the analysis of automatic images, as from the result of the segmentation depends the result of the entire analysis.
- Feature extraction: feature extraction of objects previously segmented, necessary for subsequent classification.
- Classification: assignment of segmented objects to a specific class. Through this last phase the presence of any diseases is determined.

Obviously not all the CAD systems use all the phases just mentioned. Often the phase of pre-processing is not performed, since the images captured by the new digital microscopes are of excellent quality, with a very low percentage of imperfections. On the contrary, sometimes some phases may be present several times in the system to deal with different issues regarding the type of images under consideration. For example, in the images of peripheral blood smear multiple stages of segmentation are performed, useful to identify firstly the cells and then to separate the components of the cells themselves. In the following chapters, not only the most used techniques in the various phases of a CAD system will be deepened, but the basic concepts useful to the comprehension of the proposed methods for histological image and for peripheral blood image analysis, described in part II and III respectively, will be illustrated.

Chapter 2

Image Pre-processing

Methods of pre-processing are referred to as all those operations which affect the image in order to improve certain characteristics, for example, those which serve to improve its contrast or to better separate objects from the background, those which have the purpose of suppressing other unwanted characteristics such as noise or other details not interesting for the purpose of the analysis that will be performed. This kind of operations is not always performed on CAD systems, both because nowadays the image acquisition tolls are equipped with high level cameras and so they give images of good quality and both because sometimes pre-processing operations tend to alter or suppress the tissues and cells structures useful for the further steps. Typically images structures are represented with an appropriate definition with only a small quantization noise or they can be acquired with low contrast. For these reasons the most used operations of pre-processing are:

- Operations on the histogram
- Operations of local pre-processing
  - Smoothing operators
  - Sharpening operators

2.1 Operations on the histogram

The histogram of a digital image in grey levels provides the frequency of a certain intensity value within the image, thus indicating the number of pixels for each grey level. The histogram provides interesting information on the image, such as the number of modes, the presence of a predominant pick or the distribution of the grey levels along the histogram. If the image presents a low contrast all the pixels values are condensed into a part of the histogram. A bimodal histogram often denotes the presence of a fairly homogeneous brightness object on an almost constant background. Operate on the histogram means define a mapping h from the

initial space of grey levels in a new space of grey level $h : [0..255] \rightarrow [0..255]$, whose application to the image I, is the replacement of grey level $I[i, j]$ with $h(I[i, j])$. So, defining the appropriate mapping, it is possible to increase the contrast within the image, equalize the histogram and highlight or hide some image details. The *contrast stretching* operation is necessary when the histogram does not expand over the entire frequency range. To correct this situation a real stretch of the histogram is made, mapping the minimum and the maximum value of the original image with the value 0 and 255, respectively. This operation is very common for cytometric image analysis, as it allows a good separation of the objects in the foreground from the background. The *histogram equalisation* operation is very important because it is often used to make comparable images captured in different lighting conditions. The equalisation is achieved by defining a mapping that equally distributes the pixels values. This operation produces a histogram (theoretically) flat, considering $h(x)$ as the histogram of the original image, it can be changed through the use of the function $y = y(x)$, so that the histogram $g(y)$ of the resulting image is constant for each intensity value $g(y) = c$.

2.2 Operations of local pre-processing

The methods of local pre-processing, unlike the previous ones, operate on a small neighbourhood of pixels of the original image to calculate new pixel values of the resulting image. These operations are also called filtering operations as they make use of digital filters. The neighbourhood of pixels used for the calculation of the new intensity value has cardinality always odd, in such a way that the pixel in question is in the middle. Typically, the sizes are 3x3, 5x5 or 7x7. The filter used for the filtering operation will have the same size of the neighbourhood of pixels that we are considering. The values of the filter are used as weight that are multiplied by the corresponding pixel values of the neighbourhood and then added together to give rise to the new value of the pixel in the resulting image. We can distinguish two groups of local methods of pre-processing, in accordance with their ultimate goal that are the smoothing operators and the sharpening operators.

2.2.1 Smoothing operators

The smoothing operators have the purpose of suppressing noise in images or other small unwanted details, using redundancy in images. Unfortunately, as said previously, these operators tend to flatten also interesting details such as the edges of objects and the tissue or cell structures but generally they produce good results in the removal of impulsive noise. The most used smoothing operator is the *averaging filter*, that stores in the actual pixel the average value of its neighbourhood. The results in this case are acceptable if the noise present in the image has a minor dimension of the objects of interest of the image, but the contours of the objects are

in any case heavily altered. Average filters can also be constructed with different weight values, in order to better reflect the characteristics of the Gaussian noise, in fact, they are called Gaussian filters, since they simulate the trend of a Gaussian curve. Another common smoothing operator is the *median filter*, which behaves like the average filter but this time storing in the actual pixel the median value of its neighbourhood.

2.2.2 Sharpening operators

The sharpening operators have the purpose of highlighting the interesting details of the image, such as the edges of objects. Unfortunately, these operators tend to highlight also the noise present in the image, but in any case improve the perceived picture detail. These operators are based on the use of local derivatives of the image. Since the image is a discrete function the traditional definition of derivative can not be applied. In digital images the operator used for the first derivative is the intensity difference between adjacent pixels. In most cases the sharpening operators use the second derivatives, since they are more sensitive to intensity variations. The most used sharpening operator is the *Laplacian filter* that brings the desired effect of sharpening by subtracting to the original image the resulting image after applying the Laplacian filter. Another useful sharpening operator is the *gradient operator* that makes use of the gradient of the image to detect and improve the edges. An edge is a set of connected pixels (4 or 8 connected) with sharp changes in brightness, so it can be detected as any transition of grey levels, where the slope of this transition is proportional to how the edge is sharp. Such a change of the image function can be described by the gradient pointing in the direction of greatest increase of the function. Different operators are able to correctly determine the direction of the gradient by the use of first derivative. They are operators of Prewitt, Sobel, Kirsch and Robinson. Another common sharpening operator used in most of the commercial products to make the image noticeably sharper is the filter of *Unsharp*. This filter is a simple sharpening operator which derives its name from the fact that it enhances edges and other high frequency components in an image via a procedure that subtracts a smoothed (or unsharp) version of the image from the original image.

2.3 Colour and Colour Spaces

Colour is very important in CAD systems since the biologists stain blood and tissues to highlight spacial structures. The colour is the brains reaction to a specific visual stimulus. It is extremely subjective and personal, thus trying to attribute numbers to the brains reaction to visual stimuli is very difficult. The aim of colour spaces is to aid the process of describing colour, either between people or between machines or programs. The presence of more than one colour space is due to the fact that different colour spaces are better for different applications, for example some devices

have limiting factors that dictate the size and type of colour space that can be used. Thus, some colour spaces are tied to a specific piece device (device dependent) while others are equally valid on whatever device they are used (device independent). In order to classify colour spaces into a few categories with respect to their definitions and their properties, the classical colour spaces are classified into three main families [VMP03, BVMP04]: primary spaces, luminance-chrominance spaces and perceptual spaces.

2.3.1 Primary Spaces

The primary spaces are based on the trichromatic theory, assuming that is possible to match any colour by mixing appropriate amounts of the three primary colours. Example of primary spaces are the real RGB, the subtractive CMY(K), and the imaginary XYZ. The most widely used colour space is the RGB colour space, where a colour point in the space is characterized by three colour components of the corresponding pixel which are red (R), green (G), and blue (B). In general, colour images are acquired through the RGB colour space, called the image acquisition colour space. So, all the colour spaces are expressed thanks to transformations of the R, G and B channels. As previously said, CMY(K) is the subtractive colour space and can be obtained easily from a set of RGB values by subtracting the individual RGB values from 1, since a pure cyan (C) surface does not contain R, a pure magenta (M) surface does not contain G and a pure yellow (Y) surface does not contain B. So, the equation is:

$$\begin{bmatrix} C \\ M \\ Y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.1}$$

with the assumption that in general all colour values have been normalised to the range $[0, 1]$. In general, this colour space is used in colour printers and copiers that perform this conversion internally. According to the equation, equal amounts of the colour channel should produce black. In practice instead their combination for printing produces only a dark colour far from the real black. Thus, in order to produce a true black a fourth colour has been added, giving rise to the CMYK colour space. In this case the conversions start from the just computed CMY, by finding the black (K) channel and then correcting the complementary colours based on the value of K.

$$\begin{aligned} K &= minimum(c, m, y) \\ C &= c - K \\ M &= m - K \\ Y &= y - K \end{aligned} \tag{2.2}$$

The XYZ colour space is obtained from the RGB colour space using the following equation:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412 & 0.357 & 0.180 \\ 0.212 & 0.715 & 0.072 \\ 0.019 & 0.119 & 0.950 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.3}$$

2.3.2 Luminance-Chrominance Spaces

The luminance-chrominance spaces are used in that case where it is useful to separate the colour definition into luminance, represented by one component, and chrominance represented by the two other components. Among these colour spaces there are the television transmission colour spaces, sometimes known as transmission primaries, YIQ and YUV for analogical standard and YCbCr for digital standard. For this reason only the YCbCr have been taken into account, which could be obtained from the RGB using the following equation:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ \text{-}0169 & \text{-}0.331 & 0.500 \\ 0.500 & \text{-}0.419 & \text{-}0.081 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.4}$$

The International Commission of Illumination (CIE) has defined a system that classifies colour according to the human visual system in order to specify any colour in terms of its CIE coordinates. There are two main CIE based colour spaces, CIELUV (Luv) and CIELAB (Lab). They are nearly linear with visual perception with the L parameter that has a good correlation with perceived lightness and the other two parameters that express the chrominance. They are based on the XYZ colour space so they could be obtained by a conversion from the XYZ colour space using the following equations:

Luv

$$\begin{aligned} L &= \begin{cases} 116\sqrt[3]{y_r} - 16, & y_r > 0.008856 \\ 903.3 y_r, & y_r \leq 0.008856 \end{cases} \\ u &= 13L\,(u' - u'_r) \\ v &= 13L\,(v' - v'_r) \end{aligned} \tag{2.5}$$

where

$$
\begin{aligned}
y_r &= \frac{Y}{Y_r} \\
u' &= \frac{4X}{X + 15Y + 3Z} \\
v' &= \frac{9Y}{X + 15Y + 3Z} \\
u'_r &= \frac{4X_r}{X_r + 15Y_r + 3Z_r} \\
v'_r &= \frac{9Y_r}{X_r + 15Y_r + 3Z_r}
\end{aligned}
$$

Lab

$$
\begin{aligned}
L &= 116 \cdot h\left(\frac{Y}{Y_r}\right) - 16 \\
a &= 500\left[h\left(\frac{X}{X_r}\right)\left(\frac{Y}{Y_r}\right)\right] \\
b &= 200\left[h\left(\frac{Y}{Y_r}\right)\left(\frac{Z}{Z_r}\right)\right]
\end{aligned}
\tag{2.6}
$$

where

$$
h(q) = \begin{cases} \sqrt[3]{q}, & q > 0.008856 \\ 7.787q + 16/116, & q \leq 0.008856 \end{cases}
$$

Both equations require a reference white X_r, Y_r and Z_r.

2.3.3 Perceptual Spaces

The perceptual spaces try to quantify the subjective human colour perception by means of the intensity, the hue and the saturation. This family represents a wealth of similar colour spaces, alternative names include HSI (intensity), HSV (value), HSL (Hue Saturation Lightness), HCI (chroma), etc. Most of these colour spaces are linear transforms from RGB and are therefore device dependent. Their advantage lies in the extremely intuitive manner of specifying colour, selecting the desired hue and then modifying it slightly by adjustment of its saturation and intensity. Furthermore, the separation of the luminance component from colour information is stated to have advantages in image processing. Here only the HSV colour space has been taken into account, since it is the most used perceptual space and it can be obtained from the RGB colour space with the following equation:

$$
\begin{aligned}
&Max = max(R, G, B) \\
&Min = min(R, G, B) \\
&\Delta = Max - Min \\
&V = Max \\
&S = \frac{(\Delta)}{Max} \\
&H = \begin{cases} 0, & \text{if } \Delta = 0 \\ 60^\circ \times \left(\frac{G-B}{\Delta} mod 6\right), & \text{if } R = Max \\ 60^\circ \times \left(\frac{B-R}{\Delta} + 2\right), & \text{if } G = Max \\ 60^\circ \times \left(\frac{R-G}{\Delta} + 4\right), & \text{if } B = Max \end{cases}
\end{aligned} \tag{2.7}
$$

Chapter 3

Segmentation

Segmentation is one of the most important steps in image analysis, identifying and separating, according to certain criteria of homogeneity and separation, the regions of the image. Its main objective is to divide the image into parts that have a strong correlation with objects or areas of the real world contained in the image. The commonly used segmentation methods operate essentially based on characteristics such as the value of brightness, colour and reflection of the individual pixels, identifying groups of pixels that correspond to spatially connected regions. As for many problems of image processing, there is no standard solution valid in general, so, depending on the characteristics of the images to process, and especially of the objects to be segmented, different segmentation techniques can be applied, some simple, that often lead to unsatisfactory results, and other more powerful but with a very high computational cost. For the analysis of medical images two main levels of segmentation are used: the first segmentation level which aims to separate whole cells or tissues from the background and the second level of segmentation which aims to separate the tissue structure in different regions or the cell in their components as the nucleus from the cytoplasm or intracellular parasites. The latter case is commonly used in applications in which the class of the cell depends on the morphological characteristics of its components. The segmentation techniques can be divided into three main categories: Pixel Based, Edge Based and Region Based.

3.1 Pixel Based or Thresholding

Thresholding is the most elementary and computationally cheaper technique for image segmentation, making use of a threshold operator that directly involves the histogram of the image. The hypothesis underlying this segmentation technique is that the pixels of an object have more or less the same brightness, and therefore, they can be separated from the background by a threshold brightness values. The problem becomes more difficult if the histogram presents more than two peaks, in fact in this case more threshold values will be necessary to separate the objects of the

image. This technique presents some drawbacks, in particular if the threshold is not chosen accurately, the detected objects can shrink or grow. This change in size can be crucial in applications where size is an important parameter for the classification of the object itself. Moreover, a wrong threshold value can cause the partial fusion of two or more objects between them, making it impossible for subsequent classification and identification of the same. When the intensity distribution of objects and background is sufficiently distinct it is possible to use a single global threshold [GW, GWE04] applicable to the entire image. The value of this global threshold is calculated starting from an initial threshold value T between the minimum and maximum value of the histogram, that allows to make a first segmentation. This produces two groups of pixels, their values are averaged and used to calculate the new threshold value T. The process is then iterated until the difference between successive values of T is less than a predetermined parameter. This simple algorithm works well in situations where there is a clear valley between the various fashions histogram, relative to the background and objects.

3.1.1 Otsu Algorithm

The method of Otsu [Ots75] also performs a global threshold, but differently from the previous one allows to obtain an optimal threshold value, as it maximizes the variance between the classes. In fact the classes if well segmented will be differentiated from the intensity value of their pixels. A threshold that gives the best separation between the classes in terms of intensity is an optimal threshold. The method of Otsu, moreover, can be extended to the segmentation of images that need more threshold values, since the measure of separability on which is based also extends to an arbitrary number of classes. It begins to lose meaning when the number of classes increases excessively, since it works only with one variable which is the intensity. Typically, however, applications that require more than two threshold values are resolved with the use of other values in addition to the intensity, such as the colour or the entropy present in the histogram [KSW85].

3.1.2 Zack Algorithm

The Zack algorithm [ZRL77], known also as triangle method, differently from the other methods, doesn't work directly on the intensity value of the histogram, but it works on the image obtained from the histogram plot. It is called triangle method because it draws a sort of triangle, constructing a straight line that connects the highest histogram value $h[b_{max}]$ and the lowest histogram value $h[b_{min}]$, where b_{max} and b_{min} indicate the values of the grey levels where the histogram $h[x]$ reaches its maximum and minimum, respectively. The distance d between the marked line and the histogram values between b_{min} and b_{max} is then calculated. The intensity value, where the distance d reaches its maximum, defines the threshold value. This

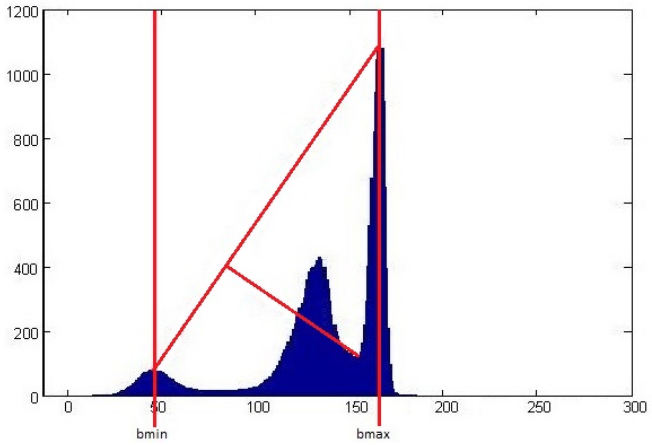

Figure 3.1: Example of Zack algorithm.

algorithm is particularly effective and fast, indeed, differently from the ones seen before, it is computed directly on an image without iteration.

3.1.3 Fuzzy Threshold

The algorithm for calculating the threshold value just seen are also called crisp techniques. They produce excellent results with well defined images and regions, but the segmentation process becomes complex in presence of noise or imprecision. The nature of this imprecision in the image arises from the presence of uncertainty, that can lead to ill defined regions. In this case it is appropriate to avoid crisp segmentation and to prefer a fuzzy segmentation. Fuzzy threshold approaches are based on Fuzzy Sets (FSs) theory, in fact regions may be viewed as fuzzy subsets of the image. Several researchers have worked on fuzzy based thresholding techniques, in particular in order to identify the best fuzzy measure able to separate the fuzzy subset, such as the fuzzy compactness [PR88], the fuzzy similarity [RAS$^+$00] or the fuzzy divergence and gamma membership [CR03, MPCB$^+$13]. Even if the segmentation performances of fuzzy threshold approaches are better than many crisp methods, their computational performances are not comparable with the crisp methods mentioned before, entirely based on computations performed on the histogram, a 1-D array easily obtainable from the image, while fuzzy approaches are based on computations performed on the image, a 2-D array of size $M \times N$.

3.1.4 Local Thresholding

When the background is not constant and the contrast of the objects varies unevenly, the global thresholding can work properly in an area, but produces unsatisfactory results in other areas. In this case the use of local thresholding could be a better solution. The image is subdivided into rectangular overlapping sub-images and the

histogram of each of them is calculated. The sub-images must be big enough to include both the object and the background pixels. If the sub-image has a bimodal histogram, the minimum value between the two peaks is precisely the threshold. Instead, if the histogram is unimodal, the threshold value must be calculated by interpolation of the thresholds of the adjacent sub-images. In general, the local thresholding is computationally more expensive than the global, although it is very useful to segment objects from the background and to extract very small and scattered variable regions.

3.2 Edge Based

This segmentation technique is not based on the intensity value of the pixels, but on the fact that an object to be identified must have a closed edge that surrounds it. This assumption is not always true, but often verified. The edges of objects are preliminarily identified by applying suitable operators of edge detection. As said previously, an edge is a set of connected pixels (4 or 8 connected) that lie on the border of two regions, and therefore presenting sharp changes in brightness. There is a little difference between edge and boundary, since the edge is a local concept while the boundary is an extended concept. So, it can be said that the boundary of an object is composed of a series of edges. In order to detect only the edge, it is possible to carry out a threshold applied to the first derivative, an operation that takes the name of non-maxima suppression, as it resets all the values in the first derivative which are not maximum. Actually, the threshold value used allows to select more than one maximum points, as to have a contour wider but visible. Instead, if the operation of threshold is applied to the second derivative, this operation takes the name of zero crossing and has the aim to seek points of zero crossing necessary for the location of the edge points without distortion, avoiding the appearance of the double border. As previously mentioned, there are several implementations of filters on the first derivative, that can be applied directly to the search of the edges. In general, before the edge detection a smoothing filter is applied to the image, in order to reduce both the noise and the thickness of the edge, so that the detection is more effective. In fact, the process works perfectly if the signal is not noisy and if the edge is quite localized in space and with small amplitude. An evolution of this approach are the Laplacian of Gaussian (LoG) filter with zero crossing and the Canny filter with non-maxima suppression.

3.2.1 LoG Operator

This edge detection approach, also known as Marr-Hildreth algorithm [MH80], is based on the second derivative of a function. The Laplacian operator is rarely used by itself for edge detection, as it is very sensitive to noise, but it is used in conjunction with the Gaussian, known precisely as Laplacian of Gauss, LoG. The

fundamental characteristics of LoG are the Gaussian smoothing filter, used to reduce noise and to enlarge the edge, the Laplacian in two dimensions, the zero crossings in the second derivative and, finally, the edge location estimated with sub-pixel using the linear interpolation. The Gaussian filter is preferred because it applies an action of smoothing both in space and in frequency. Furthermore the derivative of the Gaussian filter is independent from the considered image and can be pre-calculated analytically by reducing complexity. After applying the Laplacian filter you just have to search points in the image in which there is a zero crossing, considering only those zeros for which there is a change of sign in all possible directions around zero.

3.2.2 Canny Operator

Also the Canny algorithm [Can86] uses a Gaussian filter for the smoothing. Then the magnitude and direction of the gradient is calculated using different but finite approximations of partial derivatives. It applies the non maxima suppression to the magnitude of the gradient and finally it uses the double threshold algorithm to find and link the edges. The use of the double threshold is necessary, given that a single threshold would not lead to satisfactory results. For example if the threshold is too low it'll detect too many false edges, instead if the threshold is too high some real edge will be lost. The use of two thresholds produces two different images and, of course, the image with the higher threshold will present a much smaller number of edges. Starting from this image each edge is examined and compared with that one of the other image in search of edges that have been lost and that can be linked to create a continuous boundary.

3.2.3 Deformable Models

In medical image analysis, in many cases, the boundaries between the tissue structures and cell components are not clearly defined. The use of edge detection on these images produces poor results, in particular due to the presence of small structures or particles this approach produces a huge number of false edges. Moreover, in general the edge approaches based on filters do not yield to a closed contour. As a result, these techniques either fail completely or require some kind of post-processing step to remove invalid object boundaries in the segmentation results or to close the contour extracted. To address these difficulties, deformable models or snakes [KWT] have been extensively studied and widely used in medical image segmentation, with promising results. Deformable models are curves or surfaces defined to match a contour as an energy minimization problem, where the optimal solution constitutes an equilibrium of internal and external energy. The deformable model can move under the influence of internal forces, which are defined within the curve or surface itself to keep the model smooth during deformation, while external forces, which are computed from the image data, are defined to move the model toward an object boundary or other desired features within an image. By constraining extracted

boundaries to be smooth and incorporating other prior information about the object shape, deformable models offer robustness to both image noise and boundary gaps.

3.3 Region Based

These segmentation techniques introduce more information than previous concerning the connectivity of the pixels forming the entire object, avoiding in this way that individual points of a same region, having the right colour or the right contrast, are classified as separate objects. Unlike the pixel and edge based segmentation methods, region based approaches aim to identify objects and regions, working directly on the space occupied by the pixels instead of identifying objects from their properties, such as brightness or edges. Considering R as the entire spatial region occupied by the image, the segmentation process can be seen as the partitioning of R into n sub-regions $R1, R2, ..., Rn$, with the constraint that, the union of all regions returns R and the intersection between any set of regions is equal to 0. The pixels belonging to a region must be connected (8 or 4 connected) and they must be related by a similarity criterion. The most common techniques of segmentation region based are divided into region growing and split and merge techniques.

3.3.1 Region Growing

The region growing is a procedure that allows to select regions connected and homogeneous of an image, whose selection is effected from a single pixel and is based on a criterion of similarity or growth which imposes a maximum difference, defined a priori, between the value of the initial pixel and the pixel values of the region. The basic approach involves the selection of a set of starting pixels, called seed points, and from these seeds add the pixels of their neighbourhood that have certain properties of similarity with the seeds, such a specific intensity range or colour. The selection of the set of starting pixels is based on the nature of the problem, instead the selection criteria of similarity depends both on the problem and the type of the input image, since it must be ensured independence between the result of the segmentation and the scan direction of the image or the seed points selected.

Algorithms of region growing have been widely used for the analysis of peripheral blood images, in particular for segmentation of cells in which the nucleus is easily identifiable and thus can be used as a seed. Instead for histology images none of the segmentation approaches proposed is based on region growing algorithm, since the nature of the images doesn't help for the selection of proper seed points.

3.3.2 Split and Merge

Segmentation can be performed also by partitioning (splitting) recursively an image, until uniform regions are obtained. Then the aggregation (merge) can be performed

on adjacent regions that may be compatible on the basis of a criterion of similarity. In the simplest way an image can be partitioned recursively repeating a division into four quadrants, until regions composed also by a few pixels, uniform according to the defined criterion of similarity have been obtained. Such division into quadrants is represented by a tree called quad-tree, in which the root node contains the information of the whole image and each of the four children nodes contains the information about a quadrant. If a quadrant is sufficiently uniform will not be further partitioned. The step of splitting partitions inevitably also homogeneous regions of the image and makes necessary a subsequent phase to merge the adjacent and homogeneous regions of the image into a region that meets the defined criterion of similarity.

3.3.3 Watershed

The watershed segmentation [Mey94] is a mixed approach based on the pixel aggregation (flooding) with the use of the gradient of the image as a barrier for flooding. In this approach the gradient image can be considered as 3D image (topographical), in which can be identified points that come from a regional minimum, points that surely will fall in a local minimum called basins and points of local maximum called watershed lines. The flooding applied to this image leads to a state where only the watershed lines are visible, that correspond to the contours of the objects in the image. Actually a direct application of the watershed algorithm induces an over-segmentation due to the presence of too many basins which can never merge, or the presence of noise and other irregularities of the gradient. In order to avoid this problem often in digital microscopy images analysis different strategies that include the use of markers are used. These markers can be extracted directly from the original images intensity value and combined with the gradient to obtain a stronger result.

3.4 Post-Processing

After segmentation, an image can be represented as a map of binary objects, background and foreground. In some cases the initial segmentation is not satisfactory, as it can present holes or artefacts. Some improvement to the segmentation results can be made directly on the binary image using a series of operations based on a priori knowledge. The morphological operators are commonly used for this purpose, to reduce the number of artefacts, to fill the holes present in some regions, to remove some objects not completely enclosed in the image or other objects that do not interest the analysis. Mathematical morphology is based on set theory for binary images [Ser83a, Ser83b] and on lattice theory for gray level images. It provides some approaches for image processing that are useful to extract image components and for the representation and description of the shape of the objects. The sets in this

case represent the objects contained in the image. The operations of mathematical morphology are based in the use of structuring elements, that are small sets of sub-images used to investigate and study the properties of interest in the input image. They can have any shape defined according to the problem to be treated and are represented as binary matrices.

Given an image or set A and a structuring element B, the operations are realized by sliding B over A so that the origin of B visit all elements of A. The *erosion* operation creates a new set by considering all the location of B for which B is fully contained in A. The result is that the contour of the set A has been eroded. Such a property for which B must be fully contained in A is equivalent to the property for which B must not share any elements with the complement of A. The erosion and dilation are dual operations, thus it is possible to obtain the dilation of A through the use of a structuring element B eroding the complement of A. In a more direct way the *dilation* operation creates a new set by considering all the location of B for which at most one element of B is contained in A. The result is that the contour of the set A has been dilated. A simple operation that arises from the erosion is the *boundary extraction*. In fact the contour of a set A can be obtained from the difference between the original set A and the erosion of A with an appropriate structuring element B.

Also the opening and closing arise from the composition of erosion and dilation. The *opening* of an image or set A with a structuring element B is defined as the erosion of A with B followed by a dilation of the result. This is useful to flatten the contours of an object, breaking the thin lines and remove the sharp contour. The *closing* of an image or set A with structuring element B is defined as the dilation of A with B followed by the erosion of the result. Even the closing flattens the contours but in a different way, in fact it eliminates small holes and fills the gap in the contours. To fill bigger holes instead the operation of *hole filling* that consists of a more laborious process is used. Assuming to have a set A with a hole inside, the process starts taking the complement of A, that is composed by the background pixels and therefore also the hole pixels. Then iteratively a set containing only a pixel for each hole is dilated using an appropriate structuring element and making every time the intersection with the complement of A, so as to exclude pixels outside the contour of A. In a similar way an iterative procedure of dilatation is used for the *extraction of connected components*. This time instead the procedure starts from the points of the connected components in A, that are dilated until the connected components have been filled. At each iteration the intersection is performed with A, so as to exclude pixels outside of the connected component.

Chapter 4

Feature Extraction

Once the image has been segmented into regions, the collection of resulting segmented pixels is represented and described appropriately for further processes. The representation of a region may be based on external characteristics, such as the contour and shape or internal characteristics, such as the colour and displacement of the pixels inside the region. The next step is to describe the regions according to the chosen representation. The most used representation for histology images is based on internal characteristics, in order to describe the structure of tissues and cells using the colour and the displacement of the pixels inside the region. While in the case of the peripheral blood images, it may be necessary to use both representations, as it is important to analyse characteristics such as shape and area of a cell and also regional characteristics such as colour and texture. The features must be extracted from the object in order to describe it. The ideal descriptors are those independent to transformation such as the orientation of the object, size and position and that are sufficiently discriminatory. The purpose of the phase of feature extraction is to obtain a set of descriptors, that will be further separated into different classes by a classification procedure. Features are classified into two distinct groups:

- general features: application independent features such as colour, texture and shape. They can be further divided into features calculated at each pixel, like colour and location (*pixel-level features*), features calculated over the results of segmentation or edge detection (*local features*) and features calculated over the entire image or sub-image (*global features*).

- domain-specific features: application dependent features such as human faces, fingerprints and conceptual features.

Moreover, all features can be coarsely classified into low-level features and high-level features. Low-level features can be extracted directly from the original images, whereas high-level feature extraction must be based on low-level features.

4.1 Contour Descriptors

As previously mentioned, many descriptors may be extracted directly from the segmentation result. For example, one of the simplest descriptors is the *contour length*. A good approximation of the contour length can be easily obtained by counting the pixels of the contour. The value of the *diameter* can be obtained just as easily by computing the maximum distance between two points of the contour. The segment that connects the end points of the diameter is called the *major axis*, while the *minor axis* is that segment perpendicular to the major axis, of such a length that a rectangle, passing through the four points of intersection of the axes with the contour encloses completely the contour. This rectangle just described is called a *bounding box*, having sides parallel to the two axes. The ratio between the sides of the rectangle or the ratio between the two axes measures the value of *eccentricity* (4.1). The *elongation* measures how an object is elongated (4.2), while *rectangularity* represents how rectangular a shape is, or better how well it fills its minimum bounding box (4.3).

$$eccentricity = \frac{\sqrt{(majoraxis^2 - minoraxis^2)}}{majoraxis} \tag{4.1}$$

$$elongation = 1 - \frac{minoraxis}{majoraxis} \tag{4.2}$$

$$rectangularity = \frac{area}{majoraxis * minoraxis} \tag{4.3}$$

These descriptors are really useful to discriminate the shape of abnormal objects to normal ones, in particular peripheral blood cells given their almost circular shape should be distinguished easily from cells with abnormal shape.

4.2 Regional Descriptors

Regional descriptors are the most used, since they provide an overall characterization of the object in exam. Regional descriptors comprise shape descriptors, colour descriptors and texture descriptors.

4.2.1 Geometric Descriptors

Geometric descriptors are the most used for peripheral blood cells analysis, since cells differ considerably in size or shape, thus using these descriptors it is possible to discriminate them easily. The simplest geometrical features are area and perimeter, from which it is possible to compute other more complex descriptors. The *area* of a region is defined as the number of pixels that constitute the region. This descriptor can be useful if the visual geometry is fixed and the objects are always

analysed approximately at the same distance. Often it is also used the value of the *convex area*, which is the area of the *convex hull*, the minimum convex polygon that completely encloses the region. The *perimeter* of a region is defined as the number of pixels of its outline. Even in this case the value of the *convex perimeter* can be used, although it is not often used as descriptor, but its most common application is the computation of other descriptors, such as the compactness, the circularity and the convexity. The *compactness* of a region is defined as the ratio between the area of an object and the area of a circle with the same perimeter (4.4). The circle is used as a benchmark because it is the most compact form, in fact its value of compactness is 1. Also the *roundness* calculates the ratio between area and perimeter, however, it excludes the presence of small irregularities, in fact, it is calculated from the ratio between the area of the region and that of a circle with the same convex perimeter (4.5). The *convexity* instead expresses the relative amount that an object differs from a convex object. This value is obtained through the ratio between the convex perimeter and the perimeter of the object itself (4.6). The value of *solidity* instead describes the density of an object by comparing the area of an object and the area of its convex hull (4.7).

$$compactness = \frac{4 * \Pi * area}{perimeter^2} \tag{4.4}$$

$$roundness = \frac{4 * \pi * area}{convex\_perimeter^2} \tag{4.5}$$

$$convexity = \frac{perimeter_{convex}}{perimeter} \tag{4.6}$$

$$solidity = \frac{area}{convex\_area} \tag{4.7}$$

All the descriptors just mentioned, compactness, circularity, convexity and solidity, have a maximum value equal to 1, indicating that the region is compact, circular, convex and solid, respectively. The main drawback of the geometric descriptors is that their application requires accurate segmentation of the region of interest and, therefore, they are commonly used with other descriptors less influenced by segmentation errors, such as chromatic descriptors or texture descriptors.

4.2.2 Chromatic Descriptors

Chromatic descriptors delineate the grey level or colour distribution of images. These descriptors are calculated directly from histograms of the region, which may be considered as functions of colour density. The most used descriptors are the *mean* 4.8, *standard deviation* (4.9), *smoothness* (4.10), *skewness* (4.11), *kurtosis* (4.12), *uniformity* (4.13) and *entropy* (4.14), that describe the shape of the normalised histogram h_N, obtained from the histogram h by dividing each histogram value by the total number of pixels.

$$\mu = \sum_{k=0}^{N_g-1} k \cdot h_N(k) \tag{4.8}$$

$$\sigma = \sqrt{v} \tag{4.9}$$

$$s = \frac{1}{1 + v/(N_g - 1)^2} \tag{4.10}$$

$$\mu_3 = \sigma^{-3} \cdot \sum_{k=0}^{N_g-1} (k - \mu)^3 \cdot h_N(k) \tag{4.11}$$

$$\mu_4 = \sigma^{-4} \cdot \sum_{k=0}^{N_g-1} (k - \mu)^4 \cdot h_N(k) \tag{4.12}$$

$$uni = \sum_{k=0}^{N_g-1} h_N^2(k) \tag{4.13}$$

$$e = -\sum_{k=0}^{N_g-1} h_N(k) \cdot log_2(h_N(k)) \tag{4.14}$$

where v is the *variance* value (4.15).

$$v = \sum_{k=0}^{N_g-1} (k - \mu)^2 \cdot h_N(k) \tag{4.15}$$

The chromatic descriptors are the most discriminatory characteristics between different types of tissues and cells, but generally the discrimination on sub-classes requires further descriptors such as texture measures.

4.2.3 Texture Descriptors

The traditional machine vision and image processing approaches assume the presence of uniform intensity values in local regions of the image. This assumption is not always true, in fact some objects have a repeated pattern as the main visual feature, which is called texture. Texture probably represents the most used descriptor for the description of the regions of images. Although there are no formal definitions of textures, it can be viewed as a global descriptor generated from the repetition of local patterns. Texture is an any and repetitive geometric arrangement of the grey levels of an image. It provides important information about the spatial disposition of the grey levels and the relationship with their neighbourhoods. Human visual system determines and recognizes easily different types of textures but although for

a human observer it is very simple to associate a surface with a texture, to give a rigorous definition for this is very complex. Typically it is used a qualitative definition to describe textures. It can easily guess that the quantitative analysis of texture passes through statistical and structural relations among the basic elements of what we call just texture. Intuitively texture descriptors provide measures of properties such as regularity, smoothness, roughness, coarseness, thickness, etc. In medical image analysis texture descriptor has proven itself useful for distinguish some abnormal cells or the presence of parasites in the process of evolution. The most used approach for the description of the texture is the statistical approach, that is also the simplest for texture representation. There are many statistical descriptors that use statistical moments extracted from the histogram of the image or the region. The measures of texture based only on histograms, however, have many drawbacks. In particular statistical moments do not give information about the mutual position of the pixels. Thus, it is important to consider not only the intensity distribution but also the positions of pixels having similar grey level. Many different methods for managing textures have been developed that are based on the various ways texture can be characterized, including the scale-invariant feature transform (SIFT) [Low04], speeded up robust feature (SURF)[BTG06], histogram of oriented gradients (HOG) [DT05], Gabor filters [JF90] and others.

