Methodology for Automatic Classification of Atypical Lymphoid Cells from Peripheral Blood Cell Images

Edwin Santiago Alférez Baquero

Supervised by

Dr. José Rodellar

Dra. Anna Merino

Department of Applied Mathematics III

Universitat Politècnica de Catalunya · BarcelonaTech

This dissertation is submitted for the degree of

Doctor in Biomedical Engineering

February 2015

Dedicated to all my family, specially to my son Thiago.

Acknowledgements

In first place, I would like to acknowledge to Dr. José Rodellar and Dra. Anna Merino, who since I arrive to Barcelona, have supervised, supported, encouraged, and helped me during the development of this research. I am very fortunate to have them as advisors, who not only helped me to develop as a scientific researcher, but also to grow as a person. I will always appreciate what they have done for me.

I owe my deepest gratitude to the "Universitat Politècnica de Catalunya" for the granting of my pre-doctoral fellowship. Special thanks to the "Escola Universitària d'Enginyeria Tècnica Industrial de Barcelona" and the research group Control, Dynamics and Applications (CoDAlab) and all its members for their invaluable support in everyday that I work with them.

I am deeply grateful to Luis Mujica for introducing me in the research group. I thank him and Magda Ruiz for having supported me in good and bad times, and for giving me their precious friendship.

I am grateful to the Cytology Unit of the Core lab in the Hospital Clinic of Barcelona for the use of its facilities in the acquisition of the cell images and the confirmed diagnosis of the patients.

I am very grateful to Laura Bigorra for her contribution to the clinical analysis and the acquisition of the information in this thesis, and specially for giving me her friendship.

I would like to express my special appreciation and thanks to Professor Adriann Houtsmuller and all the members of his research group of the Erasmus MC in Rotterdam (Netherlands), specially to Martin van Royen, for their guidance and support during my research visit in 2013, allowing me to extend my knowledge to the image analysis of dynamics and reaction kinetics of nuclear process by using confocal microscopy.

I would like to show my greatest appreciation to international reviewers of my thesis, Professor Fabio Scotti from the University of Milan and Dr. Antonio Ermens from the Laboratory for Clin.Chemistry & Hematology of the Amphia Hospital, for their brilliant comments, suggestions and feedbacks.

I am thankful to Dr. Raimon Jané to be my academic tutor during my doctoral studies.

I would like to show my greatest appreciation to all the Professors in the Industrial University of Santander (UIS) in Colombia, who contributed to my formation as electronic engineer and physicist, and for giving me all the foundations as a researcher.

I am grateful to Diego Tibaduiza for his friendship during several years and for guided me during my first moments in Barcelona. I also thank Fahit Gharibnezhad for his friendship inside and outside the research group.

I am deeply grateful to all my friends in Barcelona, who have supported me in different moments of my stay, specially to my best friends Julian Barragan and Eduard Galvis for their valuable friendship, support and advice during the past few years.

I would like to express my greatest thanks to all my family in Colombia, specially to my mother Gladys Baquero, my brother Andrés Alférez and my aunts Ana María Baquero and Inirida Baquero, who from the distance always encouraged me to keep jumping obstacles in my life.

Many special and sincere thanks to Leila Carolina Sánchez who has sacrificed and helped me in my doctoral studies. Her patient and love has motivated me at different moments of this journey.

Finally, I apologize in advance for not remembering and including in the above lines people who really also deserve it.

Abstract

Morphological analysis is the starting point for the diagnostic approach of more than 80% of the hematological diseases. However, the morphological differentiation among different types of abnormal lymphoid cells in peripheral blood is a difficult task, which requires high experience and skill. Objective values do not exist to define cytological variables. This sometimes results in doubts on the correct classification in the daily hospital routine. Automated systems exist for digital peripheral blood cell analysis, but they operate most effectively in non-pathological blood samples.

The general objective of this thesis is to develop a complete methodology to automatically recognize images of normal and reactive lymphocytes, and several types of neoplastic lymphoid cells circulating in peripheral blood in some mature B-cell neoplasms using digital image processing methods. This objective has two directions: (1) with engineering and mathematical background, transversal methodologies and software tools are developed; and (2) with a view towards the clinical laboratory diagnosis, a system prototype is built and validated, whose input is a set of pathological cell images from individual patients, and whose output is the automatic classification in one of the groups of the different pathologies included in the system.

This thesis is the evolution of various works, starting with a discrimination between normal lymphocytes and two types of neoplastic lymphoid cells, and ending with the design of a system for the automatic recognition of normal and reactive lymphocytes, and five types of neoplastic lymphoid cells.

All this work involves the development of a robust segmentation methodology using color clustering, which is able to separate three regions of interest: cell, nucleus and peripheral zone around the cell. A complete lymphoid cell description is developed by extracting features related to size, shape, texture and color. To reduce the complexity of the process, a feature selection using theory information is performed. Then, several classifiers are implemented to automatically recognize different types of lymphoid cells. The best classification results are achieved using support vector machines with radial basis function kernel.

The methodology developed combining medical, engineering and mathematical backgrounds is the first step to design a practical diagnosis support tool in the near future.

Contents

List of Figures

List of Tables

List of Abbreviations

Roman Symbols

AML Acute Myeloid Leukemia

ANN Artificial Neural Networks

BPL B-prolymphocytes

CLL Chronic Lymphocytic Leukemia

CMI Conditional Mutual Information

CMIM Conditional Mutual Info Maximization

CMYK Cyan, Magenta, Yellow, Key (Black)

DIP Digital Image Processing

DWT Discrete Wavelet Transform

EFD Elliptical Fourier Descriptors

FCM Fuzzy C-means

FL Follicular Lymphoma

GLCM Gray Level Co-ocurrence Matrix

GVF Gradient Vector Flow

HCL Hairy Cell Leukemia

JMI Joint Mutual Information

KFCM Kernel Fuzzy C-means

List of Abbreviations

kNN k Nearest Neighbor

LDA Linear Discriminant Analysis

LVQ Linear Vector Quantization

MCL Mantle Cell Lymphoma

MGG May-Grünwald-Giemsa

MLP Multi Layer Perceptron

MRMR Minimum-Redundance Maximum-Relevance

NB Naive Bayes

PB Peripheral Blood

PCA Principal Component Analysis

QMF Quadrature Mirror Filters

RBC Red Blood Cell

RBF Radial Basis Function

ROI Regions Of Interest

sFCM spatial Fuzzy C-means

sKFCM spatial Kernel Fuzzy C-means

STD Standard Deviation

SVM Support Vector Machines

WBC White Blood Cell

WHO World Health Organization

WT Watershed Transformation

Chapter 1

Introduction

1.1 Motivation

Severe hematological diseases, especially leukemias and lymphomas, are common in all life periods. Early detection of the presence of leukemic cells in the peripheral blood (PB) and the subsequent possibility of a prompt treatment are essential for the patient survival. Morphologic analysis is the starting point for the diagnostic approach of more that 80% of these diseases. However, the morphologic differentiation among different types of abnormal lymphoid cells in B lymphoid neoplasms and among blast cells in PB is a difficult task, which requires high experience and skill. Objective values do not exist to define cytological variables. Moreover, subtle morphologic characteristics exist that are exhibited by some malignant neoplastic lymphoid cells, which are shared with reactive lymphoid cells. This sometimes results in doubts on the correct classification in the daily hospital routine.

The hematological diagnosis starts with the morphologic analysis and continues with other more complex procedures such as flow cytometry by using monoclonal antibodies and genetic and molecular studies, which are available only in highly specialized clinical laboratories due to the equipment costs and the required human skills. In this context, a methodology able to automatically analyze objective morphologic features of lymphoid cells from the images obtained through the optical microscope, could be a practical support tool to the morphologic diagnosis in clinical laboratories.

Recently, there have appeared equipments that perform automatic preclassification of PB cells based on digital image processing (DIP). They show a high efficiency in the recognition of PB normal cells such as neutrophils, lymphocytes, monocytes, eosinophils and basophils. However, these analyzers are not able to automatically identify neoplastic lymphoid cells circulating in various hematological diseases. On the other hand, in recent years some approaches have been reported in the literature trying to fill the gap in the automatic recognition between

normal and abnormal PB cells. The aim of this thesis is to contribute with new developments in this context, combining medical, engineering and mathematical backgrounds.

1.2 Objectives

The main goal of this thesis is to develop a complete methodology to automatically recognize images of normal lymphocytes and several types of neoplastic lymphoid cells circulating in peripheral blood in some mature B-cell neoplasms using digital image processing methods. To achieve this goal, the following specific objectives are proposed:

- To perform a first exploratory study of different digital image processing techniques to evaluate their usefulness for the discrimination of normal lymphocytes and some types of neoplastic lymphoid cells.
- To develop and validate a robust segmentation method to separate the regions of interest of the lymphoid cell images.
- To propose and implement the extraction of features that describe size, shape, color and texture of the regions of interest on the lymphoid cell images.
- To identify and select the best features that provide useful information about the morphologic characteristics of normal lymphocytes and different neoplastic lymphoid cells.
- To develop a methodology for automatic classification of normal and neoplastic lymphoid cells using the extracted and selected feature set.
- To validate the completed developed methodology for the automatic recognition of normal lymphocytes and neoplastic lymphoid cells, in a scenario where the methodology can be useful in clinical practice.

1.3 Research methodology

Digital image processing comprises the use of computational methods to process digital images producing other images (acquisition and preprocessing), but also includes procedures for the extraction of attributes of the image (segmentation and feature extraction) up to the recognition of individual objects (classification). The DIP framework used in this thesis mainly consists of the following steps as illustrated in Figure 1.1:

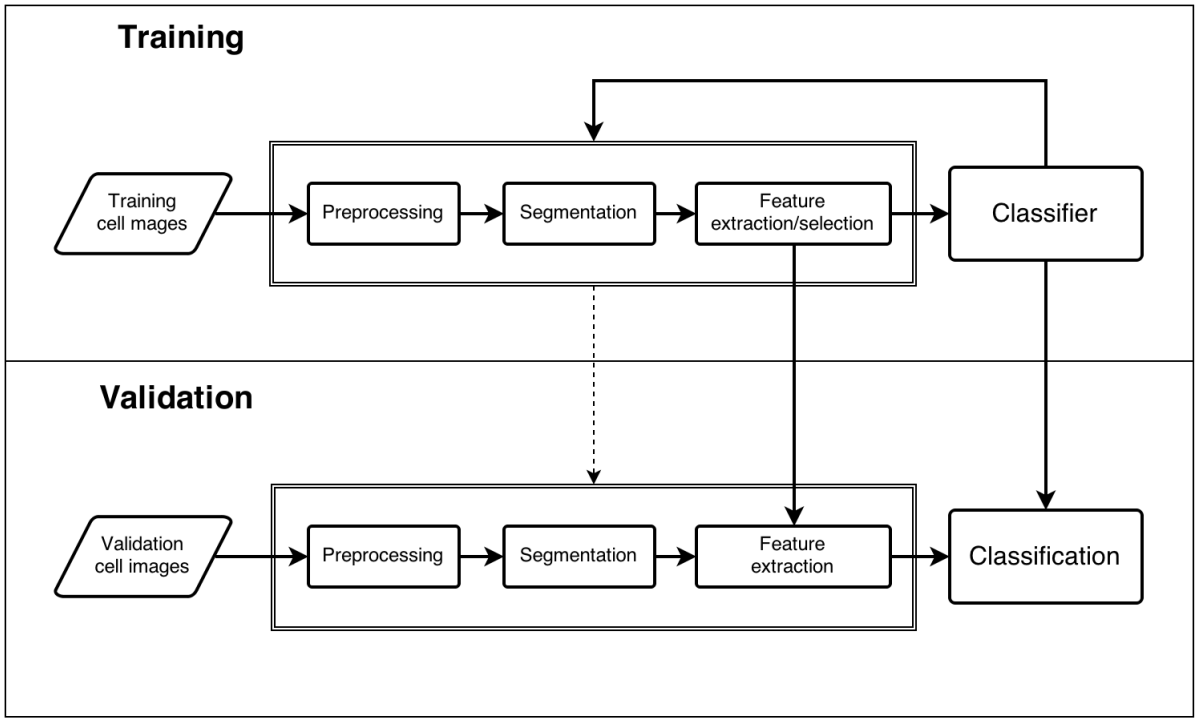


FIGURE 1.1: *Digital image processing workflow used in the research methodology of this thesis.*

Acquisition In this step the PB digital cell images are obtained from an image analyzer device which integrates a motorized microscope and a digital camera. Blood samples are obtained from the routine workload of the Core Laboratory of the Hospital Clínic of Barcelona.

Preprocessing This step groups all computational methods that attempt to improve the information present in the cell image. In this work some preprocessing techniques such as filtering process and color transformations are done .

Segmentation This is the step where the objects of interest are separated according to some similarity criteria. Color clustering and watershed transformation are used to segment the regions of interest of the cell image.

Feature extraction/selection In this step the features (size, shape, color, texture, etc.) of the segmented objects are measured by quantitative (or qualitative) parameters. If there is a high number of extracted features, it is necessary to reduce the amount of information. This work uses information theoretic feature selection to choose the best features according to their relevances and redundancies.

Classification Through the features obtained from each segmented cell in the images, pattern recognition techniques are used to classify the corresponding cell types. This thesis investigates several classification methods to develop an efficient classifier.

The workflow in this thesis follows two directions: the development and the validation of the methodology, which correspond to the top and bottom parts of Figure 1.1, respectively. The first is oriented to the development of the methodologies, in which sets of images are used in a training mode to build a robust segmentation procedure, a complete feature extraction and selection step and a sucessful classification of lymphoid cells. The second direction is devoted to the experimental validation of the methodology using sets of cell images from individual patients not included in the training mode. The classification of different lymphoid cells considering new patients allows to validate the methodology in a scenario close to the clinical practice.

In this thesis, all the algorithms for the development and validation of the methodology have been implemented using the scientific and high-level language MATLAB®.

1.4 Research framework

About four years ago a collaboration between members of the research group CoDAlab (Control, Dynamics and Applications) of the Technical University of Catalonia (UPC), led by Prof. José Rodellar and, Dr. Anna Merino, who develops her healthcare, research and teaching tasks at the Center for Biomedical Diagnosis of Hospital Clínic (HC) of Barcelona. A new research line was initiated to explore the application of digital image processing and pattern recognition techniques for automated classification of lymphoid cells in peripheral blood. All this came from the experience acquired in the HC with new laboratory equipments that incorporate software tools, but with serious limitations for recognition of malignant cells.

This thesis was proposed within this context and was made possible thanks to a pre-doctoral fellowship (FPI-UPC) granted by the UPC to Santiago Alférez. The work has been carried out in the space of CoDAlab, Department of Applied Mathematics III in the Barcelona College of Industrial Engineering (Escola Universitària d'Enginyeria Tècnica Industrial de Barcelona), with strong interaction and use of the facilities of the Cytology Unit of the Core lab in the Hospital Clinic of Barcelona.

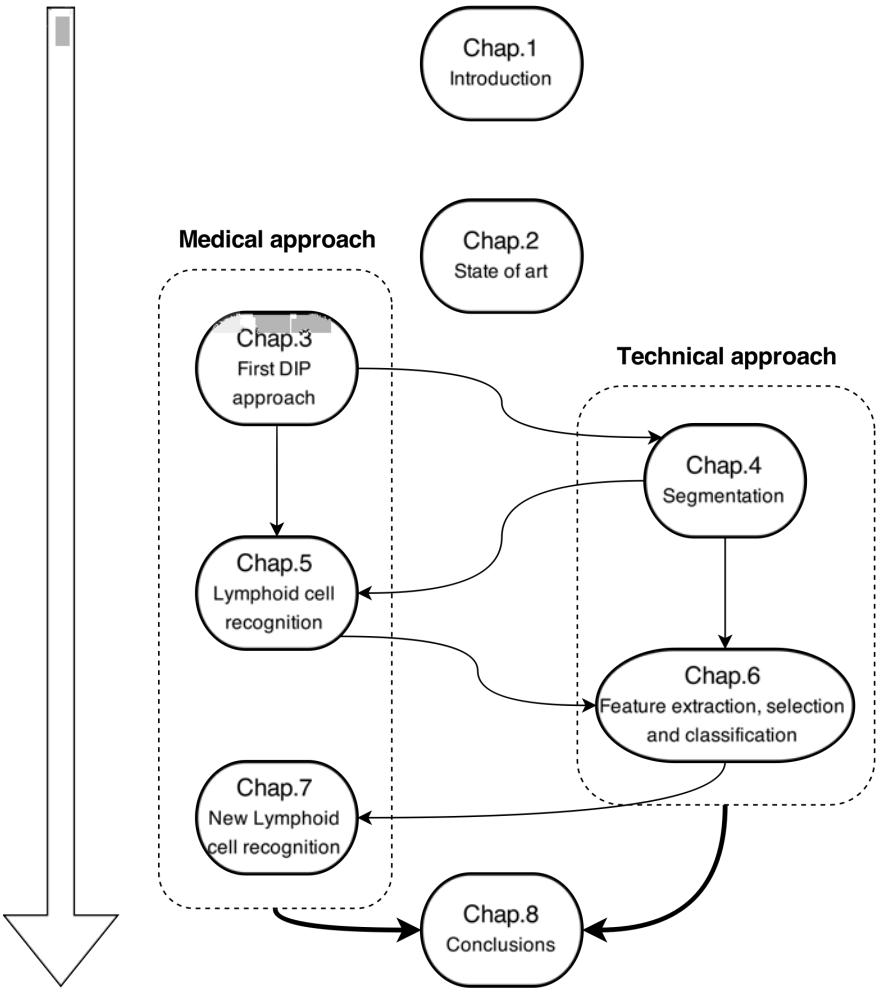


FIGURE 1.2: *Thesis outline*

1.5 Thesis outline

Figure 1.2 illustrates the reading plan of this thesis, which allows to understand the organization and the relations between the chapters. The present thesis consists of eight chapters starting with this introduction where the motivation, the objectives, the DIP workflow, the research framework and the organization are described. Chapter 2 includes a review of some basic concepts about morphologic analysis of peripheral blood, and a state of art of the digital image processing of blood cells. From Chapter 3 to Chapter 7 there is a special distribution of reading according to two main topics: Chapters 3, 5, and 7 follow a medical approach, while Chapters 4 and 6 present the main technical approaches developed in this thesis.