Gray Level Co-occurrence Matrix

One of the most powerful model for texture analysis was proposed by Haralick [HSD73]. His method involves the creation of the Gray Level Co-occurrence Matrices (GLCMs) from which features that represent some image aspects can be calculated. A GLCM represents the probability of finding two pixels i and j with distance d and orientation θ and it is denoted with $p_{d,\theta}(i,j)$. Obviously, the d and θ values can assume different values, but the most used are $d = 1$ and $\theta = [0°, 45°, 90°, 135°]$. A GLCM for an image of size $N \times M$ with N_g grey levels is a 2D array of size $N_g \times N_g$. Haralick proposed thirteen descriptors that can be extracted from these matrices.
Angular Second Moment: is the squares sum of the matrix values (4.16). It is also known as Uniformity. This feature has a range between 0 and 1. The value is 0 if the image is constant.

$$ASM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j)^2 \tag{4.16}$$

Often it is called also *Energy* but it is calculated as in (4.17).

$$Ene = \sqrt{\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j)^2} \tag{4.17}$$

Contrast: is the weighted average of all diagonals parallel to the main one which emphasizes the correlation between the different tones (4.18). The contrast is 0 if

the image is constant.

$$Con = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-j)^2 \cdot p(i,j) \tag{4.18}$$

Correlation: is the measure of how a pixel is in correlation with its neighbours across the image. It is 1 or -1 for an image related perfectly positively or negatively. It is 0 if the image is constant (4.19).

$$Cor = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \frac{(i-\mu_x) \cdot (j-\mu_y) \cdot p(i,j)}{\sigma_x \cdot \sigma_y} \tag{4.19}$$

Variance: is the measure of linear dependence of the brightness determined from the correlation (4.20).

$$Var = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-\mu)^2 \cdot p(i,j) \tag{4.20}$$

Inverse Difference Moment: is a value which measures the proximity of the distribution from GLCM elements to the GLCM diagonal. It has value 1 in the main diagonal of its GLCM (4.21).

$$IDM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \frac{p(i,j)}{1+(i-j)^2} \tag{4.21}$$

Often it is called also *Homogeneity* and is calculated as in (4.22).

$$Hom = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \frac{p(i,j)}{1+|i-j|} \tag{4.22}$$

Sum Average: is the average of the value p_{x+y} containing the sums of all the diagonal orthogonal to the main (4.23).

$$SAve = \sum_{k=0}^{2N_g-2} k \cdot p_{x+y}(k) \tag{4.23}$$

Sum Variance: is an estimation of the second order of the vector p_{x+y} centralized respect to the average (4.24).

$$SVar = \sum_{k=0}^{2N_g-2} (k - F_{SAv})^2 \cdot p_{x+y}(k) \tag{4.24}$$

Sum Entropy: provides an estimate of the vector p_{x+y} relative to entropy, which is the measure of the disorder of the vector itself (4.25).

$$SEnt = -\sum_{k=0}^{2N_g-2} p_{x+y}(k) \cdot log_2(p_{x+y}(k)) \tag{4.25}$$

Entropy: is the entropy measure for the entire matrix (4.26).

$$Ent = -\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} p(i,j) \cdot log_2(p(i,j)) \tag{4.26}$$

Difference Variance: is the variance of the vector p_{x-y} (4.27).

$$DVar = \sum_{k=0}^{N_g-1} (k - F_{DAv})^2 \cdot p_{x-y}(k) \tag{4.27}$$

where DAve, *Difference Average* or *Dissimilarity*, is the average of the vector p_{x-y} containing the differences of all the diagonal orthogonal to the main (4.45).
Difference Entropy: is the entropy measure of the vector p_{x-y} (4.28).

$$DEnt = -\sum_{k=0}^{N_g-1} p_{x-y}(k) \cdot log_2(p_{x-y}(k)) \tag{4.28}$$

Measure of correlation 1 and 2: are measures related to entropy of the matrix (4.29),(4.30).

$$MC1 = \frac{F_{Ent} - HXY1}{max(Hx - Hy)} \tag{4.29}$$

$$MC2 = \sqrt{1 - exp[-2(HXY2 - F_{Ent})]} \tag{4.30}$$

where

$$p_x(i) = \sum_{j=0}^{N_g-1} p(i,j) \tag{4.31}$$

$$p_y(j) = \sum_{i=0}^{N_g-1} p(i,j) \tag{4.32}$$

$$p_{x-y}(i-j) = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} p(i,j) \tag{4.33}$$

$$p_{x+y}(i+j) = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} p(i,j) \tag{4.34}$$

$$\mu_x = \sum_{i=0}^{N_g-1} i \cdot p_x(i) \tag{4.35}$$

$$\mu_y = \sum_{j=0}^{N_g-1} j \cdot p_y(j) \tag{4.36}$$

$$\mu = (\mu_x + \mu_y)/2 \tag{4.37}$$

$$\sigma_x = \sqrt{\sum_{i=0}^{N_g-1} p_x(i) \cdot (i - \mu_x)^2} \tag{4.38}$$

$$\sigma_y = \sqrt{\sum_{j=0}^{N_g-1} p_y(j) \cdot (j - \mu_y)^2} \tag{4.39}$$

$$HX = -\sum_{i=0}^{N_g-1} p_x(i) \cdot log_2(p_x(i)) \tag{4.40}$$

$$HY = -\sum_{j=0}^{N_g-1} p_y(j) \cdot log_2(p_y(j)) \tag{4.41}$$

$$HXY1 = -\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j) \cdot log_2(p_x(i) \cdot p_y(j)) \tag{4.42}$$

$$HXY2 = -\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p_x(i) \cdot p_y(j) \cdot log_2(p_x(i) \cdot p_y(j)) \tag{4.43}$$

To these descriptors extracted from GLCMs many others have been proposed, but only seven are widely used [ST99, Cla02], that are *mean* (4.44), *difference average* (4.45), *autocorrelation* (4.46), *maximum probability* (4.47), *cluster shade* (4.48), *cluster prominence* (4.49) and *product moment* (4.50).

$$\mu = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} i \cdot p(i,j) \tag{4.44}$$

$$DAve = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} |i-j| \cdot p(i,j) = \sum_{k=0}^{N_g-1} k \cdot p_{x-y}(k) \tag{4.45}$$

$$Aut = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} i \cdot j \cdot p(i,j) \tag{4.46}$$

$$MP = max(p(i,j)) \tag{4.47}$$

$$CS = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i + j - \mu_x - \mu_y)^3 \cdot p(i,j) \tag{4.48}$$

$$CP = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i + j - \mu_x - \mu_y)^4 \cdot p(i,j) \tag{4.49}$$

$$PM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i - \mu_x) \cdot (j - \mu_y) \cdot p(i,j) \tag{4.50}$$

Some interesting methods have also been presented in order to extend the original implementation of GLCM, computing the matrices by evaluating different distance parameters [GVB07], different windows sizes [HCx09], different colour channels [BH07], adding the colour gradient [GW12] or considering also the edge orientation [MMN+12]. Furthermore the GLCM descriptors can be extracted after computing a weighted sum of GLCM elements [WJL03] or after computing the local gradient of the matrix [CWCT09].

Gray Level Difference Matrix

Another useful tool for texture analysis is the grey level difference matrix (GLDM)[CH80], that is a particular type of matrix originated by the absolute differences between pairs of grey levels. Actually, the GLDM is defined in a manner very similar to the GLCM, using the same notions of distance and orientation to find the pairs of grey levels. The main difference arises in the construction and dimension of the matrix. In fact, the GLDM preserves the size of the original image (and not $N_g \times N_g$), collecting the absolute difference between pairs of pixel values (and not the occurrences of two grey levels). This matrix is used to calculate the histogram $h(d)$ that denotes the number of differences with value d. The histogram is then normalized $h_N(d) = h(d)/N$ with $N = \sum_d h(d)$ in order to compute easily nine descriptors that are *mean* (4.51), *angular second moment* (4.52), *contrast* (4.53), *variance* (4.54), *inverse difference moment* (4.55), *entropy* (4.56), *product moment* (4.57), *cluster shade* (4.58) and *cluster prominence* (4.59).
Mean:

$$\mu = \sum_{d=0}^{N_g-1} d \cdot h_N(d) \tag{4.51}$$

Angular Second Moment:

$$ASM = \sum_{d=0}^{N_g-1} h_N(d)^2 \tag{4.52}$$

Contrast:

$$Con = \sum_{d=0}^{N_g-1} d^2 \cdot h_N(d) \tag{4.53}$$

Variance:

$$Var = \sum_{d=0}^{N_g-1} (d-\mu)^2 \cdot h_N(d) \tag{4.54}$$

Inverse Difference Moment:

$$IDM = \sum_{d=0}^{N_g-1} \frac{h_N(d)}{1+d^2} \tag{4.55}$$

Entropy:

$$Ent = \sum_{d=0}^{N_g-1} h_N(d) \cdot log_2(h_N(d)) \tag{4.56}$$

Product Moment:

$$PM = \sum_{d=0}^{N_g-1} (d-\mu) \cdot h_N(d) \tag{4.57}$$

Cluster Shade:

$$CS = \sum_{d=0}^{N_g-1} (d-\mu)^3 \cdot h_N(d) \tag{4.58}$$

Cluster Prominence:

$$CP = \sum_{d=0}^{N_g-1} (d-\mu)^4 \cdot h_N(d) \tag{4.59}$$

Gray Level Run-Length Matrix

A different tool for texture analysis is based on information of higher order statistics that uses the Grey Level Run-Length Matrices (GLRLM) [Tan98]. In this approach the GLRLM contains information on a particular number of equal grey levels (run) in a given direction. So, a run-length matrix is defined as a set of consecutive pixels having the same grey level. The element (i, j) of a run-length matrix specifies the number of times that the image contains a run of length j composed by all pixels with grey level i. The creation of the run-length matrices is very simple and the number of operations to be done is directly proportional to the number of image points. A coarse texture will be characterized by a long run while a finer texture will be characterized by shorter run. Also, the GLRLMs are calculated by considering the main four orientations and for each matrix eleven descriptors can be extracted, that are *short run emphasis* (4.60), *long run emphasis* (4.61), *grey level non-uniformity* (4.62), *run length non-uniformity* (4.63), *run percentage* (4.64), *low*

grey level run emphasis (4.65), *high grey level run emphasis* (4.66), *short run low grey level emphasis* (4.67), *short run high grey level emphasis* (4.68), *long run low grey level emphasis* (4.69) and *long run high grey level emphasis* (4.70).
Short Run Emphasis:

$$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2} = \frac{1}{n_r} \sum_{j=1}^{N} \frac{p_r(j)}{j^2} \tag{4.60}$$

Long Run Emphasis:

$$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot j^2 = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j) \cdot j^2 \tag{4.61}$$

Grey Level Non-uniformity

$$GLN = \frac{1}{n_r} \sum_{i=1}^{M} \left(\sum_{j=1}^{N} p(i,j) \right)^2 = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i)^2 \tag{4.62}$$

Run Length Non-uniformity

$$RLN = \frac{1}{n_r} \sum_{j=1}^{N} \left(\sum_{i=1}^{M} p(i,j) \right)^2 = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j)^2 \tag{4.63}$$

Run Percentage

$$RP = \frac{n_r}{n_p} \tag{4.64}$$

Low Grey Level Run Emphasis

$$LGLRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2} = \frac{1}{n_r} \sum_{i=1}^{M} \frac{p_g(i)}{i^2} \tag{4.65}$$

High Grey Level Run Emphasis

$$HGLRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i) \cdot i^2 \tag{4.66}$$

Short Run Low Grey Level Emphasis

$$SRLGLE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 \cdot j^2} \tag{4.67}$$

Short Run High Grey Level Emphasis

$$SRHGLE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot i^2}{j^2} \tag{4.68}$$

Long Run Low Grey Level Emphasis

$$LRLGLE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot j^2}{i^2} \tag{4.69}$$

Long Run High Grey Level Emphasis.

$$LRHGLE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 \cdot j^2 \tag{4.70}$$

Local Binary Pattern

Another useful tool for texture analysis is the Local Binary Pattern (LBP), originally proposed in [OPH96] and widely used for grey level texture classification, due to its simplicity and robustness. This operator transforms the image by thresholding the neighbourhood of each pixel and by coding the result as a binary number. The resulting image histogram can be used as a feature vector for texture classification. Also, for the LBP operator two main parameters must be defined, which are the radius r and the number of neighbourhood n pixels. For example, some possible versions of this operator are the $LBP_{8,1}$ implemented with the parameters r and n equal to 1 and 8, respectively, and the $LBP_{16,2}$ implemented with the parameters r and n equal to 2 and 16, respectively. These two LBP operators are reported in Fig.4.1.

4.3 Feature Selection

A further step, not always present on CAD system, is the feature selection, a process commonly used in pattern recognition that allows to determine the most relevant features reducing the size of the vectors associated with the objects. The feature selection aims to reduce the dimensionality eliminating both the redundant features, that represent information derived from other, both the features that are irrelevant for the analysis. The "ideal" approach would be to test all the possible sub-sets of features, using them as input to the classification algorithm of interest and select the sub-set that allows to obtain the best results. Obviously this approach in most cases is not applicable. There are several techniques for the selection of characteristics, that can be grouped into three categories:

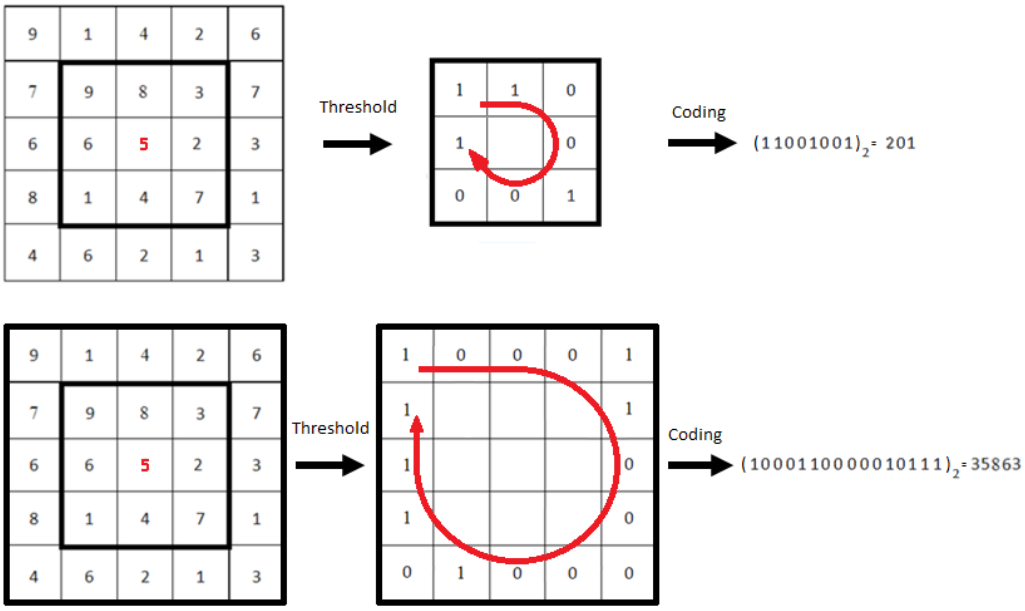

Figure 4.1: The LBP operators $LBP_{8,1}$ and $LBP_{16,2}$, respectively

- embedded methods: the selection is internal to the classification algorithm that takes advantage of internal knowledge of the classifier, such as the weight used to induce the model [DH73]
- filter methods: also known as scheme-independent selection, because the selection is made in advance, with method independent from the classification algorithm that will be applied subsequently using some measure of distance or correlation [YL04]
- wrapper methods: also known as scheme-specific selection, because the selection is performed making a comparison between different sub-sets of features, investigated with approaches of sequential forward or backward selection [Kit], according to the classification algorithm that will be applied later

Generally, the wrapper methods perform better than the other methods, as they are optimized for a specific classifier, but they are computationally eligible only for small feature vectors. On the other hand, there are other approaches called by many authors "feature selection" that do not perform properly a selection of features, but perform a dimensionality reduction through projection or combination. The Principal Component Analysis (PCA) [WEG87] is the most popular technique for the reduction of dimensionality. The purpose of the PCA is to find a set of orthogonal vectors in the feature space corresponding to the directions along which the data have the highest variance. The dimensionality reduction is performed by projecting the data from their original space to the orthogonal complement. The advantages of the PCA is that it can deal with large datasets both in objects and

variables reducing the redundancy. Moreover it doesn't make special assumptions on the data and so it can be applied on all datasets. The biggest disadvantage arises from the fact that PCA does not select the feature but it creates new ones combining the original. This greatly affects the control over the single feature, that is particularly important in image processing where features are extracted directly from the image and thus it is important to establish which are determinant for the task of classification, in order to avoid the extraction of insignificant features.

Chapter 5

Classification

Once the features have been extracted from cells or tissue they must be inserted in a process that classifies cells based on medical concepts. Given a collection of records, each one composed by a set of features x and by a label of class y, the goal is to define a function or classification model, that associates a class label y to each set of attributes x. A classification model is a tool to describe and classify the data of a specific domain. This is possible thanks to a training set, namely a set of training samples in which the values of the label of the class are known. So, through a learning algorithm the relations between attributes and class labels can be identified and encoded in a model. This model must be able not only to describe the training set, but also to correctly predict the class of new records not labelled. In literature several classification algorithms are present, but the most used on medical images are the following:

- Nearest Neighbour
- Decision Trees
- Bayesian Classifier
- Neural Network
- Support Vector Machine

5.1 Nearest Neighbour

The nearest neighbour classifier uses the concept of proximity to classify a new record, based on the examples of the training set that are similar to it. Each instance of the training set is represented as a point in a n-dimensional space, where n is the number of features. When a new record must be classified, its distance from each sample of the training set is calculated. Then, the k examples of the training set closer to the new record, called k-Nearest Neighbors (kNN) [CH67], are identified

and used to assign to the new record the class label prevailing among the kNN. There may be, however, problems in the choice of k, in fact if this value is too small, there is a high sensitivity to noise, but if it is too large, among the kNN there may be examples not similar enough to the record to be classified. One way to reduce the influence of the parameter k is to calculate the prevailing class by assigning a different weight to each of the first neighbours according to its distance from the record to be classified. This type of classifier is the simplest among those listed previously, as it does not require the induction of a model from the training set, which is used in the classification step to compare the new records with the known ones. In contrast to the saved resources for the construction of the model, the classification of a new record, however, is rather expensive. Indeed the proximity between the new record and the known examples of the training set must be calculated every time.

5.2 Decision Trees

Decision Trees [Qui86] are decision support tools that uses a tree-like graph or model of decision for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Thus during the classification of a new record the decision tree represents a flowchart-like structure in which each internal node represents a test on an attribute and each branch represents the outcome of the test. Obviously each leaf node represents a class label and the decision taken after computing all attributes. The paths from the root to the leaf nodes represents the classification rules. The greatest advantage of decision trees is that they are simple to understand and interpret, but they can be really complex with an high number of attributes, in particular the more attributes, the deeper the tree, the more complex the decision rules and the fitter the model. One disadvantage is that decision tree can create too complex trees that do not generalise the data well, generating overfitting, but the greatest disadvantage is that the same classification rules can be expressed with different decision trees. Thus finding the optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. This problem is generally mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.

5.3 Bayesian Classifier

The Bayesian classifiers are based on probabilistic relations between the class labels and feature values of the record. Considering the features of a record as of random variables a Bayesian network can be used to plot the conditional dependencies among a set of random variables. It is an acyclic and directed graph in which the nodes represent random variables and the arcs represent dependency relationships

between the variables. Each node of the network is associated with a probability table containing the a priori probability if that node does not depend to any other node, or the conditional probability if the node depends on a set of other nodes. Thus given a training sample it must be found the Bayesian network that best describes the conditional dependencies between variables. Once defined the network, the process of classification of a new record entails the calculation of the posterior probability for each class and the selection of the class for which the probability is the highest. The induction of the network that best describes a given training set involves the definition of network structure and the estimation of the values of the probabilities table associated with each node of the network. In a general case this is an intractable problem but there are algorithms that induce Bayesian classification models introducing appropriate simplifying assumptions on the network topology. The simplest among the Bayesian classifiers is the so-called Naive Bayes [DH73], [LIT92] and is based on the assumption that the features are conditionally independent, given the value of the class. In this way the a priori probability of the class and the conditional probabilities of the features of the class can be easily estimated from the training sample. With the Naive Bayes method the model is not influenced significantly by either the noise, which is mediated through the calculation of probabilities, or by any irrelevant features, for which the probabilities are distributed in an almost uniform way. Despite its simplicity, the Bayesian classifier is able to perform accurate classifications but only if the features are discriminatory.

5.4 Artificial Neural Network

Artificial Neural Networks (ANN) classifiers are the most used in medical applications. The ANNs are networks that emulate the behaviour of the human brain, composed of a set of nodes that are interconnected by links to which is associated a weight. In ANNs as in biological systems, learning corresponds to change the weight values of the connections between nodes. Given a training sample, the weights of the model are first initialised randomly and then iteratively adjusted, so that the output of the model appears consistent with the values of the label class. The simplest ANNs are composed of two levels, the input level and the output level. The input layer contains a node for each numerical features, while the categorical features, require more nodes, for example a feature with n possible values can be transformed into n binary variables. The output layer instead can contain only one node, if the problem of classification is binary and k nodes if the class label can assumes k values. There are also multilevel ANNs whose structures present additional hidden layers between the input layer and the output layer. The number hidden layers is often determined by trial and error, in fact typically the correct number of hidden layers is found starting from a network with an high number of layers and nodes and progressively decreases the complexity of the model. The training of the network involves the adjustment of the weights with the aim to minimize the internal error.

This process is really expensive, especially if the network topology is complex, even if the classification process is rapid. The model is very sensitive to noise, because the weights are adjusted for each instance of the training sample. On the contrary, the irrelevant or redundant features do not affect significantly the model, as the corresponding weights are typically very small.

5.5 Support Vector Machine

Recently the Support Vector Machine (SVM) has received a growing interest in the field of pattern recognition. This technique has been designed for binary classification problems, so with only two classes, but can also be extended to multi-class problems.The SVM binary classification is based on the mapping of input vectors in an high dimensionality features space, induced by a kernel function. The learning algorithm produces an optimal hyperplane of separation between the two classes. The SVM can perform a linear discrimination, but it can also perform a discrimination not linear due to the use of a kernel function. It is possible to find a kernel function for which the parameters of the model can be induced without an explicit mapping data. The induction of the model in this way is formulated as an optimization problem, in which it is possible to find quite efficiently a global minimum for the objective function. The SVM provides extreme flexibility both because it is possible to make use of different types of kernels and both because it is possible to define a hyperplane for separating classes which guarantees a certain tolerance with respect to noise, using a soft margin instead of an hard margin. This for an optimization problem becomes changing the constraint value with the c value, that can be increased to obtain less training error. Obviously kernel methods use other parameters for the creation of the separating hyperplane. The Gaussian Radial Basis Funcion (RBF) is one of the most used kernel performing a non-linear separation defined by the radius γ of the RBF. Other kernels are the quadratic kernel, the polynomial kernel that can be defined with different order p and the Multilayer Perceptron kernel (MLP) that instead can be defined for different slope α and the intercept constant β. The multi-class problem is solved by building many different binary classifiers and then combine them. The most used strategies are the combinations one-vs-one and one-vs-all.

5.5.1 One-vs-all SVM

The one-vs-all, also known as one-vs-rest, approach is the first and the most intuitive approach to extend the SVM to multi-class problems. The basic idea of the one-vs-all approach is very simple. In fact, for a multi-class problem having m classes, it consists on training m different binary SVM classifiers where each one of them separates one class from all the other $m-1$ classes. Then, in testing phase the class label is assigned taking into account the decision of each m classifiers. The most

common practice is to assign the class label i, where i is the classifier that maximizes the separation between the class i from the rest. Another common practice is to use binary trees to arrange the $m-1$ binary SVM; the path from the root node to a leaf determines the class label. In the best scenario only one comparison is needed, while in the worst case $m-1$ comparisons are needed. The problem encountered with binary tree SVM is that there are $\prod_{i=3}^{m} 2i-3$ possible ways to construct a tree for a multi-class problem. Thus, for a multi-class problem with a very large value of m, it's impossible to analyse all the possible solutions.

5.5.2 One-vs-one SVM

The one-vs-one approach is another commonly used SVM extension but, differently from the one-vs-all approach, for a multi-class problem having m classes, it consists on training $\frac{m(m-1)}{2}$ different binary SVM classifiers, where each one of them separates one class from another one. In this case the combination of the results from individual binary SVM classifiers becomes more complex. The most simplest way to obtain the predicted class label is to use the majority voting, counting the votes given by each binary classifier and assigning the class label that has the highest number of votes. However, voting process could produce ambiguous results (eg. tie cases). A common practice is to use the majority voting strategy combined with the maximum separation strategy, assigning the class label i, if the result of the binary classifier produces the highest number of i votes with the maximum separation between i class and each of the other classes. Another common practice consists in arranging all the $\frac{m(m-1)}{2}$ binary SVM classifiers in a Directed Acyclic Graph (DAG) structure with the same number of nodes. The test phase starts at the root node and continues until a leaf node, that represents the predicted class label, is reached. Thus, only $m-1$ comparisons are needed, avoiding completely tie cases, but the problem is encountered just during the construction of the DAG. In fact, also in this case, there are several ways to construct the DAG structure and each one of them may produce different classification results and when the number of classes is high it is impossible to test all possible orders.

5.6 Model Evaluation

The performance of the classification models are then evaluated on the basis of percentage of records correctly classified on a test set with a known class label. Therefore by comparing the class labels known and the labels predicted by the classifier it is possible to calculate the *accuracy* and the *error rate* of the model. For a binary problem in which a class considered positive and the other class negative we can consider the True Positive (TP) measure that indicates the number of positives records correctly classified positives, the True Negative (TN) measure that indicates the number of negatives records correctly classified, the False Positive (FP) measure

that is the number of negative records misclassified as positive and False Negative (FN) measure that is the number of positive records misclassified as negative. In this way the accuracy of the error rate can be written as in (5.1) and (5.2):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$errorrate = \frac{FP + FN}{TP + TN + FP + FN} \tag{5.2}$$

For classification problems in which there are uncommon classes that are often the most interesting, the accuracy and the error rate are not good measures of model performance. In this case the most used measures are the *True Positive Rate* (TPR) also called *sensitivity* or *recall* (r) (5.3), the *True Negative Rate* (TNR) also called *specificity* (5.4), the *False Positive Rate* (FPR) (5.5), the *False Negative Rate* (FNR) (5.6) and the *precision* (p) (5.7). Precision and recall are used when it is considered more important or interesting the correct classification of positive instances. Indeed a good model should be able to maximise both these measures. For this reason an other important metric is the *F-measure* or *F-score* (5.8).

$$TPR = r = \frac{TP}{TP + FN} \tag{5.3}$$

$$TNR = \frac{TN}{TN + FP} \tag{5.4}$$

$$FPR = \frac{FP}{TN + FP} \tag{5.5}$$

$$FNR = \frac{FN}{TP + FN} \tag{5.6}$$

$$p = \frac{TP}{TP + FP} \tag{5.7}$$

$$F\text{-}score = \frac{2rp}{r + p} \tag{5.8}$$

Note that the same measures are also used for the evaluation of segmentation, in fact, having some manually segmented images or ground-truth, a pixel-wise evaluation can be made to asses if a pixel has been correctly included in the region TP, it has been correctly excluded from the region TN, it has been erroneously included in the region FP or it has been erroneously excluded from the region FN.

The performance of a model may not depend only on the type of classification algorithm but also by other factors such as the size or distribution of the classes in the training and test set. In particular, if the dataset has reduced size the performances are more related to the specific composition of the samples and they are characterized

by a higher variance. Some methods useful to extract from the original dataset a representative test set able to assess the performance of the model are:

- Holdout: a part of the available samples, the training set, is used to train the model while another part, the test set, is used for its evaluation. This involves a decrease of examples available for training and an addiction due from specific partition created. To ensure that the training set and the test set are equally representative the partition can be made by using a process stratified sampling.

- Repeated Holdout: the holdout method is iterated k times, in order to avoid the control on the number of times that each record is used for training and testing. The accuracy and error rate are calculated averaging the k iteration results.

- Cross-Validation: the examples available are divided into k sub-sets of equal dimension. The process of training and evaluation of the model is repeated k times, each time using $k-1$ different sub-sets for training and one sub-set for the test. As for Repeated Holdout, the accuracy and error rate are calculated averaging the k iteration results, but in this case the final results is more stable since the test sets are mutually exclusive and cover the entire initial sample.

- Leave-one-out: special case of cross-validation in which k is equal to the number of records.Thus during the training phase the largest possible number of examples is used and each test set contains only one record. This approach provide an exhaustive results but it is computationally very expensive.

Part II

CAD for Histology Images

Chapter 6

Background

6.1 Histology

Histology is the study of the microscopic structure of cells and tissues of organisms. The knowledge of biological microscopic structures and their functions at the sub-cellular, cellular, tissue and organ levels is essential for the study of grading and prognosis of disease. The biopsy sample is processed and its sections are spread on slides and observed under a microscope for analysis. Pathologists examine the tissue under a microscope slide observing it at various levels of magnification, such as 10X, 20X, 40X, 100X, etc. to see the cells, glands, heart, and identify the morphological characteristics of the tissue that indicate the presence of diseases, such as cancer [BM12]. But this diagnosis of the pathologist is subjective. In addition, about 80% of the biopsies of 1 million samples are found to be benign in the US every year; pathologists are spending a lot of time in search of the diseased tissue that otherwise could be used for the treatment of patients [BMN+14]. Then, a quantitative evaluation of these images is really essential for an objective diagnosis. In the last ten years, many CAD systems have been developed to help histological cancer diagnosis and to reduce subjectivity. All these systems attempt to mimic pathologists by extracting features from histological images [KPYW13]. Image analysis involves complex algorithms which identify and characterize cellular colour, shape and quantity of the tissue sample using image pattern recognition technology. Such systems have been widely employed for applications in cancer detection and grading [HLAT12], including prostate [MTF+09, NDF+07], breast [DAM+08, NDA+08], cervix [GASM+05, KDM+00], lung [KRW+, SAGG06] and follicular lymphoma grading [CSK+09, SKL+08]. These different applications share similar computer techniques to support clinicians in automatic extraction of histological image features and classification, as in radiology and cytology image analysis. However, the use of these systems in the histology analysis is not simple because of the different imaging techniques and the characteristics of the image itself. In fact, histological images differ from radiological images since they have lots of objects of

interest (cells and cell structures, such as nuclei) widely distributed and surrounded by various types of tissue. On the contrary, the analysis of radiology images usually focuses on some organs in the image. A histology image usually has a size much larger than that of a radiology image. In addition, histological tissues are generally stained with different colours while X-ray images usually contain only grey intensities. Other aspects make more difficult histological images, including the types of overlapped tissues and boundaries of cells and nuclei corrupted by noise; some structures, such as the boundaries of the cells, can appear blurred or connected. The computer-based systems for histological analysis generally utilize an amount of image characteristics to obtain clinically significant information much greater than similar systems for radiology and cytology. A typical CAD system for histology image analysis consists of conventional image processing and analysis tools, including preprocessing, image segmentation, feature extraction, feature dimension reduction, feature-based classification and post-processing. Image feature extraction and selection are crucial for many image processing and computer vision applications, such as image retrieval, registration and matching, and pattern recognition. For a CAD system, traditional features include morphometrics with object size and shape (e.g. compactness and regularities), topological or graph-based features (e.g. Voronoi diagrams, Delaunay triangulation and minimum spanning trees), intensity and colour features (e.g. statistics in different colour spaces) and texture features (e.g. Haralick entropy, Gabor filter, power spectrum, co-occurrence matrices, and wavelets). In addition, besides using the image in the spatial domain, many features can also be extracted from other transformed spaces, e.g. frequency (Fourier) domain and wavelet transforms. Despite the progress made in this area thus far, this is still a large area of open research due to the variety of imaging methods and disease-specific characteristics [GBC+09]. Due to the importance of histopathology images in cancer diagnosis and grading, benign and malignant cancer differentiation and cancer cell classification, the main research progress in this area lies in developing computer based automatic histology image classification systems targeting at specific clinical fields [MLSC10]. Mainly, the research direction focuses on extracting specific features of certain data sets to represent the characteristics of the objects better in specific clinical fields. As consequence, though the aforementioned applications provide acceptable results for some data sets, they are clinical domain dependent. Developing general applications not dependant on specific histology data sets is still a challenge open problem. The main effort was to develop a general classification system for histology images. A colour texture based histology image classification framework is proposed and tested using five very different public biological image data sets. In this system, no object detection or segmentation method is needed and a wide range of features including colour and texture features are used in combination to avoid the dataset dependency. Another aspect that is considered of great importance for analysing histology images is colour. Indeed, the biopsy samples are usually prepared with some chemical solution that enhances the contrast and stains in very distinctive colours specific parts of the cells or tissue that have been sam-

pled. Common staining processes include H&E (hematoxylin and eosin) or Massons trichrome. As consequence, colour based features are important in histology applications since the biologists stain tissues to highlight special structures [BMN+14]. But there are many colour spaces used to apply in analysis of histology images. So, another challenging task consists in analysing histology images in different colour spaces in order to individuate a general representation able to solve the classification problem efficiently without any dependence on specific image dataset or specific clinical field. Summarizing, the main goal was to build a broad and general histology image classification system for assisting the pathologists in any clinical field for disease diagnosis by extracting significant features from the image representation that contains the disease-independent characteristics of the objects better and more discriminant for classification.

6.2 Related Works

There is an extensive literature in automatic histology image analysis. In [NBG+13] various types of descriptors have been tested for different image classification problem. In detail, the classical co-occurrence matrix descriptors proposed by Haralick have been extended to 3D co-occurrence matrix and the co-occurrence matrix projected onto a lower dimensional subspace by principal component analysis. These descriptors have been extracted by computing co-occurrence matrices using a multiscale approach. Moreover, the features are extracted not only from the entire co-occurrence matrix but also from sub-windows. In [MLSC10] a framework based on the novel and robust Collateral Representative Subspace Projection Modelling (C-RSPM) has been used for histology image classification. The cell image is first divided into 25 blocks to reduce the spatial complexity of computation. For each image block a C-RSPM has been built, in order to classify firstly each block separately and then, applying a fusion algorithm with a weighted majority voting strategy, it's possible to decide the final class label of the whole image. In [KPYW13] the authors examined the utility of biologically interpretable shape-based features for classification of histological renal tumour images. They proposed the use of Fourier shape descriptors to capture the distribution of stain-enhanced cellular and tissue structures in each image. They also identified the most informative shape-features for each renal tumour subtype, that are not only accurate diagnostic features, but also correlated with known biological characteristics of renal tumours. In [MTF+09] the authors compared the efficacy of the so-called probabilistic pairwise Markov models to the prevalent Potts model by incorporating both into a novel CAD system for detecting cancer on whole-mount histological sections of radical prostatectomies. In [NDF+07] an automatic method for detecting and segmenting glands in digitized images of prostate histology has been proposed. A Bayesian classifier is used to detect candidate gland regions by utilizing low-level image features to find the lumen, epithelial cell cytoplasm, and epithelial nuclei of the tissue. Then, the features

derived from gland morphology have been used to train a support vector machine in order to describe prostate cancer malignancy. In [DAM+08] a novel image analysis methodology for automatically distinguishing low and high grades of breast cancer from digitized histopathology has been presented. A set of over 3.400 features, including textural and nuclear structure based, are extracted from the breast biopsy tissue images. Spectral clustering is used to reduce the dimensionality of the feature set and the support vector machine is used to distinguish between non-cancer, low and high grades of cancer. In [NDA+08] a methodology for automated detection and segmentation of structures of interest in digitized histopathology images has been proposed. The scheme integrates low-level, high-level and domain-specific image information for the automated grading of prostate and breast cancer and the distinction between cancerous and benign breast histology specimens. In [KDM+00] an automated machine vision grading system for cervical squamous epithelium has been developed. This system segments and marks the centres of all nuclei within the epithelium, used to construct a Delaunay triangulation mesh. Finally, the mesh is analysed to compute triangle dimensions giving an individual quantitative profile of measurements for each case. In [KRW+] an approach to extract information from slides of non-small cell lung carcinomas and to automatically classify the images into tumour versus non-tumour cases has been proposed. Texture features are extracted from small areas of the images and those that best fit with the original images are selected as the teaching set of automated recognition of images containing tumour areas. In [SKL+08] a hybrid approach for follicular lymphoma detection that combines information from several slides with different stains has been developed. Thus, follicles are first detected from digitized microscopy images with immunohistochemistry and then mapped to H&E-stained counterparts. In [SAGG06] a method to discriminate the main subsets of lung carcinomas by nuclear chromatin texture feature analysis has been presented. After staining, cell nuclei are automatically measured using a high-resolution image analyser. Then, texture features describing the granularity and the compactness of the nuclear chromatin are extracted for calculation of classification rules, which allow the discrimination of different tumour groups. In [CSK+09] a method for the automatic non-rigid registration of histological section images with different stain types has been proposed. This method is based on matching high level features that are representative of small anatomical structures and by establishing local groups of coherent features through geometric reasoning. The proposed method is validated on a set of follicular lymphoma images representing different disease stages. In [ACRA+15] the authors used an unsupervised feature learning framework for the automatic detection of basal cell carcinoma in histopathology images. In particular this framework for histopathology image analysis comprises three main stages: local representation learning using patches extracted with different strategies, global representation learning using a bag-of-features or a convolutional neural network and a visual interpretation layer to highlight the most discriminant regions detected by the model. In [dSPN+15] the authors realised a new ensemble of descriptors for the classification of transmission

electron microscopy images of viruses that is based on texture analysis. The ensemble of features is composed of six different local descriptors, including local binary pattern and rotation invariant local binary pattern, that have been combined using the Bag of Features or an Edge approach which extracts the textural information from specific regions of the image instead of from the original image. Although the high number of works dealing with histology images, only few of them propose a method that uses the colour information of these images. In fact, most of the works previously appointed operate on the grey level images obtained through a conversion from the original colour image. Other works, such as [LNHD07], use different kind of conversion averaging the three colour channels in order to obtain an intensity image in which cell structures are more visible. In some works instead [KPYW13, HST+11] the colour information of histological images has been used for the extraction of statistical texture features. Finally, in [GRCC+13] an exhaustive comparison of colour texture features and classification methods for histological image has been made, but the authors validated their method using only one database dealing with only one medical problem, that consists on discrimination of cells categories in histological images of fish ovary.