Chapter 3 presents the first exploratory work when the research was initiated in 2011, using a first segmentation algorithm and a non-supervised classification technique with only three groups of cells: normal lymphocytes and two types of neoplastic lymphoid cells. It is

our first published paper in the Journal of Laboratory Hematology (2013). In Chapter 4, the segmentation methodology is fully developed, which uses color clustering methods to separate three regions of interest: the nucleus, the cytoplasm and the peripheral zone around the cell. This segmentation method is used in the subsequent chapters of the thesis. Then, Chapter 5 represents a step forward in the developments since it incorporates the segmentation method from Chapter 4 and improves the feature extraction step described in Chapter 3, developing a methodology for the supervised classification of normal lymphocytes and four types of neoplastic lymphoid cells. To do that, Chapter 5 involves the extraction of geometric and color-texture features for each cell and the implementation of a linear discriminant classifier. The work in Chapter 5 has been published in the American Journal of Clinical Pathology (2015).

The addition of more color and texture features and the exploration of other classification methods will allow to classify other types of atypical lymphoid cells, which is important in view of a potential use as a diagnosis support tool in clinical practice. With this aim, Chapter 6 goes back to the technical direction and presents a complete feature extraction/selection and classification methodology through different techniques and results from several classification (training) experiments to develop the learning process. Chapter 6 integrates the segmentation method presented in Chapter 4 with the new feature extraction/selection and classification steps, so that the methodology developed in this thesis is completed.

Chapter 7 implements the methodology for the automatic recognition of normal lymphocytes, reactive lymphocytes and five neoplastic lymphoid cells, using a big training set of peripheral blood cell images from different patients. Then, the methodology is validated considering a new group of patients in a horizon where it could be useful in clinical practice. Chapter 8 summarizes the end conclusions of the thesis and some perspectives on possible improvements and future works.

Chapter 2

State of the Art of Digital Blood Cell Image Processing

2.1 Introduction

This thesis addresses the need for automated methods for the classification of neoplastic lymphoid cells associated to mature B-cell neoplasms by using digital image processing (DIP) techniques in PB. This chapter presents a review of some basic concepts about morphologic analysis and a state of the art of the digital image processing of blood cells. The Chapter is organized as follows. Section 2.2 introduces the concept of Hematopoiesis, Section 2.3 describes the the mature B-cell and their classification, Section 2.4 explains some basics about morphologic analysis of the peripheral blood cells, Section 2.5 shows the trend of automated digital morphology systems, and Section 2.6 presents a state of art of digital image processing of peripheral white blood cells.

2.2 Hematopoiesis

Hematopoiesis is defined as the production, development, differentiation, and maturation of all blood cells [1], see Figure 2.1. The blood circulating through the blood vessels or PB contains the following cells: leukocytes (basophilic, eosinophilic and neutrophilic segmented granulocytes, monocytes and lymphocytes), erythrocytes and platelets (or thrombocytes). They are created in the bone marrow and are derived from the maturation of myeloid or lymphoid lineages of stem cells. These stem cells have two basic functions: (1) self-renewal, and (2) differentiation and maturation. The lymphoid stem cell produces lymphoblasts, which can differentiate in B and T lymphoid cells and Natural killer cells. On the other hand, the

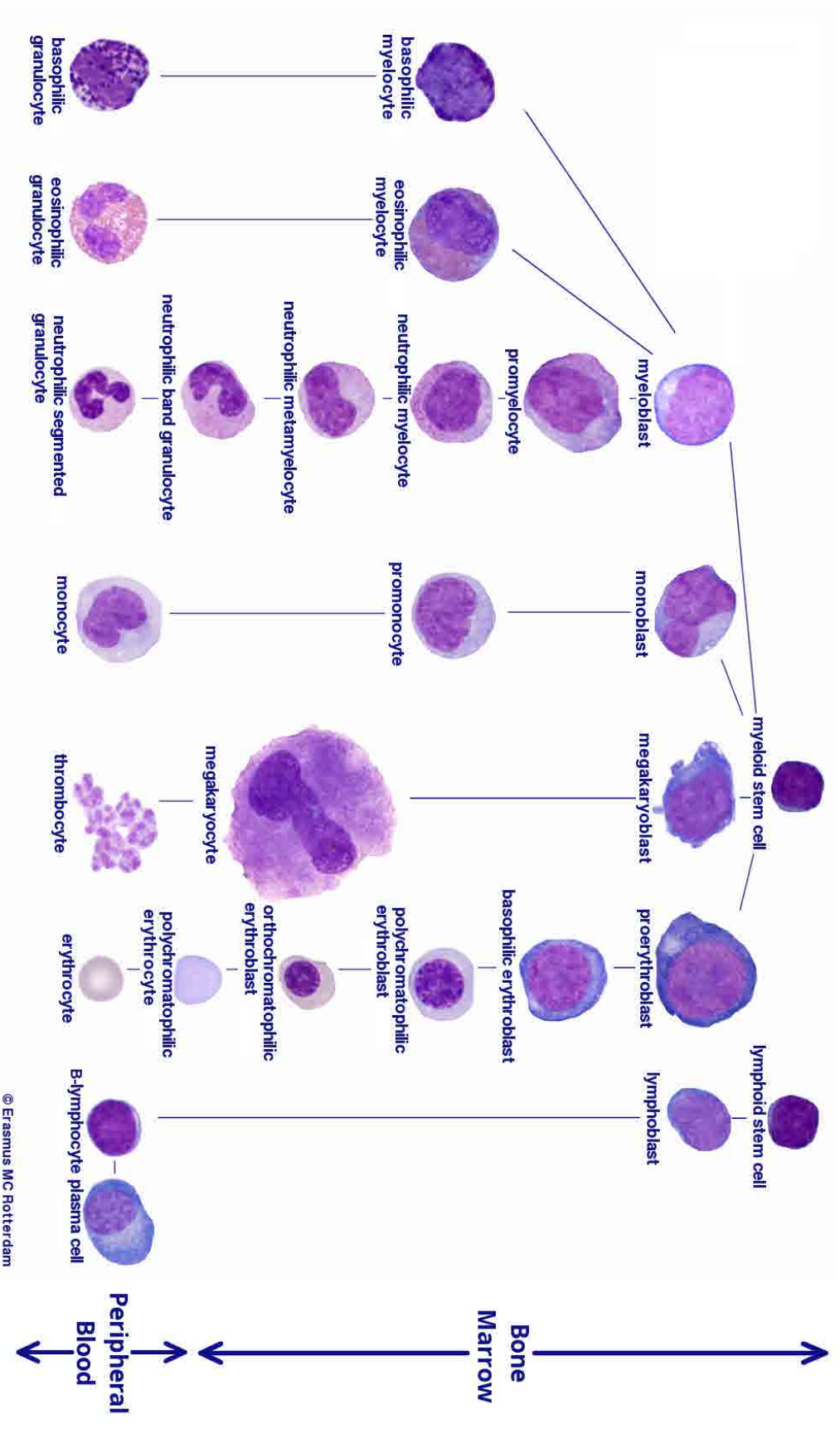


FIGURE 2.1: *Scheme of the Hematopoiesis of the various lineages of all the mature cells from a pluripotent stem cell (bone marrow).*

myeloid stem cell derives a transitional (differentiated) stem cell that originates the erythroid, megakarycytic, myeloid, monocytic, eosinophilic or basophilic lineages. Cell proliferation involves amplifying the number of mature cells produced from a cell that has been differentiated to a particular cell lineage. Normally, there is a balance between the quiescence and the ability for self-renewal of hematopoietic stem cells, with a control of the proliferation, the apoptosis and the differentiation of the progenitors to mature cells. Moreover, cell differentiation implies the progressive development of biochemical, functional and structural characteristics for a specific cell type [2–4]. This thesis deals specifically with B lymphoid cells.

2.3 Mature B-cell neoplasms and their classification

Lymphoid neoplasms are cancers that proceed from lymphoid cells of the immune system in various stages of differentiation. Which derive in a broad spectrum of immunological, morphologic and clinical findings. These neoplasms can occur in the form of leukemia (involvement of bone marrow and blood) and/or lymphomas (solid tumors) [5].

The medical classification is the language of medicine, allowing to describe, define and name a disease before this can be diagnosed, treated and studied. Then, there must be a consensus of definitions and terminology which are essential for both clinical practice and research of these type of pathologies. In 2008, the World Health Organization (WHO) presented a classification of the lymphoid neoplasms, considering the cell of origin, in the following groups [6, 7]:

1. Precursor B-cell lymphoid neoplasm (immature B-cells)
2. Precursor T-cell neoplasms (immature T-cells)
3. Mature B-cell neoplasms
4. Mature T-cell and natural killer cell neoplasms
5. Hodgkin lymphoma

Precursor B-cells that mature in the bone marrow may undergo apoptosis or develop into mature naïve cells that, following exposure to antigen and blast transformation may develop into plasma cells or enter the germinal center of the lymphoid follicle (see Figure 2.2). In other words, normal B-cell differentiation begins with precursor B-cells known as progenitor B-cells which differentiate into mature naïve B-cells in the bone marrow. These naïve B-cells circulate in PB and also occupy primary lymphoid follicles and follicle mantle zones. Naive B lymphoid cells maturation occurs in the lymphoid follicle. Most cases of mantle cell

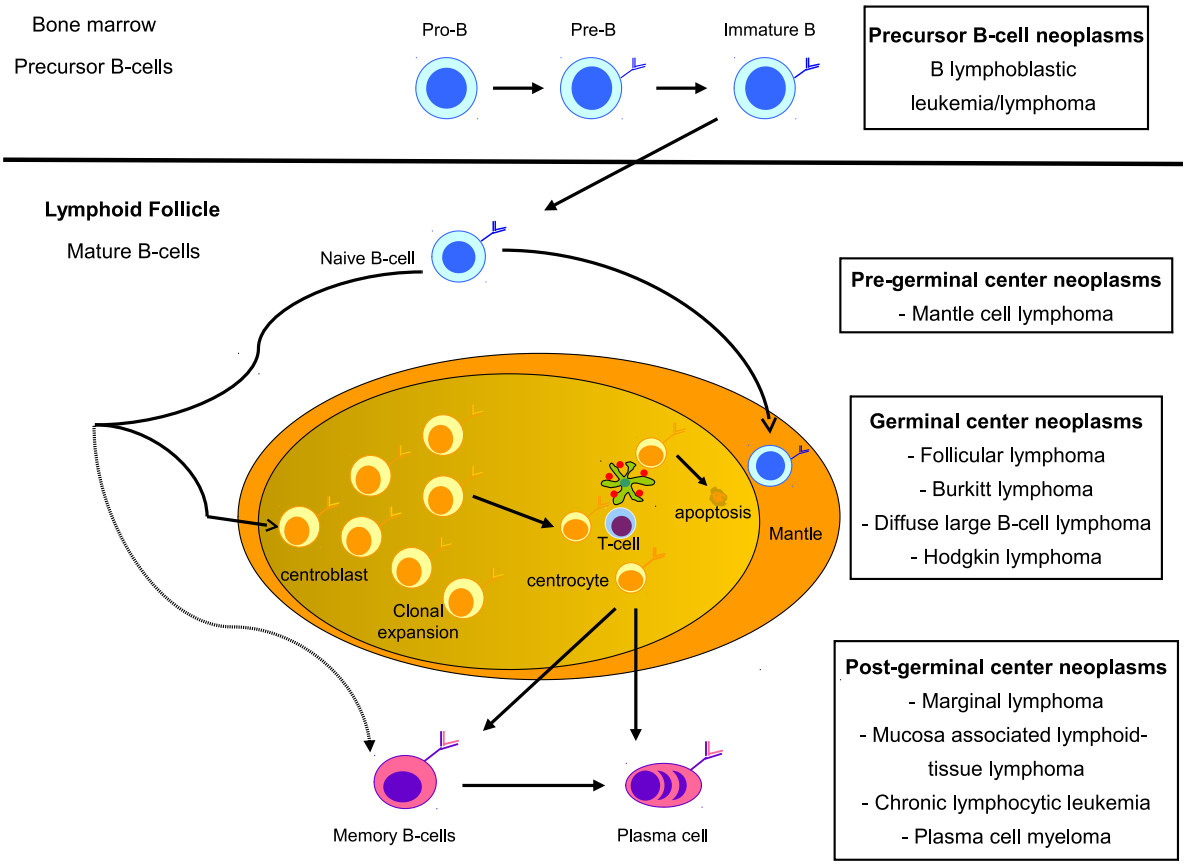


FIGURE 2.2: *Diagrammatic representation of B-cell differentiation and relationship to B-cell neoplasms. B-cell neoplasms correspond to stages of B-cell maturation. Precursor B-cells that mature in the bone marrow may undergo apoptosis or develop into naïve B-cells in the lymphoid follicles where, following exposure to antigen enter to the germinal center (GC). Centroblasts, the transformed cells of the GC either undergo apoptosis or develop into centrocytes. Post GC cells included both long-lived plasma cells and memory/marginal zone B cells.*

lymphoma are thought to correspond to naïve B-cells. Naïve cells undergo transformation, proliferate and mature to antibody secreting plasma cells and memory B-cells. Transformed cells that have encountered antigen may mature directly into plasma cells that produce IgM antibody response to antigen [6].

B-cell and T/NK neoplasms are clonal tumors of mature and immature B-cells, T-cells or natural killer (NK) cells at various stages of differentiation. Most of the B-cell neoplasms arise from follicle center cells. When these different types of neoplastic lymphoid cells in B-cell neoplasms reach the PB it means that we have leukemic cells from mature lymphoid neoplasm origin and we have the possibility to analyze the morphology and the immunophenotype or these abnormal cells. Morphology and immunophenotype are sufficient for the diagnosis of most lymphoid neoplasms.

Mature B-cell neoplasms comprise over 90% of lymphoid neoplasms worldwide. The classification of lymphoid neoplasms is based on utilization of all available information to define the different diseases. The multiparameter approach to classification adopted by the WHO classification has been validated in international studies as being highly reproducible and enhancing the interpretation of clinical studies.

2.3.1 WHO classification of mature B and T cell neoplasms

The following list contains the B and T-cell mature neoplasms in which the circulation of leukemic cells in PB more frequently can be detected. This thesis is focused in some of the groups of mature B-cell neoplasms included in this list.

Mature B-cell neoplasms

- Chronic lymphocytic leukemia (CLL)
- B-cell prolymphocytic leukemia (B-PL)
- Splenic marginal zone lymphoma
- Hairy cell leukemia (HCL)
- Lymphoplasmacytic lymphoma
- Plasma cell myeloma
- Extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma)
- Follicular lymphoma (FL)
- Mantle cell lymphoma (MCL)
- Diffuse large B-cell lymphoma (DLBCL)
- Marginal zone lymphoma
- Burkitt lymphoma

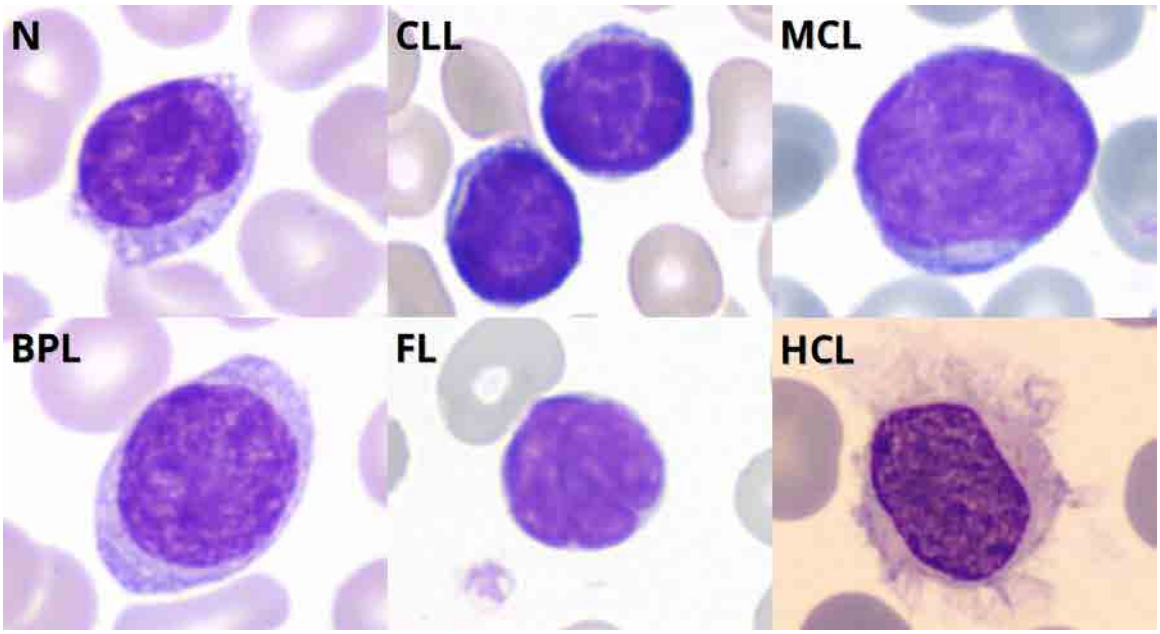


FIGURE 2.3: *Several examples of neoplastic lymphoid cells images. N, normal lymphocytes; CLL, chronic lymphocytic leukemia; MCL, mantle cell lymphoma; BPL, B-prolymphocytes; FL, follicular lymphoma; HCL, hairy cell leukemia.*

Mature T-cell and NK neoplasms

- T-cell prolymphocytic leukemia
- Aggressive NK cell leukemia
- Adult T-cell leukemia/lymphoma
- Sézary syndrome

2.4 Morphologic analysis of peripheral blood

The White Blood Cell (WBC) differential count measures the percentage of each of the subpopulations of leukocytes. This test can be requested directly for the clinician with a specific diagnostic suspicion, subject to lymphadenopathy and/or splenomegaly. It is also a task of the physician of the laboratory to indicate the realization of the WBC differential when the results of the complete blood count have some quantitative anomaly, or any alarm occurs in the autoanalyzer. The manual count under the microscope have several protocols with particular criteria because it is a laborious test [4]. However, a careful observation of this procedure provides a lot of information, being a valuable tool in both diagnostic

orientation (which guides the complementary tests) and performing some definitive diagnoses. Moreover, since this morphologic study is the first step in the diagnosis, the recognition of neoplastic lymphocytes in blood smears can contribute to a fast diagnosis of B-cell diseases, enabling rapid therapeutic intervention influecing the prognosis [8]. An altered number of WBCs in PB, leukocytosis (high counts) or leukopenia (low counts), can indicate different pathologies. Often the diagnosis of CLL is made by the accidental finding of a leukocytosis with lymphocytosis on a routine blood test in a patient of advanced age and the detection of abnormal lymphocytes in the smear. The final diagnosis is accomplished through the integration of clinical findings, bone marrow aspirate, immunophenotypic, cytogenetic and molecular biology studies [6, 9].

Morphologic distinction between various types of lymphoid cells requires experience and skill and, moreover, objective values do not exist to define cytological variables. For example, some qualitative descriptions of various types of neoplastic lymphoid cells are presented above. CLL cells are typically small lymphocytes with clumped chromatin and scant cytoplasm. HCL cells are larger than normal lymphocytes and they have abundant weakly basophilic cytoplasm with irregular *hairy* margins. In FL neoplastic lymphocytes have an irregular nuclear membrane and condensed chromatin without nucleoli. In MCL, the cells have an irregular, cleaved nuclear contour and may resemble those of FL. Prolymphocytes are slightly larger cells with nucleoli. Figure 2.3 shows several images corresponding to the above neoplastic lymphoid cells.

2.5 Automated digital morphology systems

At the beginning of XXI century, a new trend started in the digital morphology with the development of new equipment able to preclassify different types of normal WBCs in a fast and efficient way. These systems use motorized microscopy, digital image processing and pattern recognition techniques to automatically identify nucleated cells as well as perform a morphologic evaluation of the Red Blood Cells (RBCs). Subsequently, the images are displayed on a screen so that the physician can confirm or reclassify the cells.

Medica EasyCell® assistant Cell Image Analysis System (Sysmex America, Inc, Mundelein, IL) [10] automatically detects WBCs on a blood smear, then it reaches classify normal WBCs, smudge cells and nucleated RBCs. This system employs image processing and artificial intelligence to do that work. Next, the cell images are grouped and displayed for review.

The HemaCAM® system (Fraunhofer-Gesellschaft, Germany) [11, 12] is able to evaluate the PB samples and make WBC differential. This equipment can automatically preclassify different types of leukocyte as: neutrophils, basophils, eosinophils granulocytes, lymphocytes,

monocytes, nuclei shadows, and large platelets. Since 2010, Hemacam has been available on the market as a certified medical product.

Bloodhound® is developing the *cobas m 511* system [13], which combines a digital morphology analyzer, a cell counter, a classifier in only one equipment that include an auto-stainer and a microscope analyzer of blood smears. This technology analyzes the cell morphology, counts and classifies each cell providing a 5-part differential of WBCs, platelets and reticulocyte count. Currently, this instrument is not available in the USA and is not approved by the Food and Drug Administration (FDA).

CellaVision® DM96 (CellaVision AB, Lund, Sweden) [14] is an automated device for the differential preclassification of WBCs, evaluation of RBC morphology, platelet estimation on the blood smears and it can also analyze body fluids. This equipment includes a motorized microscope, a camera and a computer containing the acquisition and classification software. This system preclassifies WBCs, then a medical operator can confirm or reclassify the suggested cell classification. The Cellavision DM96 has been studied in several works which evaluate the concordance between automated differential count of this device and the manual differential count made by the physician with the microscope:

- Kratz et al. [15] evaluates the CellaVision DM96 obtaining values of correlation between the automated preclassification and the usual microscopy of WBCs between 0.67 and 0.96. The sensitivity values is between 95% and 100% and the specificity values between 88% and 97%, depending on the abnormality. It concludes that this automated analyzer showed a similar performance to the manual clinical microscopy.

- Cornet et al. [16] analyzes samples from 440 patients using the CellaVision DM96. This system obtains an efficiency of 95% for the automated preclassification, which is increased up to 98% after a manual reclassification by the medical technologist. The correlation of the manual and automated methods is excellent for neutrophils, lymphocytes and eosinophils, acceptable for immature granulocytes, erythroblasts and basophils, and low for monocytes. In this paper some B-cell chronic lymphoproliferative disorders are evaluated, but the DM96 equipment is not able to classify correctly the abnormal samples. The correct classification is done manually by the medical operator trough the screen and the CellaVision software.

- Briggs et al. [17] compares the differential done by different operators using the CellaVision DM96 with the manual differential under the microscope. The accuracy of the automated preclassification is 89.2%, while precision is similar to the 100-cell manual differential. Then, it concludes that DM96 accuracy depends on both blood pathology and the experience of the medical operator.

- Merino et al. [18] evaluates the automated preclassification performed by the CellaVision DM96 and the subsequent reclassification by the medical operator. It gets excellent values of correlation (preclassification) between this device and the conventional microscopy for segmented neutrophils, lymphocytes, monocytes and blasts, and good values for eosinophils, basophils and plasma cells. Furthermore, the correlation for the reclassification is very good for promyelocytes and myelocytes, intermediate for reactive lymphocytes and erythroblasts, and low for metamyelocytes. Neoplastic lymphoid cells are reclassified only by the medical user, because the DM96 is not able to do that.

The conclusions from the above studies are the good correlation and concordance of the application of the digital image analysis over the traditional method by direct microscopy, but emphasizing the need of subsenquently validation and review by a skilled physician. They also note the reduction of the time spent in the analysis by improving the workflow, the efficiency and the quality. Finally, several of these papers highlight that the DM96 is not able to automatically classify neoplastic lymphoid cells. Thus, the automatic neoplastic lymphoid cell recognition was the problem that gave rise to the approach of this thesis.

2.6 Peripheral blood digital cell image processing: state of the art

As it was mentioned above, several devices have been developed which use DIP techniques to achieve an automated preclassification of nucleated cells from PB involving a lot of calculations based on morphologic characteristics of the cell such as color, size, shape, texture, among others [9, 19]. However, although theses analyzers represent a technological advance of great interest because they are able to preclassify the majority of normal WBCs, they cannot automatically identify the different types of neoplastic lymphoid cells such as: centrocytes, centroblasts, prolymphocytes, hairy cells, Sézary cells and other ones circulating in PB in some lymphoid neoplasms.

Due to the difficulty of the correct automatic classification of neoplastic lymphoid cells, few studies have been published using different methods of DIP with satisfactory results.

FIGURE 2.4: *Simplified digital image processing framework.*

The problem has been addressed by the extraction of a large number of measurements and parameters that describe the morphologic features of interest in cells, along with pattern recognition techniques used to classify different types of cells [20–22].

Figure 2.4 shows a simplified scheme of the DIP, which is the same for all applications including the automated recognition of cells. The main stages of the DIP are: acquisition, preprocessing, segmentation, feature extraction and classification. The state of the art on these DIP stages relevant for this thesis is reviewed below.

2.6.1 Acquisition

The acquisition is the first process in the DIP, in which images are obtained. In the analysis of PB cells the preparation of an adequate blood smear and its staining are essential to generate quality images. Various techniques are used for the staining process, but the May-Grünwald-Giemsa (MGG) stain is widely used due to its properties, which highlight the basic components of the blood cell [3]. It is very important to develop an accurate and automated methodological process for the blood smear of high quality in terms of staining,

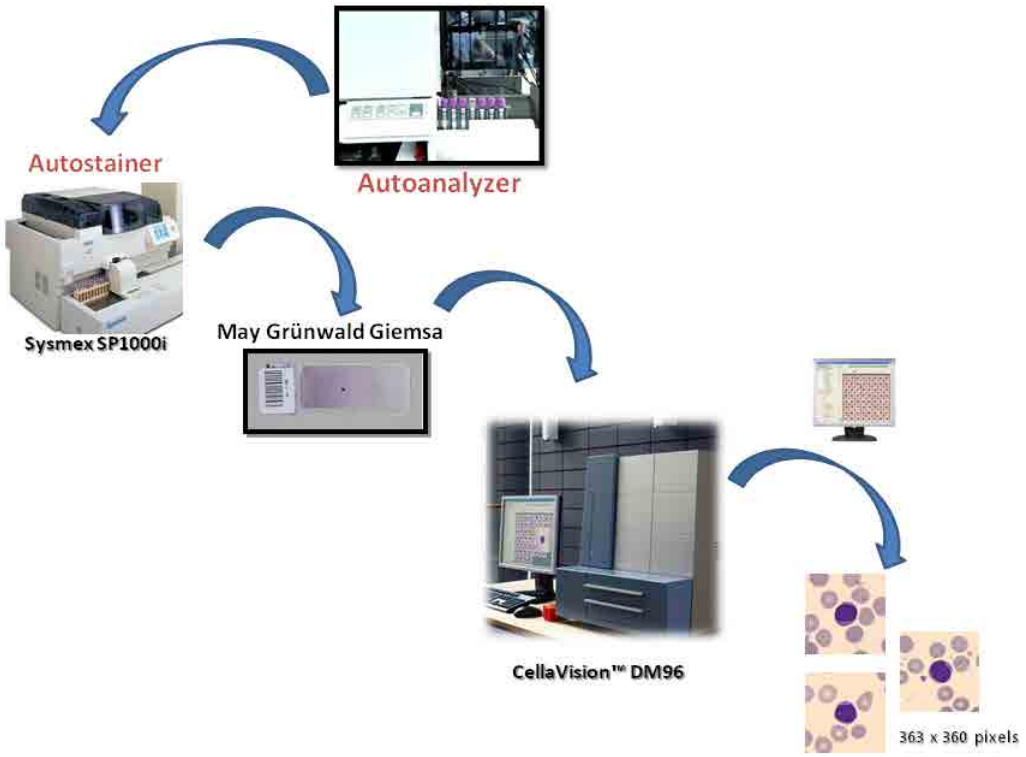


FIGURE 2.5: *Acquisition process protocol of the blood cell images. The blood smears from pathological samples from the autoanalyzer are automatically stained with May-Grünwald-Giemsa and the the images are obtained by the Cellavision DM96.*

since it is the first step in the acquisition process and it is essential for the development of post-processing algorithms. Thus, extensions of low quality will result in degraded image, i.e. with the inclusion of noise in the entire system performance.

To acquire blood cell images, the most widely used method is by optical microscopy with immersion objectives of 50 to 100 magnifications (500 to 1000 in total), followed by a digital camera with CCD sensors mounted in the optical path [23]. However, automated digital morphology devices incorporate a motorized microscope with a digital camera such as the CellaVision DM96.

Figure 2.5 shows the standardized acquisition protocol used in this work, where the pathological samples of the peripheral blood smears are automatically stained with MGG and then they are passed through the DM96 to get different images with a specified resolution. This device is used because it allows to have a big database of pathological samples.

2.6.2 Preprocessing

There are different factors that may affect the quality of the acquisition of images derived from the blood smear: staining process, variations in lighting, optical geometric distortions due to the type of lens and microscope, format in which the images are stored (not recommended compressed formats and thus lose data), random noise, lack of contrast between tone levels, etc.

Some preprocessing techniques have been described to improve the quality of the image before the segmentation process, either to select the absolute area of the blood smear [24] or to describe the individual characteristics of each cell [25, 26]: contrast enhancement, spatial and frequency filtering, color transformations, and other manipulations of the histogram.

Different contrast enhanced procedures have been used: contrast stretching to improve the possible separation between nucleus and cytoplasm of WBCs [27]; local, global, dark and bright contrast stretching to improve the interpretability of the WBC images of patients with acute leukemia [28]; and partial contrast, bright stretching and dark stretching to enhance the morphologic features helping to the recognition between two types of acute leukemia [29].

The filtering of the images aims to reduce the noise level using various types of filters. Eom et al. [30] uses a Gaussian filter of 3 x 3 to prepare WBC images for the segmentation process. Scotti et al. [24] applies a Gaussian low-pass filtering with no zero-padding or the average of multiple WBC images, to reduce the noise. Ushizima et al. [31] utilizes Gaussian filtering to the gradient of the image before some transformation for automated cytoplasm detection. Angulo et al. [32] filters the green color component of the image to remove the noise and small digitalization mistakes, and segments mainly the erythrocytes of a peripheral

blood smear. Ghosh et al. [33] uses a Wiener filter to reduce the noise of the background of WBCs images to improve the segmentation methodology of the WBCs.

The images stained with MGG are bluish and purple. Therefore, a color transformation or treatment of WBCs images which may separate, highlight or extract the best profiles can be applied in both the preprocessing and segmentation. Sinha and Ramakrishnan [34] applies color transformation from the *RGB* (red, green and blue) to *HSV* (hue, saturation and value) color space prior to the segmentation procedure. Würflinger et al. [25] implements an adaptive color space transformation by a cluster analysis of the *RGB* histogram and it applies the Fisher transform reducing the difference between three different staining process of the same blood cell, to combining the respective images in only one image by a process named coregistration.

2.6.3 Segmentation

The segmentation stage separates the different objects of an image. This procedure is essential within the DIP framework because a robust segmentation is instrumental to be successful in the remaining stages of feature extraction and classification [35]. PB cell segmentation is a complex procedure due to the complicated morphology of the cell and the problems caused by variations of the staining methodology. Several publications exist about the cell segmentation process with different works either for differential count of WBCs, selection of the work area of the blood smear, morphologic description of RBCs or recognition of neoplastic lymphoid cells.

The thresholding technique is the simplest segmentation, wherein a threshold is chosen from the approximate probability distribution function or histogram, i.e. separate in two objects. For its simplicity and efficiency, this technique has been used with multiple thresholds [36], applying the method of Otsu to segment WBCs [37], using morphologic adaptive thresholding to find the optimal working area of the blood smear [32], and a combination of automatic contrast stretching, image arithmetic operations, minimum filtering and global threshold techniques to localize the nucleus of WBCs [38]. Furthermore, Piury and Scotti [27, 39] uses Canny edge detection and some mathematical binary morphology to segment the whole cell (membrane) and cropping of the cell, contrast stretching, low-pass morphologic filtering and automated thresholding to separate the nucleus of WBCs. Ritter et al. [40] combines an automatic thresholding with connected components and an adaption of Dijkstra's shortest path algorithm and graph representation to the standard adjacency list to segment cells and identify the border on images from PB smear. Ghosh et al. [33] presents a nucleus segmentation of WBCs using fuzzy divergence by Gamma, Gaussian and Cauchy type of membership functions of the image pixels and some modified thresholding methods.

A basic way to treat color segmentation is separating the components of the *RGB* color space of the image. It has been shown that a thresholding about 100 (between 0 to 255) applied over the green component leads to a good separation of cell nucleus [41]. Also, various works use several transformations to other color spaces inside the application of different segmentation methods: scale space filtering and watershed transformation over a 3D histogram of the *HSV* color space to separate the cytoplasm [42]; Bayes classification over *L*a*b** (*CIELAB 1976, L** is the luminance, *a** and *b** are the chromaticities) color space to determine different regions depending on the color tones [31]; thresholding of the S component to obtain the nucleus region of blasts [43]; and a methodology to localize and segment lymphoblasts by extracting and binarizing Hue and Saturation components of the *HSV* color space and applying mathematical morphology operations [44].

The snake technique works as a string which is deformed according to the edges of the object (in this case the cell or the nucleus), influenced by an external force field. This active contour model has been applied to segment WBCs using different variations of the technique: Ongun et al. [45] proposes the fast snake-balloon method using morphologic operators for the initial positioning of the snakes to segment firstly PB normal WBCs, and secondly different types of cell in the bone marrow [46]; Yang et al. [47] develops and implements a robust color contour active algorithm to segment lymphoid cells, operating in the *L*u*v** (*CIELUV 1976, L** is the luminance, *u** and *v** are the chromaticities) color space by introducing a color gradient and a L_2E robust estimation into the classic gradient vector flow (GVF) active contour model [48]; Eom et al. [30] performs a level set active contours method using GVF where the region information is estimated using a statistical analysis with the expectation-maximization algorithm (EM-algorithm) and Bayes probabilities to segment nucleus and cytoplasm of normal WBCs. Sadeghian et al. [26] also uses another methodology with GVF active contours on edge images of the leukocyte processed by Canny edge detection to extract the WBC nucleus (it also segments the cytoplasm through automatic thresholding).

Segmentation by clustering is an unsupervised method for grouping of pixels according to some information, usually the color. In that case, each pixel is a vector of three components (e.g. in the *RGB* color space: r, g, b) and the problem is solved by clustering of the similar vectors according to a criterion. Foran et al. [21] segments lymphoid cells using the *L*u*v** color space transformation of the cell image to implement the fast non-parametric clustering method developed in Comaniciu and Meer [49], which is a stable and efficient algorithm. Sinha and Ramakrishnan [34] segments nucleus and cytoplasm of WBCs by the Expectation-maximization algorithm with an initial estimation using k-means clustering of the *HSV* color space of the cell image. Ramoser et al. [50] applies k-means clustering technique of the *HSV* color space to segment the nucleus and probabilistic segmentation to separate the leukocyte.

Scotti [24] implements Fuzzy k-means clustering of the *RGB* space color to segment the WBCs. Mohapatra et al. [51–53] uses a variation of fuzzy clustering named Gustafson Kessel clustering and the nearest neighbor classification of the *L*a*b** color space to segments the nucleus of WBCs from the other blood components. González et al. [54] uses the k-means clustering technique with the standard Euclidean distance on the *L*a*b** color space but omitting the luminosity component to separate the image in four groups: background, cytoplasm, nucleus of WBCs and RBC from bone marrow cell images.

Mathematical morphology is one of the most used techniques of DIP, which analyses the image in a non-linear way extracting information or modifying the objects of an image by the formal description of their geometry. Dorini et al. [55] proposes a methodology to segment the nucleus and the cytoplasm of WBCs using some simple operations of mathematical morphology and Watershed Transformation (WT) [56, 57]. Angulo and Flandrin [32] presents a technique to automatically detect the working area of blood smears stained with MGG, utilizing mathematical morphology to extract the RBCs, their centers and the RBCs without center. These cells are counted and two parameters about their overlapping and the spreading are evaluated. Angulo et al. [58, 59] develops a methodology to extract the erythrocytes and WBCs by automatic thresholding and binary filtering of the green component, and subsequently the WT with markers is applied over a morphologic gradient of the green and saturation components (*RGB* and *HSV*) to segment the nucleus and cytoplasm of lymphocytes. Markiewicz et al. [60] applies the WT on a distance transform of a gray scale version of the bone marrow cell image to segment the whole leukocyte.

2.6.4 Feature extraction

In the feature extraction step, the characteristics of the object are obtained by quantitative measures. In the lymphoid cell recognition problem these features are calculated for the entire cell, cytoplasm, the nucleus and the region around the cell. They can represent morphologic qualitative features usually employed by the hematologist [3] or abstract quantitative parameters [35]. Since one of the main objectives of this thesis is to analyze PB lymphoid cells of patients with lymphoid neoplasms, and the description methods of normal WBCs and blood cells present in leukemia can be significant for the DIP process, the state of the art is divided into three principal items corresponding to the feature extraction of normal WBCs, cells from leukemias and neoplastic lymphoid cells, respectively.