6.3 Datasets

The experimentation has been carried out on eight of the most famous colour histology image databases: HistologyDS, Pap-smear, Lymphoma, Liver Aging Female, Liver Aging Male, Liver Gender AL, Liver Gender CR and GLOMDB that represent a set of really different computer vision problems.

HystologyDS (HIS) database [CRCG11] is a collection of 20,000 histology images for the study of fundamental tissues. It is provided in a subset of 2828 images annotated by four fundamental tissues: connective, epithelial, muscular and nervous. Each tissue is captured in a 24-bit RGB image of size 720x480. Some sample tissue images from HIS database are showed in Fig.6.1. The database is available at the following link: $http://168.176.61.90/histologyDS/$

Pap-smear (PAP) database [JD06] is a collection of pap-smear images acquired from healthy and cancerous smears coming from the Herlev University Hospital (Denmark). This database is composed by 917 images containing cells, annotated

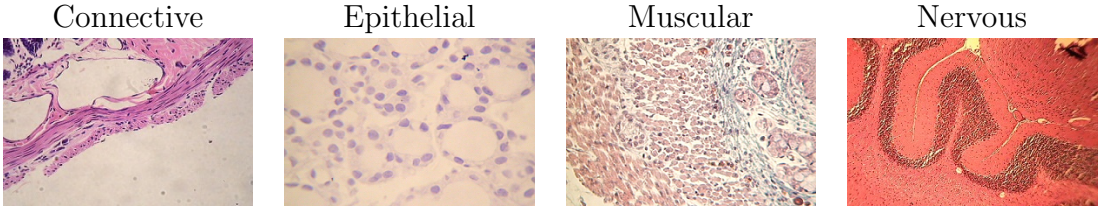

Figure 6.1: Four different tissues from HistologyDS database.

into seven classes, four representing abnormal cells and three representing normal cases. Nevertheless, from the medical diagnosis viewpoint the most important requirement corresponds to the general two-class problem of correct separation between normal from abnormal cells. For this reason in the further experiments only the binary case has been considered. Each cell is captured in a 24-bit RGB image without a fixed size that ranges from about 50x50 to about 300x300. Some examples are showed in Fig.6.2. The database is available at the following link: $http://labs.fme.aegean.gr/decision/downloads$

Lymphoma (LYM) database [SOD+08] is a collection of tissues affected by malignant lymphoma, a cancer affecting lymph nodes. Three types of malignant lymphoma are represented in the set: Chronic Lymphocytic Leukemia (CLL), Follicular Lymphoma (FL) and Mantle Cell Lymphoma (MCL). This dataset presents a collection of samples from biopses sectioned and stained with (H&E), prepared by different pathologists at different sites. Only the most expert pathologists specialised in these types of lymphomas are able to consistently and accurately classify these three lymphoma types from H&E-stained biopsies. This slide collection contains significant variation in sectioning and staining and for this reason it is more representative of slides commonly encountered in a clinical setting. This database contains a collection of 374 slides captured in a 24-bit RGB image of size 1380×1040. The database is available at the following link: $http://ome.grc.nia.nih.gov/iicbu2008/lymphoma/index.html$. In Fig.6.3 a randomly selected image from each class is showed.

AGEMAP Atlas of Gene Expression in Mouse Aging Project [ZPO+] is a study by the National Institute on Aging, National Institutes of Health, involving 48 male and female mice, of four ages (1, 6, 16, and 24 months), on ad-libitum or caloric restriction diets. Liver organs were extracted from sacrificed mice, sectioned, stained with H&E and captured by a bright-field microscope. Fifty colour images per liver were manually acquired using a Carl Zeiss Axiovert 200 microscope and 40x objective. Not all livers were suitable for imaging. Ultimately 1500 images

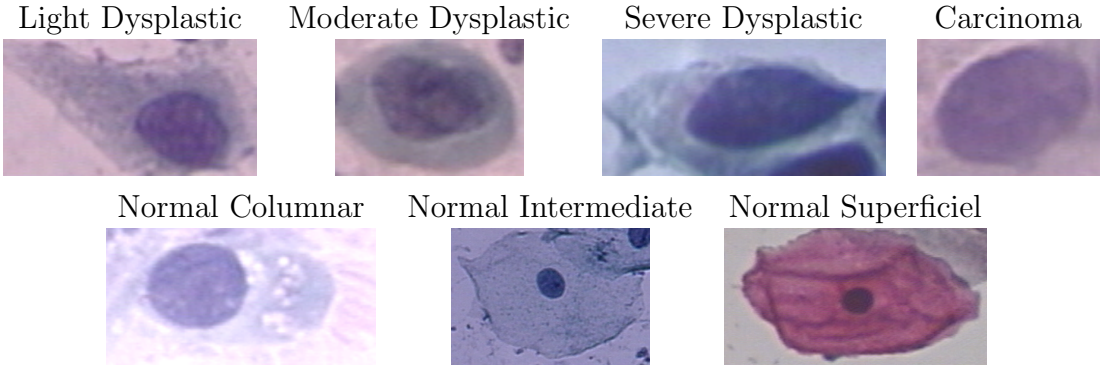

Figure 6.2: The seven classes of cells belonging to Pap-smear database: first four abnormal and last three normal.

from 30 livers were collected. Each image is of size 1388x1040 in TIFF format with a 36-bit RGB colour depth. The database is available at the following link: *http : //ome.grc.nia.nih.gov/iicbu*2008/*agemap/index.html*. As the acquisition was done using 12 bits of quantization per colour channel, the histograms have been compressed so as to cover the 8 bits encoding. Staining variability is a major challenge impeding quantitative analysis of H&E slides. All AGEMAP slides were prepared by the same person, following a set protocol over a short time-period, thus staining variability in this dataset is very limited. AGEMAP images can be analysed across multiple axis of differentiation: age, gender, diet, or individual mice to construct a variety of classification problems. Classifiers can be trained on livers of the same age and diet to predict the mouse gender or to predict the mouse diet. For these reasons the authors of the dataset proposed three different experiments using three different subsets of the original images:

- **Liver Aging Female** (LAF) experiment consists on a 4-way classification problem using the four classes (1, 6, 16 and 24 months) of images of female mice on Ad-libitum diet. This set is composed by 529 images.
- **Liver Gender AL** (LGAL) experiment consists on a 2-way classifier which classifies the gender of the mouse based on the images of 6-month old male and female mice on ad-libitum diet. This set is composed by 265 images.
- **Liver Gender CR** (LGCR) experiment consists on a 2-way classifier which classifies the gender of the mouse based on the images of 6-month old male and female mice on caloric restriction diet. This set is composed by 303 images.

To these three experiments in this thesis a fourth experiment has been added **Liver Aging Male** (LAM) that like the first one consists on a 4-way classification problem, but this time using the four classes (1, 6, 16 and 24 months) of images of male mice on Ad-libitum diet. This set is composed by 499 images.

GlomDB (GLOM) dataset has been specifically designed by the authors of [HST+11] to test colour and texture descriptors. This collection is created in order to analyse renal biopsies and to quantify the interstitial fibrosis. This dataset has been created from 15 different biopsies stained using Massons trichrome and then collected and processed by different operators at different times in the same hospital.

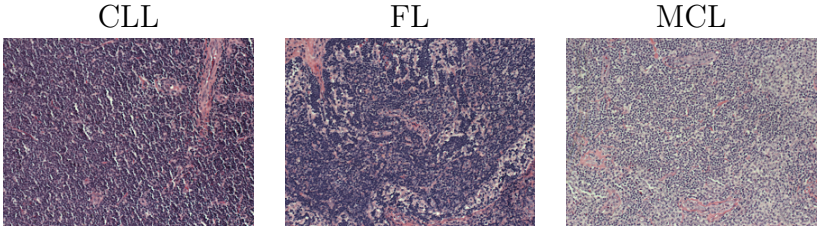

Figure 6.3: Three different kinds of lymphoma belonging to Lymphoma database.

However, all the imaging has been done using the same microscope (Zeiss Mirax

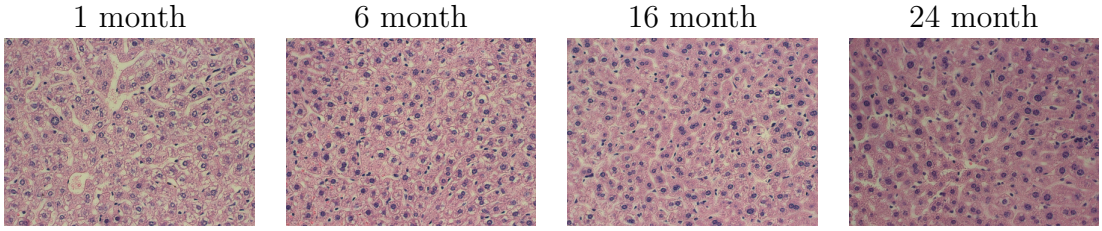

Figure 6.4: Four liver images representing the female mice of the different ages (from LAF dataset).

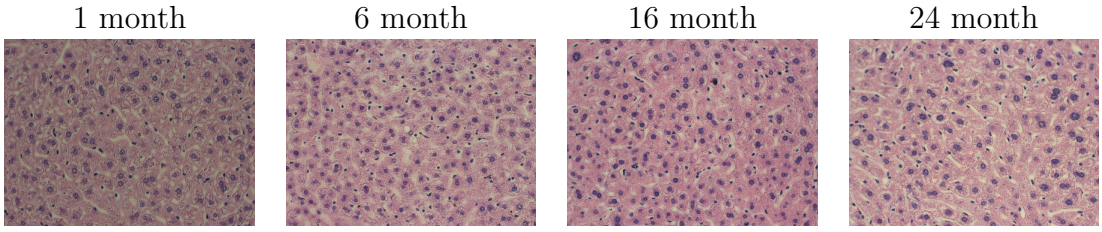

Figure 6.5: Four liver images representing the male mice of the different ages (from LAM dataset).

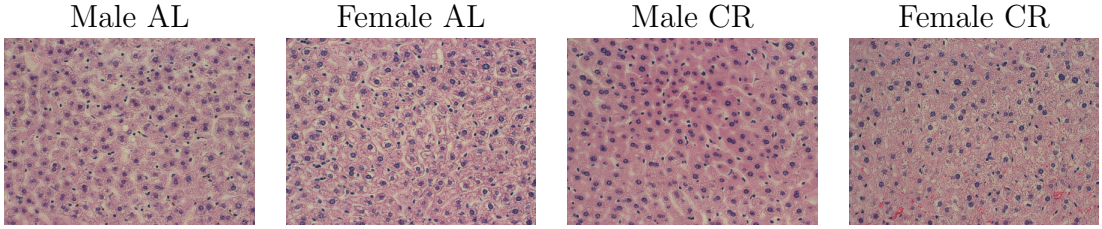

Figure 6.6: Two liver images representing the male and female mice on Ad-libitum diet (from LGAL dataset) and two liver images representing the male and female mice on caloric restriction diet (from LGCR dataset).

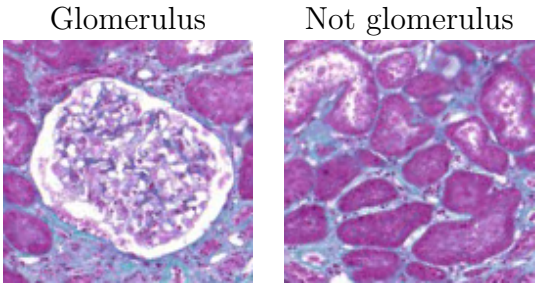

Figure 6.7: Two different samples of renal biopsy image.

Scan, 20X objective, NA 0.8) under identical illumination context. The acquired images have been manually segmented in order to properly detect the glomeruli structures of the kidney and then automatically cropped in non-overlapping patches of size 16×16. The number of total patches present in the dataset is 1976, half of which are glomeruli and the other half are biopsies parts that don't present glomeruli.

Chapter 7

Proposed Framework

Thus a CAD system could be very useful to speed up the process and to reduce subjectivity. In particular, the automated analysis by computer that uses image processing techniques is performed quickly, requires only one image of the sample and provides precise results. Developing general applications not dependant on specific histology datasets is still a challenge open problem. The main effort was to develop a general classification system for histology images able to properly classify medical images presenting different classification problems. A colour texture based histology image classification framework is proposed. In this system, no object detection or segmentation method is needed and a wide range of features including colour and texture features are used in combination to avoid the dataset dependency. Furthermore, different colour spaces have been investigated in order to individuate a general representation able to solve the classification problem efficiently without any dependence on specific image dataset or specific clinical field.

7.1 Feature Extraction

As said previously, many methods have been proposed for texture analysis. Here the attention has been given on improving some of the earliest methods used for the analysis of grey level texture based on statistical approaches GLCM (4.2.3), GLDM (4.2.3) and GLRLM (4.2.3) and one of the most recent and most used approach that is the LBP (7.3). Motivated by the wide diffusion of these methods and by the increasing numbers of medical datasets presenting colour images has been investigated the possibility to improve the accuracy of these methods using the colour information. Although the colour information from RGB images has already been used to extract GLCM descriptors [BH07], one of the goals of this work is to evaluate the performance improvement that can be obtained in histology image analysis from the computation of GLCM using the colour information from different colour spaces and furthermore, evaluate the performance improvement that can be obtained from the computation of GLDM, GLRLM and LBP using the colour information. In

order to extend the classical grey level texture features to colour texture features the starting point was the decomposition of the colour image into the three channels Ch_1, Ch_2 and Ch_3, obtaining three different images. The most intuitive way to take into account colour information for the computation of texture feature is to use the classical implementation and pass to them every time a different colour channel. This approach could be very useful thanks to a higher number of significant descriptors extracted and passed to the classifier. An improvement to this approach belongs to the combination of the colour channels in pairs (Ch_k, $Ch_{k'}$) with $k, k' = [1, 2, 3]$. This improvement is necessary in order to take into account not only repeated pattern inside the same colour channel, but also the correlation between the colour channels. The results of this combination is a feature vector nine time longer than the classical feature vector, composed by three intra-channel feature vectors (Ch_1, Ch_1), (Ch_2, Ch_2) and (Ch_3, Ch_3) and six extra-channels feature vectors (Ch_1, Ch_2), (Ch_2, Ch_1), (Ch_1, Ch_3), (Ch_3, Ch_1), (Ch_2, Ch_3) and (Ch_3, Ch_2) (see Fig.7.1).

However, not all these combinations make sense. In fact, for features extracted from GLCM and GLDM, combining the channels in pairs means that the occurrences and the differences for (Ch_k, $Ch_{k'}$) are calculated by storing on each (i, j) the number of occurrences (or differences) of $i \epsilon Ch_k$ and $j \epsilon Ch_{k'}$ having distance=d and orientation=θ and the number of occurrences (or differences) of $i \epsilon Ch_{k'}$ and $j \epsilon Ch_k$ having distance=d and orientation=θ. So, the vice versa produces the same result. Thus, for GLCM and GLDM only three extra-channels combination have been used, that are (Ch_1, Ch_2), (Ch_1, Ch_3) and (Ch_2, Ch_3). Therefore, from these six combinations the occurrences with $d = 1$ and $\theta = [0°, 45°, 90°, 135°]$ have been computed, producing 24 GLCMs. From each matrix have been extracted the 13 features proposed by Haralick that are: *angular second moment* (4.16), *contrast* (4.18), *correlation* (4.19), *variance* (4.20), *inverse difference moment* (4.21), *sum average* (4.23), *sum variance* (4.24), *sum entropy* (4.25), *entropy* (4.26), *difference variance* (4.27), *difference entropy* (4.28), *measure of correlation 1* (4.29) and *measure of correlation 2* (4.30). The total number of GLCM descriptors used is 312. In the same way, from the six channel combinations, the differences with $d = 1$ and $\theta = [0°, 45°, 90°, 135°]$ have been computed, producing 24 GLDMs. From each matrix have been extracted 9 features that are: *mean* (4.51), *angular second moment* (4.52), *contrast* (4.53), *variance* (4.54), *inverse difference moment* (4.55), *entropy* (4.56), *product moment* (4.57), *cluster shade* (4.58) and *cluster prominence* (4.59). The total number of GLDM descriptors used is 216.
Obviously, the GLRLMs can be computed by using the three classical bands only, that have been used to compute run-lengths in the four main directions, producing 12 GLRLMs. From each matrix have been extracted 11 features that are: *short run emphasis* (4.60), *long run emphasis* (4.61), *grey level non-uniformity* (4.62), *run length non-uniformity* (4.63), *run percentage* (4.64), *low grey level run emphasis* (4.65), *high grey level run emphasis* (4.66), *short run low grey level emphasis* (4.67), *short run high grey level emphasis* (4.68), *long run low grey level emphasis* (4.69) and

long run high grey level emphasis (4.70). The total number of GLRLM descriptors used is 132. The only one features extraction method that could benefit from all the extra-channel combination is the LBP. In fact, the coding obtained using the central

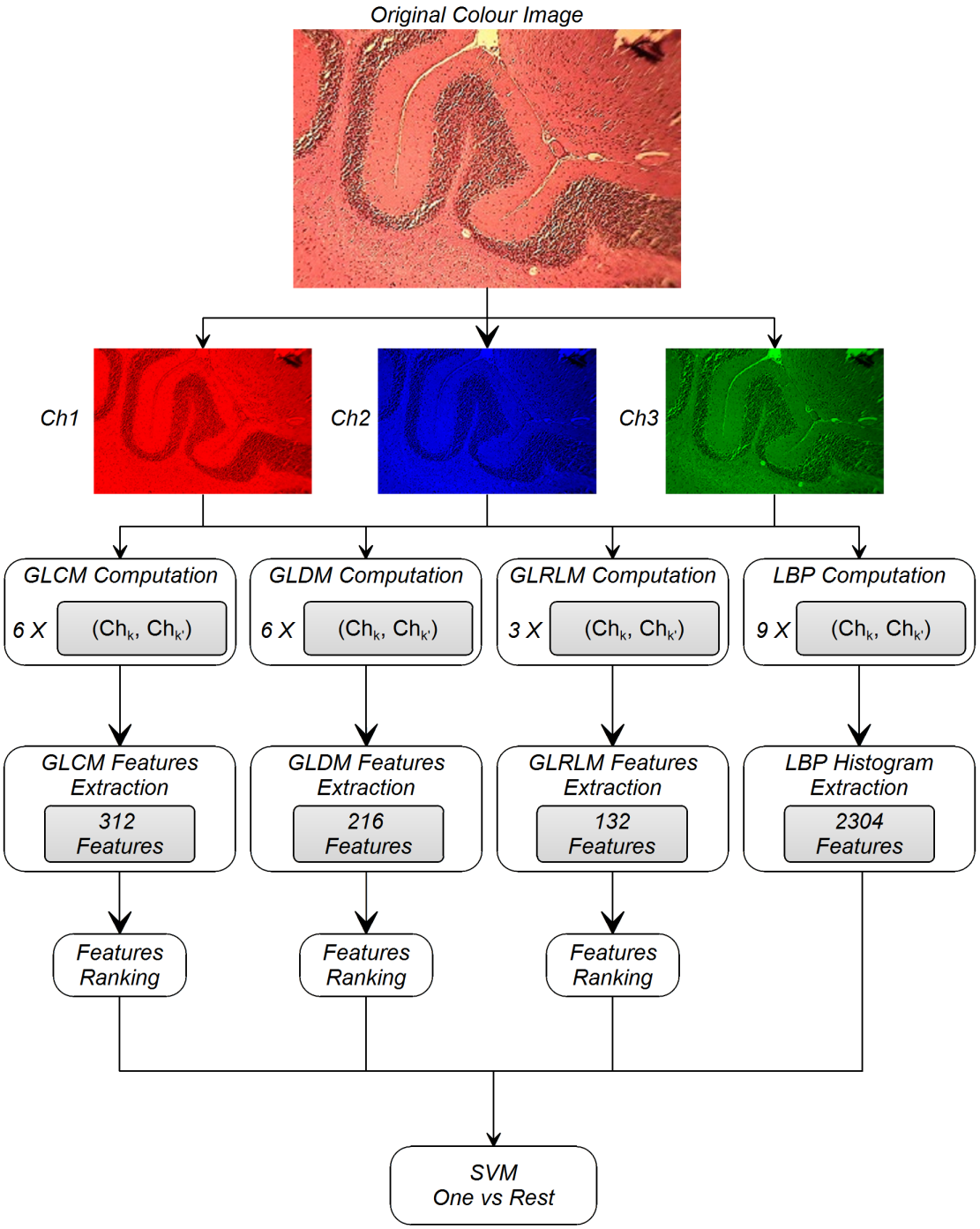

Figure 7.1: Diagram of the proposed system for histology image classification.

pixel from Ch_k to threshold the neighbourhood from $Ch_{k'}$ and the one obtained using the central pixel from $Ch_{k'}$ to threshold the neighbourhood from Ch_k with $k \neq k'$ will be different. Thus the nine channel combination have been used to extract the $LBP_{8,1}$, bringing to a feature vector of size 2304 (256 bin per 3 channels). The overall system is showed in Fig.7.1.

7.2 Classification

Each extracted feature subset has been used to train a multi-class SVM (5.5), that has been considered one of the best classification model for colour medical image application, as showed in [PR13a], where different classification model performances have been compared in classification of white blood cells affected by leukaemia. The SVM classifier is trained using the *one-VS-all* approach. This approach has been chosen both because it is the fastest to create a classification model during the training phase and because it is the fastest in testing phase. In fact, compared to the *one-vs-one* it requires a smaller number of binary classifiers. Furthermore, since we are working on different image databases with different number of classes, we cannot explore all the possible DAG structure or Binary tree in order to find the one leading to better results. Once the approach to train the multi-class SVM has been chosen a first experimentation has been performed in order to identify the most suitable kernel and parameters. Thus, through a 10 fold cross-validation each time the original training set has been divided into two subsets, the first one was used to train the SVM and the second one was used to test the obtained model. The kernel and parameters that permitted to obtain an ideal average accuracy value was the RBF kernel that uses a Gaussian Radial Basis Function with c parameter equal to 1e3 and γ equal to $1e2$. Once the kernel function and its relative parameters have been selected, the experimentation and validation of the proposed method have been performed. Given the variable size of the datasets the validation has been performed using a 5 time repeated stratified holdout (5.6), which guarantees that each class is properly represented both in the training set and in the test set and at the same time it averages the roles of each subset. In the experiments training and test sets are represented respectively by the 80% and the 20% of the samples. The performances of the classification models have been evaluated by calculating the *accuracy* (5.1), which gives a good indication of the performance since it considers each class of equal importance.

Before starting the evaluation on different colour spaces a comparison between the original features, extracted from images in grey level, and the features extracted using the colour information has been performed, in order to see if and how much the classification can benefit from this approach. Then, in order to evaluate the usefulness of the extra-channels, a comparison between the performance of the features extracted using the intra-channels only (Ch_k) with $k = [1, 2, 3]$ and the performance of the features extracted using the extra-channels too (Ch_k, $Ch_{k'}$) with

$k, k' = [1, 2, 3]$ have been made. The results of this comparison are reported in Table 7.1. Note that for GLRLM features the results reported for the extra-channel computation are the same reported for the intra-channel computation, since with the proposed method the GLRLM can be computed only considering a single channel.

In the last column the average accuracy obtained with the different features subset on each database is reported to better highlight the performance improvements. From the results it is possible to observe how the colour information improves the performances obtained with the classical grey level computation. In fact in that case the average accuracy value for all the database reaches only the 78.8 ± 15.3, while using the single colour channels computation the average accuracy value reaches the 89.1 ± 11, finally adding the combination of different channels the average accuracy value reaches the 94.9 ± 5.5.

7.2.1 Colour Space Analysis

As pointed out previously, the first goal is to compare all the feature subsets extracted with the proposed approach in order to individuate the colour space leading to better performances. As said previously, the most widely used colour space is the RGB colour space, since colour images are acquired through the RGB colour space.

Table 7.1: Accuracy values for each feature subset using grey level images and colour images (using intra-channels, RGBic, and extra-channels, RGBec).

Colour Space	Database	GLCM	GLDM	GLRLM	LBP	Average results
gray	LYM	58.9 ± 3.7	47.1 ± 5.6	58.1 ± 5.8	77.1 ± 3.1	60.3 ± 12.4
	PAP	80.4 ± 2.4	73.2 ± 1.9	82.7 ± 2.0	85.7 ± 1.6	80.5 ± 5.3
	HIS	61.0 ± 1.6	46.4 ± 2.6	51.7 ± 2.1	69.6 ± 1.7	57.2 ± 10.2
	LAF	93.5 ± 2.0	65.4 ± 6.0	81.7 ± 3.6	91.5 ± 2.0	83.1 ± 12.8
	LAM	84.4 ± 3.2	77.6 ± 2.9	71.9 ± 6.4	94.9 ± 1.3	82.2 ± 9.9
	LGAL	98.3 ± 2.1	89.4 ± 4.4	91.3 ± 4.9	97.2 ± 2.4	94.1 ± 4.4
	LGCR	99.0 ± 1.6	73.6 ± 6.3	95.6 ± 3.0	98.4 ± 1.3	91.7 ± 12.1
	GLOM	87.9 ± 1.1	85.7 ± 1.4	80.1 ± 2.4	74.4 ± 2.1	82.1 ± 6.0
RGBic	LYM	78.0 ± 4.2	60.4 ± 5.7	84.4 ± 2.3	90.3 ± 3.3	77.5 ± 12.0
	PAP	86.0 ± 3.3	76.3 ± 2.2	88.9 ± 1.4	85.4 ± 1.0	84.2 ± 5.5
	HIS	73.1 ± 1.3	55.8 ± 1.5	86.9 ± 1.6	87.1 ± 1.0	73.5 ±13.0
	LAF	98.8 ± 0.9	82.6 ± 2.9	99.2 ± 0.8	97.2 ± 1.1	94.6 ± 8.0
	LAM	98.3 ± 0.7	86.6 ± 3.4	99.1 ± 1.0	96.3 ± 0.9	95.4 ± 5.9
	LGAL	100.0 ± 0.0	99.6 ± 0.8	99.8 ± 0.6	98.5 ± 1.9	99.5 ± 0.7
	LGCR	100.0 ± 0.0	93.6 ± 2.5	99.2 ± 0.9	98.4 ± 1.7	99.7 ± 2.9
	GLOM	95.4 ± 0.6	90.9 ± 1.4	92.5 ± 1.1	92.0 ± 1.7	90.0 ± 5.8
RGBec	LYM	94.7 ± 2.9	87.6 ± 5.9	84.4 ± 2.3	97.2 ± 2.5	**91.0** ± 5.9
	PAP	92.5 ± 1.8	87.2 ± 1.8	88.9 ± 1.4	85.6 ± 2.3	**88.5** ± 2.9
	HIS	92.5 ± 1.3	84.6 ± 1.0	86.9 ± 1.6	89.0 ± 1.2	**88.2** ± 3.4
	LAF	99.8 ± 0.3	98.6 ± 0.6	99.2 ± 0.8	99.6 ± 0.5	**99.3** ± 0.5
	LAM	99.6 ± 0.5	97.3 ± 1.3	99.1 ± 1.0	99.3 ± 0.8	**98.8** ± 1.0
	LGAL	100.0 ± 0.0	100.0 ± 0.0	99.8 ± 0.6	100.0 ± 0.0	**99.9** ± 0.1
	LGCR	100.0 ± 0.0	100.0 ± 0.0	99.2 ± 0.9	99.7 ± 0.7	**99.7** ± 0.4
	GLOM	96.6 ± 0.9	94.3 ± 0.7	92.5 ± 1.1	92.9 ± 1.2	**94.0** ± 1.8

Also, for histology image analysis the RGB colour space is the most common, but recently the Luv and Lab have been used for this kind of applications. To realise a relevant comparison in these experiments five colour spaces that are representative of the three colour space families have been taken into account : RGB, HSV (2.7), Lab (2.6), Luv (2.5) and Ycbcr (2.4). So, the original features subsets have been tested separately on each database to asses their classification performances applied on different medical problems. The results of these experiments are reported in Table 7.2 that presents in the last column the average results for each database in each colour space. In this table all the entries related to the AGEMAP project have been removed, since their results are not so significant to this analysis, in fact they produce almost always results close to 100%, that have been still considered in the calculation of the average value. The best average results obtained for each database has been represented in bolt.

From the results it is possible to observe that there is not a great difference between the results obtained with the different colour spaces. In fact, the average accuracy value for all the tested database reaches almost the 95.0% with a standard deviation of ± 5.0. It is also interesting to see that in general for each feature subset the RGB colour space is not the best choice, except for the LBP demonstrating an higher correlation between channels in the RGB colour space. Motivated by this

Table 7.2: Accuracy values for each feature subset in each colour space.

Colour Space	Database	GLCM	GLDM	GLRLM	LBP	Average results
RGB	LYM	94.7 ± 2.9	87.6 ± 5.9	84.4 ± 2.3	97.2 ± 2.5	91.0 ± 5.9
	PAP	92.5 ± 1.8	87.2 ± 1.8	88.9 ± 1.4	85.6 ± 2.3	88.5 ± 2.9
	HIS	92.5 ± 1.3	84.6 ± 1.0	86.9 ± 1.6	89.0 ± 1.2	88.2 ± 3.4
	GLOM	96.6 ± 0.9	94.3 ± 0.7	92.5 ± 1.1	92.9 ± 1.2	**94.0** ± 1.8
	Average	97.0 ± 3.5	94.0 ± 6.6	93.5 ± 6.7	94.9 ± 5.6	
HSV	LYM	94.9 ± 2.7	88.1 ± 4.3	88.1 ± 3.8	92.7 ± 2.9	91.0 ± 3.4
	PAP	93.6 ± 2.0	88.3 ± 1.9	90.4 ± 1.2	86.9 ± 1.9	**89.8** ± 2.9
	HIS	93.1 ± 1.4	87.7 ± 1.6	89.9 ± 1.8	89.0 ± 1.9	**89.9** ± 2.3
	GLOM	97.4 ± 0.5	94.5 ± 1.0	93.8 ± 0.8	82.0 ± 1.2	91.9 ± 6.8
	Average	97.3 ± 2.9	94.9 ± 5.6	95.2 ± 5.1	93.9 ± 7.1	
Lab	LYM	92.9 ± 2.9	83.7 ± 3.7	86.1 ± 2.8	91.1 ± 2.4	88.4 ± 4.3
	PAP	93.9 ± 2.0	87.3 ± 2.0	90.1 ± 1.4	87.0 ± 3.1	89.6 ± 3.2
	HIS	93.4 ± 0.8	86.1 ± 1.2	90.7 ± 0.7	87.9 ± 0.8	89.5 ± 3.2
	GLOM	97.4 ± 0.8	96.2 ± 0.9	95.8 ± 0.6	88.4 ± 1.6	94.9 ± 4.1
	Average	97.0 ± 3.3	93.9 ± 6.8	95.1 ± 5.1	94.4 ± 5.7	
Luv	LYM	92.8 ± 2.7	82.5 ± 4.1	88.7 ± 2.1	83.7 ± 1.9	86.9 ± 4.7
	PAP	93.6 ± 1.7	87.2 ± 2.5	90.4 ± 2.1	85.0 ± 2.0	89.0 ± 3.7
	HIS	92.3 ± 1.0	83.6 ± 1.4	90.4 ± 0.9	80.3 ± 1.6	86.6 ± 5.6
	GLOM	97.4 ± 1.1	96.2 ± 0.9	96.1 ± 1.1	91.1 ± 1.1	**95.2** ± 2.8
	Average	97.3 ± 3.1	93.6 ± 7.8	95.8 ± 4.5	92.2 ± 8.0	
Ycbcr	LYM	95.1 ± 2.4	84.8 ± 3.7	91.5 ± 2.3	95.3 ± 1.7	**91.7** ± 4.9
	PAP	93.2 ± 1.4	87.8 ± 2.2	89.9 ± 1.3	86.2 ± 2.4	89.2 ± 3.0
	HIS	93.2 ± 0.8	86.1 ± 1.6	90.8 ± 1.0	86.9 ± 1.9	89.2 ± 3.3
	GLOM	97.9 ± 0.5	96.3 ± 0.6	96.1 ± 0.7	88.5 ± 1.4	94.7 ± 4.2
	Average	97.2 ± 3.3	93.5 ± 7.6	95.3 ± 5.1	94.4 ± 5.8	

Table 7.3: Accuracy values for LBP using only intra-channels in each colour space.

Database	RGB	HSV	LAB	LUV	YCBCR
LYM	90.3 ± 3.8	94.4 ± 2.7	89.3 ± 2.4	91.9 ± 1.5	86.3 ± 4.5
PAP	85.4 ± 2.0	85.0 ± 1.7	85.7 ± 2.1	87.1 ± 1.7	86.4 ± 1.4
HIS	87.5 ± 1.5	91.7 ± 1.1	89.4 ± 1.1	91.5 ± 1.2	89.6 ± 0.8
LAF	97.2 ± 1.7	99.8 ± 0.4	99.3 ± 0.6	98.7 ± 1.1	98.0 ± 2.0
LAM	96.3 ± 2.5	99.1 ± 0.9	97.8 ± 1.5	98.4 ± 1.2	98.2 ± 1.5
LGAL	98.9 ± 1.3	100.0 ± 0.0	99.2 ± 1.0	99.6 ± 0.8	99.8 ± 0.6
LGCR	98.4 ± 1.1	99.7 ± 0.7	99.7 ± 0.7	99.8 ± 0.5	99.8 ± 0.5
GLOM	82.1 ± 1.2	85.4 ± 1.8	87.1 ± 1.9	88.4 ± 2.0	88.4 ± 1.2
Average	92.0 ± 6.5	95.1 ± 5.9	93.4 ± 6.1	94.4 ± 5.3	93.3 ± 6.2

result another experiment totally based on the LBP descriptor has been performed. In fact, it is important to understand if the LBP presents a real correlation between channels in the other colour spaces, or if for this particular descriptor it is better to use only the single channels without considering the extra-channels. The results of this experiment are reported in Table 7.3, where it can be observed how the performance obtained with the LBP extracted considering the extra-channels or not are comparable. This leads to the conclusion that the LBP don't present a real correlation between colour channels, except for the RGB colour space that in Table 7.2 demonstrate better results. In general, it can be also confirmed that the HSV is the colour space that outperforms the others, considering both the results obtained in Table 7.3 and the results obtained with the other feature subsets and reported in Table 7.2. Thus, since the performance with the intra-channels and the extra-channels LBP are comparable and considering that the intra-channels version produce a quite smaller feature vector, of size 768 (256 bin per 3 channels) from now on only the intra-channels LBP will be computed.