2.6.4.1 Feature extraction of normal WBCs

To recognize WBCs, Ongun et al. [46, 61] calculates 57 features mainly grouped in two categories: shape based features (moments and affine invariants, length of the cell boundary, curvature and boundary energy), and color/texture based features (mean and standard deviation for cell, cytoplasm and nucleus in the *L*a*b** color space and *HSV* color histograms features). Sinha and Ramakrishnan [34] presents a methodology for automatically differentiate WBCs, extracting several parameters: shape features (eccentricity of the nucleus and cytoplasm, nuclear compactness, area-ratio and number of nucleus lobes), color features (means of each component of the *RGB* color space), texture features based on computations of the gray level co-ocurrence matrix (GLCM) and the autocorrelation matrix (energy, entropy and correlation for GLCM, and coarseness and busyness for the autocorrelation matrix). Sanei and Lee [62] uses principal component analysis (PCA) in similar way to face recognition but extended to the *YIQ* (*Y* is the luma, *I* and *Q* are the chromaticities) color space to obtain the eigencells and after a linear transformation use them as features to describe WBCs. In order to describe WBCs, Piuri and Scotti [27, 39] extracts various geometric features for the nucleus and cytoplasm such as: area, perimeter convex area, solidity, major axis length, orientation, filled area, eccentricity, rectangularity, circularity, and number of nucleus lobes. The best features are selected by a technique named forward selection. Ramoser et al. [50] proposes the extraction of 18 color statistical features for the nucleus and cytoplasm (mean, standard deviation and skewness of each component of *HSV* color space), five nucleus shape features (convexity, principal axes ratio, compactness, circular variance and elliptical variance) and three geometric features (sizes of nucleus and cytoplasm, and number of detected nucleus regions), to obtain information of the WBCs. Pan et al. [63] employs image-based features rather than concrete features, thus it uses *RGB* color histogram of the whole cell, the intensity histograms of the nucleus and the cytoplasm to make an only feature vector; then this one is reduced by Kernel PCA to represent a quantitative characterization of blood and bone marrow cells. Rodrigues et al. [64] utilizes shape descriptors by spatial moments with invariance for translation and rotation (row and column moments of inertia, aspect, spread and Hu descriptors), texture features (mean, standard deviation, skewness, kurtosis, first and second neighbor contrasts), and some geometric features (mean, area, perimeter of the cytoplasm and nucleus, and circularity) to recognize WBCs.

2.6.4.2 Feature extraction of blast cells from acute leukemias

Markiewicz et al. [60, 65] and Siroic et al. [66] present the extraction of four groups of features of WBCs from bone marrow smear of patients suffering from acute leukemia: texture

(applied to the three *RGB* color components on the nucleus and cytoplasm), geometric (of the cell), statistical (color distribution of the cell image) and morphologic (mathematical morphology operations); then the most relevant features are selected by a genetic algorithm feature selection or a Support Vector Machines (SVM) feature selection. Mohapatra et al. [51–53] calculates various nuclear features to detect acute lymphoblastic leukemia from PB smear images: perimeter roughness (by fractal geometry), contour signatures, shape features (area, perimeter, compactness, solidity, eccentricity, elongation, form factor), color features (means of each component of *RGB* and *HSV* color spaces) and four second statistical features based on GLCM of a gray version of the image.

After a segmentation process, González et al. [54] extracts several features corresponding to the geometry (perimeter, area, major and minor axis, orientation, Euler number, among others), texture (gray threshold of the segmentation, sum of the histogram, maximum and minimum of the histogram, mean, standard deviation and variance), and another type of feature obtained from PCA of the bone marrow cells with the purpose of identify possible leukemias. In order to differentiate between normal lymphocytes and abnormal lymphoblast cells, Madhloom et al. [67] obtains shape features (area, eccentricity, perimeter, circularity, elliptical features of the nucleus, cell area and ratio of the nucleus to the whole cell) and texture features (first and second statistical features); subsequently Fisher's discrimination is utilized to select the best and uncorrelated features. Aimi et al. [68] gets different quantitative parameters to describe WBCs from patients with acute leukemia: size based features (cell area, nucleus area, cytoplasm area, nucleus-cytoplasm ratio, cell perimeter and nucleus perimeter), shape based features (roundness, compactness, central moment and affine invariant moment of the nucleus), and color based features (mean and standard deviation of intensity and *RGB* color space for nucleus and cytoplasm).

2.6.4.3 Feature extraction of neoplastic lymphoid cells

Comaniciu et al. [49] and Foran et al. [21] extract the following features of the nucleus: area, shape (elliptical Fourier descriptors) and texture features based on a multiresolution simultaneous autoregressive model within the develop of a image-guide decision support for pathology to characterize neoplastic and normal lymphoid cells. Benattar et al. [69] proposes a scoring system for lymphocytes in B-cell neoplasm using various morphometric parameters: nuclear shape, cell shape, cell area, nucleus-cytoplasm ratio, nuclear red/blue ratio, cytoplasmic green/blue ratio and the proportion of cell with nucleolus. Angulo et al. [58, 59] extracts several quantitative parameters from the lymphocytes to define some qualitative morphologic features: nuclear and cell sizes, nucleus - cytoplasm ratio, nuclear excentration, chromatin density (texture) by granulometric curves, regular and irregular nu-

clear shapes through some simple parameter (form factor, circularity, eccentricity) and the specific analysis of the nuclear lobes and others irregularities, number of big or medium nucleolus, cytoplasmic basophilia (mean of the each color component of $L^*a^*b^*$ color space), and cytoplasmic granulations, cytoplasmic shape by binary granulometry. Ushizima et al. [70] studies the leukocyte recognition problem by calculating shape and size features (perimeter, area, circularity, bending energy, nucleus - cytoplasm ratio, etc.) and texture features based on the GLCM applied for different block sizes over a grayscale version of the cell image. Afterwards, it uses feature selection by an exhaustive search, and heuristic search with forward sequential selection and backward sequential selection. In a subsequent work, Ushizima et al. [71, 72] extends its proposed method by applying the texture features to the components of the RBC color space and adding some statistical features, later it utilizes again the feature selection method to choose the most important and independent features. Jahanmehr et al. [73] analyzes quantitative and qualitative cytological parameters of lymphocytes from B-cell neoplasms (CLL, MCL and B-PL). Particularly, cell area, cell diameter, cytoplasm area, nuclear area, nuclear/cell ratio and nuclear density are evaluated and it demonstrates that these features can be useful to differentiate the lymphoid neoplasms. For the purpose of describing neoplastic lymphoid cells and blast cells, Tuzel et al. [74, 75] performs a cell representation characterizing its texture structure using textons inside both nucleus and cytoplasm by the construction of two texton histograms.

2.6.5 Classification

The classification step consists of the application of different algorithms (supervised and/or unsupervised) to recognize the different cells from their extracted features. The state of the art is divided into three parts, which correspond to the classification of normal WBCs, neoplastic lymphoid cells, and acute leukemia cells, respectively.

2.6.5.1 Classification of normal leukocytes

With the purpose of automatically differentiate 258 WBCs from bone marrow and PB, Ongun et al. [46, 61] extracts quantitative parameters to be used as inputs for different classification algorithms: k Nearest Neighbor (kNN), linear vector quantization (LVQ), multi layer perceptron (MLP) neural networks and support vector machines (SVM). It obtains the best result using SVM with a training accuracy of 100% and a testing accuracy of 91.03%. Sinha and Ramakrishnan [34] implements an automated method for differential of WBCs by several classification methods of 50 samples for training and 34 for testing, obtaining the best results with 97% accuracy for Artificial Neural Networks (ANN) and 94% for SVM. Sanei and Lee

[62] proposes an automatic method to classify 15 classes of mature cells and blasts, excluding mature platelets, and other variations by Bayes classification, resulting in accuracy of 96.5% for mature cells on PB, and 85% for inmmature cells or blasts on PB, but with an missed classification of 21% for myelocytes, promyelocytes, monocytes and metamyelocytes from bone marrow. Piuri and Scotti [27, 39] utilizes kNN, feed-forward ANN, radial basis function neural network, and a parallel classifier built using feed-forward ANN to identify the following classes of WBCs: basophil, eosinophil, lymphocyte, monocyte and neutrophil. The best result was achieved for the parallel ANN with a mean error of 0.08. Ramoser et al. [50] evaluates a set of 1166 cell images (13 classes) by a polynomial SVM classifier using the extracted feature database. It results in accuracies around 90%, but for neoplastic lymphocytes (grouped in only one class) the accuracy is 73%. Mircic and Jorgovanovic [76] proposes a methodology for automatic classification of WBCs by the analysis of blood smear images using artificial neural networks (ANN), specifically a feed-forward ANN with 19 inputs neurons, 5 output neurons and two hidden layers with 120 and 70 neurons; this procedure results in a high sensitivity and a classification accuracy of 86%. Pan et al. [63] develops a classification system using SVM classifiers with polynomial kernel for 10 classes of WBCs from bone marrow smears involving monocytic and granulocytic series (15 samples), all of them labeled by the pathologist. It obtains an accuracy of 90.5% (it divides randomly the data into five sub-test-sets). Rodrigues et al. [64] proposes two approaches for the classification of five types of normal WBCs, the first one uses feed forward neural networks trained by a backpropagation algorithm (BPNN) and the second method is a hybrid model between SVM and pulse-coupled neural networks (PCNN). The BPNN approach achieves an accuracy of 81.31% and the PCNN-SVM method gets 86.9%. Colunga et al. [77] implements a method to classify three types of WBCs: band neutrophils, eosinophils and lymphocytes from PB smear. It uses PCA to project the cell image to a lower dimensional subspace, employing the Gaussian mixture EM-algorithm with the maximum posterior decision rule.

2.6.5.2 Classification of cells from acute leukemias

After feature selection, Markiewicz et al. [60, 65] achieves an automatic recognition of 10 types of WBCs from bone marrow smear samples of patients suffering of acute myeloid leukemia (AML), acute lymphocytic leukemia (ALL) and CLL by the implementation of SVM with Gaussian kernel. The respective results confirm good efficiency, achieving the agreement of almost 87% with the human expert score and around 13% of missclassification. Reta et al. [78] proposes an automatic method for morphologic classification of abnormal WBCs images from bone marrow smears into different subtypes of acute leukemia. This classification is done using instance based classifiers, decision trees, regression functions and

metaclassifiers, resulting in an overall accuracy of 92% for acute leukemia types, 84% for lymphoblastic subtypes and 92% for myeloblastic subtypes. In various works, which differ mainly in the segmentation, Mohapatra et al. [51–53] uses the extracted features to classify WBCs from (108) blood smears of patients with ALL by SVM, obtaining a classification accuracy of 95%. González et al. [54] presents a data mining strategy to identify leukemias from bone marrow smears by the extraction of features that feed several machine learning algorithms. This method achieves accuracies above 95.5% to differentiate between AML and AML. After that, it obtains an accuracy of 90% among five leukemia subtypes. Madhloom et al. [67] describes a method for the automated classification between normal lymphocytes and abnormal lymphoblast cells (ALL) from PB smear, which utilizes a kNN with Euclidean distance classifier, resulting in 92.5% of accuracy. Aimi et al. [68] presents a methodology for automatic recognition of WBCs inside ALL and AML blood samples by using the MLP and Simplified Fuzzy ARTMAP (SFAM) neural networks, obtaining the best classification with the MLP trained by Bayesian Regulation algorithm with a testing accuracy of 95.7%.

2.6.5.3 Neoplastic lymphoid cell classification

Comaniciu et al. [21, 49] implements a content-based image retrieval (CBIR) system to support the decision making in clinical pathology using a database of 261 specimens, which belong to three types of lymphoid neoplasm (CLL, FL and MCL) and the normal cell type. The performance of the system is evaluated by 10-fold cross validation classification, obtaining satisfactory results compared with human experts. Ushizima et al. [72] evaluates the use of SVM classification of five types of normal WBCs and only one type of neoplastic lymphoid cell (CLL). It obtains an average accuracy around 94%, however the CLL accuracy is 88%.

Angulo et al. [58] firstly presents a classification of five types of normal WBCs from PB smears using the full cell image, then it combines statistical techniques, granulometries and color histograms to identify the cell type, obtaining an accuracy of 95%. Secondly, it develops a methodology withing the CBIR framework (Ontology) to classify neoplastic lymphoid cells, in which the extracted features from the cell are classified into categories using decision trees [59].

Tuzel et al. [74] addresses a classification among four types of malignancies (CLL, MCL, FL and acute leukemia) and normal cells by SVM with linear kernel using only texture features, and a combination between shape, area and texture features. The best results are obtained with the full set of features within a leave-one-out cross-validation classification test (to evaluate independently each case without cell mixing), obtaining classification rates of 84.62% for cell test and 91.42% for case; while for a 10-fold cross validation test (where the cell were mixed between cases) the best classification rate is 93.18%. In a related work,

Yang et al. [75] presents another 10-fold cross-validation classification test with a new and independent smaller database but with the SVM trained with the old bigger database, resulting in 87.22% of accuracy due to that the new interclass similarities and intraclass variations were never seen during the training.

Chapter 3

A First Digital Image Processing Approach for Neoplastic Lymphoid Cell Classification

Based upon: S. Alférez, A. Merino, L.E. Mujica, M. Ruiz, L. Bigorra, and J. Rodellar, Automatic classification of atypical lymphoid B cells using digital blood image processing, International Journal of Laboratory Hematology, vol. 36, no. 4, pp. 472-80, Aug. 2014. doi: 10.1111/ijlh.12175

Abstract

There are automated systems for digital peripheral blood (PB) cell analysis, but they operate most effectively in non-pathological blood samples. The objective of this chapter is to design a first approach to explore the automatic classification of abnormal lymphoid cells. 340 digital images of individual lymphoid cells from PB films obtained in the CellaVision DM96 were analyzed: 150 Chronic Lymphocytic Leukemia (CLL) cells, 100 Hairy Cell Leukemia (HCL) cells and 90 normal lymphocytes (N). Afterwards, the Watershed Transformation was implemented to segment the nucleus, the cytoplasm and the peripheral cell region. From these regions, 44 features were extracted and then the clustering Fuzzy C-Means (FCM) was applied in two steps for the lymphocyte classification. Then, the images were automatically clustered in three groups, one of them with 98% of the HCL cells. The set of the remaining cells was clustered again using FCM and texture features. The two new groups contained 83.3% of the N cells and 71.3% of the CLL cells, respectively. The developed approach has been able to automatically classify with high precision three types of lymphoid cells. The addition of more descriptors and other classification techniques will allow extending the classification to other classes of neoplastic lymphoid cells.

3.1 Introduction

Peripheral blood (PB) is an organic fluid easily accessible and its study is the initial analytical step in the diagnosis of most of the hematological and non hematological diseases [79]. Frequently, the blood smear provides the primary or the only evidence of a specific diagnosis, remaining an important diagnostic tool even in the age of molecular analysis [9]. Morphologic evaluation of leukemia and lymphoma cells is essential for their diagnosis and classification. In the World Health Organization (WHO) classification, neoplastic cell morphology, along with immunophenotype and genetic changes, remains essential in defining lymphoid neoplasms [6].

Despite the significant improvements during the last years in hematology analyzers, no significant progress has been made in terms of automatic classification of neoplastic PB cells.

These devices are limited to identifying normally circulating leukocytes and flagging abnormal cells, without being able to classify the abnormal leukocytes.

The close collaboration between cytologists, mathematicians and engineers over the last few years has made possible the development of automatic methodologies for digital image processing of normal blood cells. Some equipment are able to pre-classify cells in different categories by applying neural networks, extracting a large number of measurements and parameters that describe the most significant cell morphologic characteristics [19]. These systems, when integrated in the daily routine, represent an interesting technological advance since they are able to pre-classify most of the normal blood cells in PB [17].

Neoplastic lymphoid cells are the most difficult pathological cells to classify using morphology features only [80], so that few studies of automatic classification of these cells with satisfactory results have been published. In most of the previous studies, the lymphoid cell classification has been addressed with pattern recognition systems to separate the cells into categories [20, 59, 70, 74]. Nevertheless, the image processing techniques used in some of the papers are not useful for the current digital images, since the present acquisition technology is based on charge-coupled device sensors [20].

Morphologic distinction between various types of lymphoid cells requires experience and skill and, moreover, objective values do not exist to define cytological variables. Chronic Lymphocytic Leukemia (CLL) cells are typically small lymphocytes with clumped chromatin and scant cytoplasm. Hairy cell leukemia (HCL) cells are larger than normal lymphocytes (N) and they have abundant weakly basophilic cytoplasm with irregular *hairy* margins. This chapter explores a starting methodology for lymphocyte recognition to allow the automatic classification of abnormal lymphoid cells circulating in PB in some B lymphoid neoplasms, such as CLL and HCL cells.

3.2 Material and methods

3.2.1 Blood sample preparation and digital image acquisition

Samples from patients with CLL and HCL were included in the study. The diagnoses were established by clinical and morphologic findings as well as characteristic immunophenotype of the lymphoid cells. Specifically, CLL cells had the phenotype CD5+, CD19+, CD23+, CD25+, weak CD20+, CD10-, FMC7- and dim surface immunoglobulin (sIg) expression. All the patients with HCL had lymphoid cells with the phenotype CD11c+, CD25+, FMC7+, CD103+ and CD123+.

Blood samples were obtained from the routine workload of the Core Laboratory of the Hospital Clínic of Barcelona. Venous blood was collected into tubes containing K3EDTA as anticoagulant. The samples were analyzed by a cell counter Advia 2120 (Siemens Healthcare Diagnosis, Deerfield, USA) and PB films were automatically stained with May-Grünwald-Giemsa in the SP1000i (Sysmex, Japan, Kobe) within 4 hours of blood collection.

The quality of the smears and cell morphology was assessed by hematologists prior to the image study. 340 lymphoid cell images from PB films were selected, where 90 images were lymphocytes from healthy patients, 100 were lymphoid cells from patients with HCL and 150 were lymphoid cells from patients with CLL. Each individual cell image had a resolution of 360 x 363 pixels and they were obtained by the CellaVision DM96 system (Lund, Sweden).

3.2.2 Novel method for lymphocyte classification

In this chapter a novel method for lymphocyte recognition was developed based on 3 steps: 1) color segmentation; 2) feature extraction; and 3) classification. They are shortly described in the remaining of this section.