7.3 Features Ranking

After comparing the performances of each feature subset separately, it has been decided to create a new feature set grouping togheter all the feature subsets or some combinations of different feature subset, in order to test if the classifier could benefit from a higher number of descriptors and from which group of feature subset it could benefit the most. But, since the number of features is really high, another step has been added to the system that consists on a selection of features. Obviously the feature selection step cannot be applied for the LBP descriptor. In fact, since the feature vector obtained from the LBP is nothing more than the histogram generated from the LBP image, selecting some descriptors from this feature vector means use only some bins of the histogram. Although an approach for LBP histogram selection has already been proposed in [PVH13] for colour texture classification, it seems too specific for single use cases and so not useful for a general framework with the aim

of being independent from the individual classification problem. The step of feature selection used consists on a combination of feature selection techniques, that have been applied separately to each feature subset in order to trace the best features for each approach and then combining them for the classification process. Combining different feature selection techniques means that various approaches have been used to obtain a stronger result. This is because in general all the classification models benefit from the process of feature selection, but in particular each classifier has better performance associated with feature selection based on the same classifier, for example the k-NN improves more with feature selection based on k-NN, Naive Bayes improves with feature selection based on Naive Bayes and so on [RFP15b]. So, by combining different methods of feature selection it is possible to obtain a result that is strong enough to be used in each condition and with the majority of the classification models. The adopted methods of feature selection make use of sequential forward feature selection (4.3) and they are based on k-Nearest Neighbour (FS-kNN) and Decision Trees (FS-tree). Using the features selected from these methods it is possible to establish a sort of ranking that is finally combined with the ranking provided by the ReliefF (FS-Rel) algorithm and then used to extract the final feature set. As previously said, the main effort is to develop a general classification system for histology images able to properly classify medical images presenting different classification problems, and thus not dependent on a specific dataset. For this reason the feature ranking step has been applied separately to each database and then combined in order to provide a vector in which the features are sorted by relevance according to all the dataset. Obviously this experimentation is not exhaustive of all the medical problems that can be diagnosed by histology image analysis, but with the use of 8 very different datasets it is possible to guarantee a strong result. Two main experiments using the feature ranking have been performed. The first one is devoted on the analysis of the singular descriptors, in order to find which produces good results, or better, which one produces bad results and could be leaved out from the proposed method. This is done by making a further ranking starting from that one already obtained, that lists the descriptors sorted by relevance. For comparison purposes in Table 7.4 have been reported the descriptors sorted by relevance for each subset. So, in order to consider all the most relevant descriptors, a series of experiments have been performed by selecting each time a different number of descriptors ranging from 1 to the maximum number of descriptors. Table 7.5 shows some of the performed experiments. Also in this table all the entries related to the AGEMAP project have been removed. The results obtained show that the ranking is correct for all the tested database. Moreover, using only the descriptors that brought to better performances it is possible to reduce significantly the dimensionality for each feature subset. In fact from now on the proposed framework will use only the first 6 GLCM descriptors (*correlation* (4.19), *measure of correlation 1* (4.29), *measure of correlation 2* (4.30), *sum average* (4.23), *difference entropy* (4.28), *sum entropy* (4.25)), the first 5 GLDM descriptors (*mean* (4.51), *entropy* (4.56), *inverse difference moment* (4.55), *angular second moment* (4.52),

Table 7.4: Ranking of the descriptors for each features subset.

Ranking	GLCM	GLDM	GLRLM
1	Cor	μ	LRE
2	MC1	Ent	HGLRE
3	MC2	IDM	SRE
4	SAve	ASM	RLN
5	DEnt	Con	SRLGLE
6	SEnt	Var	GLN
7	IDM	PM	RP
8	Ent	CS	LGLRE
9	Con	CP	LRHGLE
10	ASM		LRLGLE
11	DVar		SRHGLE
12	Var		
13	SVar		

contrast (4.53)) and the first 4 GLRLM descriptors (*long run emphasis* (4.61), *high grey level run emphasis* (4.66), *short run emphasis* (4.60), *run length non-uniformity* (4.63)).

Table 7.5: Accuracy values for each feature subset after feature selection.

Descriptors Number	Database	GLCM	GLDM	GLRLM
2	LYM	93.3 ± 1.7	90.7 ± 3.4	84.5 ± 4.0
	PAP	92.3 ± 2.4	90.6 ± 2.5	87.5 ± 2.8
	HIS	89.5 ± 1.1	87.1 ± 0.6	78.4 ± 1.2
	GLOM	96.2 ± 0.8	96.5 ± 1.2	91.3 ± 1.4
	Average	96.3 ± 4.1	95.5 ± 5.2	92.1 ± 8.0
3	LYM	93.7 ± 3.0	90.7 ± 2.7	85.1 ± 4.4
	PAP	92.0 ± 2.9	91.2 ± 1.6	87.0 ± 1.9
	HIS	91.5 ± 1.0	90.1 ± 1.5	80.0 ± 1.5
	GLOM	96.3 ± 0.4	96.5 ± 1.1	91.5 ± 1.4
	Average	96.6 ± 3.8	96.0 ± 4.6	92.6 ± 7.8
4	LYM	92.8 ± 3.2	91.5 ± 2.1	85.6 ± 2.7
	PAP	92.9 ± 1.4	91.2 ± 2.2	88.1 ± 1.3
	HIS	92.9 ± 0.9	90.5 ± 1.4	83.0 ± 1.8
	GLOM	97.3 ± 0.9	96.1 ± 1.3	93.6 ± 1.4
	Average	97.0 ± 3.5	96.1 ± 4.3	93.6 ± 7.1
5	LYM	95.5 ± 1.9	92.5 ± 3.5	84.0 ± 3.3
	PAP	93.3 ± 1.9	90.7 ± 1.8	88.7 ± 2.4
	HIS	92.8 ± 0.8	91.6 ± 1.5	84.0 ± 1.1
	GLOM	97.1 ± 0.6	96.7 ± 1.0	92.6 ± 1.1
	Average	97.3 ± 3.1	96.5 ± 4.0	93.3 ± 7.0
6	LYM	94.3 ± 2.3	94.1 ± 2.5	82.9 ± 3.1
	PAP	94.5 ± 2.4	90.3 ± 2.0	89.4 ± 1.7
	HIS	92.8 ± 1.0	91.4 ± 1.1	84.6 ± 0.6
	GLOM	97.3 ± 0.7	96.6 ± 1.0	92.0 ± 1.6
	Average	97.3 ± 3.0	96.2 ± 4.1	93.4 ± 7.2

7.4 Features Aggregation

Once the most relevant features have been obtained and the dimensionality has been reduced, it is possible to perform an aggregation of feature sub-set in order to find a complete feature set able to classify correctly images belonging to different database. The initial idea was to create a feature set aggregating all the feature sub-set, but with the hope to find a smaller feature set, various test have been performed also grouping the feature sub-set first in pairs and then in triple. To summarise in Table 7.6 only the groups of feature sub-set that brought to better results have been reported.

Table 7.6: Some successful grouping of feature subsets.

Database	GLCM GLDM	GLDM GLRLM	GLCM GLRLM	GLCM GLDM GLRLM
LYM	94.8 ± 2.0	95.1 ± 2.5	96.0 ± 1.5	**97.4** ± 0.9
PAP	93.7 ± 2.5	93.6 ± 1.2	94.2 ± 1.3	**96.6** ± 0.3
HIS	93.2 ± 0.9	92.6 ± 1.0	92.7 ± 1.0	**96.0** ± 0.9
LAF	100 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	**100.0** ± 0.0
LAM	99.9 ± 0.2	99.7 ± 0.5	99.9 ± 0.3	**100.0** ± 0.0
LGAL	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	**100.0** ± 0.0
LGCR	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	**100.0** ± 0.0
GLOM	97.3 ± 0.6	96.3 ± 0.9	97.4 ± 0.8	**97.7** ± 0.7

As it can be noted from this table none of the groups of features sub-set that obtained the best performances presents the LBP. In fact with this descriptor the performances are excellent only for some databases, so it cannot be used for a general framework. On the contrary the other descriptors provide good results for each database, up to excellent results if combined together. Finally, in Table 7.7 a comparison between the results obtained and others present in literature has been reported. As it can be seen, the proposed approach outperforms the results proposed in literature for all the tested databases. In particular, it is important to note that only few authors have analysed more than one database and that none used all the 8 datasets which contain different histological images. As a consequence it's really difficult to identify an approach able to achieve good classification performances for different medical problems.

7.5 Discussion

The average accuracy value reported in Table 7.2 confirmed what has been already brought to light in [RFP15a], namely that co-occurrence features extracted using the colour information are the most accurate in classification of histological images. In fact, the comparison with the other features is totally in favour of co-occurrence features. This is confirmed also by the comparison with newer features like the LBP,

Table 7.7: Comparison of the results obtained with the state of the art.

Database	Nanni [NBG+13]	Shamir [SOD+08]	Meng [MLSC10]	Arevalo [ACRA+15]	Herve [HST+11]	Dos Santos [dSPN+15]	Proposed approach
LYM	-	85.0	92.7	-	63.3	-	**97.4**
PAP	92.5	-	-	-	-	91.4	**96.5**
HIS	92.4	-	-	94.1	-	92.4	**96.0**
LAF	-	51.0	96.4	-	-	-	**100**
LAM	-	-	-	-	-	-	**100**
LGAL	-	69.0	99.4	-	97.3	-	**100**
LGCR	-	99.0	-	-	-	-	**100**
GLOM	-	-	-	-	97.4	-	**97.7**

that on the contrary demonstrate a lower correlation between colour channels, even if with a great improvements related to the use of colour. Furthermore, thanks to this comparison it is also confirmed what has been already established in [RFP15b] that is the HSV colour space outperforms the others in extraction of features with colour information. Thanks to the feature ranking step it is now possible to say which are the best and the worst descriptors for histology image classification that from now can be completely excluded from further extraction. In this way it is possible to reduce not only the feature extraction time but also the classification time. In fact, you have to note that the exclusion of one descriptor from the proposed framework means that such descriptor will not be computed in each direction and orientation and in each colour combination. For example, for co-occurrence matrix excluding just one descriptor means reducing the feature vector of 24 elements. Thus, since 7 descriptors from the GLCM subset have been removed, the number of features changed from 312 to 144. Again removing 4 descriptors from the GLDM subset, the number of features changed from 216 to 120 and removing 7 descriptors from the GLRLM subset, the number of features changed from 132 to 48. After the performed experiments the final feature vector of size 312 have been created, whereas LBP descriptors have been completely excluded, as they are too dependent on the single database. Finally, it is important to report some consideration on the execution time for the whole framework. Despite the speed of a CAD system of histology images depends on many factors, such as the size and resolution of the image, the complexity of texture and last but not least the configuration of the computer (the computers used were configured respectively with a processor Intel(R) Core(TM) i7 CPU, 960@3.20 GHz, RAM 4.00 GB size and with a processor AMD Phenom(TM) II x6 1090T, RAM 4.00 GB size), on average, for the final feature subset, the feature extraction step is completed in about 1.6 ± 0.04 seconds per image, considering the worst case with the biggest images belonging to Lymphoma. The training phase for the SVM model is completed in the worst case, with HistologyDS that presents the highest number of images, in 2.6 seconds, while the test phase is performed on average in 0.018 ± 0.002. Thus, considering a real case of application in which we can assume that the dataset has already been acquired and the model already

created, since this is an operation performed only once, a computer aided diagnosis can be performed in less than two seconds. Considering also that the code has not yet been optimised and that it runs on single core the computing time for the proposed method seems already excellent.

Part III

CAD for Peripheral Blood Images

Chapter 8

Background

8.1 Haematology

Haematology is the branch of medicine concerned with the study, diagnosis, monitoring, treatment, and prevention of diseases related to the blood and blood-forming organs. Haematology studies the blood in health and in pathologic conditions, not only to identify the patient's haematologic condition correctly but also to predict how the bone marrow may have contributed to that condition. Thus haematology studies the relationship of the bone marrow to the systemic circulation. In fact, there are many diseases, disorders, and deficiencies that can affect the number and type of blood cells produced, their function, and their lifespan. Usually, only normal, mature or nearly mature cells are released into the bloodstream, but certain circumstances can induce the bone marrow to release immature and/or abnormal cells into the circulation. One of the most frequently ordered test to monitor the proportion of the cell components into the blood stream is the Complete Blood Count (CBC). This evaluation consists of nine components and offers various haematologic data to interpret and review that directly relate to the health of the bone marrow, represented by the numbers and types of cells in the peripheral circulation. The percentages of cells are compared with the reference ranges in order to determines if the cells are present in normal proportion to another, if one cell type is increased or decreased, or if immature cells are present. Reference ranges for blood tests are sets of values used to interpret a set of diagnostic test results from blood samples. Since it is difficult to prove that subjects that are defined as healthy may not have infections, parasitic infection and nutritional deficiency, it is more feasible to talk of reference ranges rather than normal ranges. A reference range is usually defined as the set of values in which 95% of the normal population falls within. It is determined by collecting data from vast numbers of laboratory tests result from a large number of subjects who are assumed to be representative of the population. With the automatic counter or the flow cytometry previously mentioned, an automated CBC can be performed quickly. However, if the results from an automated cell count indicate

the presence of abnormal cells or if there is reason to suspect that abnormal cells are present, then a blood smear will be performed. A blood smear is often used to categorize and/or identify conditions that affect one or more type of blood cells and to monitor individuals undergoing treatment for these conditions. The results of a blood smear typically include a description of the appearance of the cells as well as any abnormalities that may be seen on the slide.

8.2 Peripheral Blood Images

A typical blood image usually consists of three components: platelets, red blood cells (RBCs) and white blood cells (WBCs).

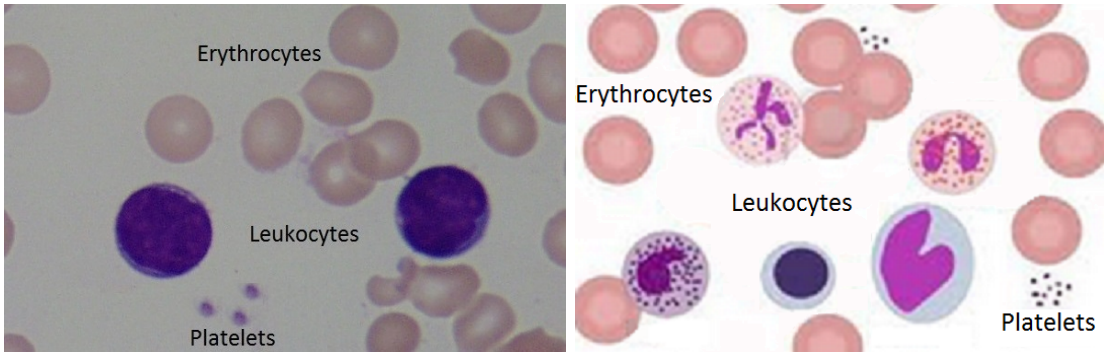

Figure 8.1: Peripheral blood smear components: a real image and a schematic representation.

Platelets or thrombocytes are small non-nucleated disc shaped cells with a diameter between 1 and 3 um. Upon release into the peripheral blood from the bone marrow, they appear as fragments. They play a major role in haemostasis leading to the formation of blood clots when there is blood vessel injury or other bleeding, starting to clump together to form aggregates. There must be a sufficient number of platelets to control bleeding. If there are too few, or if they don't function properly, the ability to form a clot becomes impaired and can be a life-threatening situation [Cie11]. Normal, mature RBCs or erythrocytes are uniform in size, 7-8 m in diameter, and do not have a nucleus as most other cells do. They are round and flattened like a donut with a depression in the middle instead of a hole (biconcave). Due to the haemoglobin inside the RBCs, they appear pink to red in colour with a pale centre after staining the blood smear. While not every RBC will be perfect, any significant number of cells that are different in shape or size may indicate the presence of disease [EA13]. WBCs or leukocytes instead have a nucleus surrounded by cytoplasm and also for this reason they are easily identifiable, as their nucleus appears darker than the background. However, the analysis and the processing of data related to the WBCs are complicated due to wide variations in cell shape, dimensions and edges. The generic term leukocyte refers to a set of cells that are quite different from each

other. Indeed, although they are all derived from bone marrow stem cells, in the bone marrow, they differentiate into two main groups: cells containing granules, called granulocytic or myelocitic, and cells without granules called mononuclear or lymphoid. Thus, we can distinguish between these cells according to their shape or size, the presence of granules in the cytoplasm and the number of lobes in the nucleus, as it can be seen in Fig. 8.2. The lobes are the most substantial part of the nucleus, and thin filaments connect them to each other. WBCs mature into five distinct types of cells, that include for the granulocytic neutrophils, basophils and eosinophils and for the non-granulocytic lymphocytes and monocytes. Neutrophils compose the majority of WBCs in a healthy adult, present in the human blood at a percentage ranging between 50 and 70%. They range in size from 10-15 microns, and present a cytoplasm with pink or purple granules. They are distinguishable also due to the number of lobes present in the nucleus, which can range from 1 to 6 according to the cell maturation. They are involved in the defence against infections. Basophils represent only 0-1% of all lymphocytes in human blood, and they have a diameter of approximately 10 microns. Generally, basophils have an irregular, plurilobated nucleus that is obscured by large and dark granules. Eosinophils are easily recognized in stained smears due to the presence of large, red-orange granules, which include para-crystalline structures in the shape of a coffee bean. They are round, 10-12 microns in size, and have a nucleus with two lobes. Generally low in number, present at 1-5% in human blood, they most often increase in number in individuals with allergies and parasitic infections. Monocytes are usually the most voluminous WBCs, with a diameter of 12-20 microns and are often referred to as scavenger cells (phagocytes). They can ingest particles such as cellular debris, bacteria, or other insoluble particles. They represent 3-9% of circulating leukocytes. Their nucleus is large and curved, often in the shape of a kidney. Lymphocytes are usually the smaller WBCs, with a diameter of 7-12 microns. They are characterised as having a smooth, round nucleus and a small amount of cytoplasm and often a smooth. They are very common in human blood, with a percentage of 20-45%.

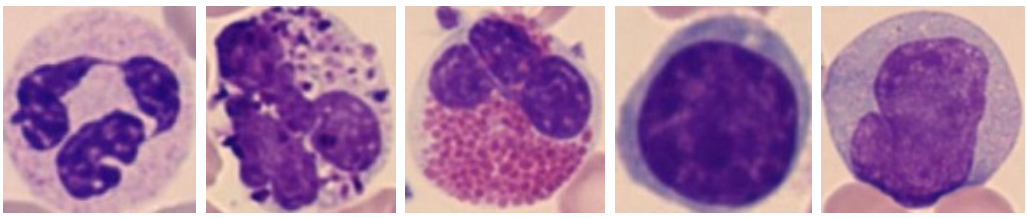

Figure 8.2: A comparison between different types of WBCs: neutrophils, basophils, eosinophils, monocytes and lymphocytes.

Numerous diseases and conditions can affect the absolute or relative number of WBCs and their appearance on a blood smear. More details of the conditions that affect the number and the morphology each kind of cells are listed in Appendix A. Examples of the most common diseases that involve variation in shape and number of

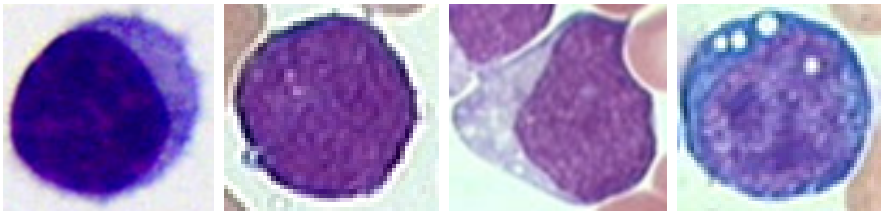

Figure 8.3: A comparison between lymphocytes suffering from ALL: a healthy lymphocyte, followed by lymphoblasts classified as L1, L2 and L3, respectively, according to the FAB [BCMT+76].

blood cells include anaemia, haemophilia, general blood clots and bleeding disorders while more serious cases that need to be diagnosed are leukaemia, myeloma, and lymphoma. This thesis focused on the detection of a particular type of leukaemia.

8.3 ALL - Acute Lymphoblastic Leukaemia

Leukaemia is a blood cancer that can be detected through the analysis of WBCs. There are two types of leukaemia: acute and chronic. According to the French-American-British (FAB) classification model [BCMT+76], acute leukaemia is classified into two subtypes: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Here, only ALL has been considered, which affects a group of leukocytes called lymphocytes. ALL primarily affects children and adults over 50 years of age. The risk of developing ALL is highest in children younger than 5 years of age, and it declines and begins to rise again after age 50. Due to its rapid expansion into the bloodstream and vital organs, ALL can be fatal if left untreated [BCPP00]. Therefore, early diagnosis of this disease is crucial for a patients' recovery, especially for children. Diagnosis of ALL is based on the morphological identification of lymphocytes suffering from ALL, called lymphoblasts, by microscopy and the immunophenotypic assessment of lineage commitment and developmental stage [IGM13]. Lymphoblasts present morphological changes that increment with increasing severity of the disease. In particular, lymphocytes are regularly shaped and have a compact nucleus with regular and continuous edges, whereas lymphoblasts are irregularly shaped and contain small cavities in the cytoplasm, termed vacuoles, and spherical particles within the nucleus, termed nucleoli [LPS11] (Fig. 8.3).

The observation of blood smears by skilled operators is one diagnostic procedure available to initially recognise the ALL, where the automatic counter fails due to the presence of abnormal cells. Human visual inspection is tedious, lengthy and repetitive, and it suffers from the presence of a non-standard precision because it depends on the operator's skill; these disadvantage limit its statistical reliability. The use of image processing techniques can help count the cells in human blood and at the same time to provide information about cell morphology, making them less

expensive and providing more accurate standards. One of the goal of this thesis is to provide a fully automatic procedure based on the analysis of blood smear images, to support medical activity. This procedures counts the WBCs present in the smear through a process of segmentation and detection. The detected WBCs are then classified as suffering from ALL or not.

8.4 Related Works

The progress of automated methods for the classification of blood cells from digitized images is a current problem in pattern recognition. These techniques can help to count the cells in human blood and, at the same time, be able to provide information on the morphological cells themselves. Unfortunately, for the analysis and processing of images, there are no standard techniques to apply to all types of images, but the processing must be adapted to the context. In particular regarding the blood smear images the processing techniques vary according to the type of blood cell to be analysed. In the literature, few attempts of automated systems, based on techniques of image processing, able to identify and classify peripheral blood cells have been proposed. Moreover, the existing systems are only partially automated, combining manual steps to automated ones. Furthermore, most of them do not work on the entire analysis process, but on individual phases of the whole analysis process. This section will show the techniques most commonly used by different authors at different stages of the process.

The used methods of pre-processing depend on many variables, such as lighting conditions, the duration of the dye, defects caused by visual artefacts or not uniform background. The images should be processed in order to improve certain characteristics or to reduce further operations that could be required in the later stages of the analysis. The main issues addressed at this stage include noise reduction and enhancement of some structures of the images. Mohapatra et al. [MP10b, MPS10, MP10a, MPS14] used a median filter to remove the noise followed by a Unsharp filter. The median filter has been preferred to the average filter since it preserves details of edges and then are enhanced with the use of the Unsharp filter. Other authors instead to enhance the quality of the images preferred operations based on the histogram such as contrast stretching or the histogram equalisation in order to redistribute the grey level values. Often these two transformations have been used in combination between them or with other techniques, as the algorithm proposed in [MKA$^+$10] that starting from the grey scale image performs separately a contrast stretching and a histogram equalisation. Then a series of arithmetic operations between the two images just obtained is performed in order to highlight the nuclei of leukocytes. The results is very impressive because not only it enhances the nuclei of leukocytes, but it drastically reduces the number of the other blood components, making the further steps much more simple.

As said previously in the analysis of medical images different levels of segmen-

tation are used. In particular for what concerns peripheral blood images two main levels are used: the level cells segmentation, which aims to separate whole cells from the background or plasma and the level of segmentation that tries to separate the various components inside the cell, such as the nucleus from the cytoplasm or intracellular parasites. Several authors have proposed methods for effective segmentation of the nucleus of leukocytes, while there are few attempts of segmentation of the cytoplasm. The characteristic generally used for segmentation is the intensity value of grey level images. However, many authors showed how the use of single channel from different colour spaces could be used to highlight differences between blood components. Indeed the nuclei of the white blood cells are more in contrast on the Green component of the RGB colour space [Cse92], while the cytoplasm of the white blood cells is more evident on the Hue component [WZZO06] or the Saturation component [HMH11] of the HSV colour space. The Saturation component of the HSV colour space is also useful to identify and separate erythrocytes [RDKJ00] when present in complex agglomerates of cells, otherwise they can be easily detected and counted with a threshold value computed using Zack algorithm from grey level images [BTKD11]. Based on this knowledge many algorithms of region growing have been proposed [KGHS96, LEC$^+$98, LC02] that use the pixels within the nuclei, identified previously, as seeds in order to segment the whole white blood cells. The cytoplasm is detected through iterative aggregation of the pixels surrounding the region of interest according to the homogeneity of the colours and the information of the gradient. Edge based segmentation methods are rarely used in this contest, since the boundaries between cells are not clearly defined. However, the performances of edge detection operations can be improved using morphological operators [PS04, Sco05] being able to connect in a better way the detected edges and restore the complete boundary of the cells.

A further problem in the analysis of peripheral blood cell images is the presence of cells grouped together or adjacent. This is an important problem since it doesn't allow an analysis of the single cells, such as the computation of shape descriptors or the proportion of cytoplasm and nucleus. A priori knowledges about the average size and shape of the cells, allow to work on sub-images extracted from the original image, by cutting a square around the nucleus previously segmented [KGHS96, SR03]. Thus, assuming that each sub-image has only one white blood cell using some restrictions on the shape and the colour information it is possible to perform a clustering around the nucleus. Unfortunately this assumption is not always true, in fact it is possible to find more than one nucleus on a sub-image that affects the result of the clustering. An improvement of this approach works on the whole images without using the a priori knowledges about the size and shape. This is possible thanks to the distance transform [JRQ03] that associates to each pixel of the binary image its distance from the border. Thus, the maximum distance obtained can be used as a marker for a subsequent segmentation step [MdSV$^+$97], or the distance image can be used directly as a shape delimiter for a watershed segmentation [Lin]. The main drawbacks of these approaches is that the cells should be segmented

perfectly since they work directly on the binary images. Furthermore, since the distance transform in this case works like a shape delimiter, it is able to separate only small agglomerates of cells that should have an almost circular shape.

The step of feature extraction is very important in the analysis of peripheral blood cells, as only with significant features, the system will be able to discriminate the various types of cells, the cells affected by diseases from the normal ones or the presence of other abnormalities in the blood. The idea is to extract the descriptors that best approach to visual patterns to which pathologists refers to. The colour and texture descriptors are certainly the most discriminatory features of blood cells. Generally this kind of images are acquired using the RGB color space that allows a good discrimination of platelets, red blood cells and white blood cells, especially if the analysis is extended to all RGB channels [AS02]. However, the discrimination of subclasses of cells such as various types of leukocytes, usually requires the use of other colour spaces, taking into account all the colour channels or only the most discriminatory such as the H channel of the HSV colour space [HSP02]. Furthermore, geometrical features can be used to discriminate cells with abnormal size or with irregular shape. Other geometrical features have been specifically proposed to classify the type of nucleated cells, since they can be distinguished not only using the ratio between cytoplasm and nucleus [PS04, Sco05, Sco06] but also extracting the number of lobes of the nucleus. Extract the number of lobes means count the more substantial parts of the nucleus connected by thin filaments. This is not an easy task, since the connection between the lobes sometimes can appear larger than normal and thus, also a skilled operator, could count two small lobes as it was only one. A good approximation of the number of lobes can be obtained by iterative erosion of the nucleus. The number of lobes correspond to the connected components with an area bigger than a prefixed parameter [PS04].

Once the features have been extracted they must be inserted in a process which classifies cells based on haematological concepts [BCPP00, SDH$^+$91]. Different learning methods have been used to classify blood cells, but above all very different choices about the number of classes have been taken. The choice taken more frequently is the binary classification to distinguish healthy white blood cells from abnormal ones making use of the SVM classifier, which is excellent in the separation of binary classes with pattern very close in space [MP10b, MPS10, MP10a, MPS14]. Instead, when the number of classes is higher the most used classifier have been the Neural Networks, performing a separation into the five types of leukocytes [Sco06] or performing a separation into 7 classes, lymphocytes, neutrophils, eosinophils, other (monocytes and basophils), lymphocytic leukaemia L1, lymphocytic leukaemia L2 and non-lymphocytic leukaemia [BSa08].

In the following chapter you will see the algorithms proposed to create a full automated system for peripheral blood images. For clarity each step of the whole process has been discussed separately following the order mentioned before and used for most of the methods presented in literature. Furthermore in order to make a comparison with the state-of-the-art, each step of the proposed method has been

tested using the same dataset presented below.

8.5 Dataset

One problem encountered during the testing of the proposed approaches is the absence of public datasets. In fact, many authors have tested their system with only a few samples of images, or with their own datasets which are not publicly available. This disadvantage does not allow a direct comparison with the results obtained by various proposed systems, limiting the reproducibility of the innovations proposed in other similar systems. For these experiments the Acute Lymphoblastic Leukaemia Image Database (ALL-IDB), proposed by Donida Labati in [LPS11] has been used. ALL-IDB is a public image dataset of peripheral blood samples of normal individuals and leukaemic patients and it contains the relative supervised classification and segmentation data. So, this dataset allows not only to assess the quality of the algorithms for cell counting but also to assess the ability to discriminate the white blood cells affected from leukaemia from healthy ones. The sample images have been collected by the experts of the M. Tettamanti Research Centre for childhood leukaemia and haematological diseases, Monza, Italy. The ALL-IDB database has two distinct versions. The first version the ALL-IDB1 contains full size original images that can be used both for testing segmentation capability of algorithms, as well as the classification systems and image preprocessing methods, while in the second version the ALL-IDB2 is a collection of cropped area of interest of normal and blast cells that belong to the ALL-IDB1 dataset, so it can be used only for testing the performances of classification systems.

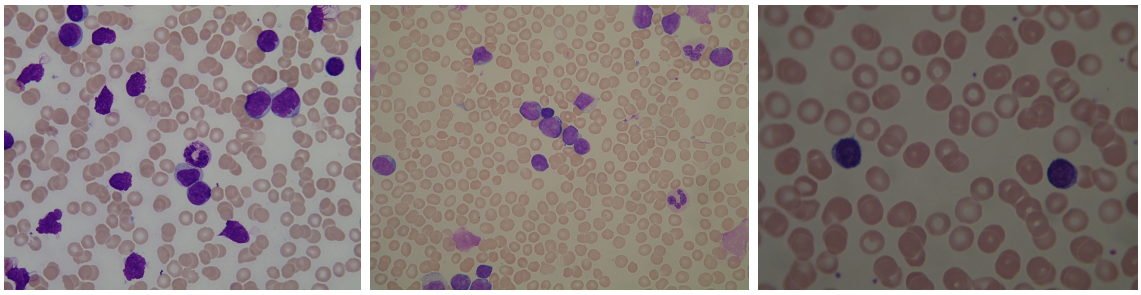

Figure 8.4: Sample images from the ALL-IDB1

In both versions of the dataset, each image has an associated text file containing the coordinates of the centroid of each candidate lymphoblast, which was manually labelled by a skilled operator and can be used as a ground truth for classification. The dataset ALL-IDB1 includes 108 images in JPG format with 24 bit colour depth. Most of the images in the dataset was captured with an optical laboratory microscope, with different magnifications ranging from 300 to 500, coupled with a Canon

PowerShot G5 camera and their resolution is 2592x1944. The remaining images were acquired with a microscope at a constant magnification, coupled with an Olympus C2500L camera and their resolution is 1712x1368. Some images belonging to the ALL-IDB1 are showed in Fig. 8.4.

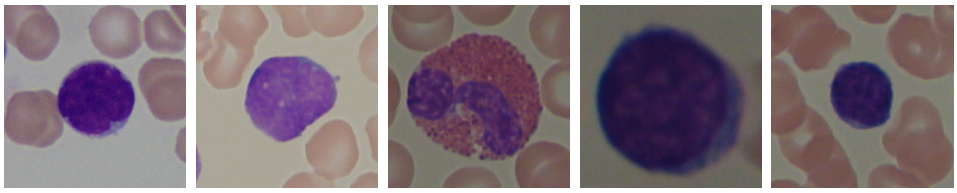

Figure 8.5: Sample images from the ALL-IDB2

As it can be seen from the sample images there are many differences both in terms of colour and illumination and both in terms of resolution and cells dimension. The dataset ALL-IDB2 includes 260 images in TIFF format with 24 bit colour depth. As said previously, these images are cropped areas of interest, containing a single leukocyte per image, belonging from the first version of the database. These images, differently from the first ones, have a standard size of 257x257, but being cropped area of them they present the same issues about colour, illumination and cells dimension, as can be seen in Fig. 8.5.

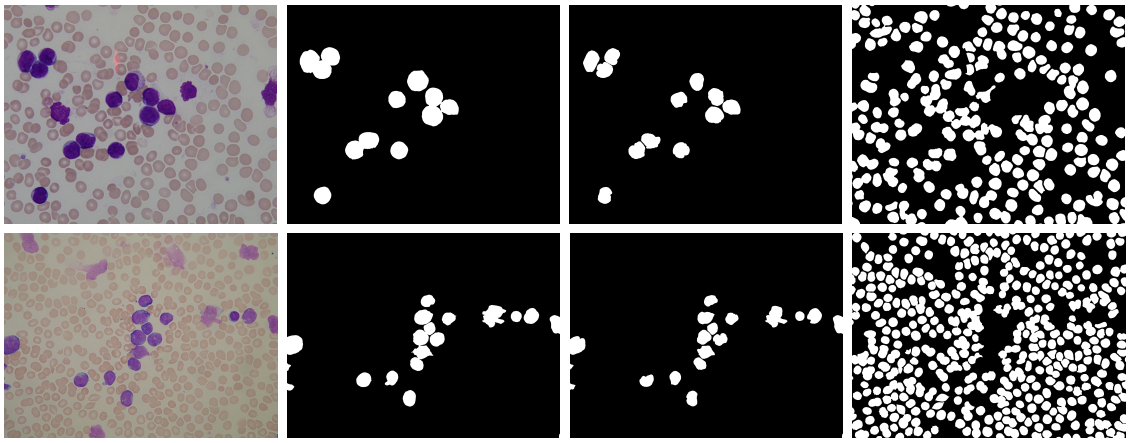

Figure 8.6: From left to right: original images from the ALL-IDB1 database, ground-truth for whole leukocyte, only nuclei and RBCs

In order to evaluate the segmentation performances of the proposed method, a subset of 10 random samples images belonging to the ALL-IDB1 have been manually segmented by skilled operator, creating two ground-truth images for each sample. These images display respectively each blood cell present in the image and the white blood cell nuclei. Fig. 8.6 shows some images belonging to the ALL-IDB1 and their relative ground-truth images. Ground-truth images have also been extracted for images belonging to the ALL-IDB2, but in this case the manual segmentation is

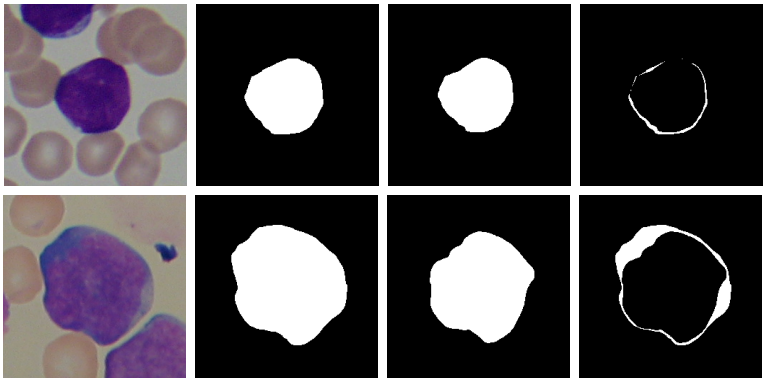

Figure 8.7: From left to right: original images from the ALL-IDB2 database, ground-truth for whole leukocyte, only nucleus and only cytoplasm

only devoted to the analysis of leukocytes, so the ground truth images display only the cytoplasm and the nucleus of the leukocyte, as it can be seen in Fig. 8.7.