3.2.2.1 Color segmentation

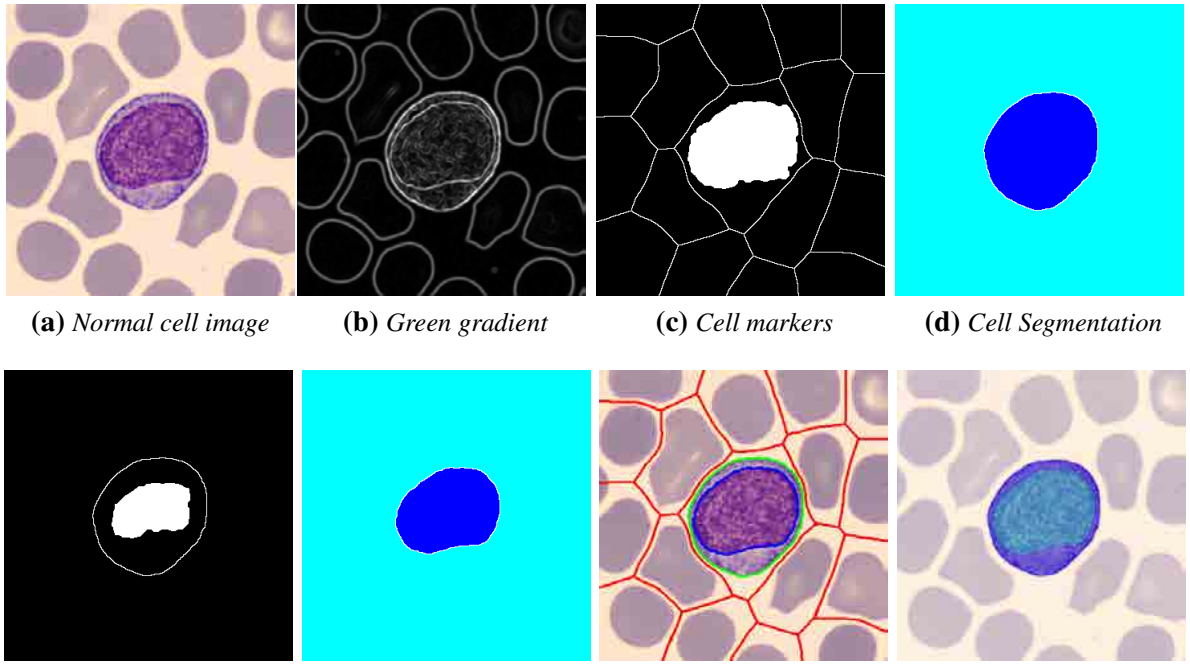
A digital blood image is composed of a finite number of pixels. Each one has a particular location and color value, which can be represented in several spectral components or color spaces: RGB, HSV, Lab, among others [35]. The goal of the segmentation procedure is to separate lymphoid cells captured by microscope from other objects in the image [26, 48, 81].

In this chapter, lymphoid cell segmentation was obtained using the Watershed transformation (WT), which was applied only on the gradient of the green component from RGB color space [57]. As a final result, 3 different regions of the cell were identified (segmented): the cytoplasm, the nucleus and the peripheral zone around the cell.

3.2.2.2 Feature extraction

The objective of this stage is to obtain information about the objects in the image under analysis. A number of 44 features were used in this chapter, which are related respectively to: geometry (10), texture (30), basophilia intensity (3), and cytoplasm external profile (1). They are summarized as follows:

Geometric features These features are quantitative geometric interpretations of the cell and nucleus shapes. For each cell and nucleus were calculated: *area, diameters, perimeters* and *conic eccentricities*. Then, *nucleus/cytoplasm ratio* was calculated by dividing the respective



(a) *Normal cell image* **(b)** *Green gradient* **(c)** *Cell markers* **(d)** *Cell Segmentation*

(e) *Nucleus markers* **(f)** *Nucleus segmentation* **(g)** *Watershed lines* **(h)** *Nucleus and cytoplasm*

FIGURE 3.1: *Different stages of the Watershed segmentation (WT): The original cell (a) was processed to obtain the external and internal markers (c). The WT was calculated on the gradient of green component (b). The markers limit the WT to segment the cell (d). Once the lymphoid cell was separated, its edges are used as the new external marker and the thinned mask of the nucleus as the new internal marker (e) in the WT to segment the nucleus (f). Finally the watershed lines (g) showed the regions of interest: the nucleus, the cytoplasm (h) and the peripheral zone around the cell.*

areas. The *nucleus eccentricity relative to the cell center* was calculated as the distance between the center of the cell and the nucleus [59].

Texture features Several statistical measures were used to describe the texture of the cytoplasm and nucleus regions [82]. The skewness measures the asymmetry of the shape; the kurtosis, the relative flatness; the energy, the uniformity; and the entropy, the variability. In addition, the mean and the standard deviation were calculated. Other second-order statistical features were considered: contrast, homogeneity, correlation, energy, entropy, variance and difference variance [83].

Granulometric features of the nucleus Four features were calculated from the granulometric curve of the lymphoid cell (mean, standard deviation, skewness and kurtosis), to discriminate the different types of nuclear texture and improve chromatin description [84].

Basophilia features of the cytoplasm Cytoplasmic basophilia can be estimated by color analysis. The Lab color space is characterized by its approximation to human perception. Therefore, the means of the intensities for each color component are appropriate to represent the basophilia degree of the cytoplasm [59].

Cytoplasmic profile feature In this chapter a novel method to characterize the cytoplasmic profile was proposed. It estimates the projections of the cytoplasm using the peripheral region around the cell segmented by WT. This feature is obtained by using thresholding segmentation to the green component and counting the pixels of this region.

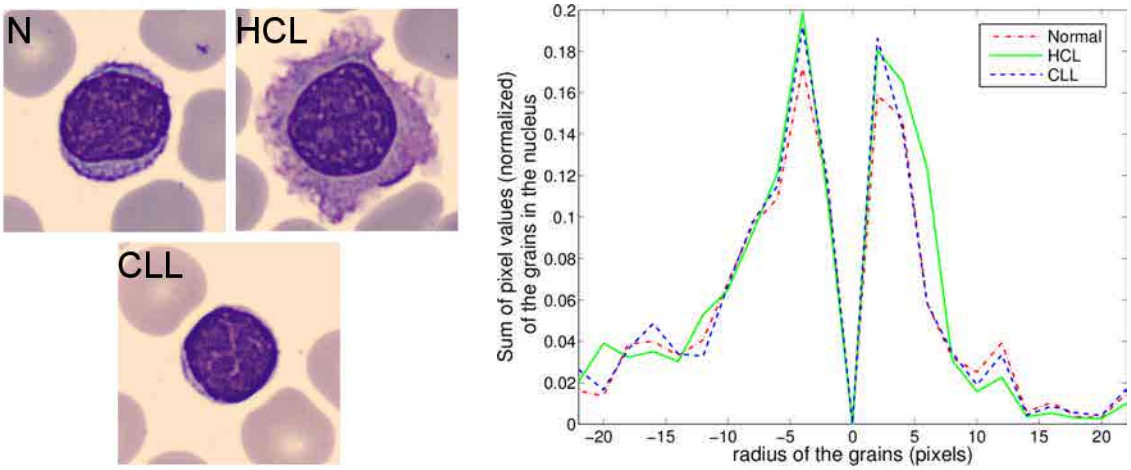
3.2.2.3 Classification

In this chapter, the main objective of the classification step was to obtain an automatic clustering using the features extracted from each image to analyze how they can provide relevant information for the detection of normal, HCL or CLL lymphoid cells.

All features were stored in a data matrix, which was used as the input data for the classification. The unsupervised classification methodology Fuzzy c-mean (FCM) was applied. Similar input data were grouped in each cluster with certain membership degree [85]. Finally, the maximum membership value was considered to select the cluster for each lymphoid cell.

3.3 Results

In the first step, WT was effective in separating the nucleus of the cell. Besides, it allowed segmenting more regions, specifically the outer profile of the cytoplasm, which is crucial to extract the useful information to discriminate different types of lymphocytes. Moreover, its computational cost was low. Figure 3.1 shows the images corresponding to the different stages that were obtained by applying the WT segmentation to the lymphoid cells. The lymphoid original cell stained with MGG is shown in Figure 3.1a. The WT was applied only on the gradient of the green component from RGB color space (Figure 3.1b). Since the gradient highlights the edges (high intensity changes) of the objects, some external and internal markers were included as minimum values over the gradient image to improve the delimitation of the different regions as shown in Figure 3.1c. Thereby, the over-segmentation was avoided and only the entire lymphoid cell was separated (the darkest region on Figure 3.1d). Once the entire lymphoid cell was separated, new markers were imposed (Figure 3.1e) and the WT was applied again to segment the nucleus (Figure 3.1f). Afterwards, mathematical morphology operations were performed to improve the quality of the regions from the nucleus



(a) *Cell images* **(b)** *Cell granulometric curves*

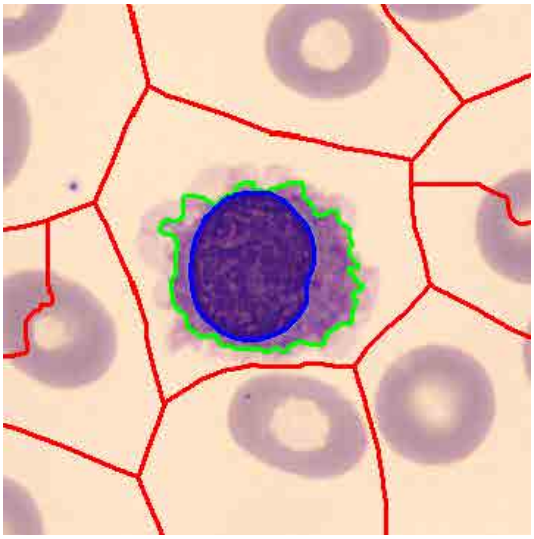
FIGURE 3.2: *Normal (N), HCL and CLL lymphoid cells (a) and their corresponding granulometric curves (b), which place the information from the dark spots on the left (negative coordinates) and the information from the bright spots on the right (positive coordinates).*

and cytoplasm. Finally, 3 different regions of the cell were identified: the cytoplasm, the nucleus and the peripheral zone around the cell (Figures 3.1g and 3.1h).

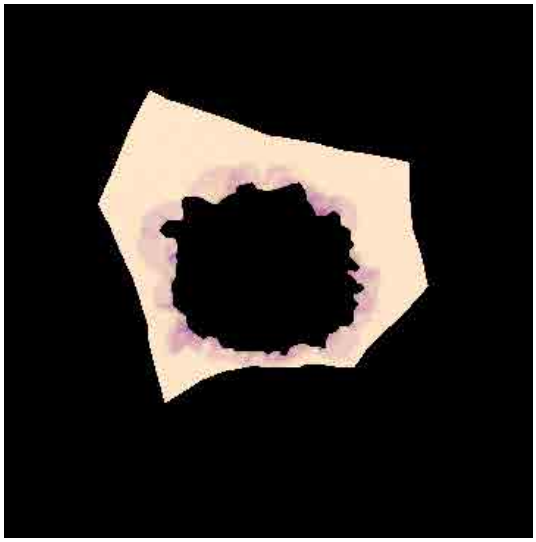
Corresponding to the second step (feature extraction), Figure 3.2 shows an example of N, HCL and CLL lymphoid cell images (Figure 3.2a) and their granulometric curves. Figure 3.2b shows how these curves discriminate the types of nuclear texture in the different lymphoid cells, improving chromatin description. In order to obtain information from each curve, four features were calculated: mean, standard deviation, skewness and kurtosis.

Figure 3.3 displays an example of cytoplasmic profile feature extraction obtained in one of the hairy cell images. After the segmentation of the cell (Figure 3.3a), the peripheral zone around the cell was selected (Figure 3.3b). The histogram representation of this region showed an intermediate lobe that contained most of these "hairy" projections (Figure 3.3c). Then, the presence or absence of these projections was determined (Figure 3.3d). Finally, this area was quantified. The novel cytoplasmic profile feature proposed in this chapter was decisive for the detection of the hairy cells. Figure 3.4 shows the characteristic cytoplasmic profile feature for all the cells. HCL cells showed very high values of this feature compared with CLL and N lymphoid cells.

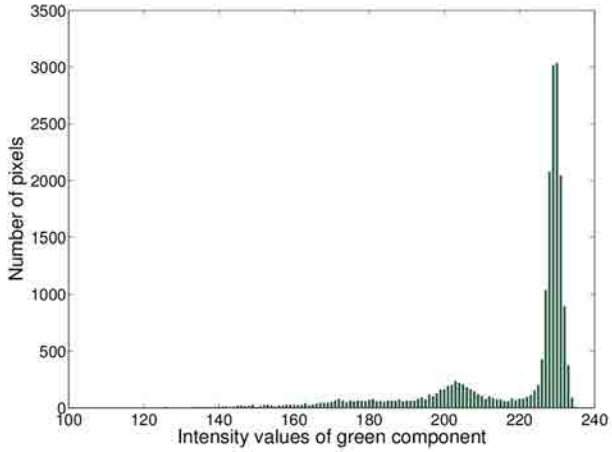
Afterwards, for the classification step, the 44 features of the 340 available cells were used to create the data matrix. It was automatically clustered into 3 groups using FCM, producing 3 membership functions. Since every cell pertains to one of the 3 groups with different degrees of membership, the criterion that each lymphoid cell belongs to the group with the highest membership value was used. The left part of Table 3.1 gives a summary of the whole data



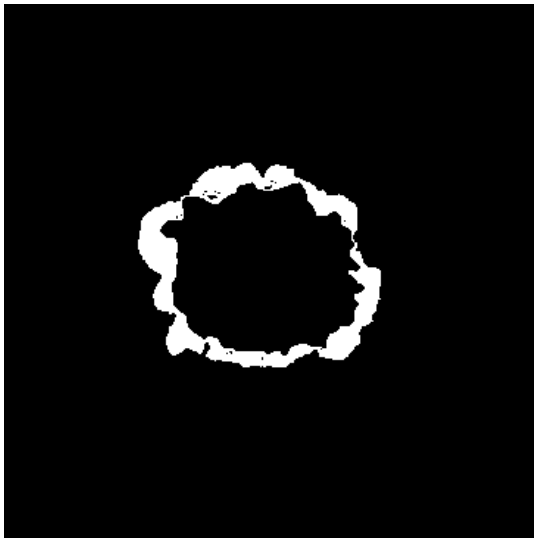
(a) *Cell watershed lines*



(b) *Cell peripheral zone*



(c) *Cell peripheral zone histogram*



(d) *Quantified* hairy *projections*

FIGURE 3.3: *Stages to calculate the HCL cell cytoplasmic feature. After the cell segmentation (a), the peripheral zone around the cell was selected (b). The histogram representation of this region showed an intermediate lobe that contained most of the* hairy *projections (c). Then, the presence of these projections was determined (d). Finally, this area was quantified.*

obtained in the first FCM classification step. This shows an excellent classification on the group 3 because it included 98% of the HCL cells. However, the groups 1 and 2 contained 75.6% of normal lymphoid cells and 62.7% of CLL cells, respectively. Figure 3.5 contains 3 plots corresponding to each group. The horizontal axis represents each individual cell, while the vertical axis gives its membership value. These 3 values represent the probability of

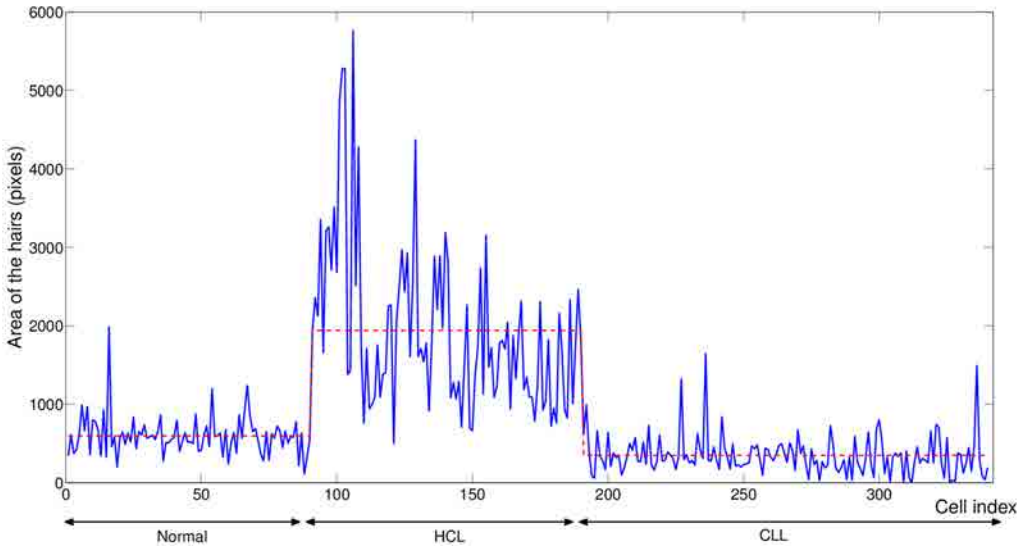


FIGURE 3.4: *Cytoplasmic profile feature in N, HCL and CLL lymphoid cells. HCL cells showed very high values of this feature compared with CLL and N lymphoid cells.*

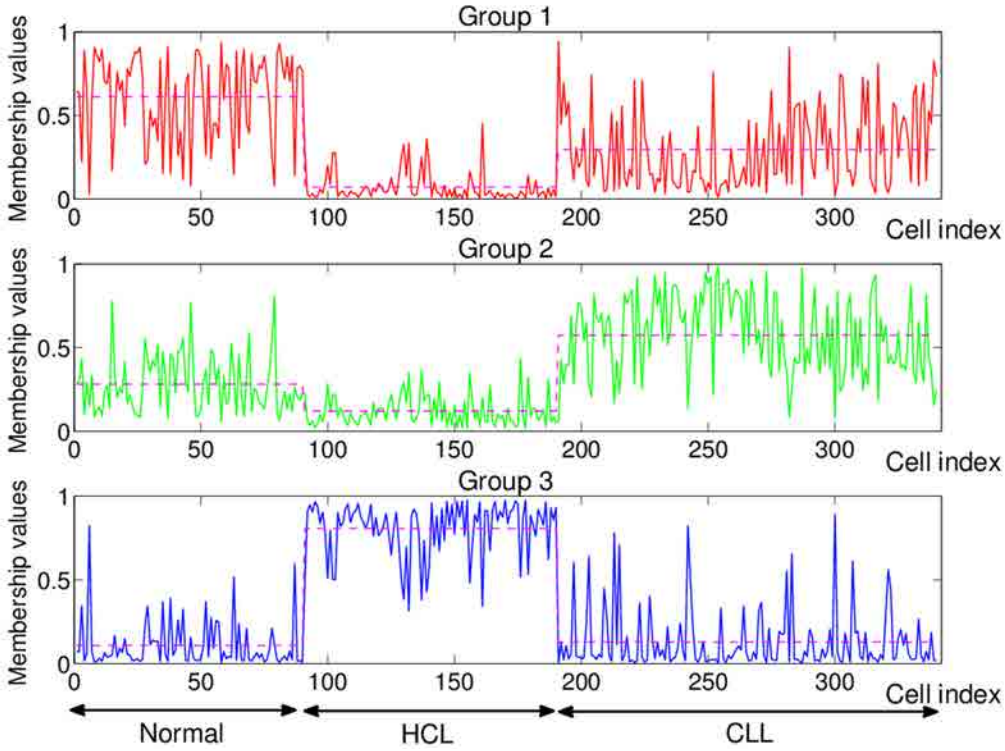


FIGURE 3.5: *Membership function of each type of cell: normal lymphocytes (Normal), hairy cells (HCL) and lymphoid cells from chronic lymphocytic leukemia (CLL). The horizontal axis represents the cells, while the vertical axis represents the probability of belonging to each group. The horizontal line for each type of cell represents the mean of their membership values in each group.*

TABLE 3.1: *Two steps classification process. First, 3 different classes of data were obtained. Each group has cells of the 3 types, i.e. the group 3 has 98% of the HCL cells. A second Fuzzy C-means (FCM) was applied using the texture features only. It resulted in two new groups with 83.3% of N lymphoid cells and 71.83% of CLL cells, respectively.*

Type	FCM step 1			FCM step 2	
	Group 1	Group 2	Group 3	New group 1	New group 2
N	75.6%	20.0%	4.4%	**83.3%**	12.2%
HCL	1.0%	1.0%	**98.0%**	2.0%	0.0%
CLL	30.0%	62.7%	7.3%	21.3%	**71.3%**

N, normal cells; HCL, Hairy Cell Leukemia cells; CLL, Chronic Lymphocytic Leukemia cells

belonging to each group, their sum being equal to 1. In every plot, the data set was sorted in this way: the first 90 images belong to N lymphocytes, the following 100 to HCL cells and the last 150 to CLL cells. From Figure 3.5, it was clear to assure that both the normal cells and the CLL cells do not belong to group 3 due to their low membership values. It was also clear that HCL cells had high probability of belonging to group 3. On the contrary, from Figure 3.5 it was difficult to infer to which group (1 or 2) belong the normal and the CLL cells due to the high variance of their membership values in groups 1 and 2.