Chapter 9

WBCs Segmentation

The proposed method of cell analysis starts with a segmentation step. Although many authors proposed methods of WBCs counting based on detection, using the circular Hough transform or texture analysis of the regions of interest, in this work the final goal is not only the cells counting but also the classification of cells for the diagnostic task. Thus, it is important to segment the cells regions accurately, in order to extract meaningful information concerning the boundary and the shape. As said previously, in the analysis of medical images different levels of segmentation are used. In particular for what concerns peripheral blood images two main levels can be distinguished: the level cells segmentation, which aims to separate whole cells from the background or plasma and the level of segmentation that tries to separate the various components in the cell, such as the nucleus from the cytoplasm or intracellular parasites. Several authors have proposed methods for effective segmentation of the nucleus of leukocytes, while there are few attempts of segmentation of the cytoplasm. In this chapter the segmentation techniques used to segment the whole white blood cells will be illustrated. Since during the analysis of the images and the segmentation results many issues have been observed, different approaches have been proposed in order to make further improvement and to obtain better result. Most of the proposed approaches are based on threshold operations. Although in many cases threshold is not considered a robust approach for segmentation, here special effort has been devoted to improve this kind of approaches to be robust against uneven illumination, local imprecision, different acquisition devices and staining. Threshold approaches have been chosen mostly because they are computationally not expensive, being able to segment the image quickly and independently from the number of cells, as opposed to algorithms such as deformable models, in which the number of computations heavily depends on the number of cells and the convergence is not always guaranteed.

9.1 Double Segmentation

The first realised method for white blood cells segmentation makes use of two segmentation steps performed on two different images. In contrast to other reported methods, this approach does not require separate steps of pre-processing and segmentation; it uses pre-processing steps inserted among the various stages of segmentation to make the latter simpler and more robust. As said previously, other methods aim first to identify the nuclei, which are more prominent than other components [MKA+10], and then to detect the entire membrane (i.e., by region growing [KGHS96, CLAS11, LC02]). In contrast, this approach detects the membrane first, in order to perform the further step of separation of the adjacent cells more accurately. The algorithm consists of the following phases:

- Conversion from RGB to CMYK colour model
- Histogram equalisation
- Segmentation by threshold using Zack algorithm
- Background removal operation

As said previously, the WBCs nuclei are more in contrast in the Green component of the RGB colour space. But to segment adequately the whole leukocyte this colour space is not the best choice. In fact it has been observed that leukocytes are more contrasted in the Yellow component of CMYK colour space [PR13c], because the yellow colour is present in all elements of the image, except leukocytes (Fig. 9.1 shows two examples). So, the first phase consists in converting the original RGB image into the CMYK image (2.2). The Yellow channel presents an high contrast between leukocytes and the other image components, but at the same time it presents an histogram with intensity values too close between them. For this reason a redistribution of the intensity values is necessary to make the subsequent segmentation process easier. Then, a histogram equalisation (2.1) has been used to redistribute the intensity value equally (see Fig. 9.1). The result of this operation seems to present some noise, as histogram equalisation enhances the non uniformity of the background, typical of this kind of images. At the same time histogram equalisation creates clear valley between the modes of the histogram preserving also the shape of the white blood cells. Segmentation is achieved using a threshold automatically calculated by the Zack algorithm (3.1.2), since this algorithm is particularly effective when the histogram displays clear valleys between high and weak peaks. The results of the segmentation with threshold are very good as it can be seen in Fig. 9.1. Then the complement image is calculated to obtain WBCs on a dark background.

Considering that the images captured at the microscope suffer from uneven lighting, it becomes necessary to remove the background because the segmentation methods based on threshold may suffer heavily for this problem. Some approaches for background extraction have been described in literature, but they use a collection

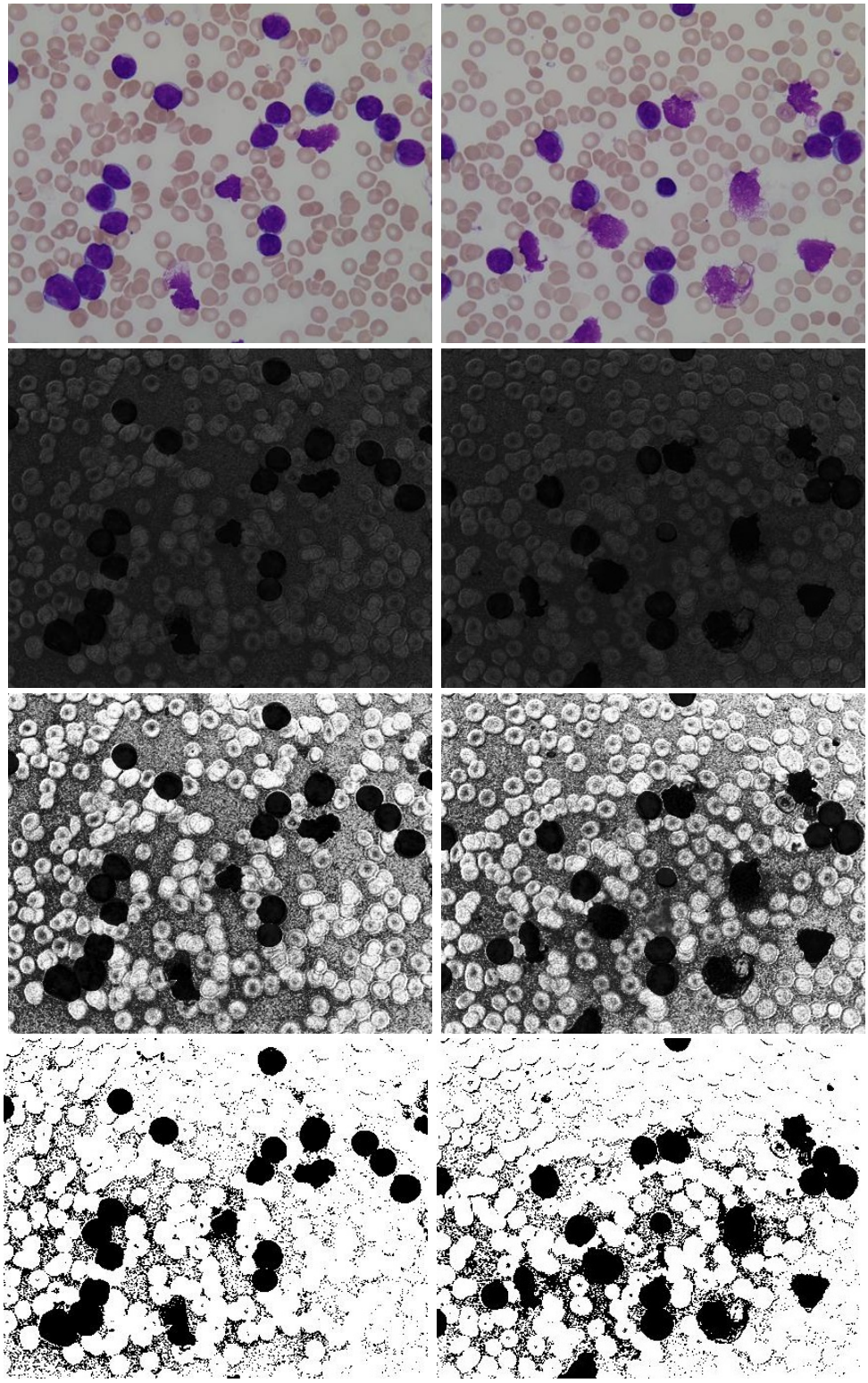

Figure 9.1: Top to bottom: original blood sample images, Y component images, histogram equalisation results and segmentation results.

of images captured with the same camera and the same microscope for estimation of the pixels belonging to the background [Sco05], while others have a very high computational cost, not necessary for this application. The proposed approach involves the use of an automatic threshold applied to a single image. This threshold could be applied to the grey level image or to the Green channel of the RGB colour space. The choice is made automatically by the algorithm that analyses the contrast of the two images looking for the one presenting a less condensed histogram, in order to avoid further operations of grey level redistribution. The threshold value is calculated again using the Zack algorithm (3.1.2). Fig. 9.2 shows how the result may not be accurate in all aspects, in fact even the centre of red blood cells can be detected as background. However, this does not preclude the achievement of effective background removal because the goal now is to preserve only the WBCs present in the image. Background removal can be performed using simple arithmetical operations, by subtracting the background just extracted from the image segmented starting from the Y component. Obviously, the background removal process does not produce a clean result for the whole image. To clean up the image, the area opening operation (3.4) has been used, which allows to delete all the objects with a size smaller than the average size of the objects in the image.

As it can be seen in Fig. 9.2 the segmentation results of this approach are really effective, being able to individuate and segment all the WBCs.

9.2 Fuzzy Threshold

Nevertheless, when the image presents noise or imprecision the segmentation process becomes complex, leading to ill defined regions. So, in this case it is appropriate to avoid a crisp threshold and to prefer a fuzzy threshold. Furthermore, images acquired from digital microscope are affected from uneven lighting, with a central area very bright, actually caused by the light of the microscope and the presence of shading area more marked towards the corners. Very few papers are designed to achieve a robust segmentation performance under uneven lighting conditions. Low pass filter have been used for background removal [Sco06] in order to improve the segmentation results based on threshold operations. In [Cha10] instead a local threshold based on FSs have been used in order to calculate a different threshold value for each windows avoiding problems related to uneven illumination. The obtained results are very encouraging, but they have been obtained making some arithmetical operation between the threshold value in order to obtain good segmentation results. Moreover the optimal threshold computation is particularly expensive. This approach have been realised taking inspiration from [Cha10], thus using the local fuzzy threshold to improve the segmentation accuracy. To better understand this approach a brief introduction on FS theory is needed.

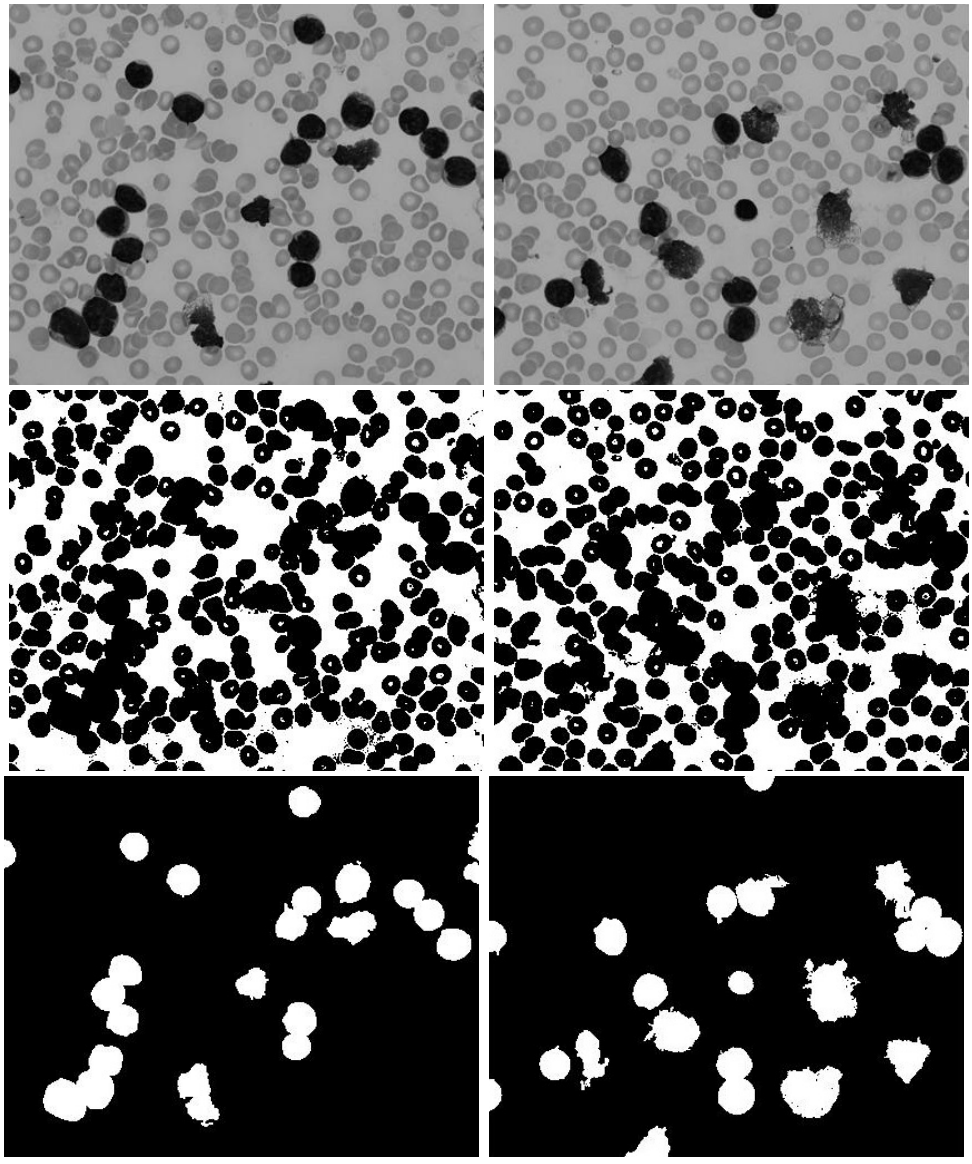

Figure 9.2: Top to bottom: grey level images, background identification results and background removal results.

9.2.1 On Fuzzy Sets

Fuzzy Sets (FSs) theory was proposed in 1965 by Zadeh [Zad75a, Zad75b, Zad75c] as a novel way of representing indeterminateness, in order to provide a tool able to describe the characteristics of a too complex or ill-defined system to admit precise mathematical analysis. This theory reflects also the human reasoning which is not based on traditional two-valued or even multi-valued logic, but it is based on fuzzy logic truths. In traditional set theory, if we have a set A such that $\{0,1\} \in A$, an element has only two options: it belongs totally to the set A or it doesn't belong at all to the set A. This set is called crisp set. According to Zadeh the concept of set is more flexible introducing the idea of membership. So, if we have a FS A in a universe X, and x is some element in X, the degree of membership μ is some real number between $[0,1]$, indicating how much x belongs to A. A FS A in a universe X is a set of ordered pairs

$$A = \{(x, \mu_A(x)) | x \in X\} \tag{9.1}$$

where the function $\mu_A(x) : X \to [0,1]$ defines the degree of membership of the element $x \in X$. It is obvious that an element presents also a degree of non-membership, which is automatically set to 1 minus the degree of membership. Human being who expresses the degree of membership of a FS not always can be sure that the degree of non-membership of an element in a FS is equal to the complement to 1. That is to say, there may be some hesitation degree. Based on this idea, Atanassov [Ata86] in 1983 introduced the concept of Intuitionistic Fuzzy Sets (IFSs) that contains both a membership and a non-membership degree. An IFS A in a universe X may be represented as

$$A = \{(x, \mu_A(x), \upsilon_A(x)) | x \in X\} \tag{9.2}$$

where the functions $\mu_A(x), \upsilon_A(x) : X \to [0,1]$ define, respectively, the degree of membership and the degree of non-membership of the element $x \in X$, with the condition $0 \leq \mu_A(x) + \upsilon_A(x) \leq 1$. Using the definition (9.2) it can be observed that every FS is a particular case of IFS

$$A = \{(x, \mu_A(x), 1 - \mu_A(x)) | x \in X\}.$$

Afterwards, a third parameter for the hesitation degree $\pi_A(x)$ has been proposed [SK00], which gives the possibility to take into account also lack of knowledge or personal error that can arise in computing distances among FSs. Using this parameter the representation of an IFS A can be rewritten as

$$A = \{(x, \mu_A(x), \upsilon_A(x), \pi_A(x)) | x \in X\} \tag{9.3}$$

with the condition $\mu_A(x) + \upsilon_A(x) + \pi_A(x) = 1$. So, it is obvious that $0 \leq \pi_A(x) \leq 1$ for each $x \in X$. As important contents in fuzzy mathematics, similarity measure and distance measure between IFSs have attracted many researchers. The intuitionistic fuzzy divergence (IFD) between two IFSs measures how much two sets differ

between them. Many divergence formulae have been proposed in literature, such as a generalization of the distance measures from classical FS theory, Hamming distance, normalized Hamming distance, Euclidean distance and normalized Euclidean distance, by incorporating the non-membership function [Ata86] and then adding the hesitation degree [SK00]. In [CR03] a fuzzy divergence measure between two FSs, A and B has been introduced. It is defined as the sum of the discrimination of A against B, and the discrimination of B against A:

$$\begin{aligned} D_{AB} &= D^1_{AB} + D^2_{BA} \\ &= \sum_{x\in X}[2 - (1 - \mu_A(x) + \mu_B(x))e^{\mu_A(x)-\mu_B(x)} \\ &- (1 - \mu_B(x) + \mu_A(x))e^{\mu_B(x)-\mu_A(x)}]. \end{aligned}$$

In a more recent work [Cha10] this divergence measure has been extended to IFS, including the hesitation degree. So, for two IFS A and B the divergence measure can be defined as

$$\begin{aligned} D_{AB} = \sum_{x\in X}[&2 - (1 - \mu_A(x) + \mu_B(x))e^{\mu_A(x)-\mu_B(x)} \\ &- (1 - \mu_B(x) + \mu_A(x))e^{\mu_B(x)-\mu_A(x)} \\ &+ 2 - (1 - \mu_A(x) - \pi_A(x) + \mu_B(x) + \pi_B(x)) \\ &\cdot e^{\mu_A(x)+\pi_A(x)-\mu_B(x)-\pi_B(x)} \\ &- (1 - \mu_B(x) - \pi_B(x) + \mu_A(x) + \pi_A(x)) \\ &\cdot e^{\mu_B(x)+\pi_B(x)-\mu_A(x)-\pi_A(x)}]. \end{aligned} \tag{9.4}$$

9.2.2 Global IFS threshold

Many approaches for image thresholding based on fuzzy methods have been proposed and, recently, IFSs have been used to determine the optimal threshold value for grey level image segmentation [Cha10, MPCB+13]. In fact the hesitation degree value can represent the hesitance of an expert in determining whether a pixel of the image belongs to the background or to the object of the image. Furthermore, using IFS we can transform an ill-defined segmentation problem into a precise and well-defined optimization problem, by minimizing the divergence between the actual image and the ideally thresholded image in order to find the optimal threshold. An ideally thresholded image [CR03] is an image where the background and the foreground regions are precisely segmented and all pixels present the values $\mu_B(x) = 1$ and $\pi_B(x) = 0$. So, the divergence measure in (9.4) can be rewritten as

$$\begin{aligned} D_{AB} = \sum_{x\in X}[&2 - (2 - \mu_A(x))e^{\mu_A(x)-1} - \mu_A(x)e^{1-\mu_A(x)} \\ &+ 2 - (2 - \mu_A(x) - \pi_A(x))e^{\mu_A(x)+\pi_A(x)-1} \\ &- (\mu_A(x) + \pi_A(x))e^{1-\mu_A(x)-\pi_A(x)}]. \end{aligned} \tag{9.5}$$

The main disadvantage of using a IFSs formulation is the relatively high computational complexity. In fact, IFSs are used to create an intuitionistic fuzzy, by computing the membership, non-membership and hesitation degree for each pixel. Then the final threshold can be obtained by minimizing the IFD between the original image and the ideally thresholded image. Therefore, the idea is to modify again the divergence function in order to directly process the grey levels, rather than pixels values, and to use the histogram values as weights. So, let $\{0, 1, 2, ..., L-1\}$ denote the L distinct intensity levels in a digital image, the histogram is a discrete function $h(l)$ that denotes the number of pixels with intensity l. Using this definition the (9.5) can be rewritten as

$$\begin{aligned} D_{AB} = \sum_{l=0}^{L-1} [&2 - (2 - \mu_A(l))e^{\mu_A(l)-1} - \mu_A(l)e^{1-\mu_A(l)} \\ &+ 2 - (2 - \mu_A(l) - \pi_A(l))e^{\mu_A(l)+\pi_A(l)-1} \\ &- (\mu_A(l) + \pi_A(l))e^{1-\mu_A(l)-\pi_A(l)}]h(l) \end{aligned} \tag{9.6}$$

Since the membership function denotes the belongingness of a grey level to a region, then the smaller the difference between the grey level and the mean of the region to which the pixel belongs, the greater the membership value and vice versa. The membership function μ can be written as

$$\mu(x) = 0.582(e^{1-|x-y|} - 1) \tag{9.7}$$

where the value 0.582 comes from $1/(exp(1) - 1) = 0.582$, x is related to the grey level value and y is related to the histogram mean for a certain threshold t. So, given a certain threshold t that separates the background and foreground region, the membership function can be rewritten as

$$\mu_A(l) = 0.582(e^{1-|l-m_0|} - 1) \text{ if } l \leq t, \text{ background} \tag{9.8}$$

$$= 0.582(e^{1-|l-m_1|} - 1) \text{ if } l > t, \text{ foreground} \tag{9.9}$$

where m_0 , m_1 are the average grey levels of the background and foreground (or object) region and may be written as

$$m_0 = \frac{\sum_{l=0}^{t} l\ h(l)}{\sum_{l=0}^{t} h(l)} \quad m_1 = \frac{\sum_{l=t+1}^{L-1} l\ h(l)}{\sum_{l=t+1}^{L-1} h(l)} \tag{9.10}$$

According to this definition, the membership value is higher when the difference between the grey level and the mean of the region is smaller. In order to create IFSs from fuzzy set, intuitionistic fuzzy generators are used. In this work, intuitionistic fuzzy generator is created taking into account Sugeno generating function [RS96], according to which the hesitation degree function may be written as

$$\pi(l) = 1 - \mu(l) - \frac{1 - \mu(l)}{1 + \lambda\mu(l)} \tag{9.11}$$

Under this interpretation of $\pi_A(x)$, if the expert is certain that a grey level belongs to the background or the object, then the hesitation degree must be zero. The value of $\pi_A(x)$ increases as the hesitance of the expert grows, that is when the membership degree is close to 0.5. However, the hesitance must have the least possible influence on the choice of the membership degree. Using the equation (9.11) the hesitation degree is related to the adopted λ value. In [CR03] the λ value was chosen during experiments, observing that with a too small λ value some noise was still present in the thresholded images. So, a default value of 0.8 was chosen and remained the same for each analysed image. In the same way here a default value has been taken in order to avoid further computation and observing that a too small λ value is irrelevant for the computation of the divergence function, while a too big λ value tends to increase too much the hesitation degree, moving the higher hesitance far from $\mu = 0.5$, as it can be seen in Fig. 9.3. For all these reasons, in the proposed approach $\lambda = 0.5$ has been chosen as a default value. Accordingly, in the worst case, the hesitance will have a maximum influence of 0.1. Finally, the divergence function is calculated for each possible threshold value, in the range $[L_{min}\ ,\ L_{max}]$, the minimum and maximum grey level value present in the image, respectively. Since the analysis is limited to the range $[L_{min}\ ,\ L_{max}]$ the computations of the equations (9.6) and (9.10) are optimized with summations that operate only in that range (summations that operate on all the grey levels in the histogram have been reported for clarity). The grey level corresponding to the minimum divergence value is selected as the optimal threshold value.

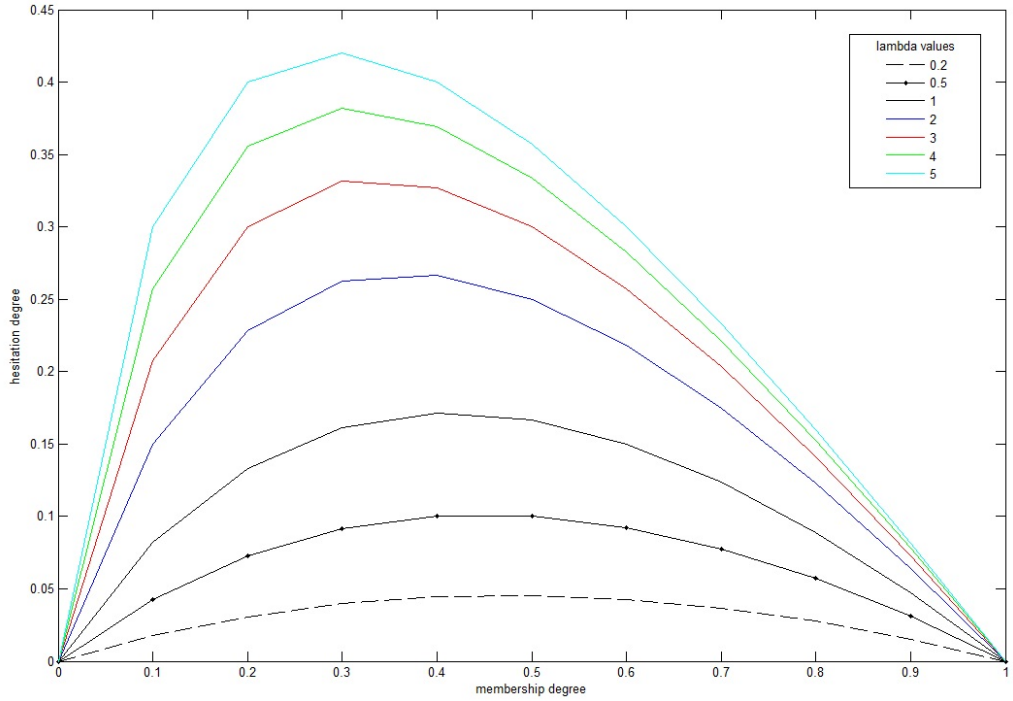

Figure 9.3: Influence of lambda value on the hesitation degree calculation.

9.2.3 Local IFS threshold

The proposed IFS threshold approach has showed excellent results if used on the whole image as a global threshold. Nevertheless images could be complex and can present histograms with more than two modes or an uneven illumination. So, local threshold (3.1.4) must be used in order to take into account local variations and multiple thresholds must be used in order to subdivide images into more meaningful regions. To combine these two approaches is not always simple and it can lead to a considerable increase of the computational cost. For all these reasons, in the proposed system, the original image has been divided into a prefixed number of equal sized sub-images. The number of sub-images chosen is 16, in order to deal with the peripheral windows differently from the central windows and to take into account the reduction of brightness or saturation at the periphery caused by camera settings or lens limitations (vignetting effect). Each sub-image is segmented separately and in order to consider also histograms with more than two modes a multiple threshold has been performed. The simplest way to realise a multiple threshold is to iterate a single threshold procedure until the number of needed threshold is reached. For example, if there are three regions in the sub-image, two threshold t_1 and t_2 must be selected, such that $L_{min} \leq t_1 < t_2 \leq L_{max}$. As showed previously, using an iterative single threshold procedure the first threshold can be easily obtained using the whole histogram from L_{min} to L_{max}. The second threshold could be obtained using only the biggest part of the segmented histogram, from L_{min} to t_2 or from t_1 to L_{max}. This approach is simple and computationally not expensive since for each iteration after the first it uses a smaller histogram. To determine how many threshold values must be selected for a single sub-image a preliminary histogram analysis has been introduced. In fact it must be noted that a sub-image could present only two modes histogram and thus a single threshold must be selected or even a single mode histogram, thus the threshold operation is no more required. An example of the sub-images variability is shown in Fig.9.4, where an original image has been divided into 16 sub-images of the same size and for each one the histogram has been computed.

As it can be noted some sub-images present a single mode histogram and many other present a two mode histogram but usually a typical blood image presents three modes identifying, respectively, white blood cells, red blood cells and plasma. Thus the histogram analysis is necessary to obtain further knowledges on the single sub-images, to be able to select appropriate threshold values and at the same time avoid iterations when they are not necessary. The histogram analysis performed is based on the average grey level (4.8) and the standard deviation (4.9). The standard deviation value is used to determine the modes number in the histogram, while the average value is used to choose the portion of the histogram for searching the second threshold. The standard deviation is firstly computed for the whole image in order to get a reference value std_{tot} that is used to define a minimum standard deviation value that a sub-image should have when it presents three modes. This

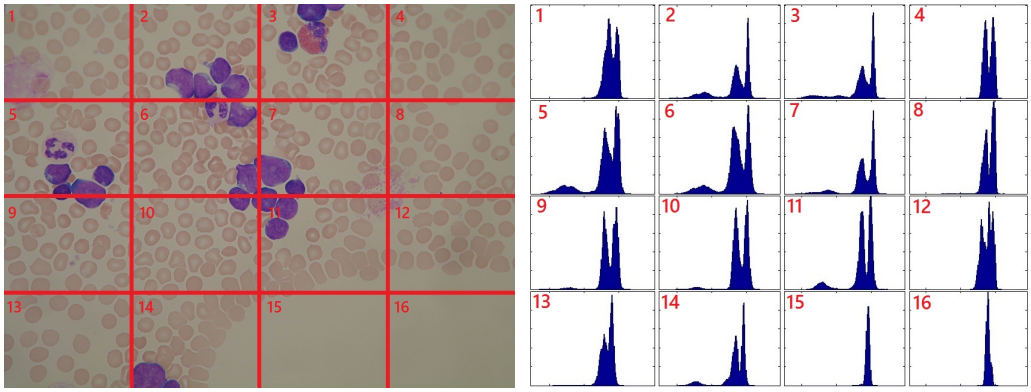

Figure 9.4: An example of the sub-images variability and histogram differences.

value, calculated as $0.6 \cdot std_{tot}$, has been selected after many experimental results, showing strong accuracy in identifying three modes histogram. At this point the first threshold value is computed in the same way for all the sub-images and only those that present a standard deviation higher than the reference value are submitted to the second step of threshold selection. The average grey level value is now used to identify which portion of the histogram has already been thresholded. In fact, depending on the cells proportion in a sub-image and so the size of the single modes, the first threshold could be selected between the first two modes of the histogram, as it happens for the fifth and sixth sub-images of Fig.9.4, or between the last two modes of the histogram, as it happens for the second and third sub-images of Fig.9.4. Thus, if the threshold value already obtained is less or equal to the average grey level, it is assigned to t_1 and used as starting point (from t_1 to L_{max}) otherwise it is assigned to t_2 and used as ending point (from L_{min} to t_2) for searching the second threshold value.

9.2.4 Experimental evaluation

The results of the proposed approach are firstly compared with two threshold approaches, that are Zack algorithms and with the fuzzy threshold method proposed in [CR03] in order to show how the use of the proposed local threshold approach can improve the segmentation accuracy. Fig.9.5 shows two original blood sample images and the results after thresholding with each technique. Both the original images are affected by an uneven illumination, in particular they present shadows in the corners and a strong illumination in the centre, a typical effect of optical microscope. In particular, in the first case, it's possible to observe that with each threshold techniques the image centre is segmented quite accurately while the image corners are badly segmented, preserving the shadows presents in the original image. Instead, in the second case, it's possible to observe that none threshold technique is able to preserve all the blood cells showed in the original image, segmenting only

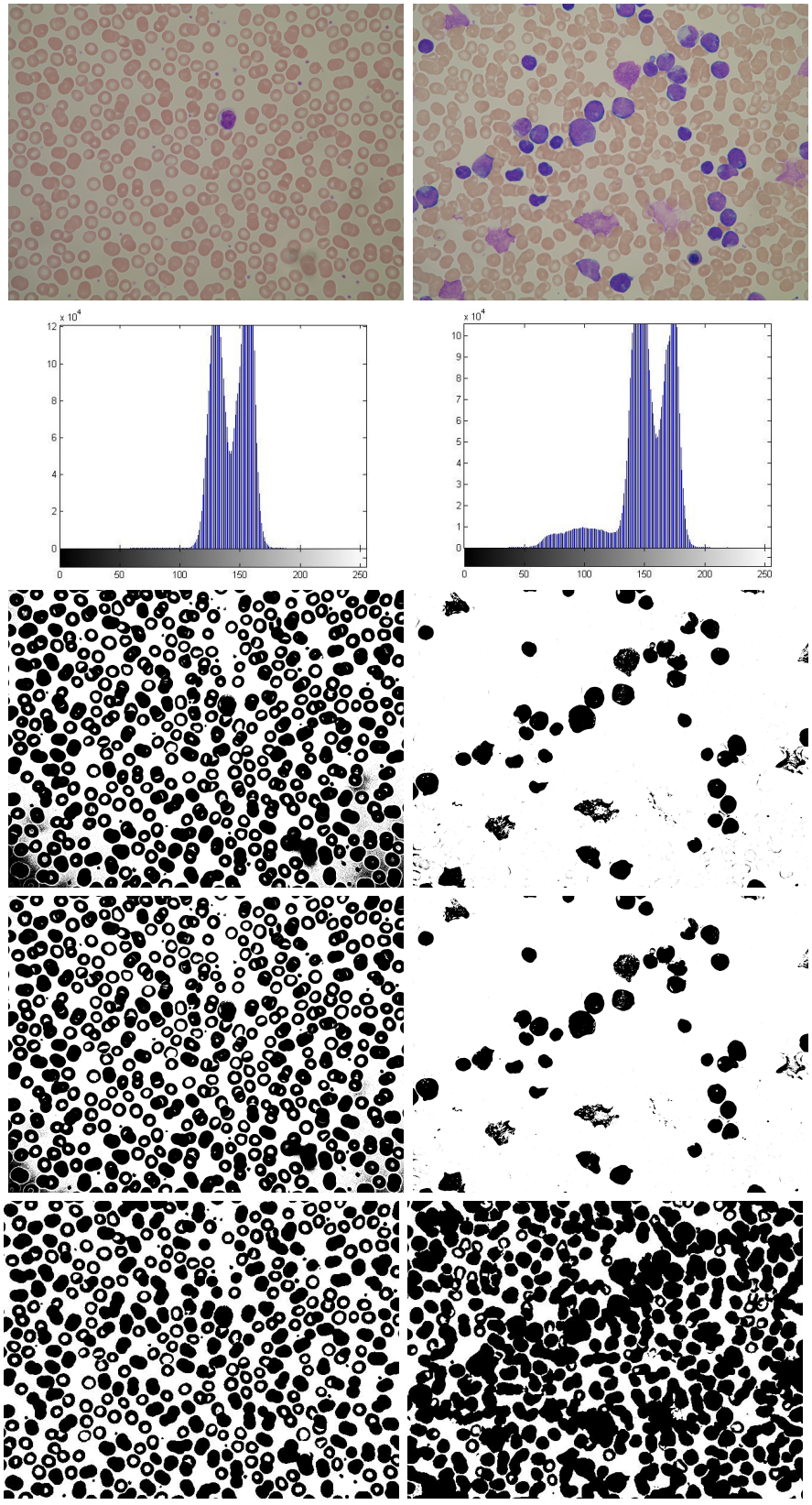

Figure 9.5: From top to bottom: original RGB images, grey level histograms, segmentation results with Zack, fuzzy and the proposed IFS approaches.

white blood cells and adding noise in the whole image, especially in the image corners. The main difference compared to the first image is the high presence of white blood cells. In this case the histogram is very different from the previous one, as it can be seen in Fig. 9.5 (the two histograms present a different number of modes). In both cases it is possible to note that using a single global threshold the segmentation result is not accurate, while using the local IFS threshold approach this problem is resolved and it is possible to segment correctly all the cells present in the images. Furthermore, as it can be seen in Fig. 8.4 there are many differences between the images present in the ALL-IDB1, not only in terms of lighting but also in terms of resolution and magnification. For all these reasons in [PR13b, PR13a] a subset of 33 images acquired from the same camera and under the same lighting conditions has been considered. With the approach proposed in this work it is possible to extend the experimentation to the whole dataset, being able to deal with local variations and with multiple modes. Some results of the whole process of multiple adaptive local threshold are showed in Fig.9.6. After thresholding operations, all the images have been cleaned using an area opening as showed previously (3.4). Then in order to highlight the results accuracy the contours of the segmented regions have been superimposed on the original images. As it can be seen with the proposed approach the segmentation results are very accurate, able to identify correctly each blood cell present in the image and performing a further segmentation level in order to detect the white blood cell nuclei. Finally, the ground-truth images previously described, have been compared with the segmentation result images in order to calculate the most common metrics for segmentation evaluation, that are: accuracy (5.1), sensitivity (5.3), specificity (5.4), FPR (5.5), FNR (5.6), precision (5.7) and F-measure (5.8). The proposed segmentation approach has been compared with the algorithms of Otsu (3.1.1), Zack (3.1.2) and Chaira [Cha10] properly modified in order to obtain two threshold values. Table 9.1 shows the average performances obtained with the ten tested samples. As it can be seen the most important values obtained with the proposed approach are higher for both the threshold values.

Table 9.1: Segmentation performances.

	Otsu 1st th	Otsu 2nd th	Zack 1st th	Zack 2nd th	Fuzzy 1st th	Fuzzy 2nd th	Proposed 1st th	Proposed 2nd th
Accuracy	0.7701	0.7832	0.7176	0.7889	0.9873	0.9129	0.9935	0.9491
Sensitivity	0.9970	0.5381	0.9907	0.6746	0.7224	0.8884	0.9119	0.9761
Specificity	0.7680	0.9550	0.7127	0.9268	0.9920	0.9279	0.9942	0.9311
FPR	0.2320	0.0450	0.2873	0.0732	0.0080	0.0721	0.0057	0.0688
FNR	0.0030	0.4619	0.0093	0.3254	0.2776	0.1116	0.0880	0.0238
Precision	0.2355	0.8054	0.1912	0.8577	0.8351	0.9037	0.8803	0.9115
F-measure	0.3118	0.5924	0.2597	0.6515	0.6486	0.8943	0.8798	0.9417

It must also be noted that the whole process of adaptive local threshold using the proposed equation is performed in less than a second and that a single global threshold could be performed on about 0.03 seconds (on average), so comparable with the

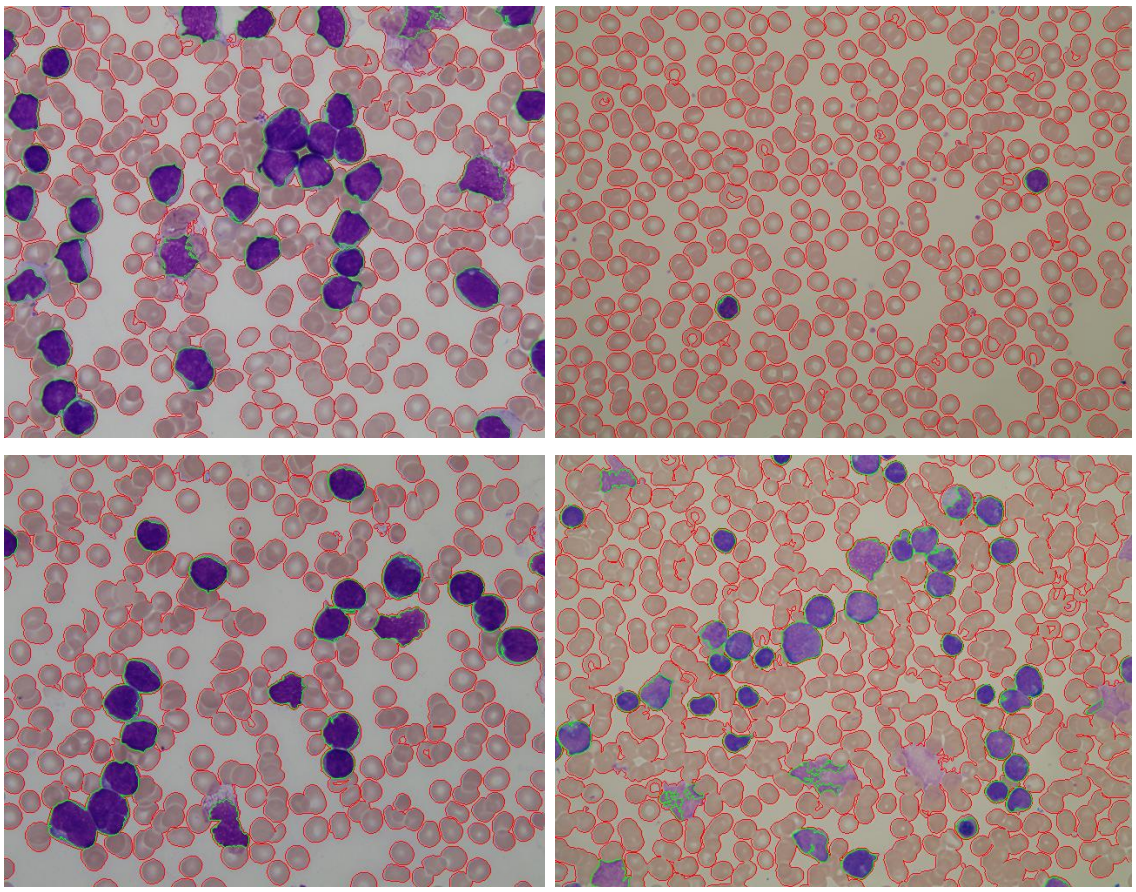

Figure 9.6: Original images superimposed with the contours of the segmented region with the first threshold (in green) and the second threshold (in red).

crisp methods mentioned before. The results obtained show that this method is able to segment correctly the image components offering excellent computational performances.

9.3 Segmentation by Samples

The approach of fuzzy threshold provides excellent segmentation results, resolving two main issues related to digital microscopy images, that are the presence of uncertainty and the presence of uneven lighting. Nevertheless, a further issue is not yet resolved, the presence of different acquisition devices and in particular the use of different staining techniques. In fact different staining procedures could produce very different blood smears and consequently very different images. Although standardization is useful to avoid superfluous differences in the features of similar

images, a robust segmentation approach can cope with the described issues. Thus, despite for comparison purposes all the proposed approaches have been tested using a single dataset, the main goal is to be able to extend the segmentation procedure to all datasets and thus to all kind of images. The main purpose is to develop an automatic machine learning approach to perform image segmentation. As for all the approaches that involve the use of machine learning techniques a training set is needed in order to create a model or to make a comparison with the unknown samples. Thus considering that the ALL-IDB dataset presents two distinct versions, one with original images and the other with singular leukocyte only, the idea is to use a part of this dataset as a training set for the learning by sampling algorithm. Firstly, the images of training set have been segmented with classic segmentation method to obtain pure samples related to the regions of white blood cells nucleus and cytoplasm, mature erythrocytes and background. As a comparison, also a method based on three Mean Shift [FH75] procedures have been realised in order to search the clustering modes corresponding to the colour of the above defined three regions. The pixels obtained from this region have been reduced in number through a Nearest Neighbour Search (NNS) by removing any duplicates or elements with distance next to zero. Then the training samples have been prepared by adapting sampling from the regions obtained from the classic segmentation phase so as to perform the training process of a multi-class SVM in order to correctly classify all the pixels of a given image. Finally, the SVM is used to segment the image for extracting whole white cells, using a classification phase by means of a classification model. Since the size of training set could be controlled and reduced in sampling, SVM training is really fast.

9.3.1 Sample Preparation

The first step of the algorithm is the creation of the training samples. It is important to note that this step is crucial in order to provide to the machine learning technique very clean samples but at the same time representative of the whole range of variation that are enclosed on a region. To evaluate the best approach to select these samples and their behaviour in relation with the SVM, both classic methods of segmentation and both the Mean Shift have been tested. Mean Shift technique was originally proposed in 1975 for general clustering problem[FH75] and then adapted and generalized for image analysis purposes [PLC09]. More recently it has been extended to low-level vision problems [FCMG00], including segmentation, adaptive smoothing and visual tracking. Even if it was developed in order to perform mode finding on clustering procedures it has been adapted to become a very effective image segmentation technique. Indeed, in contrast to the classic K-means clustering approach, it doesn't need a priori knowledges about the number of modes and clusters, that are computed by the Mean Shift procedure itself. This procedure finds the modes of an image through an iterative substitution of the pixel value with the average of its neighbourhood, according to a chosen kernel function and a distance

parameter. Thus, it attenuates shape and colour inside the images, working as a local homogenization technique. While with the classic segmentation methods the results are really clean, with a sharp separation between white blood cell nucleus and cytoplasm, with the Mean Shift method, instead, the produced results are very smooth, being just a representation of the dominant colour of the image. The Mean Shift produced a real homogenization between leukocyte nucleus and cytoplasm. Furthermore the regions belonging to the cells are really dilated, needing further steps of processing. For this reason the classic segmentation has been preferred [RLP15a, RLP15b]. Thus, in order to provide to the SVM the most accurate pixels related to white blood cell nuclei and cytoplasm all the images belonging to the ALL-IDB2 have been manually segmented by skilled operators, creating two binary masks, the first one contains the white blood cells segmented in their entirety while the second one contains only the white blood cells nuclei. From these images, the segmented cytoplasm region could be easily obtained performing a difference operation between the first image and the second one and remembering that the cytoplasm region is always placed around the white blood cell nucleus. Using these binary masks it is possible to perform the sampling of the pixel colour from the original RGB images. Once the pixels belonging to each region have been obtained, it is necessary to accurately choose these pixels with a uniform sampling, in order to consider every single image available in the group of images given for the training set. Statistics theory has revealed that, through uniform or Monte Carlo sampling, a subset could be produced to represent the entire data set approximately while retaining the distribution of data effectively [Caf98]. Since the chosen pixels must be the most various possible all over the regions, Nearest Neighbour with Euclidean distance has been used in order to provide to the SVM a smaller but more effective set of pixels. In this way the SVM should realize a more robust classification model during the training phase. In particular a NNS all over the pixels belonging from the same region has been performed, in order to remove duplicates (pixels with distance $= 0$) or too close values (pixels with distance close to 0). Also pixels that present a distance higher than the others have been removed, in fact they can be considered as outliers or noise. The distances computed with the NNS have also been used to select pixels that compose the training set. In fact not all the pixels will be used, but only a small portion that permits to obtain a fast but accurate segmentation. Pixel selection uses distances so as to consider all the possible variations in colour inside a region. So, a uniform sampling has been made taking N pixels from each region equally spaced between them.

9.3.2 SVM for segmentation

Once obtained the training sets a series of experiments have been performed in order to individuate the best SVM (5.5) implementation able to complete the segmentation task. Here the main implementations are reported. The **first implementation** works as a normal binary SVM classifier, hence there are exactly two classes in which

the pixels will be classified: the positive class groups together the white blood cell nuclei and cytoplasm pixels, instead the negative class represents pixels belonging to erythrocytes or background. Fig. 9.7 shows the segmentation result using this solution, in which the WBC is exactly recognized and segmented, but the lighter region of erythrocytes is misclassified as WBC region. The **second implementation** substantially works like the first one, with the main difference that all the pixels belonging to the cytoplasm have been excluded from the training samples, in order to avoid misclassification due to similarities with the lighter region of erythrocytes. Fig. 9.7 shows the segmentation result using this solution. Again the nucleus is well detected but for some classes of WBCs the cytoplasm is not well detected. The **third implementation** is based on the results obtained with the two previous versions. In fact the classifier needs more valid training samples for cytoplasm only. So, in last version a three-class SVM with one-vs-all approach has been performed, using both the pixels belonging from WBCs nuclei (class 1) and both pixels belonging to the WBCs cytoplasm (class 2). Thus, pixels belonging to erythrocytes or background are labelled with class 3. Fig. 9.7 shows the segmentation result using this solution in which both nucleus and cytoplasm are finally well detected.

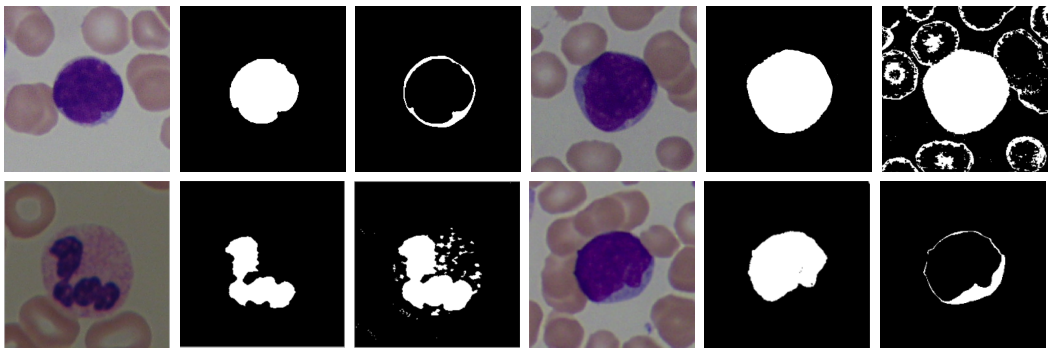

Figure 9.7: (Top) From left to right: training original image from ALL-IDB2, manually segmented nucleus and cytoplasm; test original image, segmentation result for nucleus an cytoplasm with the first strategy. (Bottom) From left to right: test original image, segmentation result for nucleus an cytoplasm with the second strategy; test original image, segmentation result for nucleus and cytoplasm with the third strategy.

Once the first (visual) results have been obtained another experimentation to assess the various features that can be used to train the classifier have been performed. In fact, even though we are talking of a segmentation technique, pixels are used as features for the SVM classifier. Until now the only descriptors used are the colour values. Although in many cases these features are enough to reach a good segmentation result, in other cases a poor feature set like this one is not able to discriminate pixels belonging to regions with wide variations in colours. Thus the first intuition has been to add the average colour values of the pixels neighbourhood, in order to

use also information about the pixels regions. The average values have been tested for neighbourhood of size 3×3, 5×5 and 7×7. For the same neighbourhood also other statistical features, that are often used for segmentation purposes, have been computed: standard deviation (4.9), uniformity (4.13) and entropy (4.14). While the segmentation accuracy highly benefits from the use of these new features, the overall system became to slow, both in training and segmentation phase. Furthermore, the step of samples selection, used to train the classifier, became too complex, due to a higher number of samples with different values. For all these reasons not all the features previously mentioned have been finally used to train the SVM. A good compromise between performances and segmentation accuracy has been found using all the statistical descriptors only for the neighbourhood of size 3×3. A further experiment has been performed at this point in order to identify the most suitable kernel and parameters for the multi-class SVM. Thus, through a 10 fold cross-validation each time the original training set has been divided into two subsets, the first one was used to train the SVM and the second one was used to test the obtained model. The kernel and parameters that permitted to obtain an ideal average accuracy value was the RBF kernel 5.5 with c parameter equal to $1e3$ and γ equal to $1e1$. The final experiment has been realised to verify the segmentation performances of the proposed method. This time the whole original training set has been used to create the SVM model and the 10 random samples images previously described have been used as test set and to check the method applied to a natural image composed of several white blood cells of many different classes, showing excellent performances as showed in fig. 9.8, outperforming previous results.

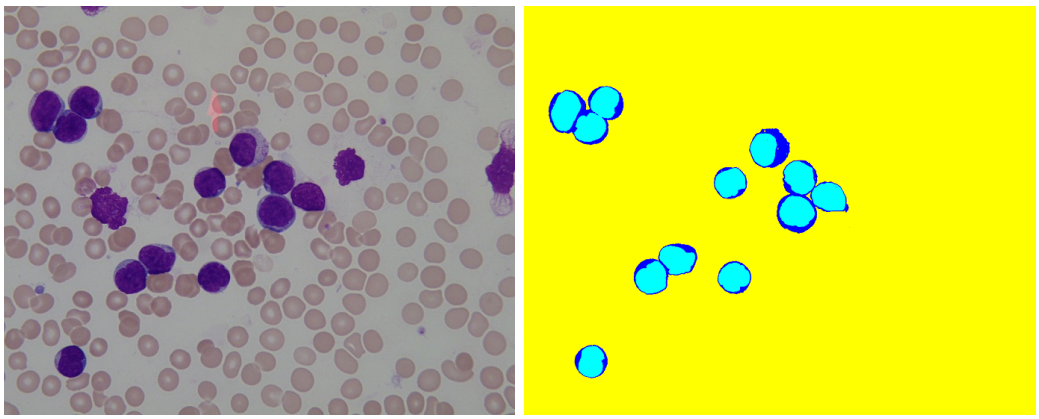

Figure 9.8: Original image from the ALL-IDB1 database and the segmentation result.

Finally, the ground-truth images have been compared with the automated segmented images in order to calculate the most common metrics for segmentation evaluation, that are: accuracy (5.1), sensitivity (5.3), specificity (5.4), precision (5.7) and F-measure (5.8). The proposed segmentation approach has been also compared

with some well know segmentation algorithms, like Otsu (3.1.1) and Zack (3.1.2). Table 9.2 shows the average performances obtained with the ten tested samples. As it can be seen the most important values obtained with the proposed approach are higher than the other segmentation approaches, obtaining an average accuracy of 97.61% that in many cases reaches the 99%.

9.3.3 System extension

Since getting manually segmented images is not so simple and cheap, an extension of this system, to be applied to each dataset of peripheral blood images, acquired in each illumination condition and with different combinations of cameras and microscopes has been proposed. The proposed extension is based on the selection of Region of Interest (ROI). Thus, making use of few original images the objects of interest (WBCs) could be selected and used as positive example for the multiple classifier. Considering that we are talking of a segmentation method based on classifiers also negative instances are needed, so the background region, that comprises red blood cells and plasma, must be selected. An example of ROI selection for positive and negative example is showed in Fig. 9.9.

Obviously for the negative example the selected regions mustn't present WBCs. In fact in this case the NNS is performed also over pixels belonging to different region, in order to avoid errors committed during the ROI selection and in order to remove pixels with close values. In this way the obtained training set should present again uniform pixel values. Note that with this approach the WBC cytoplasm and

Table 9.2: Segmentation performances.

	Otsu Method	Zack Method	Proposed Approach
Accuracy	77.67 ± 0.6	74.76 ± 3.6	97.61 ± 1.7
Sensitivity	76.70 ± 8.3	85.43 ± 6.8	98.45 ± 0.3
Specificity	85.62 ± 4.6	81.27 ± 5.4	97.56 ± 1.2
Precision	53.55 ± 12.8	79.12 ± 8.6	70.45 ± 5.8
F-measure	45.18 ± 5.3	55.15 ± 3.3	82.13 ± 2.3

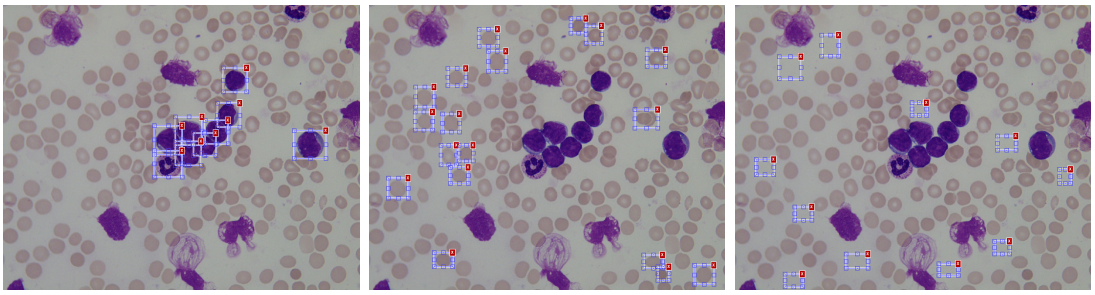

Figure 9.9: Examples of ROI selection for WBCs, RBCs and plasma.

nucleus are managed as a unique region, both because the ROI selection is not so suitable for adjacent region and both because they can be easily separated in a further step by using a simple threshold. Differently with this approach it is possible to take into account also RBCs, considering them as a different class. Thus it is possible to perform a binary segmentation or a multiple segmentation as showed in Fig. 9.10.

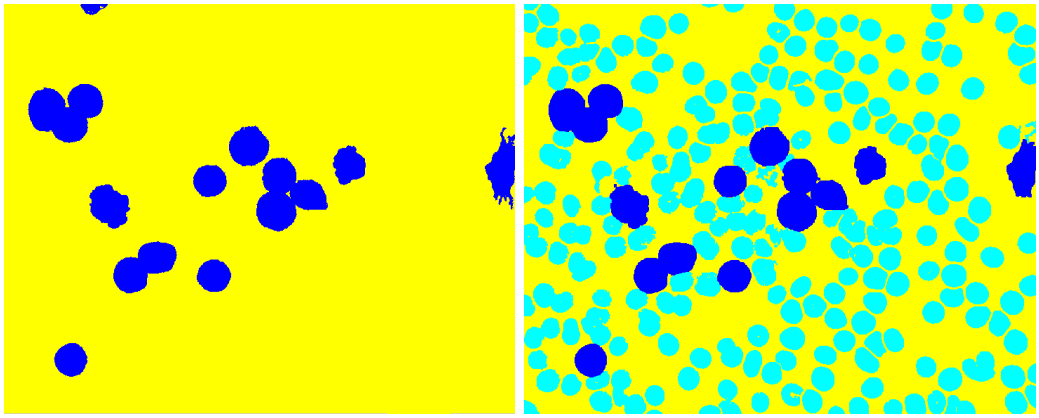

Figure 9.10: Segmentation results after ROI selection for two and three classes.

As it can be seen also in this case the segmentation is really accurate, being able to properly segment WBCs and also RBCs. Again using the 10 ground-truth images previously described, the segmentation accuracy has been computed, indicating that the segmentation by sampling is a valid approach for blood image segmentation, reaching again the 99% . The proposed approach is very accurate and robust in relation to some traditional methods, but in particular it is important to highlight the possibility to tune this approach to each dataset using only few image samples.

9.4 Discussion

In this chapter the segmentation techniques used to segment white blood cells have been illustrated . Different approaches have been proposed, since during the development phase and mostly during the analysis of the results, other issues not yet resolved have been observed and thus the need of further improvement. In fact the first proposed method, which uses two thresholds on two different colour channels, was used as a starting point. From the obtained results, very good for most of the images of the dataset used, it has been observed how the results could be influenced by the presence of different lighting condition and especially by the presence of uneven lighting within the same image. Precisely for this reason it has been chosen the use of a double local fuzzy threshold, thus able to take into account the uncertainty present in the image itself and at the same time to address the problems of local variations of light. Despite the segmentation accuracy of the fuzzy method can be

considered already excellent, a further segmentation approach has been proposed, in order to be able to manage images from different datasets and thus, able to deal not only with images that present different illuminations, but also with images acquired with different acquisition devices or from blood smears stained with different procedures. This method uses the SVM, a machine learning technique providing extreme flexibility, both because it is possible to make use of different types of kernel and both because it is possible to define a hyperplane for separating classes which guarantees a certain tolerance with respect to noise. Some visual results have been showed for each of the proposed approach and finally the most common metrics for segmentation evaluation have been computed and compared with the results obtained with some well know segmentation algorithms. To make a comparison with the state-of-the-art is not easy, since no ground truth images for segmentation are available. So, each author that proposed a new segmentation approach, tested his method with only few samples creating manually some ground truth images. Thus a direct comparison is not possible, but an overall idea of the segmentation performances can be made by comparing the approaches that used the ALL-IDB. Can be mentioned for example the method proposed in [RSBK14], that uses a k-means clustering, obtained an average accuracy of 85%. Or the method proposed in [AK13], that uses a rectangular detection using Gray Level Co-occurrence Matrix to firstly find the region containing the WBCs later segmented with a reshaping procedure on the region detected, achieved an overall accuracy of 91%. Again, the method proposed in[KAG], that uses a vector quantization technique to segment the white blood cells, obtaining an accuracy of 92%. As it can be seen the results obtained with the SVM approach outperforms the state-of-the-art, with an average accuracy of 97.6%. It must also be noted that most of the algorithms proposed in literature are focused only on WBCs segmentation, thus none of them performs a whole segmentation of peripheral blood images.

Chapter 10

WBCs Identification and Counting

As said previously, peripheral blood images present a further problem, that is the presence of cells grouped together or adjacent, in particular the presence of leukocyte agglomerates. These is an important problem since agglomerates of cells can be identified as single cells and then misclassified, considering that all the shape descriptors belonging to this regions will be misleading. Only in this phase, after a proper segmentation, it is possible to detect and separate leukocyte agglomerates. Finally, the single leukocytes can be decomposed and segmented into their components that are nucleus and cytoplasm. This process can be summarised in the following basic steps:

- Agglomerates identification
- WBCs separation
- Image cleaning

10.1 Agglomerates Identification

Once the segmented image has been obtained, it is important to find a good method to identify the agglomerates of cells. As mentioned earlier, many authors used the a priori knowledge about the typical size and shape of the cells, assuming that only agglomerates of cells should present a size very different from the average size of leukocytes. Unfortunately, this assumption is respected only for images acquired with the same camera resolution and the same microscope magnification, thus again limiting the usefulness of these approaches to a single dataset, or even a few sets of images. The main goal here is to find an approach able to identify agglomerates of leukocytes properly on each set of images, this means using one or more descriptors in order to analyse directly each region of the image just segmented. Many shape descriptors could be used to verify if a region is a single cell or an agglomerate [GW], but knowing that a WBC is typically round shaped, making an analysis of the region

roundness it is possible to individuate the agglomerates. In fact the roundness (4.5) is a measure of circularity (area-to-perimeter ratio) that is relatively insensitive to irregular boundaries and its value is equal to 1 for a circular object and is less than 1 for an object that departs from circularity. During the experimentation, it has been observed that this descriptor is really effective in shape discrimination and that roundness value lower that 0.80 indicates the presence of groups of leukocytes. The roundness value is computed for each connected component of the segmented images and using the threshold value just mentioned two different images have been created (Fig. 10.1). The first one contains only single leukocytes and thus it proceed directly to the next step of the process, while the second one contains only grouped leukocytes and thus it proceeds with the WBCs separation process. It is important to note that in some cases the second image may be empty and so the phase of WBCs separation will not take place.

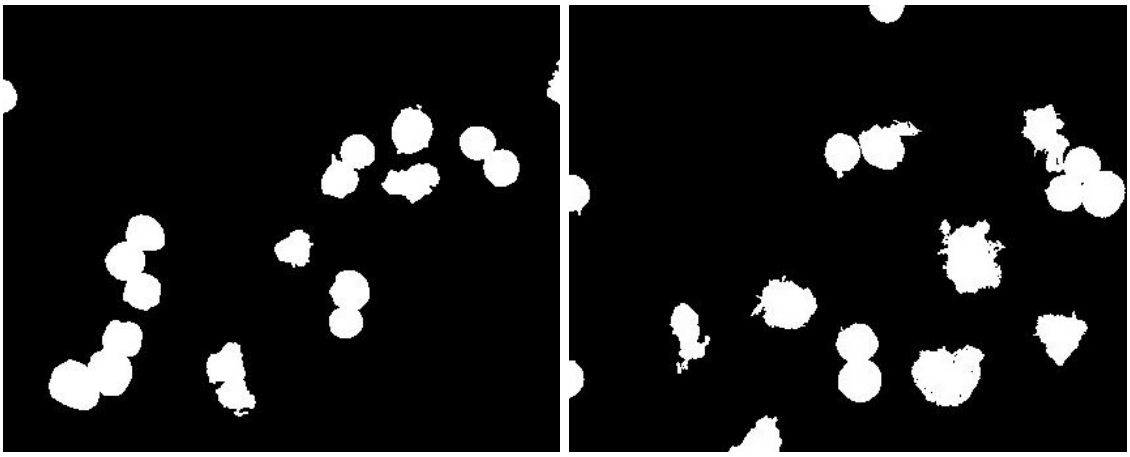

Figure 10.1: Examples of leukocytes identified as grouped.

10.2 WBCs separation

The separation between grouped cells is a crucial step in peripheral blood cells analysis, as a precise separation allows an extraction of more meaningful features. As mentioned previously, this step has be faced by other authors in two different way. The first one consists in analysing the regions of interest in the original image, while the second one is based on the analysis of the binary image just segmented. The proposed approach is divided into two parts taking into account both versions in order to realise a stronger result. The first part of the proposed approach is based on the method proposed by Lindblad [Lin], which uses the distance transform that associates to each pixel of the binary image its distance from the border. The distance image just obtained represents the cells as round shaped mountains, with the peak located in its centre. At this point the best choice is to use the distance

image directly as a segmentation function for a watershed segmentation (3.3.3). Despite a first separation between adjacent leukocytes has been made, in this way the contour of the cells tends to be inaccurate, because, using only the distance transform, none information about the single pixel has been taken into account. Thus, it only performs well in the presence of adjacent leukocytes with a nearly rounded shape, but it does not perform well in the presence of multiple complex forms, as displayed in the last image of Fig. 10.2.

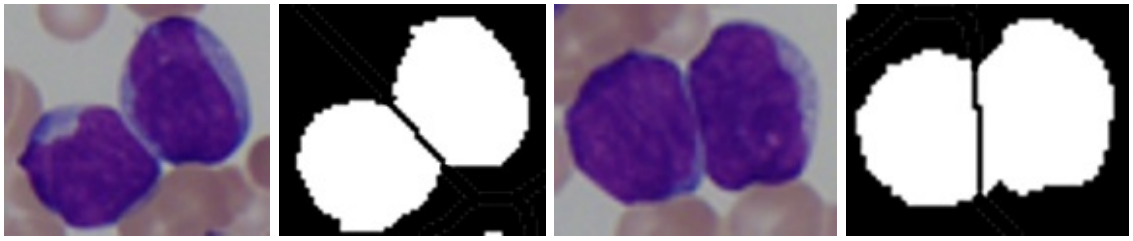

Figure 10.2: Two original blood sample sub-images and their respective watershed results.

Therefore, a second phase is needed to refine the contours extracted through the watershed transform. This phase should use other information from the grey level image, in particular a useful information comes from the local maxima image as can be seen in Fig. 10.3. The local maxima image doesn't present clean object or details, since that some maxima can be observed also on flat regions such as the background, but it can give an useful idea on the shape of the regions to be separated, in detail it provides a path for which the line of exact separation must pass. Thus the watershed lines already extracted have been adjusted in order to fit to the path described by the local maxima. Obviously many line can cut the region without override any local maxima, but by exploiting the information regarding the points of concavity it is possible to obtain a cutting line that best fits the contour of the leukocytes, that is also the shortest line that override the region. Fig. 10.3 illustrates the final separation of two adjacent leukocytes, and Fig. 10.4 shows the final separation results for the whole sample image.

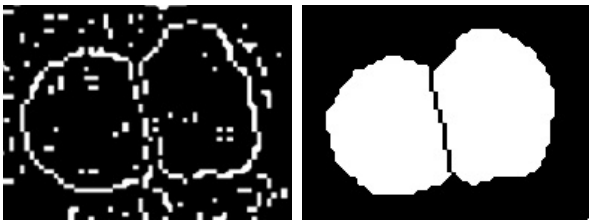

Figure 10.3: Local maxima image and final separation results.

10.3 Image Cleaning

Before being able to count the leukocytes a last step is necessary. Indeed not all the objects can be considered but only the object that are real leukocytes and only those leukocytes completely enclosed in the image. This is necessary in order to prevent errors in the later stages of the analysis process. Deleting the leukocytes that are not completely enclosed in the image is an easy task, since it can be completed by a search of the element touching the border of the image.

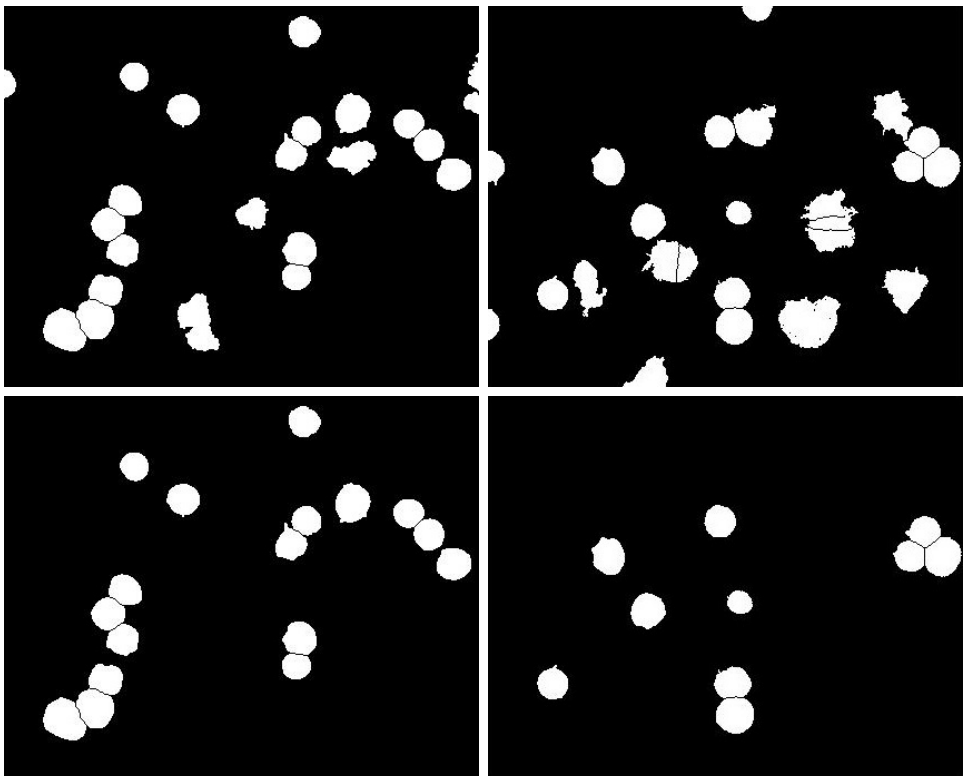

Figure 10.4: Final separation results and image cleaning results.

With this procedure many WBCs will be lost but it ensures that only those cells that can be analysed accurately will pass to the next step. The removal of abnormal components instead is a more complex task. Also in this case it is important to find a good method to identify the abnormal cells and as mentioned earlier no assumption about the typical size or shape of the cells can be made based on previous knowledge. Since the size is discriminatory for WBCs, the area is computed for each object in the image, in order to have a reference value, that is the average area. The average area is useful to determine the presence of objects with irregular size. For example, a very small area might indicate the presence of an artefact that was not removed. Alternatively, a very large area may indicate the presence of adjacent leukocytes that were not adequately separated. Thus a reference range value for the area

has been established in order to remove all these anomalies, preserving only those objects for which $0.8 * avg\_area \leq area \leq 1.2 * avg\_area$. Unfortunately abnormal objects could present a size close to the reference value and thus can bypass this check. Typically, this object are WBCs that have been altered by the staining process, so they don't present the typical morphology even a separated nucleus and cytoplasm. Area is the used in combination with another shape descriptor that is the *solidity* (4.7). This descriptor measures the density of an object. A solidity value of 1 signifies a solid object while a value less than 1 signifies an object with an irregular boundary or containing holes. The reference value for solidity is again computed by averaging the solidity value of all the object in the image, when present in a number greater than 5, otherwise a default value is used. During the experimentation it has been observed that a solidity value lower that 0.90 is able to adequately discriminates abnormal components. Fig. 11.1 shows some results after border cleaning and the removal of abnormal components. To better highlight the performance of segmentation and separation of leukocyte agglomerates Fig. 10.5 shows also some results after these two main steps on different images belonging to ALL-IDB1, superimposing the segmented leukocyte borders on the original images. As it can be seen, although the images are really different between them, both in terms of resolution and both in terms of colours, the results are really precise.

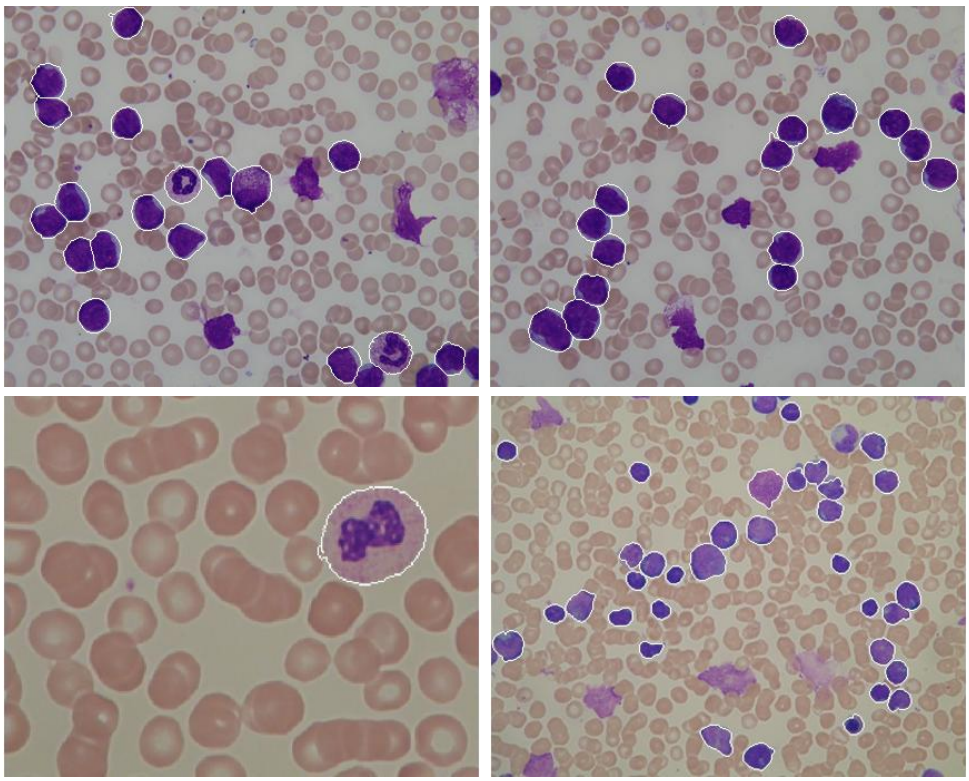

Figure 10.5: Original images superimposed with the contours of the leukocytes identified.