In order to improve the classification, once the HCL group was identified, the set of the remaining cells was clustered again using FCM. In this new clustering process, only two types were considered (N and CLL cells) and only the texture features were used. The right part of Table 3.1 gives a report of the results in this step: the percentage of normal lymphoid cells increased to 83.3% in the new group 1 and the percentage of CLL cells increased to 71.3% in the new group 2. In this case, this was clear enough to distinguish these two types of cells, because their membership values were quite different for a significant percentage of cells as observed in the right part of Table 3.1.

3.4 Discussion

In this chapter, a group composed by normal and two types of neoplastic lymphoid cells (HCL and CLL) has been analyzed. HCL and CLL cells were selected in this work for their representative morphology and the large number of these cells found in the routine workload in the laboratory.

Cell morphology is subject to variability in slide making and staining procedures. In order to minimize this variability, the images used in this chapter were obtained in a standard and reproducible way using automatic staining and the Cellavision DM96 analyzer. The system scanned the slides identifying different types of white blood cells (WBC). It takes digital

cell images and uses artificial neural networks to analyze them [17, 19]. The analyzer pre-classifies WBC but is not able to separate the different abnormal lymphoid cells circulating in PB in some B cell lymphoid neoplasms [18].

Since neoplastic lymphoid cells are the most difficult ones to be classified using only morphology features [80], in this chapter a starting methodology is proposed combining segmentation, feature extraction and classification algorithms. It was shown that this automated image-based methodology extracted granulometric, basophilia and cytoplasmic profile features in an objective and reproducible way. Then, this work explores a methodology that could help to provide a new generation of automated systems to assist in the diagnosis through hematological cytology.

The results showed that texture descriptors were the most relevant in CLL lymphoid cell discrimination. Moreover, nuclear characteristics are important features in morphologic diagnosis. The nuclear staining pattern reflects chromatin organizations and, in addition, the CLL cells typically contain clumped chromatin [73]. Therefore it supplies a good descriptor.

In a previous work [59], granulometry was used to describe cytoplasmic profile feature. Although that work showed good segmentation and description results, it was not completed with further studies towards the discrimination among different groups of similar diagnosis.

In this chapter, a novel cytoplasmic profile feature is proposed based on a simple thresholding of the peripheral zone around the cell. As it was expected, this feature was crucial for the HCL cells detection, since in PB stained with MGG they show a soft, blue-gray cytoplasm with hair-like cytoplasmic projection [86]. On the other hand, this feature could be used for the detection of another neoplastic lymphoid cells with cytoplasmic villous, such as the splenic marginal zone lymphoma.

Concerning to the classification process, 26 features (geometric and second-order statistical features) were used in [70] to automate the classical microscopic diagnosis, obtaining good results in the classification of CLL cells but only with respect to the different abnormal types of leukocytes from PB. In this chapter, 44 features are used adding other geometric and second-order statistical features as well as basophilia, granulometric, first-order statistical and cytoplasmic profile features. In addition, 3 types of lymphoid cells were distinguished: normal, CLL and HCL cells. It is relevant to remark that hairy cells have never been automatically classified before in the literature.

3.5 Conclusion

One of the main contributions of this chapter is the segmentation of the peripheral region of the cell, in addition to the nucleus and cytoplasm, which allows proposing a new cytoplasmic

profile feature describing information about the villous. This enables to achieve an accurate classification of HCL cells.

In summary, the approach presented in this chapter has been able to discriminate between 3 groups of lymphoid cells with encouraging results. It goes in the direction of combining medical, engineering and mathematical backgrounds to provide more objective and reproducible estimation of the neoplastic lymphoid cell morphology than the standard microscopy analysis. The development of a more robust segmentation, the addition of more features and other classification techniques will allow extending the methodology to other classes of neoplastic lymphoid cells. In this respect, the subsequent chapter develops the full segmentation methodology that will be used in the remaining of this thesis.

Chapter 4

Color Clustering Segmentation of Lymphoid Cell Images Using Spatial Kernel Fuzzy C-means

Based upon: S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, Color clustering segmentation of lymphoid cell images using spatial kernel fuzzy c-means, manuscript in preparation.

Abstract

The study of the characteristics of the peripheral blood cells under the microscope provides very useful information and it is the first analytic step in diagnosing most of the hematological diseases. In this chapter, a robust segmentation methodology is developed using mainly color clustering by the spatial kernel fuzzy c-means technique (sKFCM), to obtain three regions from lymphoid cell images: nucleus, cytoplasm and the peripheral zone around the cell. The methodology consists principally in two parts: (1) the cropping of the lymphoid cell from the remaining components of the images by the sKFCM clustering of the Y and K color components of the CMYK space, and (2) the application of the sKFCM clustering of the XYZ color space on the cropped image to separate the nucleus. Then, the Watershed transformation and other operations of mathematical morphology are applied to complete the segmentation. The general methodology was tested using a database of 3394 cell images containing normal lymphocytes and five types of neoplastic lymphoid cells corresponding to different pathologies, showing an overall efficiency of 92.81%. This segmentation technique will be part of the neoplastic lymphoid cells automatic classification system developed in this thesis.

4.1 Introduction

The hematological diagnosis is an integrated process involving clinical information and different complementary tests such as morphologic analysis, genetic studies, inmunophenotype tests, molecular biology, among others. But cytology is still the first analytical step in the diagnosis of most of the diseases [79]. However, the morphologic differentiation between different types of neoplastic lymphoid cells in the peripheral blood (PB) is a challenging task that requires extensive experience and skill. Furthermore, there are no objective values to define cytological variables, so that the slight differences in morphologic characteristics present in several pathologies can lead to doubts on the classification of malignancies in the hospital daily routine as well as false negatives [80].

In recent years, various automated methods have been developed for digital image processing (DIP) of white blood cells (WBCs), specially for five types of them: basophils, eosinophils, monocytes, neutrophils and lymphocytes. This normal leukocyte recognition problem has been well studied and the differences between the types are significant, allowing effective solutions. On the other hand, the automated classification of neoplastic subtypes of lymphoid cells, corresponding to the neoplasms defined by the World Health Organization (WHO) [6], is a more complicated problem, which has been scarcely studied in the literature.

The central focus of this thesis is the automated identification of neoplastic lymphoid cells. To solve that within the DIP framework, one of the most important steps is the segmentation, which consists in separating the different objects of an image according to similar characteristics of the pixels. This procedure is fundamental to obtain good results in the subsequent stages of feature extraction and classification. The segmentation has been studied in various works in the context of the identification of normal WBCs, neoplastic lymphoid cells and blood cells from leukemia (from PB smears and bone marrow smears), by applying various methods such us: automatic thresholding, color clustering, mathematical morphology, active contours, etc. These works are detailed in the state of the art in Chapter 2 in Section 2.6.3.

The usual leukocyte classification methodology is not able to automatically recognize neoplastic lymphoid cells, because their segmentation and subsequently description should be more detailed. Then, although there have been several approaches for segmentation of normal leukocytes and blood cells from leukemias, neoplastic lymphoid cells can be seen as new types of cells requiring a particular treatment. This segmentation problem has been scarcely studied in the literature (most are about normal leukocytes), and it is always limited to a few types of neoplastic lymphoid cells. In this chapter, a novel methodology is proposed using the color information of the image through fuzzy clustering of different color components and the application of WT with markers, to obtain three regions of the cell: the nucleus, the cytoplasm and the peripheral zone around the cell.

The remainder of the chapter is organized as follows. Section 4.2 explains the original Fuzzy C-means (FCM) clustering technique, its kernelized version Kernel Fuzzy C-means (KFCM), the spatial Fuzzy C-means (sFCM), and the combined technique spatial Kernel Fuzzy C-means (sKFCM). Section 4.3 introduces the principles of the marker-controlled WT. In Section 4.4 is explained the developed methodology for the color segmentation of the lymphoid cells using sKFCM and WT. Section 4.5 provides the experimental results of the sKFCM clustering and the completed segmentation applied over several cell types. Finally, in Section 4.6 the conclusion and future perspectives are presented.

4.2 Fuzzy clustering techniques

Fuzzy c-means (FCM) clustering is an unsupervised method that partitions the data into clusters, minimizing the distance between each data point in the cluster and its center, and maximizing the distance between cluster centers. However, under certain restrictions each data point can belong to several groups at the same time in a fuzzy way [87]. FCM has been widely in medical applications such as analysis of magnetic resonance images, microarray data, cell identification, among others. In this work, a combination of a kernelized fuzzy clustering with a simple improvement using spatial information was made to segment color PB cell images.

4.2.1 Original fuzzy c-means

The original FCM is introduced by Bezdek [88], extending the hard c-means method to a fuzzy perspective. Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$ be a set of n data samples, where each $\mathbf{x}_k$ is a vector defined by p features. For example, $\mathbf{X}$ can be an image, and $\mathbf{x}_k$ a pixel defined by three values corresponding to the RGB color components. $\mathbf{X}$ can be partitioned into c clusters, minimizing the objective function

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} \mu_{ik}^m \|\mathbf{x}_k - \mathbf{v}_i\|^2 \tag{4.1}$$

where $\mathbf{v}_i$ is the ith cluster center, $\|\cdot\|$ is a norm metric (to calculate the distance between each vector and the cluster center) and m is a constant parameter that controls the fuzziness of the partition (the best choice is between 1.5 and 2.5). The membership matrix is defined as

$$U = \left\{ \mu_{ik} \in [0,1] \,|\, \textstyle\sum_{i=1}^{c} \mu_{ik} = 1; 0 < \sum_{k=1}^{n} \mu_{ik} < n \right\} \tag{4.2}$$

where μ_{ik} is the membership of $\mathbf{x}_k$ in the ith cluster.

Although there is not a simple solution, Bezdek et al. [89] proposes an efficient and iterative algorithm to minimize the objective function:

1. *Fix* c $(2 \leq c \leq n)$ *and* m. *Initialize the membership matrix* $U^{(0)}$. *Then, each step is numbered as* r, *where* $r = 0, 1, 2...$

2. *Calculate the c centers* $\mathbf{v}_i$ *for each step using*

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{n} \mu_{ik}^m}; \qquad 1 \leq i \leq c \tag{4.3}$$

3. *Calculate the update membership matrix* $U^{(r+1)}$ *using*

$$\mu_{ik} = \left[\sum_{j=1}^{c} \left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|} \right)^{2/(m-1)} \right]^{-1} \qquad 1 \leq k \leq n; 1 \leq i \leq c; for \mathbf{x}_k \neq \mathbf{v}_i \tag{4.4}$$

4. *If* $\left\| U^{(r+1)} - U^{(r)} \right\| \leq \varepsilon$*, stop; otherwise set* $r = r+1$ *and return to step 2.*

In the last step, the algorithm compares a matrix norm $\|\cdot\|$ of two successive fuzzy iterations to a level of accuracy ε. It is also possible to use the objective function to stop the algorithm by $\left| J_m^{(r+1)} - J_m^{(r)} \right| \leq \varepsilon$.

4.2.2 Kernel fuzzy c-means (KFCM)

Several kernel learning algorithms have been proposed in the literature [90]. One of the best known applications is the Support Vector Machines (SVM) [91, 92]. However this is not the only application where kernels can transform a linear algorithm (with inner products) to a nonlinear version [93, 94]. The kernel trick can also be applied to the FCM algorithm modifying its objective function [95–97].

Consider $\Phi : \mathbf{x} \in \mathbf{X} \subseteq R^p \rightarrow \Phi(\mathbf{x}) \in \mathbf{F} \subseteq R^Q$ with $p \ll Q$, a nonlinear transformation to a higher dimensional feature space F (it may be infinite). A kernel can be defined as $K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$, which is a inner product in the new feature space $\mathbf{F}$. This method allows to calculate the inner product in the feature space, without explicitly using Φ. Then, equation (4.1) is kernelized introducing the transformation into the distance:

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} \mu_{ik}^m \|\Phi(\mathbf{x}_k) - \Phi(\mathbf{v}_i)\|^2 \tag{4.5}$$

The expression inside the norm metric can be expanded as follows:

$$\begin{aligned} \|\Phi(\mathbf{x}_k) - \Phi(\mathbf{v}_i)\|^2 &= (\Phi(\mathbf{x}_k) - \Phi(\mathbf{v}_i))^T (\Phi(\mathbf{x}_k) - \Phi(\mathbf{v}_i)) \\ &= \Phi(\mathbf{x}_k)^T \Phi(\mathbf{x}_k) - \Phi(\mathbf{v}_i)^T \Phi(\mathbf{x}_k) - \Phi(\mathbf{x}_k)^T \Phi(\mathbf{v}_i) + \Phi(\mathbf{v}_i)^T \Phi(\mathbf{v}_i) \\ &= K(\mathbf{x}_k, \mathbf{x}_k) + K(\mathbf{v}_i, \mathbf{v}_i) - 2K(\mathbf{x}_k, \mathbf{v}_i) \end{aligned} \tag{4.6}$$

There are some typical kernel functions used in different applications: Polynomial, Sigmoid, inverse multiquadric and specially radial basis function (RBF) [93, 95]. This last kernel is defined as

$$K(\mathbf{x}, \mathbf{v}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{v}\|^2}{\sigma^2}\right) \tag{4.7}$$

where σ is a customizable parameter of the RBF function kernel.

From this kernel, a convenient consequence when both variables are the same is obtained: $K(\mathbf{x}, \mathbf{x}) = K(\mathbf{v}, \mathbf{v}) = 1$. Then, substituting equation (4.7) for kernel (4.7) into equation (4.5), the objective function is simplified as

$$J_m = 2\sum_{i=1}^{c}\sum_{k=1}^{n} \mu_{ik}^m (1 - K(\mathbf{x}_k, \mathbf{v}_i)) \tag{4.8}$$

Following the same principle that FCM clustering, the minimization of the new objective function of the RBF kernel produces two expressions for the cluster centers and the membership matrix [97]:

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m K(\mathbf{x}_k, \mathbf{v}_i)\, \mathbf{x}_k}{\sum_{k=1}^{n} \mu_{ik}^m K(\mathbf{x}_k, \mathbf{v}_i)} \tag{4.9}$$

$$\mu_{ik} = \frac{(1 - K(\mathbf{x}_k, \mathbf{v}_i))^{-1/(m-1)}}{\sum_{j=1}^{c} (1 - K(\mathbf{x}_k, \mathbf{v}_i))^{-1/(m-1)}} \tag{4.10}$$

Since the above two equations are necessary but not sufficient to solve the optimization problem, it is necessary to perform a similar algorithm as the one proposed for FCM in Section (4.2.1). This procedure is accomplished by substituting equations (4.3) and (4.4) by equations (4.9) and (4.10), respectively into the algorithm. In equation (4.9) the RBF kernel function acts as a weighting, thus when the data points are far from the cluster center, the corresponding weighted sum will be smaller, making the KFCM algorithm a robust estimator. Furthermore,

the parameter for σ should have an appropriate value, neither too large nor too small, which can be found by trial and error testing [96].

4.2.3 Spatial kernel fuzzy c-means (sKFCM)

The previous development is general for any data set, but if the problem is particularized to an image it can be taken advantage of intrinsic considerations. For instance, the neighboring pixels are strongly correlated because they have related characteristics. Thus, there is a greater probability for these pixels to belong to the same cluster. Chuang et al. [98] proposes to include the spatial information inside the FCM algorithm and therefore modifying the membership matrix as summarized below.

Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$ be an image of n pixels, where each $\mathbf{x}_k$ is a vector that defines a pixel by p color components in a color space. The spatial function is defined as

$$h_{ik} = \sum_{j \in NB(\mathbf{x}_k)} \mu_{ij} \tag{4.11}$$

where $NB(\mathbf{x}_k)$ is a neighborhood region around the pixel $\mathbf{x}_k$ (centered square window). This function is a spatial likelihood for the pixel $\mathbf{x}_k$ to belong to the ith cluster, which is high if the neighboring pixels are in the same cluster. Then, after the FCM updating step of the membership matrix (step 3, equation (4.4)), a new modified membership is calculated as

$$\tilde{\mu}_{ik} = \frac{\mu_{ik}^p h_{ik}^q}{\sum_{j=1}^{c} \mu_{jk}^p h_{jk}^q} \tag{4.12}$$

where p and q are control parameters of the membership matrix and the spatial function, respectively. If $p = 1$ and $q = 0$ the original memberships remain unchanged as the classic FCM. A significant consequence of this simple method is that it reduces the noise influence due to the weighting of the spatial function.