10.4 WBC Count

To evaluate the performances in counting the test should be performed on meaningful dataset. Considering that there are many differences among the images in ALL-IDB1 in terms of resolution, magnification and lighting, as it can be seen in Fig. 8.4 and Fig. 10.5, testing was carried out using a subset of 33 images acquired from the same camera and under the same lighting conditions, in order to evaluate the performance of the proposed method both for the detection and both for the classification steps. The performance for WBC counting are reported in detail in Table 10.1, where the results of the proposed method for each image have been compared with the results of a manual counting performed by skilled operators. As it can be seen the results are excellent in most cases, while the worst results have been obtained in images with significant overlapping between leukocytes, which are difficult even for human experts. To summarise, on 33 images containing 267 leukocytes the proposed approach has properly individuated 245, for an average accuracy of 92% .

10.5 Discussion

In this chapter the segmentation, separation and counting of white blood cells that can be applied to support some existing medical methods, like the White Blood Cells Counting (WBCC) have been illustrated . After the segmentation phase the image is analysed in order to detect agglomerates, that present an abnormal shape and size. The agglomerates are then separated by means of watershed transform, that permits to obtain a first rough contour, that is refined adapting the line of separation with the grey level of the image. This step permits to separate adequately the WBCs that can be easily counted and also analysed in a further step. The accuracy obtained in counting is very good, being able to detect 92% of the cells in the images, outperforming the state-of-the-art methods, such as the method proposed in [MLMR] that uses the circular Hough transform, without any restriction on the area of interest, obtained an average accuracy of 81%, with an high number of false positive, or the method proposed in [AK13], based on a rectangular detection using Gray Level Co-occurrence Matrix, that achieved a 88% of accuracy with an not insignificant amount of false positive. It is important to note that the method proposed in this chapter does not produce any false positive, being able to exclude all the other image regions since it is based on a previous phase of segmentation.

Image No	Manual count	Auto count	Accuracy
Image1	9	5	55%
Image2	10	10	100%
Image3	12	11	91%
Image4	7	4	57%
Image5	24	19	79%
Image6	18	18	100%
Image7	7	7	100%
Image8	17	16	94%
Image9	7	7	100%
Image10	12	12	100%
Image11	15	12	80%
Image12	12	12	100%
Image13	10	7	70%
Image14	5	3	60%
Image15	17	17	100%
Image16	16	16	100%
Image17	3	3	100%
Image18	8	8	100%
Image19	12	12	100%
Image20	2	2	100%
Image21	3	3	100%
Image22	5	5	100%
Image23	6	6	100%
Image24	4	4	100%
Image25	3	3	100%
Image26	5	5	100%
Image27	3	3	100%
Image28	2	2	100%
Image29	4	4	100%
Image30	3	3	100%
Image31	2	2	100%
Image32	2	2	100%
Image33	2	2	100%

Table 10.1: Performance of the proposed method for WBCs identification.

Chapter 11

ALL Classification

Many diseases that affect blood cells do not involve only changes in the number of cells, but they also involve morphological changes in the cells themselves. In particular lymphocytes affected with ALL present not only a different shape but also holes inside the cytoplasm and nucleus. This variation could be identified directly with a machine learning approach, but typically a further step of segmentation is performed in order to be able to manage nucleus and cytoplasm separately. Then from each cell component a feature set can be extracted and submitted to the model of classification.

11.1 Nucleus and cytoplasm selection

The separation of the cytoplasm and the nucleus is a pretty simple task, since once the leukocytes have been segmented the only two remaining components of the image are just the cytoplasm and the nucleus. Considering also that the agglomerates have already been separated it is possible to perform this task on a single leukocyte. Thus an automatic image crop is performed using the bounding box, which is the smallest rectangle that completely contains a connected component, in order to isolate a single leukocyte in each sub-image (Fig. 11.1). Another border cleaning operation is necessary to preserve only the WBC under examination. By definition, the leukocyte nucleus is inside the membrane, making it possible to further simplify this step by cropping the entire portion of the image outside the leukocyte under examination (Fig. 11.1). This procedure allows for more robust nucleus selection because it completely excludes artefacts of the selection. The nucleus selection approach takes advantage of Cseke's [Cse92] observations, which demonstrated that WBC nuclei are more in contrast on the green component of the RGB colour space. However, in this colour space, the threshold operation described by Otsu (3.1.1) does not produce clean results, especially in the presence of granulocytes, because granules are selected erroneously as part of the nucleus. To avoid this issue, the binary image obtained from the green component is combined with the binary image obtained

from the a* component of the Lab (2.6) colour space via a threshold operation performed again with Otsu algorithm. The mask obtained allows to clearly extract the leukocyte nucleus. Finally, to obtain the cytoplasm, a subtraction operation is performed between the binary image containing the whole leukocyte and the image containing only the nucleus (Fig. 11.1).

Figure 11.1: Left to right: grey level sub-image, binary sub-image, whole leukocyte sub-image, nucleus sub-image and cytoplasm sub-image.

11.2 Feature extraction

In this phase, the goal is to transform the images into data and then to extract information reflecting the visual patterns that pathologists refer to, while simultaneously extracting the descriptors that are most relevant to the subsequent classification process. To this end, three different types of descriptors from the previously calculated sub-images have been extracted: shape features, colour features and texture features. Starting from binary sub-images of the nucleus and cytoplasm, shape descriptors have been extracted, such as *area*, *perimeter*, *convex area*, *convex perimeter*, *major axis*, *minor axis*, *eccentricity* (4.1), *elongation* (4.2), *rectangularity* (4.3), *compactness* (4.4), *roundness* (4.5), *convexity* (4.6) and *solidity* (4.7). Shape descriptors have been extracted for the whole leukocyte and for nucleus only, for a total of 30 shape descriptors. To these classical measures two specific measures for the analysis of leukocytes have been added: the ratio between the area of the cytoplasm and the nucleus and the number of nuclear lobes. As said previously, to extract the number of lobes, Scotti [Sco06] proposed an approach using repeated erosions until the correct number of lobes is reached. In a similar manner, the proposed approach makes use of the ultimate erosion of the binary image [Ser83a, Ser83b], which consists of the regional maxima of the Euclidean distance transform of the complement of the binary image (Fig. 11.2). Notably, the number of lobes remains unchanged. The total number of shape features is than 32.

The main disadvantage of shape features is that they are susceptible to errors in segmentation. Thus, these descriptors are used together with regional descriptors less susceptible to errors such as chromatic and texture descriptors. Both chromatic and texture descriptors have been extracted from the grey level images, using the binary image as a mask. Thus, for each segmented leukocyte only the pix-

Figure 11.2: The binary image of the nucleus and the result of the extraction of the number of lobes obtained through iterative erosion and through ultimate erosion.

els belonging to it have been taken into account for feature computation [PCR14]. Colour descriptors are generally the most discriminatory features of blood cells, so all features extractable from the histogram have been computed, that are *mean* 4.8, *standard deviation* (4.9), *smoothness* (4.10), *skewness* (4.11), *kurtosis* (4.12), *uniformity* (4.13) and *entropy* (4.14). Also chromatic features have been extracted for the whole leukocyte, for nucleus and cytoplasm only, a total of 21 chromatic descriptors. However, the descriptors based only on histograms frequently have drawbacks, as they do not provide information regarding the mutual position of the pixels. Some objects have a repeating pattern as the primary visual characteristic, so it is necessary to consider both the intensity distribution and the position of the pixels having a similar grey level. Then, the GLCM with distance $d = 1$ and angles $\theta = [0°, 45°, 90°, 135°]$ have been computed to extract the 13 features proposed by Haralick [HSD73] plus the 7 descriptors proposed by [ST99, Cla02]. The texture descriptors evaluated are: *angular second moment* (4.16), *contrast* (4.18), *correlation* (4.19), *variance* (4.20), *inverse difference moment* (4.21), *sum average* (4.23), *sum variance* (4.24), *sum entropy* (4.25), *entropy* (4.26), *difference variance* (4.27), *difference entropy* (4.28), *measure of correlation 1* (4.29), *measure of correlation 2* (4.30), *mean* (4.44), *difference average* (4.45), *autocorrelation* (4.46), *maximum probability* (4.47), *cluster shade* (4.48), *cluster prominence* (4.49) and *product moment* (4.50). The total number of texture descriptors used is 80.

11.3 Classification

The first model chosen for classification of ALL has been the SVM (5.5) [PR13b], because this model is particularly suitable for binary classification problems for which the separation between classes depends on a large number of variables. As a starting point the SVM was used with the standard configuration suggested by Hsu, Chang, and Lin in [HCL03], providing SVM novices with a recipe to rapidly obtain acceptable results. However, the proposed approach is not good enough in some situations, in particular this classifier needs a process of tuning in order to individuate the best kernel function and the optimal parameters. For this reason the SVM has been tested with the most common kernel: linear (L), quadratic (Q), polynomial (P)

and Gaussian radial basis (R). For each kernel function, the parameters were tuned using optimization techniques in order to find the maximum accuracy value. In all the configurations the SVM was trained with the one-vs-rest approach. To evaluate the goodness of SVM models, the results were compared with other classifiers such as k-NN (5.1) using the Euclidean distance measure with different values of k, Decision Trees (5.2) and Naive Bayes (NB) (5.3) by a Gaussian (G) and kernel data distribution (K). In addition to the type of algorithm used to induce the model, the performance of a model also depends on the size of the training and the test set. In particular, as the size of the training and the test sets decrease, the performance of the model depends on their specific composition, resulting in higher variance. Therefore, given the small size of the used dataset, the performance of the models is evaluated using a k-fold cross-validation re-sampling technique (5.6). Considering $k = 10$, the whole dataset is randomly divided into 10 folds. The cross-validation process is repeated 10 times, using a different sub-sample as the validation data for testing the model and the remaining k-1 sub-samples as the training data each time. Finally, the 10 performances from the folds are averaged to achieve a single estimation. Once the estimate of instances predicted by the model of classification is obtained, it is possible to evaluate performance by comparing with the real class of instances, which, in this case, compares the class predicted by the classification model for a certain WBC with the class assigned to it by an expert haematologist. In a binary problem, as in this case, the instances are subdivided in positive and negative. For this particular problem, the instances have been defined as positive when the WBCs were affected by leukaemia and negative when the WBCs were not suffering from leukaemia, and based on this definition, the *accuracy* value (5.1) has been calculated. Although accuracy is the most widely used metric, it considers each class of equal importance. Often, as in this case, it is more appropriate to use a metric that places the most importance on the correct classification of positive instances. Clearly, in this case the most importance is placed on the correct classification of WBCs affected by leukaemia. For this reason also the *sensitivity* (5.3) value is used. Considering that in the early stages of the analysis 245 leukocytes have been properly individuated, from all the sub-images containing individual leukocytes, a feature matrices of size 132x245 have been extracted containing the previously described features. Obviously to each leukocyte is associated also a class label that has been assigned by skilled operators. The class label is fundamental to test the classification performances of the system. Many experiments have been realised to test the best configuration of features and the best classifier [PR13b, PR13a, PCR14]. Here the main experiment that permitted to individuate the best feature set and to confirm the excellent performance of SVM model for leukaemia detection have been reported. In particular in Table 11.1 it is highlighted the contribute that arises from each feature set, by testing chromatic features and texture features only and finally the whole feature set.

	Colour Features	Texture Features	All Features
SVM-L	0,856 ± 0,006	0,887 ± 0,03	0,901 ± 0,006
SVM-Q	0,813 ± 0,018	0,858 ± 0,011	0,9 ± 0,005
SVM-P	0,83 ± 0,01	0,856 ± 0,009	0,901 ± 0,007
SVM-R	0,884 ± 0,006	0,906 ± 0,011	0,932 ± 0,008
k-NN	0,771 ± 0,006	0,724 ± 0,014	0,855 ± 0,009
NB-G	0,834 ± 0,006	0,852 ± 0,002	0,85 ± 0,003
NB-K	0,844 ± 0,006	0,864 ± 0,002	0,885 ± 0,006
tree	0,866 ± 0,014	0,806 ± 0,019	0,863 ± 0,02
Mean	0,852 ± 0,007	0,844 ± 0,009	0,873 ± 0,009

Table 11.1: Performance on single and whole feature sets.

11.4 Discussion

In this chapter a CAD system for ALL detection has been illustrated. This system analyses each WBCs singularly in order to detect the morphological changes that the cell presents if affected. To better highlight this variation a further step of segmentation has been performed, in order to be able to manage nucleus and cytoplasm separately. Then from each cell component the feature set can be extracted and submitted to the model of classification. In order to find the best implementation many feature sets and many classifiers have been tested. As it can be seen from Table 11.1, each classifier benefits from the combination of the feature sets, in particular the SVM with Gaussian radial basis kernel that outperforms the others, reaching an accuracy of 93.2%. Moreover, the sensitivity value obtained in the test phase using the SVM classifier was never below 0.95 and reached a maximum value of 0.987 with the SVM-R classifier. The obtained results are comparable to those obtained by Deore [DN13], one of the few authors who used the ALL-IDB to test their method for classification of leukocytes affected by ALL. In fact, at the end of the classification stage, their accuracy value reached 93.6%. Unfortunately, this work does not provide any detail about the segmentation method used for WBCs or any accuracy value for their identification, so it is not possible to determine the number of samples used to train the classifier.

Part IV

Conclusions

Conclusions

This thesis has addressed the visual analysis of bodily fluids and tissues focused on diagnosing diseases using a microscope, a crucial step to confirm if and which illness is present. The main purpose has been the analysis of the outstanding issues in a CAD from digital microscopy images, studying some possible solutions and proposing strategies and algorithms applied in two specific use cases: histology and haematology. Special efforts have been focused on strategies to represent with meaningful information the visual content of digital images. Indeed this issue is very important in artificial vision and becomes further challenging in medical imaging, considering that there is not a colour standardization for the staining and acquisition of digital slides. In fact, there are several colour differences or intensity variations between different slides, due to the quality of the biological sample and the sample preparation, such as the quantity of dye used during the staining procedure, or due to different acquisition systems and the image capturing parameters, such as the environment illumination. Furthermore, mainly in peripheral blood images, such variability may be present in the same slide, due to the presence of uneven lighting caused by the microscope light. Thus, by computing descriptors that ignore this variability, it is possible to extract from the images more general information that may be used by conventional learning models for distinguishing different biological concepts, avoiding any dependency on specific dataset.

In particular, for what concerns the histology image analysis, a general classification framework, able to manage different medical problems with an high accuracy has been proposed. The proposed framework, differently from other CAD systems for histology images, avoids the use of segmentation method or region of interest detection. This is because every segmentation algorithm, due to the complexity of histology images, produces an huge number of regions and structures, too difficult to manage singularly. The overall procedure instead is totally based on the analysis of textures, being the most suitable descriptors to analyse the tissue structure. Also, great importance has been given to the analysis of colours, considered one of the most interesting contents to be analysed in the histological images, studying not only the internal correlation of various colours, but also by analysing the correlation between different colours. For this purpose also different colour spaces have been evaluated in order to assess which one leads the better results. The feature set has been obtained by generalizing some existent grey scale approaches to colour images,

by decomposing the original colour image in separated channels and then recombining them in pairs. Since the number of features extracted was really high a feature ranking step has been performed. This step realized using an ensemble approach also permitted to highlight the best descriptors for each feature subset being able to combine them in order to build a new feature set producing excellent results and outperforming the existing methods present in literature for all the tested databases. The proposed framework has been tested using five very different public biological image databases, that present different medical problems and so they represent different classification problems. The obtained results demonstrated the effectiveness and versatility of the proposed approach, suggesting also that it can be used in the analysis of many other types of histology images.

For what concerns the peripheral blood image analysis special efforts have been made to address the counting and in particular the segmentation issues, proposing different segmentation algorithms able to isolate the cell of interest from images acquired in different illumination conditions and stained with different dye. The experimental results demonstrate that the final approach is very accurate and robust in relation to some traditional methods, being able to obtain an average accuracy of 97.6% that often reaches the 99%. The results in this phase have also permitted to correctly identify and count the WBCs, that can be directly used to support some existing medical methods, like the WBCC. The identification of single WBCs is also important for the diagnosis of leukaemia, for which the cell components must be analysed in detail, in order to find the morphological changes that can be observed in the cells affected by that disease. Since the importance of this kind of diagnosis different ensembles of descriptors and classifiers have been evaluated in order to provide a result as accurate as possible. The classification results are really good given that the proposed system is able to classify correctly almost all the positive samples, that are the most important in this context.

Finally, it is important to note that many of the proposed approaches could be also used in different medical imaging system and also for artificial vision system far from the medical field. This could be possible thanks to the generality of the proposed approaches, being designed to overcome many different issues, such as the colour differences, that make them independent from dataset and in some cases also independent from the problem itself.

Despite the good results obtained with both the case study, further extensions can be applied to the proposed approaches. In particular further research will be devoted to improve robustness and accuracy of the method in rotation invariant classification task, which is an important issue especially for medical images that can occur in different and uncontrolled rotation angles. The study of other features could be extended also for the description of other medical problems, including the possibility to study different stages of pathology, if present. For what concerns the peripheral blood image analysis, many other phases can be integrated. In particular a first improvement could arise from a detailed analysis of the WBCs in order to detect other type of disease that can affect the type of cell. Furthermore it will

be important the use of parallel phases for cells detection and counting, in order to provide other measures such as the red blood cell count, the platelet count, reticolocyte count and so on. These measures can be diagnostic by themselves, since that an overproduction or an underproduction is always symptom of problems related to the health of the bone marrow. Once single cells of each type have been detected and segmented, they can be analysed in detail, in order to detect the presence of parasite, like malaria parasite or to diagnose disease that can affect that particular cell type.

Acknowledgments

Lorenzo gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

La presente tesi stata prodotta durante la frequenza del corso di dottorato in Informatica dellUniversità degli Studi di Cagliari, a.a. 2014/2015 - XXVIII ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1 - Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell'ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali.

Bibliography

[ACRA+15] J. Arevalo, A. Cruz-Roa, V. Arias, E. Romero, and F. A. Gonzlez. An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artificial Intelligence in Medicine*, 64(2):131–145, 2015.

[AK13] M. Alilou and V. Kovalev. Automatic object detection and segmentation of the histocytology images using reshapable agents. *Image Analysis and Stereology*, 32(2):89–99, 2013.

[AS02] J. Angulo and J. Serra. Morphological color size distributions for image classification and retrieval. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 46–53, 2002.

[Ata86] K. T. Atanassov. Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 20(1):87–96, 1986.

[BCMT+76] J. M. Bennett, D. Catovsky, D. Marie-Theregse, G. Flandrin, D. A. G. Galton, H. R. Gralnick, and C. Sultan. Proposals for the classification of the acute leukaemias french-american-british (FAB) co-operative group. *British Journal of Haematology*, 33(4):451–458, 1976.

[BCPP00] A. Biondi, G. Cimino, R. Pieters, and C.-H. Pui. Biological and therapeutic aspects of infant leukemia. *Blood*, 96(1):24–33, 2000.

[BH07] M. Benčo and R. Hudec. Novel method for color textures features extraction based on glcm. *Radioengineering*, 4(16):64–67, 2007.

[BM12] A. D. Belsare and M. M. Mushrif. Histopathological image analysis using image processing techniques: An overview. *Signal & Image Processing*, 3(4):23, 2012.

[BMN+14] S. Bhattacharjee, J. Mukherjee, S. Nag, I. K. Maitra, and S. K. Bandyopadhyay. Review on histopathological slide analysis using digital microscopy. *International Journal of Advanced Science and Technology*, 62:65–96, 2014.

[BSa08] S. Buavirat and C. Srisa-an. Classification for acute lymphocytic leukemia using feature extraction and neural networks in white blood cell stained images. 2008.

[BTG06] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV European Conference on Computer Vision*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer Berlin Heidelberg, 2006.

[BTKD11] H. Berge, D. Taylor, S. Krishnan, and T. S. Douglas. Improved red blood cell counting in thin blood smears. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 204–207, March 2011.

[BVMP04] L. Busin, N. Vandenbroucke, L. Macaire, and J. G. Postaire. Color space selection for unsupervised color image segmentation by histogram multithresholding. In *International Conference on Image Processing*, volume 4, pages 203–206, October 2004.

[Caf98] R. E. Caflisch. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7:1–49, 1998.

[Can86] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, Nov 1986.

[CH67] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[CH80] R. W. Conners and C. A. Harlow. A theoretical comparison of texture algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(3):204–222, May 1980.

[Cha10] T. Chaira. Intuitionistic fuzzy segmentation of medical images. *IEEE Transactions on Biomedical Engineering*, 57(6):1430–1436, June 2010.

[Cie11] B. Ciesla. *Hematology in practice*. FA Davis, 2011.

[Cla02] D. A. Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, 28(1):45–62, 2002.

[CLAS11] J. Cheewatanon, T. Leauhatong, S. Airpaiboon, and M. Sangwarasilp. A new white blood cell segmentation using mean shift filter and region growing algorithm. *International Journal of applied biomedical engineering*, 4(1):30–35, 2011.

[Cou56] W. H. Coulter. High speed automatic blood cell counter and cell size analyzer. In *National Conference on Electronics*, pages 1034–1040, October 1956.

[CR03] T. Chaira and A. K. Ray. Segmentation using fuzzy divergence. *Pattern Recognition Letters*, 24(12):1837–1844, Aug 2003.

[CRCG11] A. Cruz-Roa, J. C. Caicedo, and F. A. Gonzlez. Visual pattern mining in histology image collections using bag of features. *Artificial Intelligence in Medicine*, 52(2):91–106, 2011.

[Cse92] I. Cseke. A fast segmentation scheme for white blood cell images. In *IAPR International Conference on Pattern Recognition. Vol. III: Image, Speech and Signal Analysis*, pages 530–533, 1992.

[CSK+09] L. Cooper, O. Sertel, J. Kong, G. Lozanski, K. Huang, and M. Gurcan. Feature-based registration of histopathology images with different stains: An application for computerized follicular lymphoma prognosis. *Computer Methods and Programs in Biomedicine*, 96(3):182–192, 2009.

[CWCT09] S. Chen, C. Wu, D. Chen, and W. Tan. Scene classification based on gray level-gradient co-occurrence matrix in the neighborhood of interest points. In *IEEE International Conference ICIS on Intelligent Computing and Intelligent Systems*, volume 4, pages 482–485, Nov 2009.

[DAM+08] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *IEEE ISBI International Symposium on Biomedical Imaging: From Nano to Macro*, pages 496–499, May 2008.

[DH73] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[DN13] S. G. Deore and N. Nemade. Image analysis framework for automatic extraction of the progress of an infection. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6):703–707, 2013.

[dSPN+15] F. L. C. dos Santos, M. Paci, L. Nanni, S. Brahna, and J. Hyttinen. Computer vision for virus image classification. *Biosystems Engineering*, 138:11–22, 2015. Innovations in Medicine and Healthcare.

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference CVPR on Computer*

Vision and Pattern Recognition, volume 1, pages 886–893 vol. 1, June 2005.

[EA13] O. Erhabor and T. C. Adias. *Hematology made easy*. AuthorHouse, 2013.

[FCMG00] D. J. Foran, D. Comaniciu, P. Meer, and L. Goodell. Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy. *IEEE Transactions on Information Technology in Biomedicine*, 4(4):265–273, 2000.

[FH75] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, Jan 1975.

[GASM+05] M. Guillaud, K. Adler-Storthz, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekirk, D. Cox, N. Poulin, M. Follen, and C. MacAulay. Subvisual chromatin changes in cervical epithelium measured by texture image analysis and correlated with HPV. *Gynecologic Oncology*, 99(3, Supplement):S16 – S23, 2005.

[GBC+09] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[GRCC+13] E. Gonzlez-Rufino, P. Carrin, E. Cernadas, M. Fernndez-Delgado, and R. Domnguez-Petit. Exhaustive comparison of colour texture features and classification methods to discriminate cells categories in histological images of fish ovary. *Pattern Recognition*, 46(9):2391 – 2407, 2013.

[GVB07] A. Gelzinis, A. Verikas, and M. Bacauskiene. Increasing the discrimination power of the co-occurrence matrix-based features. *Pattern Recognition*, 40(9):2367–2372, 2007.

[GW] R. C. Gonzalez and R. E. Woods. *Digital image processing*. Pearson Prentice Hall Pearson Education.

[GW12] R. Gong and H. Wang. Steganalysis for GIF images based on colors-gradient co-occurrence matrix. *Optics Communications*, 285:4961–4965, November 2012.

[GWE04] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital Image Processing Using MATLAB*. Pearson Prentice Hall Pearson Education, New Jersey, USA, 1st edition, 2004.

[HCL03] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification, 2003.

[HCx09] Y. Hu and Z. Chun-xia. Unsupervised texture classification by combining multi-scale features and k-means classifier. In *CCPR Chinese Conference on Pattern Recognition*, pages 1–5, Nov 2009.

[HLAT12] L. He, L. R. Long, S. Antani, and G. R. Thoma. Histology image analysis for carcinoma detection and grading. *Computer methods and programs in biomedicine*, 107(3):538–556, Sep 2012.

[HMH11] N. H. A. Halim, M. Y. Mashor, and R. Hassan. Automatic blasts counting for acute leukemia based on blood samples. *International Journal of Research and Reviews in Computer Science*, 2(4), 2011.

[HSD73] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973.

[HSP02] H. Hengen, S. L. Spoor, and M. C. Pandit. Analysis of blood and bone marrow smears using digital image processing techniques. In *Medical Imaging*, volume 4684, pages 624–635, 2002.

[HST+11] N. Herve, A. Servais, E. Thervet, J.-C. Olivo-Marin, and V. Meas-Yedid. Statistical color texture descriptors for histological images analysis. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 724–727, March 2011.

[IGM13] H. Inaba, M. Greaves, and C. G. Mullighan. Acute lymphoblastic leukaemia. *The Lancet*, 381(9881):1943–1955, 2013.

[JD06] J. Jantzen and G. Dounias. Analysis of pap-smear image data. In *Nature-Inspired Smart Information Systems 2nd Annual Symposium*, 2006.

[JF90] A. K. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 14–19, Nov 1990.

[JRQ03] C. R. Maurer Jr and V. Raghavan R. Qi. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):265–270, 2003.

[KAG] H. B. Kekre, B. Archana, and H. R. Galiyal. Segmentation of blast using vector quantization. *International Journal of Computer Applications*, 72.

[KDM+00] S. J. Keenan, J. Diamond, G. W. McCluggage, H. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN). *The Journal of pathology*, 192(3):351–362, 2000.

[KGHS96] V. A. Kovalev, A. Y. Grigoriev, and A. Hyo-Sok. Robust recognition of white blood cell images. In *International Conference on Pattern Recognition*, volume 4, pages 371–375, Aug 1996.

[Kit] J. Kittler. Feature set search algorithms. In *Pattern Recognition and Signal Processing*.

[KPYW13] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang. Histological image classification using biologically interpretable shape-based features. *BMC Medical Imaging*, 13(1), 2013.

[KRW+] G. Kayser, U. Riede, M. Werner, P. Hufnagl, and K. Kayser. Towards an automated morphological classification of histological images of common lung carcinomas. *Electronic Journal of Pathology and Histology*, 8(2):22–30.

[KSW85] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, 29(3):273–285, 1985.

[KWT] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.

[LC02] O. Lezoray and H. Cardot. Cooperation of color pixel classification schemes and color watershed: a study for microscopic images. *IEEE Transactions on Image Processing*, 11(7):783–789, Jul 2002.

[LEC+98] O. Lezoray, A. Elmoataz, H. Cardot, G. Gougeon, M. Lecluse, H. Elie, and M. Revenu. Segmentation of cytological image using color and mathematical morphology. In *European conference on Stereology*, 1998.

[Lin] J. Lindblad. Development of algorithms for digital image cytometry.

[LIT92] P. Langley, W. Iba, and K. Thompson. An analysis of bayesian classifiers. In *National Conference on Artificial Intelligence*, volume 90, pages 223–228. AAAI Press, July 1992.

[LNHD07] B. Lessmann, T. W. Nattkemper, V. H. Hans, and A. Degenhard. A method for linking computed image features to histological semantics in neuropathology. *Journal of Biomedical Informatics*, 40(6):631–641, 2007. Intelligent Data Analysis in Biomedicine.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LPS11] R. Donida Labati, V. Piuri, and F. Scotti. All-idb: The acute lymphoblastic leukemia image database for image processing. In *IEEE ICIP International Conference on Image Processing*, pages 2045–2048, Sept 2011.

[MdSV+97] N. Malpica, C. Ortiz de Solorzano, J. J. Vaquero, A. Santos, I. Vallcorba, J. M. Garcia-Sagredo, and F. del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Citometry*, 28(4):289–297, Sep 1997.

[Mey94] F. Meyer. Topographic distance and watershed lines. *Signal processing*, 38(1):113–125, 1994.

[MH80] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, 207(1167):187–217, 1980.

[MKA+10] H. T. Madhloom, S. A. Kareem, H. Ariffin, A. A. Zaidan, H. O. Alanazi, and B. B. Zaidan. An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold. *Journal of Applied Sciences*, 10(11):959–966, 2010.

[MLMR] N. H. Mahmood, P. C. Lim, S. M. Mazalan, and M. A. A. Razak. Blood cells extraction using color based segmentation technique. *International Journal of Life Sciences Biotechnology and Pharma Research*, 2(2).

[MLSC10] T. Meng, L. Lin, M.-L. Shyu, and S.-C. Chen. Histology image classification using supervised classification and multimodal fusion. In *IEEE ISM International Symposium on Multimedia*, pages 145–152, 2010.

[MMN+12] D. Mitrea, P. Mitrea, S. Nedevschi, R. Badea, M. Lupsor, M. Socaciu, A. Golea, C. Hagiu, and L. Ciobanu. Abdominal tumor characterization and recognition using superior-order cooccurrence matrices, based on ultrasound images. *Computational and mathematical methods in medicine*, 2012, 2012.

[MP10a] S. Mohapatra and D. Patra. Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images. In *ICSMB International Conference on Systems in Medicine and Biology*, pages 49–54, Dec 2010.

[MP10b] S. Mohapatra and D. Patra. Automated leukemia detection using hausdorff dimension in blood microscopic images. In *INTERACT International Conference on Emerging Trends in Robotics and Communication Technologies*, pages 64–68, Dec 2010.

[MPCB+13] P. Melo-Pinto, P. Couto, H. Bustince, E. Barrenechea, M. Pagola, and J. Fernandez. Image segmentation using atanassovs intuitionistic fuzzy sets. *Expert Systems with Applications*, 40(1):15–26, 2013.

[MPS10] S. Mohapatra, D. Patra, and S. Satpathy. Image analysis of blood microscopic images for acute leukemia detection. In *International Conference IECR on Industrial Electronics, Control Robotics*, pages 215–219, Dec 2010.

[MPS14] S. Mohapatra, D. Patra, and S. Satpathy. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing and Applications*, 24(7-8):1887–1904, 2014.

[MTF+09] J. Monaco, J. E. Tomaszewski, M. D. Feldman, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi. Probabilistic pairwise markov models: application to prostate cancer detection. In *Medical Imaging*, volume 7259, pages 725903–725912, 2009.

[NBG+13] L. Nanni, S. Brahnam, S. Ghidoni, E. Menegatti, and T. Barrier. Different approaches for extracting information from the co-occurrence matrix. *PloS one*, 8(12), 2013.

[NDA+08] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *IEEE ISBI International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287, May 2008.

[NDF+07] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information. In *MIAAB workshop*, pages 1–8, 2007.

[OPH96] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[Ots75] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[PCR14] L. Putzu, G. Caocci, and C. Di Ruberto. Leucocyte classification for leukaemia detection using image processing techniques. *Artificial Intelligence in Medicine*, 62(3):179–191, 2014.

[PLC09] C. Pan, H. Lu, and F. Cao. Segmentation of blood and bone marrow cell images via learning by sampling. In *Emerging Intelligent Computing Technology and Applications*, volume 5754 of *Lecture Notes in Computer Science*, pages 336–345. Springer Berlin Heidelberg, 2009.

[PR88] S. K. Pal and A. Rosenfeld. Image enhancement and thresholding by optimization of fuzzy compactness. *Pattern Recognition Letters*, 7(2):77–86, 1988.

[PR13a] L. Putzu and C. Di Ruberto. Investigation of different classification models to determine the presence of leukemia in peripheral blood image. In *ICIAP International Conference on Image Analysis and Processing*, volume 8156 of *Lecture Notes in Computer Science*, pages 612–621. Springer Berlin Heidelberg, 2013.

[PR13b] L. Putzu and C. Di Ruberto. White blood cells identification and classification from leukemic blood image. In *IWBBIO International work-conference on bioinformatics and biomedical engineering*, pages 99–106, Mar 2013.

[PR13c] L. Putzu and C. Di Ruberto. White blood cells identification and counting from microscopic blood images. *World Academy of Science, Engineering and Technology*, 7(1):363–370, Gen 2013.

[PS04] V. Piuri and F. Scotti. Morphological classification of blood leucocytes by microscope images. In *IEEE International Conference CIMSA on Computational Intelligence for Measurement Systems and Applications*, pages 103–108, July 2004.

[PVH13] A. Porebski, N. Vandenbroucke, and D. Hamad. Lbp histogram selection for supervised color texture classification. In *IEEE ICIP International Conference on Image Processing*, pages 3239–3243, Sept 2013.

[Qui86] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[RAS+00] K. Ramar, S. Arumugam, S.N. Sivanandam, L. Ganesan, and D. Manimegalai. Quantitative fuzzy measures for threshold selection. *Pattern Recognition Letters*, 21(1):1–7, 2000.

[RDKJ00] C. Di Ruberto, A. Dempster, S. Khan, and B. Jarra. Automatic thresholding of infected blood images using granulometry and regional extrema. In *International Conference on Pattern Recognition*, volume 3, pages 441–444 vol.3, 2000.

[RFP15a] C. Di Ruberto, G. Fodde, and L. Putzu. Comparison of statistical features for medical colour image classification. In *International Conference ICVS on Computer Vision Systems*, volume 9163 of *Lecture Notes in Computer Science*, pages 3–13. Springer International Publishing, 2015.

[RFP15b] C. Di Ruberto, G. Fodde, and L. Putzu. On different colour spaces for medical colour image classification. In *International Conference CAIP on Computer Analysis of Images and Patterns*, volume 9256 of *Lecture Notes in Computer Science*, pages 477–488. Springer International Publishing, 2015.

[RLP15a] C. Di Ruberto, A. Loddo, and L. Putzu. Learning by sampling for white blood cells segmentation. In *ICIAP International Conference on Image Analysis and Processing*, volume 9279 of *Lecture Notes in Computer Science*, pages 557–567. Springer International Publishing, 2015.

[RLP15b] C. Di Ruberto, A. Loddo, and L. Putzu. A multiple classifier learning by sampling system for white blood cells segmentation. In *International Conference CAIP on Computer Analysis of Images and Patterns*, volume 9257 of *Lecture Notes in Computer Science*, pages 415–425. Springer International Publishing, 2015.

[RS96] D. A. Ralescu and M. Sugeno. Fuzzy integral representation. *Fuzzy Sets and Systems*, 84(2):127–133, 1996.

[RSBK14] J. Rawat, A. Singh, H. S. Bhadauria, and I. Kumar. Comparative analysis of segmentation algorithms for leukocyte extraction in the acute lymphoblastic leukemia images. In *Internation Conference on Parallel, Distributed and Grid Computing*, pages 245–250, December 2014.

[SAGG06] K. Schmid, N. Angerstein, S. Geleff, and A. Gschwendtner. Quantitative nuclear texture features analysis confirms who classification 2004 for lung carcinomas. *Modern pathology*, 19(3):453–459, 2006.

[Sco05] F. Scotti. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *IEEE International Conference CIMSA on Computational Intelligence for Measurement Systems and Applications*, pages 96–101, July 2005.

[Sco06] F. Scotti. Robust segmentation and measurements techniques of white cells in blood microscope images. In *IEEE IMTC Instrumentation and Measurement Technology Conference*, pages 43–48, April 2006.

[SDH+91] S. Serbouti, A. Duhamel, H. Harms, U. Gunzer, H. M. Aus, J.-Y. Mary, and R. Beuscart. Image segmentation and classification methods to detect leukemias. In *IEEE International Conference on Engineering in Medicine and Biology Society*, volume 13, pages 260–261, Oct 1991.

[Ser83a] J. Serra. *Image analysis and mathematical morphology*, volume I. Academic Press, Inc., 1983.

[Ser83b] J. Serra. *Image analysis and mathematical morphology, Theoretical Advances*, volume II. Academic Press, Inc., 1983.

[SK00] E. Szmidt and J. Kacprzyk. Distances between intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 114(3):505–518, 2000.

[SKL+08] O. Sertel, J. Kong, G. Lozanski, U. Catalyurek, J. H. Saltz, and M. N. Gurcan. Computerized microscopic image analysis of follicular lymphoma. In *Medical Imaging*, pages 691535–691535, 2008.

[SOD+08] L. Shamir, N. Orlov, M. E. David, T. Macura, and I. Goldberg. Iicbu 2008: a proposed benchmark suite for biological image analysis. *Medical & Biological Engineering & Computing*, 46(9):943–947, 2008.

[SR03] N. Sinha and A. G. Ramakrishnan. Automation of differential blood count. In *TENCON Conference on Convergent Technologies for the Asia-Pacific Region*, volume 2, pages 547–551, Oct 2003.

[ST99] L.-K. Soh and C. Tsatsoulis. Texture analysis of sar sea ice imagery using gray level co-occurrence matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):780–795, Mar 1999.

[Tan98] X. Tang. Texture information in run-length matrices. *IEEE Transactions on Image Processing*, 7(11):1602–1609, Nov 1998.

[VMP03] N. Vandenbroucke, L. Macaire, and J.-G. Postaire. Color image segmentation by pixel classification in an adapted hybrid color space. application to soccer image analysis. *Computer Vision and Image Understanding*, 90(2):190–216, 2003.

[WEG87] S. Wold, K. Esbensen, and P. Geladi. Proceedings of the multivariate statistical workshop for geologists and geochemists principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987.

[WJL03] R. F. Walker, P. T. Jackway, and D. Longstaff. Genetic algorithm optimization of adaptive multi-scale glcm features. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(01):17–39, 2003.

[WZZO06] J. Wu, P. Zeng, Y. Zhou, and C. Olivier. A novel color image segmentation method and its application to white blood cell image analysis. In *International Conference on Signal Processing*, volume 2, 2006.

[YL04] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, December 2004.

[Zad75a] A. L. Zadeh. The concept of a linguistic variable and its application to approximate reasoning I. *Information sciences*, 8(3):199–249, 1975.

[Zad75b] A. L. Zadeh. The concept of a linguistic variable and its application to approximate reasoning II. *Information sciences*, 8(4):301–357, 1975.

[Zad75c] A. L. Zadeh. The concept of a linguistic variable and its application to approximate reasoning III. *Information sciences*, 9(1):43–80, 1975.

[ZPO+] J. M. Zahn, S. Poosala, A. B. Owen, D. K. Ingram, A. Lustig, A. Carter, A. T. Weeraratna, D. D. Taub, M. Gorospe, K. Mazan-mamczarz, E. G. Lakatta, K. R. Boheler, X. Xu, M. P. Mattson, M. S. H. Ko, D. Schlessinger, J. Firman, S. K. Kummerfeld, W. H. Wood, A. B. Zonderman, S. K. Kim, and K. G. Becker. Agemap: a gene expression database for aging in mice. plos genet 3: e201.

[ZRL77] G. W. Zack, W. E. Rogers, and S. A. Latt. Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry & Cytochemistry*, 25(7):741–753, 1977.

Appendix A

Haematopoiesis

The production of all types of blood cells including formation, development, maturation and differentiation of blood cells is called haematopoiesis. Its purpose is to ensure the constant daily production of mature cells of the peripheral blood both in normal condition and both in response to particular situations of increased demand, such as in the presence of infection or blood loss. The haematopoiesis is supported by a small number of primitive cells called *Haematopoietic Stem Cells* (HSCs) or *haemocytoblasts*, characterized by the ability of self-renewing, namely the ability to generate cells identical to themselves. At the same time the HSCs are pluripotent, having the potential to develop into all types of blood cells. Haematopoiesis occurs in bone marrow, where the HSCs are present in the ratio of one stem cell for every 1.000 nonstem cell elements. This is why the cause and effect of haematologic disease are usually rooted in the bone marrow. Usually, only normal, mature or nearly mature cells are released into the bloodstream, but certain circumstances can induce the bone marrow to release immature and/or abnormal cells into the circulation. The predominance of immature cells noted in a complete blood count is indicative of infections, inflammations and other severe illnesses. This is also the reason why it is very important to analyse the whole haematopoietic process, being able to recognise inside the blood smears any cell type at any stage of maturation. The HSCs give rise to mature cells, that enter in the peripheral circulation via the bone marrow sinuses, by firstly differentiating into myeloid (non-lymphoid) and lymphoid precursor committed cells. Myeloid precursor cells develop into monocytes, macrophages, neutrophils, basophils, eosinophils, erythrocytes, megakaryocytes, platelets, and dendritic cells, while the lymphoid precursors develop into lymphocyte T-cells, B-cells, NK-cells. Haematopoiesis can also be subdivided according to the type of cell being formed: granulopoiesis (neutrophils, eosinophils, basophils), monopoiesis (monocytes), lymphopoiesis (lymphocytes), erythropoiesis (erythrocytes) and megakaryocytopoiesis (platelets). Fig. A.1 shows a schema of the haematopoietic process.

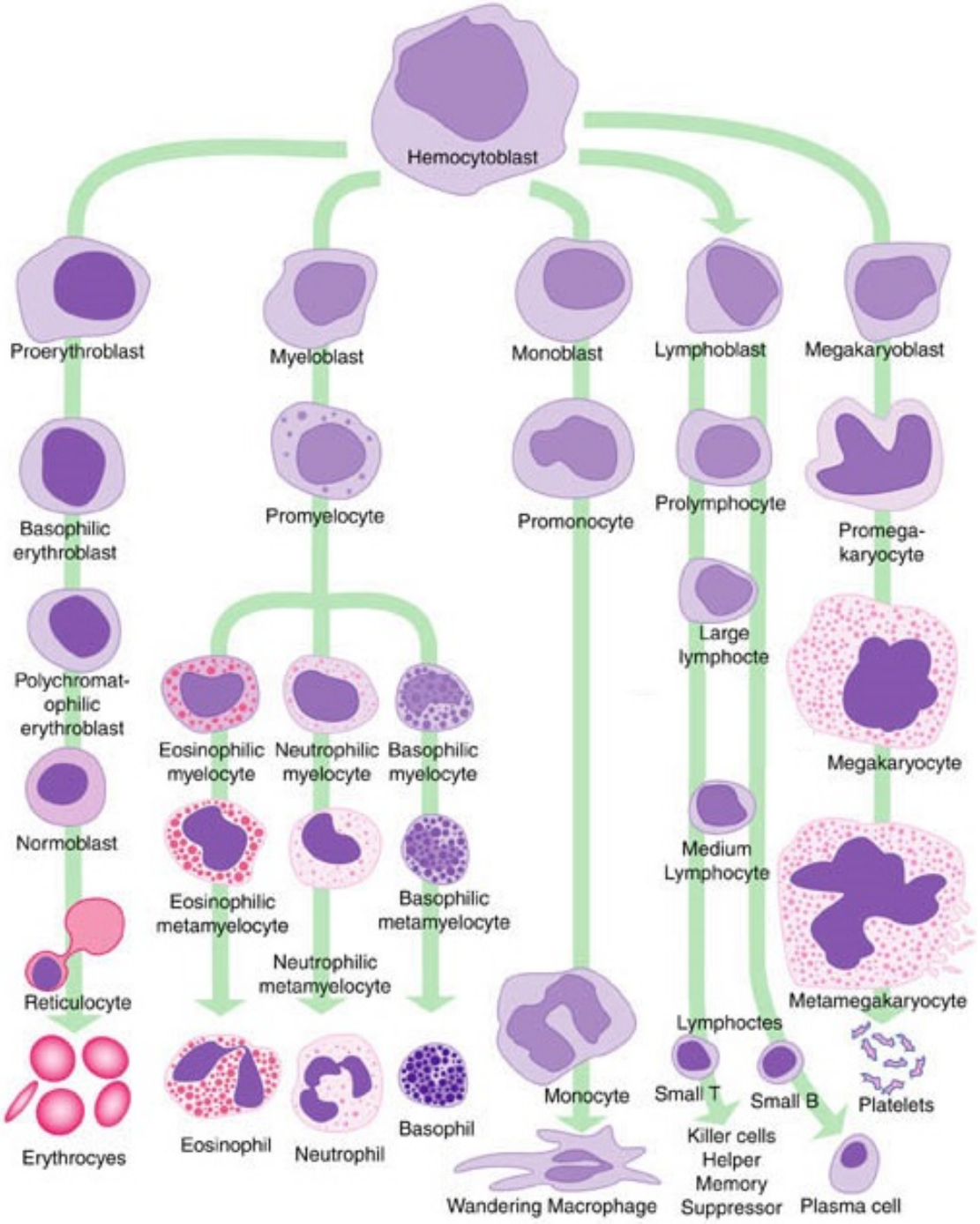

Figure A.1: Haematopoietic process.

A.1 Granulopoiesis

Granulocytes are also called *polymorphonuclear leukocytes* because of their characteristically shaped nuclei and cytoplasmic granules. Granulocytes include neutrophils, eosinophils, and basophils. A granulocyte differentiates into a distinct cell type by a process called granulopoiesis. The stages of maturation for the neutrophilic, eosinophilic and basophilic series is very similar. They start to differentiate at the third stage, so the first two stages are in common. The first five stages of granulopoiesis are illustrated in Fig. A.2.

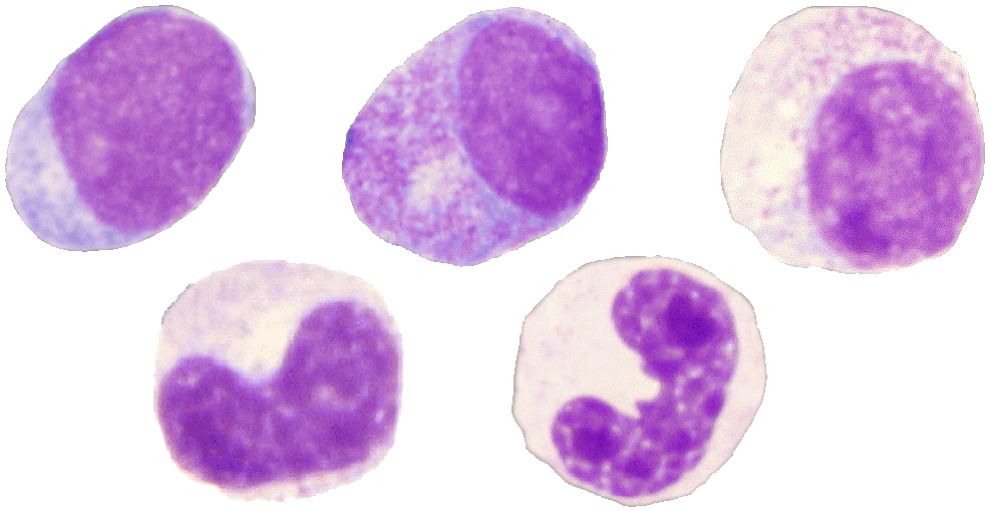

Figure A.2: Granulopoiesis.

At the first stage of maturation the myeloid progenitor, called *myeloblast*, has a size that ranges from 10 to 20 micron. The nucleus is large and central round, that could have a round or oval shape and it has an open, unclumped nuclear chromatin that is of a light red-purple colour. The nucleus contains several nucleoli, from two to five, which appear as lightened, refractile round structures, while the cytoplasm is poor and has a moderate blue colour and usually without granules. At the second stage the myeloblast transforms in a *promyelocyte* or *progranulocyte*, that has similar size that ranges from 10 to 22 micron. The nucleus is oval, round, or eccentric and the nuclear chromatin is more condensed of a light red-purple colour. The nucleus contains less prominent nucleoli while the cytoplasm presents azurophilic granules. At this stage the three series starts to differentiate, even if they preserve a similar appearance. Thus the promyelocyte gives rise to a unique myelocyte that can either be eosinophilic, basophilic, or neutrophilic. The myelocyte then differentiates further into a metamyelocyte and then into a band cell before becoming a mature neutrophil, eosinophil, or basophil. The *myelocytes* are slightly smaller than promyelocytes with a diameter of 10-18 micron. They present an eccentric, round-oval nucleus with a coarse and condensed chromatin and small, non-visible nucleoli. Some azurophilic

granules still persist in the cytoplasm, but secondary or specific granules begin to predominate, in particular the neutrophilic granules are dusty, fine, and red-blue, while eosinophilic granules are large red-orange and singular, instead basophil granules are large deep blue-purple. The *metamyelocytes* are slightly smaller than myelocytes with a diameter of 10-15 micron. They have characteristic kidney-shaped nuclei and relatively densely clumped nuclear chromatin with no nucleoli. The cytoplasm range from pale blue to pinkish and becomes filled with predominantly secondary granules, although primary granules persist, and tertiary granules begin to appear. The *band forms* have a size of 9 to 15 micron, with a curved or band-shaped nucleus but non-lobular or unsegmented. The cytoplasm is brown-pink, with many fine specific or secondary granules, that start to predominate. The *segmented forms* have a size similar to the band forms, but they present a segmented nucleus, with two to five nuclear lobes connected by thin threadlike filaments. The cytoplasm is pale lilac with blue shading and many fine secondary dust-like granules. In detail, *neutrophils* or polymorphonuclear neutrophils have a diameter of 12-15 micron filled with pink or purple granules and 2-5 nuclear lobes. The chromatin of the segmented neutrophil is coarsely clumped. The cytoplasm is faint pink and it is filled with fine pink secondary granules. They are involved in the defence against infections. Neutrophils are the most abundant white blood cells in humans and account for approximately 70% of all white blood cells. The presence of abnormally low number of neutrophils is described as neutropenia. Also the number of lobes and the extent of granulation are diagnostic. Neutrophils with more than 5 lobes are called hyper segmented neutrophils. Neutrophils with more intensely stained (large dark blue) and more granules are described as toxic granulated neutrophils. Vacuoles appear as holes in the cytoplasm and are frequently found in association with toxic granulation. *Eosinophils* instead have a diameter of 10-15 micron and they are easily recognised in stained smears because of their cytoplasm is filled with large, red-orange granules and a bi-lobed nucleus. They are generally low in number (1-3%). The presence of abnormally high number of eosinophils is described as eosinophilia. Also *basophils* have a diameter of 10-15 micron and a coarse, clumped bi-lobed nucleus and the presence of many large, specific secondary purple-black granules in the cytoplasm. Basophils are the least often seen type of WBC (1 %). Increased basophils number is called basophilia.

Because white blood cells have such a short time span in the peripheral circulation, alterations either in the quantity or in the quality of a particular white blood cell can be quite dramatic. As white blood cells increase, the peripheral smear usually shows an increased numbers of segmented neutrophils, or the presence of younger cells. In either of these cases, toxic changes, such as toxic granulation, toxic vacuolization, the presence of Dohle bodies or Auer Bodies, Pelger-Huet and Hypersegmentation may be observed. This toxic changes are illustrated in Fig. A.3.

Toxic granulation is excessive in amount and intensity, with more prominent granules in segmented neutrophils and bands. Normal granulation in the segmented neutrophils has a dust-like appearance, with the red and blue granules being difficult

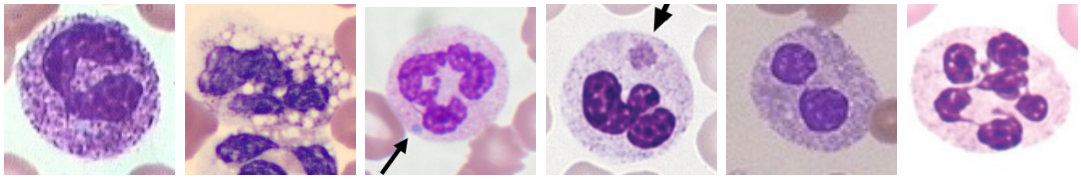

Figure A.3: Granulocyte toxic changes.

to observe, while, with toxic granulation, these granules are more frequent and have much more vivid blue-black colouration. Clusters of toxic granules usually appear in neutrophils. Sometimes the granulation is so heavy as to resemble basophilic granules. Toxic granulation can be observed during acute bacterial infections. *Toxic vacuolization* occurs in the segmented neutrophil with the appearance of small or large vacuoles in the cytoplasm. *Dohle bodies* are light blue cytoplasmic inclusions that range from 1 to 5 micron in size, are located in the peripheral cytoplasm of neutrophils and appear as a rod-shaped, pale bluish grey structure. They are nuclear remnants that are often seen in association with toxic granules and vacuoles. Dohle bodies may be present in sepsis or severe inflammatory responses. *Auer Bodies* are clumps of azurophilic granular material that form elongated needles seen in the cytoplasm of leukaemic blasts. They are unique, pink or red rod-shaped inclusions that are seen in very immature granulocytes in patients with acute non-lymphocytic leukaemias. *Pelger-Huet* is an anomaly characterized by impaired nuclear segmentation of mature granulocytes. The nucleus is often in the shape of a peanut or dumbbell, or may consist of two lobes connected with a filament. *Hypersegmentation* is defined as a segmented neutrophilic nucleus having more than five lobes, since normal segmented neutrophils have between three and five lobes in the nucleus.

A.2 Monopoiesis

Monocytes are produced by the cell precursors called monoblasts. Monocytes differentiate and mature from monoblasts into promonocytes and then to matured monocytes. The stages of monopoiesis are illustrated in Fig. A.4.

Figure A.4: Monopoiesis.

Monoblast is very similar to myeloblast, with a size of about 12-20 micron. The cytoplasm is agranular and the nucleus is large, round to oval and has fine nuclear chromatin. The main difference with myelobast is that nucleoli (one or two) are more prominent in monoblasts. *Promonocyte* has an average diameter of 14-18 micron. It has a large, convoluted nucleus, a coarse chromatin structure and one or two nucleoli. The cytoplasm is grey-blue and may contain a few fine azurophillc granules. *Monocytes* are the largest of the white blood cells with a size of 12-20 micron. The cytoplasm is grey-blue, it may have numerous vacuoles and fine azurophilic granules. Monocytes have abundant cytoplasm and a large, distinctive, kidney-shaped nucleus. They circulate in the bloodstream for about one to three days, where they move into tissues throughout the body. They constitute between 3-8% of all leukocytes in the blood. In the tissues, monocytes mature into different types of macrophages and help protect tissues from foreign substances. A decreased percentage of monocyte levels is called monocytosis.

A.3 Lymphopoiesis

Outlining the lymphocyte cell population is a complex task and beyond the scope of this thesis. Furthermore, some populations of lymphocytes appear morphologically similar on peripheral smear. For this reason, only a modified subset of subpopulation is included. Lymphocytes are produced by the cell precursors called lymphoblast, that gives rise to prolymphocyte that differentiate into large lymphocyte and small lymphocyte. Fig. A.5 shows the lymphocyte precursors.

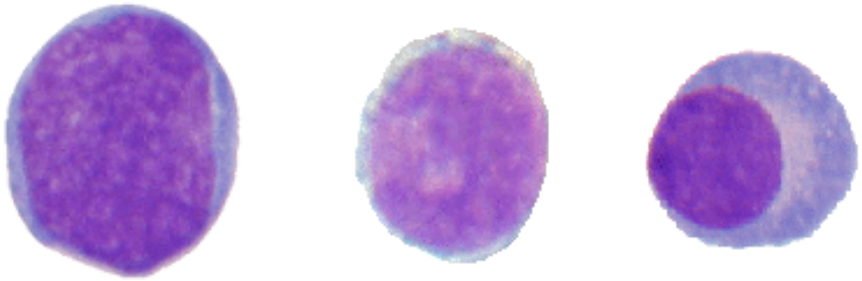

Figure A.5: Lymphopoiesis.

Lymphoblasts have a size of 10-20 micron with a little cytoplasm deep blue staining at edge. They present one or two nucleoli. *Prolymphocytes* have a size of 9-18 with a grey-blue cytoplasm that is mostly blue at edges. The nucleus is almost round with coarse chromatin and some nucleoli may be present. *Large lymphocytes* have a size of 15 to 18 micron, a chromatin more transparent and they present a larger amount of cytoplasm, lighter in colour. *Small lymphocytes* have a size of 7-12 micron. They present an oval eccentric nucleus with coarse, lumpy chromatin with

specific areas of clumping. The cytoplasm is usually just a thin border, with few azurophilic granules. Small lymphocytes consist of T cells and B cells, but it is not possible to distinguish between them in a peripheral blood smear as they appear morphologically similar. Their derivation and function, however, are quite different. B lymphocytes comprise 10% to 20% of the total lymphocyte population, whereas T lymphocytes comprise 60% to 80%. A third minor population, NK lymphocytes, constitutes less than 10% of the total lymphocyte population. Lymphocytes can also differentiate into dendritic cells, that unlike T-cells, B-cells and NK cells, arise from lymphoid or myeloid lineages. Lymphocytes normally represent 20 to 40% of circulating white blood cells and they are the cornerstones of the adaptive immune system. T lymphocytes and B lymphocytes play a role in the maintenance of cell-mediated and antibody-mediated immunity. The increase in the number or proportion of lymphocytes in the blood is termed lymphocytosis, while a decreased number of lymphocytes is termed lymphocytopenia, or lymphopenia.

A.4 Erythropoiesis

Erythropoiesis is the process by which red blood cells (RBCs) or erythrocytes are produced. This process starts with the proliferation and differentiation of HSCs into the red cell precursors. This process is composed of six stages of maturation in the red blood cell series: pronormoblast, basophilic normoblast, polychromatophilic normoblast, orthochromic normoblast, reticulocyte, and mature erythrocytes. In general, several morphological clues mark the RBC maturation series, the cell size decreases, nuclear chromatin becomes more condensed, the cytoplasm colour is altered during haemoglobin production, but the most evident are the vanishing of the nucleus and the decrease in size, as it can be seen in Fig. A.6.

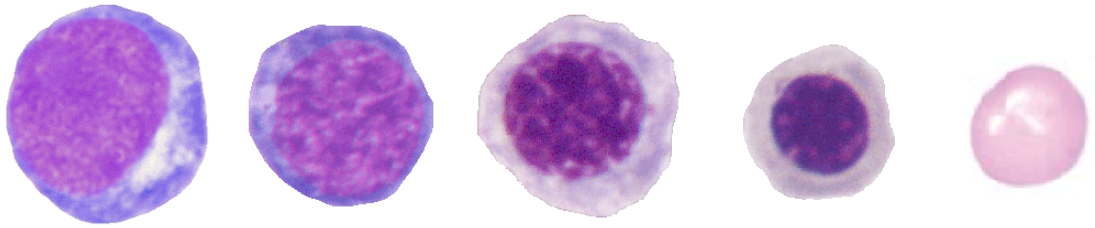

Figure A.6: Erythropoiesis.

The *proerythroblast* or *pronormoblast* is typically 14-20 micron in size. It presents a round centrally located nucleus with a coarser chromatin, more reticular, and condensed with a fine texture with deep violet colour, nucleoli may be present but are hard to visualize. The cytoplasm presents a dark marine blue colour with definitive areas of clearing. The *basophilic normoblast* or *erythroblast* is slightly smaller in size then pronormoblast, typically 12-17 micron of diameter. The nucleus is round with

crystalline chromatin appearance and it presents closed nucleoli. The cytoplasm becomes more basophilic, a cornflower blue colour with indistinct areas of clearing and a grainy and reticular textured chromatin. The *polychromatic* or *intermediate normoblast* has a size of 12-15 micron. The nuclear chromatin is condensed and moderately compacted with no nucleoli. The nucleus becomes smaller with a size of 7-9 micron and the cytoplasm colour shift from deep basophilic to grey. The *orthochromic* or *non-nucleated normoblast* has a size of 8-12 micron. The cytoplasm increases with orange-red colour tinges with slight blue tone. The nuclear chromatin condenses further and the nucleus shrinks and tends to become more peripheral and eventually extruded. An orthochromic normoblast becomes a reticulocyte once the nucleus is extruded. *Reticulocytes* or *polychromatic erythrocyte* are larger about twice than normal mature red cells, with a size of 8 micron, but the most evident difference is the presence of a reticulum in the cytoplasm. The mature cell is released from the bone marrow into peripheral circulation at the reticulocyte stage. Under normal circumstances, reticulocytes constitute about 1% of circulating red blood cells. The reticulocyte count is the most effective measure of erythropoietic activity since it reflects bone marrow healthy or injury. Low reticulocyte counts indicate decreased erythropoietic activity, or may occur in ineffective erythropoiesis, a condition in which red blood cell precursors are destroyed before they are delivered to the peripheral circulation, or if the bone marrow is infiltrated with tumour or abnormal cells. Increased reticulocyte counts indicate increased erythropoietic activity, usually as the bone marrow compensates in response to anaemia. The reticulocyte matures after one to two days in circulation into a mature and functional RBC. The mature *erythrocytes* present a significant reduction in the cell size that ranges from 6 to 8 micron and the cytoplasm changes characteristically from blue to salmon pink. Erythrocytes are disk-shaped cells, due to the presence of haemoglobin that is located peripherally, leaving an area of central pallor equal to 1-3 micron, approximately 30-45% of the diameter of the cells.

A.4.1 Erythrocyte Variations

Identifying normal and abnormal erythrocytes is really important, since automated cell counters have not yet replaced the well-trained eye with respect to the subtleties of red blood cell morphology. Erythrocytes of normal size are termed *normocytes*, while erythrocytes larger than normal, thus with a diameter greater than 9 micron, are called *macrocytes*, while smaller than normal, thus with a diameter less than 6 micron, are called *microcytes*. Erythrocyte colour, that in normal conditions is pinkish red with a central pallor, is representative of haemoglobin concentration in the cell. Under normal conditions, when the colour, central pallor, and haemoglobin are proportional, the erythrocyte is termed *normochromic*. *Hypochromic* cells exhibit an area of central pallor larger than normal, thus greater than 50% of the diameter (3 micron), that means a decreased haemoglobin concentration. *Polychromatophilic* cells exhibit a blue-grey cytoplasm and they are slightly larger than normal. Poik-

ilocytosis is a general condition associated with the presence of one or more types of abnormally shaped mature erythrocytes, some of which may indicate possible presence of a specific disease or disorder. Examples include; spherocytes, elliptocytes, sickle cells, teardrop cells, echinocytes, acanthocytes, keratocytes, bite cells, shistocytes, target cells, stomatocytes and rouleaux formation. This shape abnormalities are illustrated in Fig. A.7.

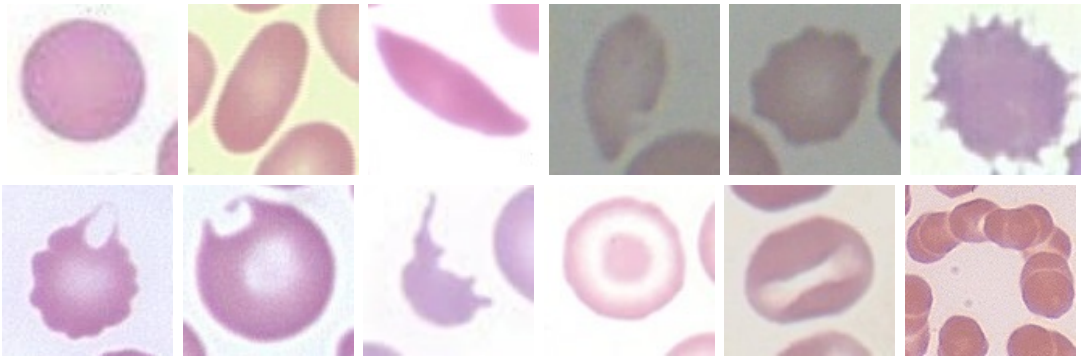

Figure A.7: Poikilocytosis.

Spherocytes are compact, round, densely staining red cells that lack central pallor. They are easily recognized among the rest of the red blood cell because they are dense, dark and small. *Ovalocytes* or *elliptocytes* are the most common red cells, in fact they appear rather than the typical biconcave disc shaped. Erytrocytes with this defect range from slightly oval to elongated cigar-shaped forms and they may appear macrocytic, hypochromic, or normochromic. *Sickle cells* are elongated or shaped like crescents or sickles. The fragile, sickle-shaped cells deliver less oxygen to the body's tissues. They can also get stuck more easily as they try to go through small blood vessels, and break into pieces that interrupt healthy blood flow. Many sickle cells may revert to normal disk shape on oxygenation, but approximately 10% are unable to revert. *Teardrop* are characterised by a smaller size and above all from they appearance that resemble a tear. *Acanthocytes* are characterised by a smaller size and above all from the presence of thorny projections distributed irregularly around the red blood cell, which lacks central pallor. The number of thorn can range from three to nine and must be distinguished from the *echinocytes*, in which the projections are typically evenly spaced on the cell surface and they are more numerous, from 10 to 30. Another difference is that echinocytes have serrated edges over the entire surface and often the membrane is smaller and much more uniform in shape and distribution. *Schistocytes* are fragmented erythrocytes that are irregular in shape and size. They are usually half the size of the normal red blood cells and have a deeper red colour. They can appear as small triangular erythrocytes, helmet cells, and normal-size erythrocytes with 2 to 3 pointed surface projections (*keratocytes*). Round erythrocytes with a single, elliptical or round surface defect are termed *bite cells*. *Stomatocytes* are characterized by a mouth-shaped area of central

pallor and a decrease in the ratio of surface area-to-volume that can be induced either by a reduction in surface area or an increase in red cell volume. Several agents can induce this morphology and often they can also be found on the peripheral smear of normal subjects, due to drying artefact. This can be distinguished since the percentage of stomatocytes in normal subjects is usually below 3% of the total red cells. *Target cells* have a centrally located disk of haemoglobin surrounded by an area of pallor with an outer ring of haemoglobin adjacent to the cell membrane giving the cell the appearance of a target. They are seen in the peripheral blood due to the presence of artefacts, because of decreased volume or increased red blood cell surface membrane. *Rouleaux* formation is a phrase denoting an agglomerate of erythrocytes, that create a stack generally in a curving pattern. The flat surface of the RBCs give them a large surface area to make contact and stick to each other forming a rouleaux.

A.4.2 Erythrocyte Inclusions

The cytoplasm of all normal red blood cells is free of debris, granules, or other structures. Inclusions are the result of distinctive conditions and their identification can be clinically helpful. Examples of inclusion bodies are: howell-jolly bodies, siderotic granules, basophilic stippling, Heinz bodies, malaria and nucleated red cells. This cell inclusion are illustrated in Fig. A.8.

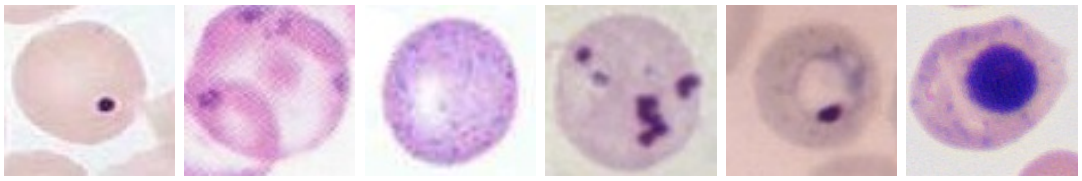

Figure A.8: Erythrocyte Inclusions.

Howell-jolly bodies represent remnants of the nucleus as it is extruded from the cytoplasm, that appear in the red blood cell as round, deep purple structures of about 1 micron in size. They are eccentrically located in the cytoplasm and seen when erythropoiesis is rushed. *Siderotic granules* or *Pappenheimer bodies* appear as small, dark blue or purple dots, located along the periphery of the red blood cells. *Basophilic stippling* refers to numerous very small coarse or fine blue granules in the periphery of the cytoplasm. They are difficult to visualize in the peripheral smear without fine focusing, but red blood cell containing basophilic stippling often is polychromatophilic. *Heinz bodies* are defined as large structures approximately 1 to 3 micron in diameter located toward the periphery of the red blood cell membrane. Although they cannot be visualized by standard stain, bite cells in the peripheral smear are evidence that a Heinz body has been formed. *Malaria* is a mosquito-borne infectious disease that results from the multiplication of a parasite within the cytoplasm of the red blood cells. Five species of this parasite can infect and

be transmitted by humans, for this reason the inclusion appearance can be very different. *Nucleated red blood cells* (NRBCs), that are red cells with a retained nucleus can be observed inside the blood smears. The average size of the NRBC is 7-12 micron in diameter, the cytoplasm is pink and the nucleus is a homogeneous blue-black mass with no structure. NRBCs detection and quantification is still based on the microscopic analysis of stained blood, since they are often counted as white cells by most haematology analysers because of the presence of the retained nucleus, and this, in particular in patients with a high nucleated cell count, could lead to misleading results. Sometimes platelets overlying erythrocytes may be mistaken for erythrocyte inclusions.

A.5 Megakaryocytopoiesis

Megakaryocytopoiesis is the process by which platelets or thrombocytes are produced. Platelet development is originated in the bone marrow from the HSC that differentiate into the megakaryocytic precursor, that then develops into the megakaryoblast, that give rise to the pro-megakaryocyte and then the megakaryocyte before developing into mature platelets. During this period the megakaryocyte nucleus undergoes extensive endomitosis, the cytoplasmic differentiation, the formation of platelet granules, and the fragmentation into mature platelets as it can be seen in Fig. A.9.

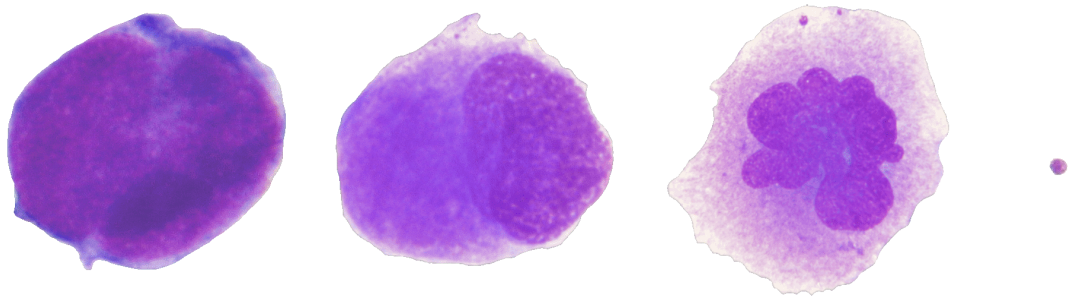

Figure A.9: Megakaryocytopoiesis.

The *megakaryoblast* has a size between 20 to 30 micron. Its nucleus is large, oval or kidney-shaped and contains several nucleoli. It has an insignificant non-granular and slightly basophilic cytoplasm. The *pro-megakaryocyte* is very similar to the megakaryoblast except from the presence of an intensely basophilic cytoplasm that contains fine azurophillc granules. The *megakaryocyte* instead is much bigger than its precursors, having a size around 50-100 micron. It presents a single, indented nucleus and a basophilic (light blue) cytoplasm that contains azurophilic granules. Each megakaryocyte fragments into thousands of platelets. *Platelets* are very small, about 3 micron, the cytoplasm is stained light blue and it contains purple-reddish

granules. Platelets play a key role in haemostasis, and they are involved in the formation of blood clot. A low number of platelets is called thrombocytopenia, while a decrease in function of platelets is called thrombasthenia. In some people, too many platelets may be produced, which may result in interferences with the flow of blood. An increase in the number of platelets is called thrombocytosis. Sometimes this problem could cause bleeding, because many of the extra platelets may be dysfunctional even though they appear normal. A platelet count is usually evaluated by preparing a blood smear to directly visualize any anomalies in shape or size. In fact blood smear could present platelets greater than 3 micron in diameter, that are called macrocytic platelets or megathrombocytes. This kind of platelets are disregarded by the modern haematology analysers since they count the platelets based of their sizing. Also the count will also be falsely low when there are platelets clumps. In such instances the instrument does not count these clumps of platelets and gives the platelet count as falsely low. In a normal person usually less than 5% of the platelets appear large. Platelet size is of diagnostic significance, particularly if considered in relation to the platelet count. Small or normal-sized platelets in association with thrombocytopenia is suggestive of a failure of bone marrow production, while thrombocytopenia with large platelets is more likely to be caused by peripheral destruction or consumption of platelets with the bone marrow responding by increasing platelet production. Platelet size is also useful in assessing the likely cause of thromhocytosis.