The previous method can be also included into the KFCM modifying the corresponding membership function. In order to clarify the FCM procedure with the kernel and spatial insertions, the algorithm for the sKFCM method is outlined below:

1. *Fix c and m. Initialize the membership matrix $U^{(0)}$. Then, each step is numbered as r, where $r = 0, 1, 2...$*

2. *Calculate the c centers $\mathbf{v}_i$ for each step using equation (4.9).*

3. *Update the membership matrix $U^{(r+1)}$ using equation (4.10).*

4. *Calculate the spatial function of equation (4.11) and update again the membership matrix using equation (4.12).*

5. *If* $\left\| U^{(r+1)} - U^{(r)} \right\| \leq \varepsilon$, *stop; otherwise set* $r = r + 1$ *and return to step 2.*

4.3 Marker-controlled watershed transformation

Intuitively, the WT is a region-based segmentation method that treats the image like a topographic relief. Then, water falling on this relief will fill the basins beginning with the local minimums. Thus, water flows moving from different basins could clash, particularly when the water reaches the maximum peaks, resulting in watershed lines that correspond to the limits of the adjacent basins of the water regions [57]. Generally, the WT is not applied over the original image, but over some function of it, e.g. the gradient image. There have been several implementations of the WT algorithms depending of its formulation [56], but this work uses one of the most commonly applied algorithms based on distance functions, which is introduced by Meyer et al. [99]. In order to avoid any over segmentation, a set of markers can be defined over the image function, considerably improving the results [100].

4.4 Color segmentation using sKFCM and watershed transformation

The segmentation methodology proposed in this work uses the color composition of the cell image due to the MGG staining applied over the blood sample, particularly the blue, magenta and pink colors, seen as a first approximation of perception. The developed technique involves the segmentation by color clustering using the sKFCM algorithm and the WT to obtain directly the nucleus region and indirectly the cytoplasm and the peripheral zone around the lymphoid cell. Figure 4.1 shows a simple scheme of the proposed segmentation methodology, which has different parts. The most important sub-processes of this methodology are described below.

4.4.1 Preprocessing and color transformation

For removing noise of an image, mean or Gaussian filters are generally used, but they also eliminate important information such as the details of the object edges. In this work, the cell image was preprocessed by a filter to smooth the image but also preserve the edges [101]. This

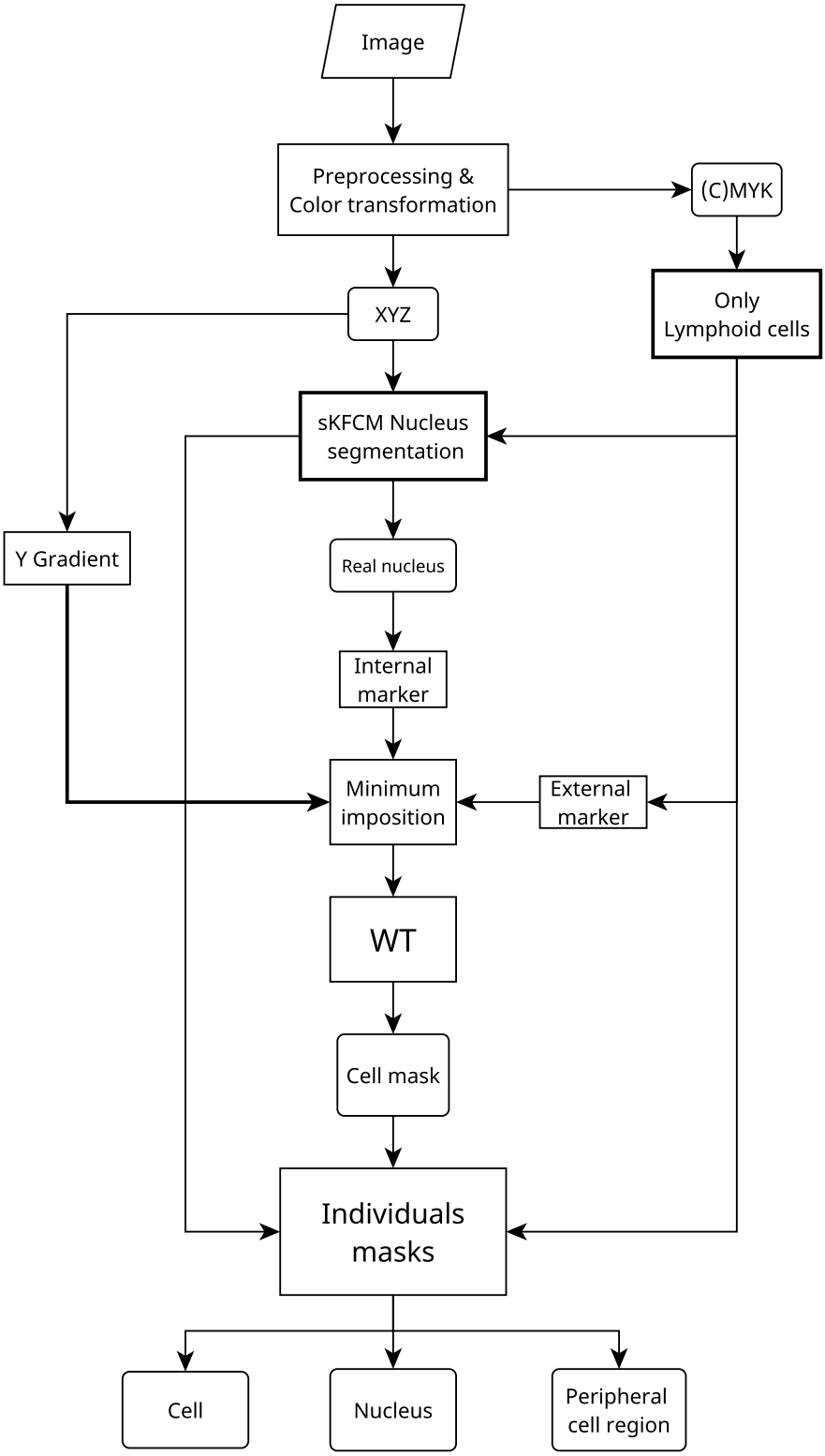


FIGURE 4.1: *General methodology of the lymphoid cell segmentation.*

filter is based on a convolution mask that uses the Manhattan color distances between a central pixel and its 8-connected neighboring pixels. The color distance is defined as

$$d_i = \frac{|r_c - r_i| + |g_c - g_i| + |b_c - b_i|}{3N} \qquad 0 \leq i \leq 8 \tag{4.13}$$

Where r_c, g_c, b_c are the color values of the central pixel in the RGB color space, r_i, g_i, b_i are the pixels in the 8-connected neighborhood, and N is the number of possible values for each color component (usually 255). Then, the convolution mask is given by

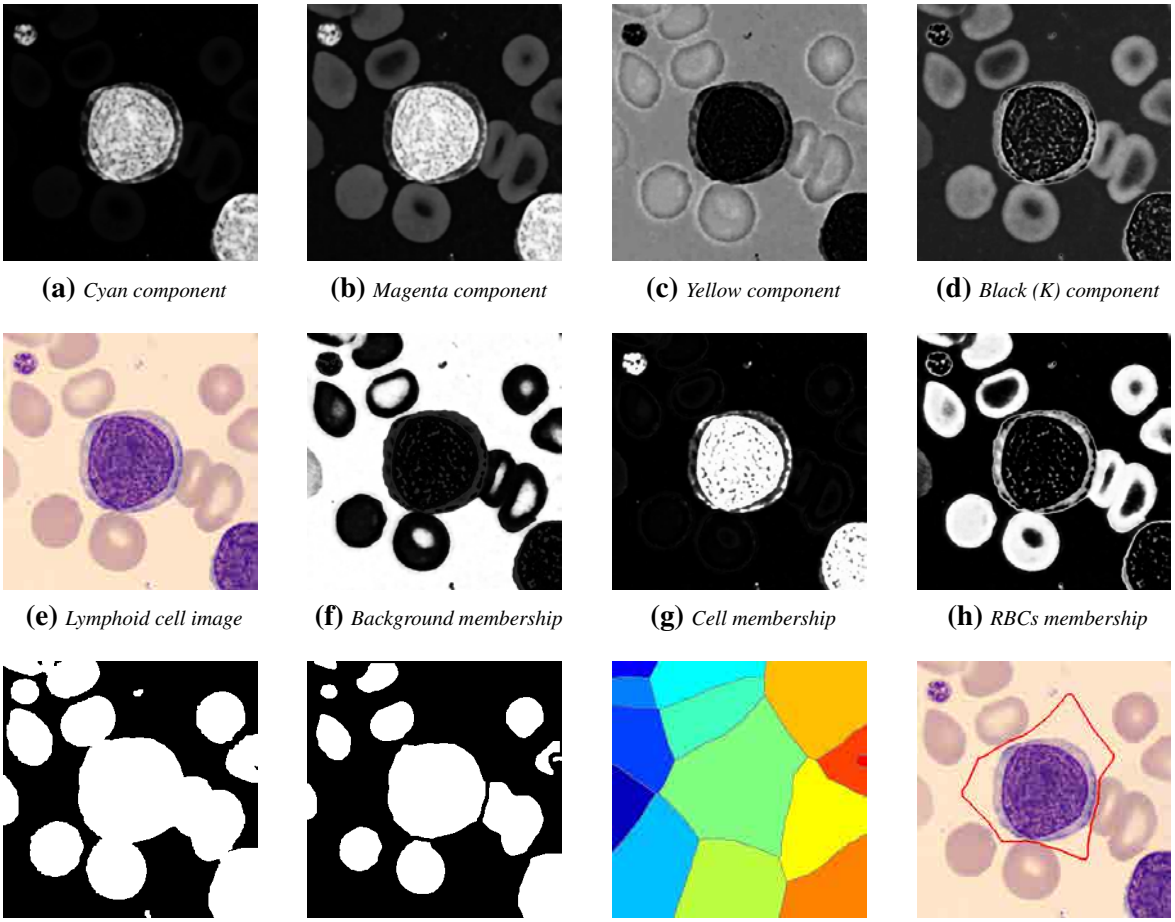
$$\frac{1}{\sum_{i=1}^{8} c_i} \begin{pmatrix} c_1 & c_2 & c_3 \\ c_4 & 0 & c_5 \\ c_6 & c_7 & c_8 \end{pmatrix} \tag{4.14}$$

where $c_i = (1 - d_i)^s$ with $s \geq 1$. Filtering is carried out through the convolution of each color component (plane) with the above mask. The parameter s controls the blurring over the image, since it scales the color differences. A value of $s = 10$ was utilized to filter the images, producing a good performance.

After the filtering process, the image was transformed to the CMYK (cyan, magenta, yellow and black) and CIE 1931 XYZ (Y is luminance, and the XZ plane contains the chromaticities), to provide the color information to the subsequent algorithms.

4.4.2 Only lymphoid cells algorithm

In this algorithm, the filtered cell image was processed to produce a binary mask which separates the lymphoid cell from the remaining blood components. The main part of this algorithm worked as a filter of RBCs producing a new image without them. This result was achieved through the clustering segmentation by sKFCM with three clusters of the components Y and K of the CMYK color space, generating three membership images: background, cell and RBCs. Since this clustering method does not detect which corresponds to each cluster, these three images were automatically identified by comparing with the green background and the magenta foreground (both were binary masks obtained by thresholding). Afterwards, a simple thresholding was applied on the RBCs membership image resulting in a binary mask, which was used to reduce the thickness of the RBCs over the foreground (a defuzzification of the background membership image). Then, the WT was performed over the distance transform of this modified foreground, obtaining a separation of the cells in a label matrix that divides the image in several regions which contain different cells. Subsequently, a first approach of the nucleus by combining the red and green components of the RGB color space allowed to identify the cropping section that encloses the lymphoid cell. All the above procedure could



(a) *Cyan component* **(b)** *Magenta component* **(c)** *Yellow component* **(d)** *Black (K) component*

(e) *Lymphoid cell image* **(f)** *Background membership* **(g)** *Cell membership* **(h)** *RBCs membership*

(i) *Foreground mask* **(j)** *Modified foreground mask* **(k)** *WT label matrix* **(l)** *Selected cropping section*

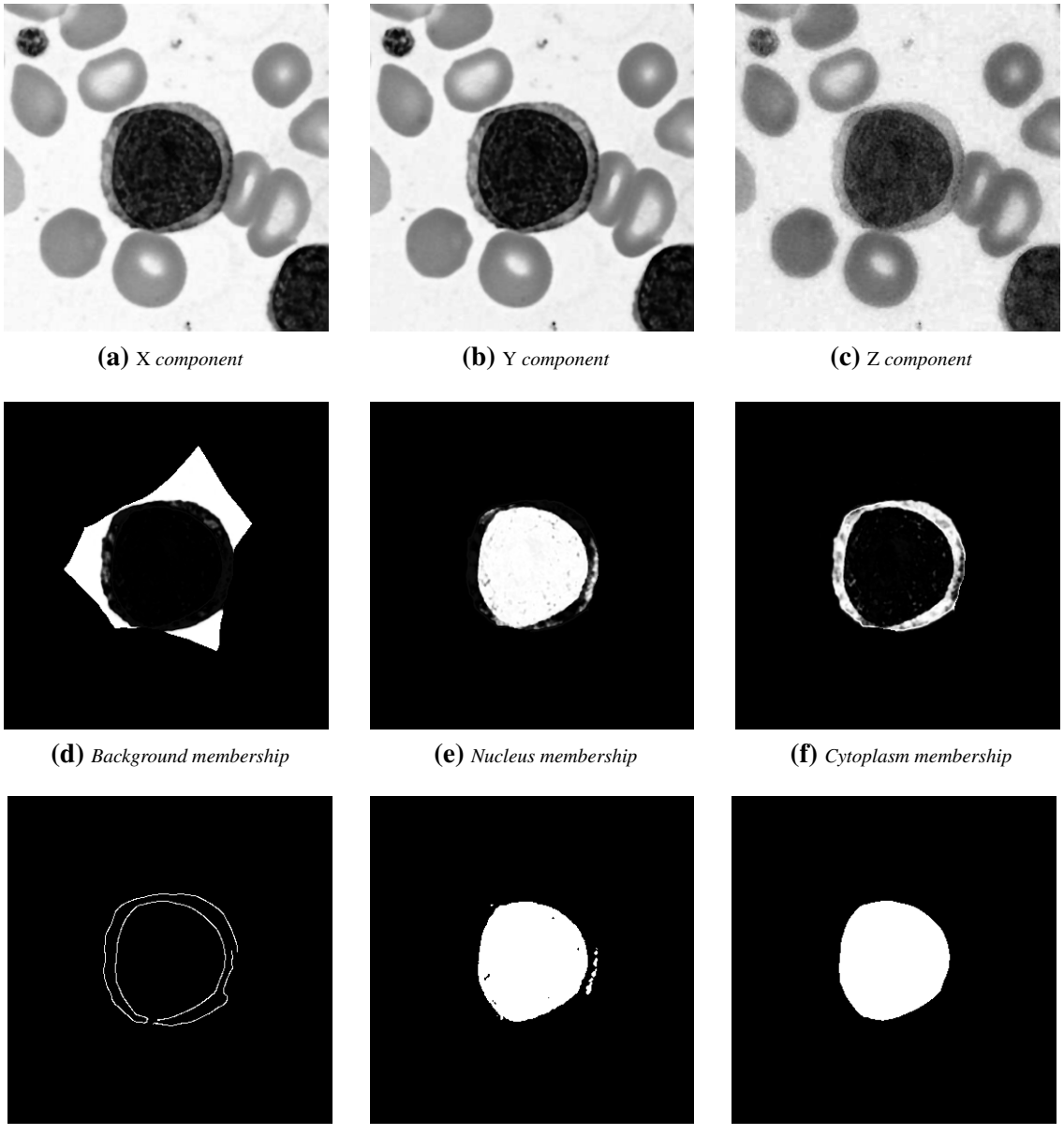
FIGURE 4.2: *Different stages for a Mantle Cell Lymphoma cell in the algorithm* only lymphoid cells.

produce various regions of interest (ROI) because an image may contain multiple lymphoid cells.

Figure 4.2 illustrates the above described procedure for a cell from a patient suffering Mantle Cell Lymphoma (MCL) : Figures 4.2a-4.2d show the CMYK color components of the lymphoid cell image in Figure 4.2e; Figures 4.2f-4.2h show the related membership images as a result of the Y-K clustering process; Figure 4.2i exhibits the modified mask respect to the original foreground in Figure 4.2j; Figure 4.2k illustrates the label matrix after the application of the WT; and Figure 4.2l shows the lymphoid cropping region delimited by the red line.

4.4.3 sKFCM nucleus segmentation

The sKFCM technique was applied on the transformed image to the CIE XYZ space color, but limited to the lymphoid cell cropping region, to obtain (at least) three membership images:



(a) X *component* **(b)** Y *component* **(c)** Z *component*

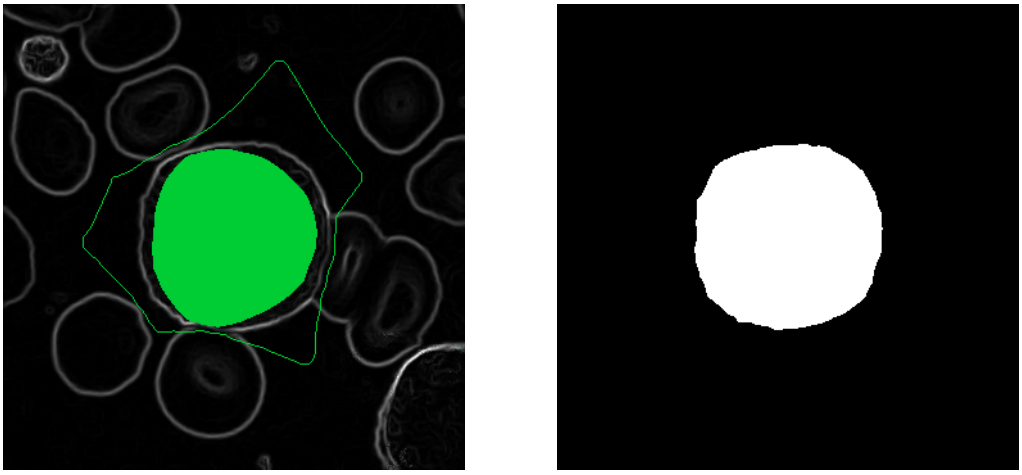
(d) *Background membership* **(e)** *Nucleus membership* **(f)** *Cytoplasm membership*

(g) *Cytoplasm membership edges* **(h)** *Defuzzified nucleus membership* **(i)** *Nucleus binary mask*

FIGURE 4.3: *Different stages during the nucleus segmentation procedure for a MCL cell.*

nucleus, cytoplasm and background. They were automatically identified by comparing with the green background and the complement of the green component[2] (maximum green value minus green value). Then, a robust but simple procedure was done by combining the binary mask of the nucleus membership image (defuzzification by thresholding) and the edges of the cytoplasm membership image, achieving an excellent nucleus segmentation. Although,

[2]It is similar to the magenta color component



(a) *Y gradient with markers* **(b)** *Cell binary mask*

FIGURE 4.4: *Application of the watershed transformation on the Y gradient.*

segmentation could be done using only the membership image of the nucleus, the last step significantly improved the results.

Continuing with the example shown in Figure 4.2, Figure 4.3 illustrates the nucleus segmentation procedure: the XYZ color components of the lymphoid cell are shown in Figures 4.3a-4.3c; the resulting membership images of the sKFCM clustering over the X, Y and Z components are exhibited in Figures 4.3d-4.3f; the edges of the cytoplasm membership image calculated by Canny edge detection are shown in Figure 4.3g; the defuzzification by thresholding of the nucleus membership image is illustrated in Figure 4.3h; and the final binary mask of the nucleus is shown in Figure 4.3i.

4.4.4 Cell segmentation by WT

This stage produced the binary mask of the whole cell by applying the WT with controlled markers. The perimeter of the lymphoid cropping region was utilized as the external marker and the binary mask of the nucleus as the internal mask. Subsequently, they were imposed as minimum over the gradient of the Y component of the XYZ color space. Then, the WT was applied on this modified gradient, obtaining the whole cell region.

Figure 4.4 shows the cell segmentation by applying the WT for the lymphoid cell in Figure 4.2e. Figure 4.4a depicts the Y gradient and the external and internal markers (green lines) and Figure 4.4b illustrates the binary mask of the entire cell after performing the WT.

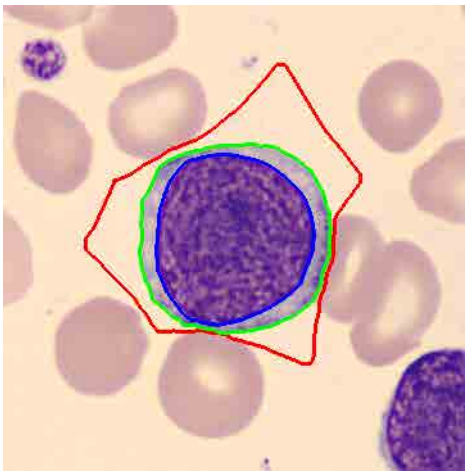


FIGURE 4.5: *Complete segmentation of the lymphoid cell in Figure 4.2e.*

4.4.5 Individual masks

At this point, if there was a single lymphocyte in the cell image, the three ROI had already been obtained: the nucleus, the entire cell and the peripheral zone around the cell. However, if there were multiple lymphoid cells in the image, the three regions obtained would be: all the nucleus, all the entire cells and the cropping section that contains all the lymphoid cells. Therefore, an extra algorithm was developed to generate the individual mask for each lymphocyte. This procedure was able to obtain masks for each cell separating the cases: (1) there was a single nucleus per cell and (2) there were multiple nucleus per cell. Finally, some post-processing operations on each final mask were done to clean several residues, such as: cell fragments in the border, cytoplasm remains that cross the boundaries, RBC residues in the peripheral zone around the cell, among others. Finally, the cytoplasm mask was obtained by the difference between the masks of the whole cell and the nucleus, and the peripheral zone around the cell was obtained by the difference between the lymphoid cropping section and the entire cell mask.

The final segmentation of the lymphoid cell in Figure 4.2e is shown in Figure 4.5, which includes three regions: peripheral zone around the cell (between red and green lines), cytoplasm (between green and blue lines) and nucleus (inside the blue line).

4.5 Experimental results

The developed algorithms were applied over peripheral blood images which come from samples of normal donors (N) and patients with different B mature lymphoid neoplasms such as:

Chronic Lymphocytic Leukemia (CLL), Hairy Cell Leukemia (HCL), MCL and Follicular Lymphoma (FL). B-prolymphocytes (BPL) images were obtained from transformed CLL.

The diagnoses were established by clinical and morphologic findings as well as characteristic immunophenotype of the lymphoid cells. Specifically, CLL cells had the phenotype CD5+, CD19+, CD23+, CD25+, weak CD20+, CD10-, FMC7- and dim surface immunoglobulin (sIg) expression. All the patients with HCL had lymphoid cells with the phenotype CD11c+, CD25+, FMC7+, CD103+ and CD123+. Patients with MCL showed lymphoid cells with the phenotype CD5+, FMC7+, CD43+, CD10- and BCL6-. FL cells showed B-cell associated antigens (CD19, CD20, CD22, CD79a) BCL2+, BCL6+, CD10+, CD5- and CD43- .

Blood samples were obtained from the routine workload of the Core Laboratory of the Hospital Clínic of Barcelona. Venous blood was collected into tubes containing K_3EDTA as anticoagulant. Samples were analyzed by a cell counter Advia 2120 (Siemens Healthcare Diagnosis, Deerfield, USA) and PB films were automatically stained with MGG in the SP1000i (Sysmex, Japan, Kobe) within 4 hours of blood collection.

Individual lymphoid cell images from PB had a resolution of 360 x 363 pixels and they were obtained by the CellaVision DM96 system (Lund, Sweden). The quality of the smears and cell morphology was assessed by hematologists prior to the image study.

The parameters of the sKFCM clustering algorithm used in most segmentation experiments in this section were: $m = 2, p = 0, q = 2, \sigma = 150$, and $n = 3$ clusters. These values will be used in the rest of this chapter, unless specific values are stated.

4.5.1 sKFCM clustering of the entire lymphoid cell image: RBCs, cell and background

As it was mentioned above, the first main part of the segmentation methodology is the separation of the lymphoid cells from the remaining components, i.e. the rough differentiation between WBCs, RBCs, and background (and other components). In this section, the results of the sKFCM clustering for different color spaces of six types of lymphoid cells are presented to show how the different parts of the image can be separated. To display results in a clear and compact way, a false colored representation was assigned for each resulting membership images of the clustering process: yellow for the cell/nucleus, magenta for the RBCs, and black for the background; all of these were integrated into a single image.

Figure 4.6 shows the clustering experiments for lymphoid cells using several color spaces. Rows represent a different cell type: (a) N, (b) HCL, (c) MCL, (d) FL, (e) CLL, and (f) BPL. Every column (from the second one) represents the clustering results for a particular color

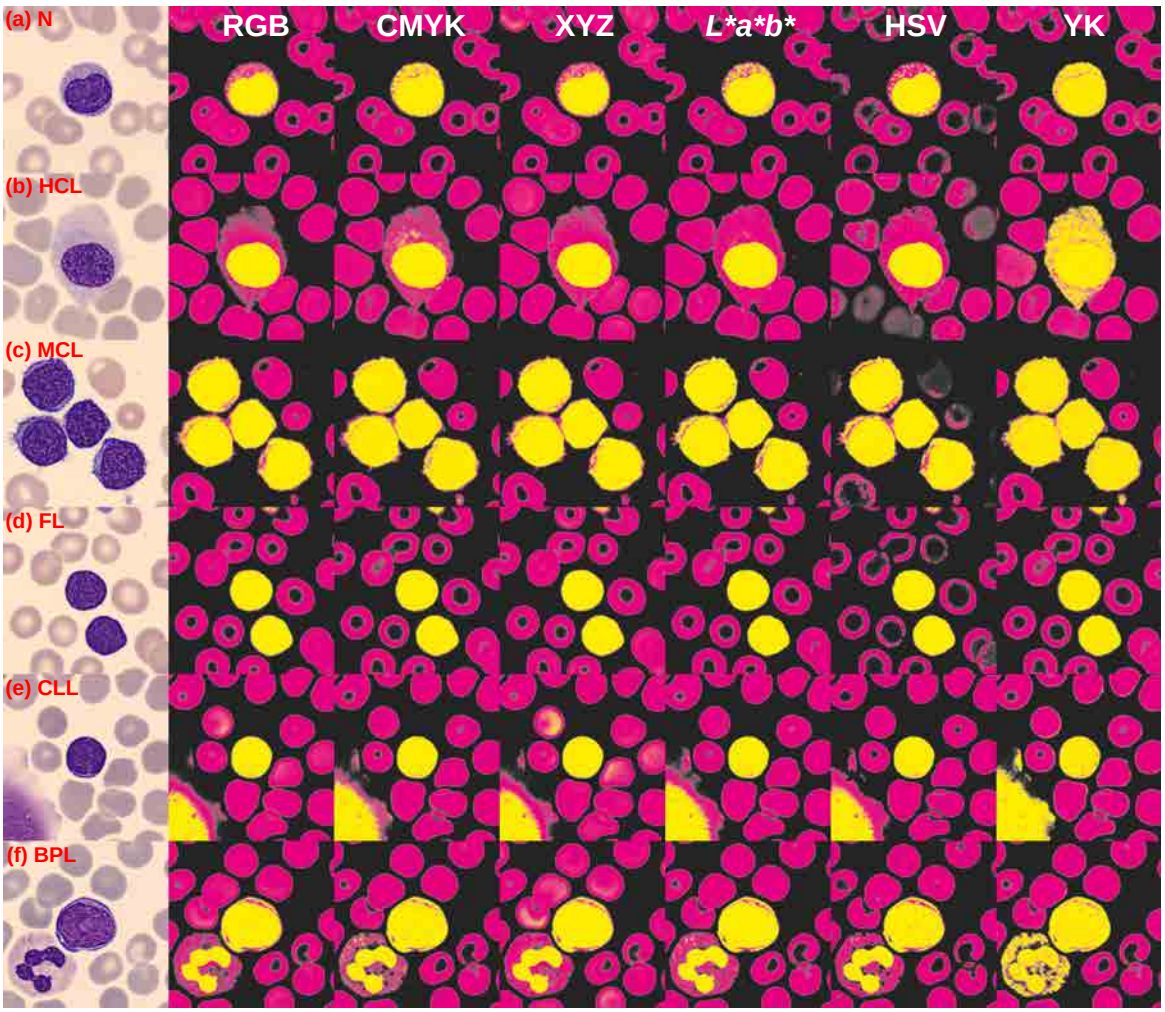


FIGURE 4.6: *sKFCM clustering experiments of six cell-type images (rows) for five colors spaces and two particular components Y and K of the CMYK space (columns). The parameters used for the algorithm were:* $m = 2; p = 0; q = 2; \sigma = 150$; *and* $n = 3$ *clusters.*

space: RGB (red, green and blue), CMYK (cyan, magenta, yellow, black), CIE 1931 XYZ (Y is the luminance, X and Z are the chromaticities), *L*a*b** CIE 1976 (L* is the luminance, a* and b* are the chromaticities), HSV (hue, saturation and value), and the last column represents the clustering using only the Y and K components of the CMYK space. RGB and XYZ clustering shows an excellent and precise definition of each object, but they may confuse the cytoplasm region with the RBCs such as for N and HCL cells in Figure 4.6a-b. CMYK and *L*a*b** clustering exhibit a good performance for most cell types, except for those with a clear cytoplasm (low basophilia) as the HCL cell in Figure 4.6b. HSV clustering does not work well because it may confuse some RBCs with the background, as shown in the respective cases in Figure 4.6a-d. Finally, although Y-K sKFCM process presents some few mixed granulations,

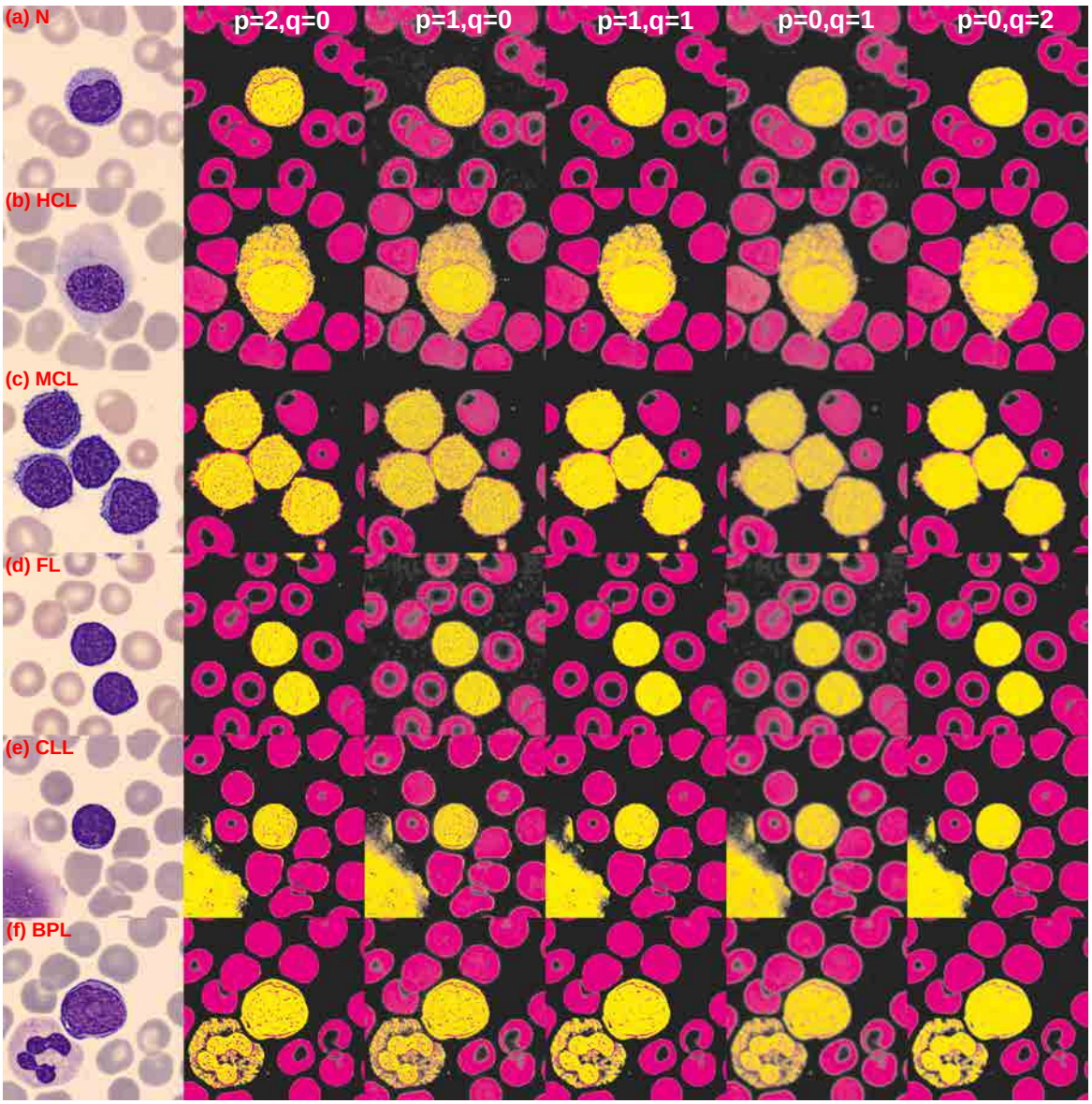


FIGURE 4.7: *sKFCM clustering experiments for the Y and K color components varying the spatial parameters p and q. The remaining parameters used for the algorithm were:* $m = 2, \sigma = 150$ *and* $n = 3$ *clusters.*

this clustering has the best performance in the differentiation between the components of the cell image. For instance, normal and HCL cell in Figure 4.6a-b, the shadow in Figure 4.6e, and the entire neutrophil in Figure 4.6f are clearly distinguished by the membership images. The latter is the main reason why Y and K clustering was chosen as an essential part of the complete methodology to isolate the lymphoid cell.

4.5.2 sKFCM clustering of the CMYK color space with different spatial parameters

In this section, the sKFCM clustering of the Y and K components of the CMYK color space with different spatial parameters p and q are presented for the same cells in Figure 4.6. The resulting color membership images for $p = 2$ and $q = 0$, i.e. the spatial part is null while the membership of the intensities is amplified by squared, are shown in the second column of Figure 4.7; they are clear, well defined and sharpened with some variations inside the cell membership similar to a noise. When $p = 1$ and $q = 0$, the technique is the same that KSFCM, and these images present a high quantity of noise variations into the background, RBC and cell memberships (e.g. Figure 4.7d). This is the worst result of all the experiments with various parameters. The experiment with $p = 1$ and $q = 1$ involves both the intensity membership as the spatial part, resulting in clearer and more defined images with less variations inside the cell membership. The clustering with $p = 0$ and $q = 1$ is equivalent to a smoothed version (low-pass filter) of the membership produced by the KSFCM case, as it can be seen comparing the third and fifth columns in Figure 4.7. Lastly, the sKFCM clustering with $p = 0$ and $q = 2$ produces the best results with well defined, slightly sharpened and moderately smoothed membership images for the background, RBCs and the cell.

4.5.3 sKFCM clustering of the limited lymphoid cell image: nucleus, cytoplasm and background

Through the Y-K clustering and the *only lymphoid cells* algorithm, the generated cropping region was used to reduce the information of the cell image. This section presents several results of the kSFCM clustering on the limited image for various color spaces applied to the same lymphoid cell types shown before in Figure 4.6. Figure 4.8 shows the clustering experiments, where rows represent the different lymphoid cell types and the columns the color spaces. The false color representation for each membership image was assigned as follows: yellow for the nucleus, magenta for the cytoplasm and black for the background. Each cell image is zoomed according to the lymphoid cropping region to improve the visualization. Due to the remarkable textures of the nuclear chromatin, most of the color membership images exhibit significant variations of the cytoplasm membership inside the nucleus region as it can seen especially for the RGB, CMYK, *L*a*b** and HSV clustering processes for the FL cell (Figure 4.8d) or the CMYK clustering of the CLL cell (Figure 4.8e). Thus, as the fourth column of the Figure 4.8 shows, the best clustering performance happens for the XYZ color space, resulting in clear definitions and small variations for the different regions. This

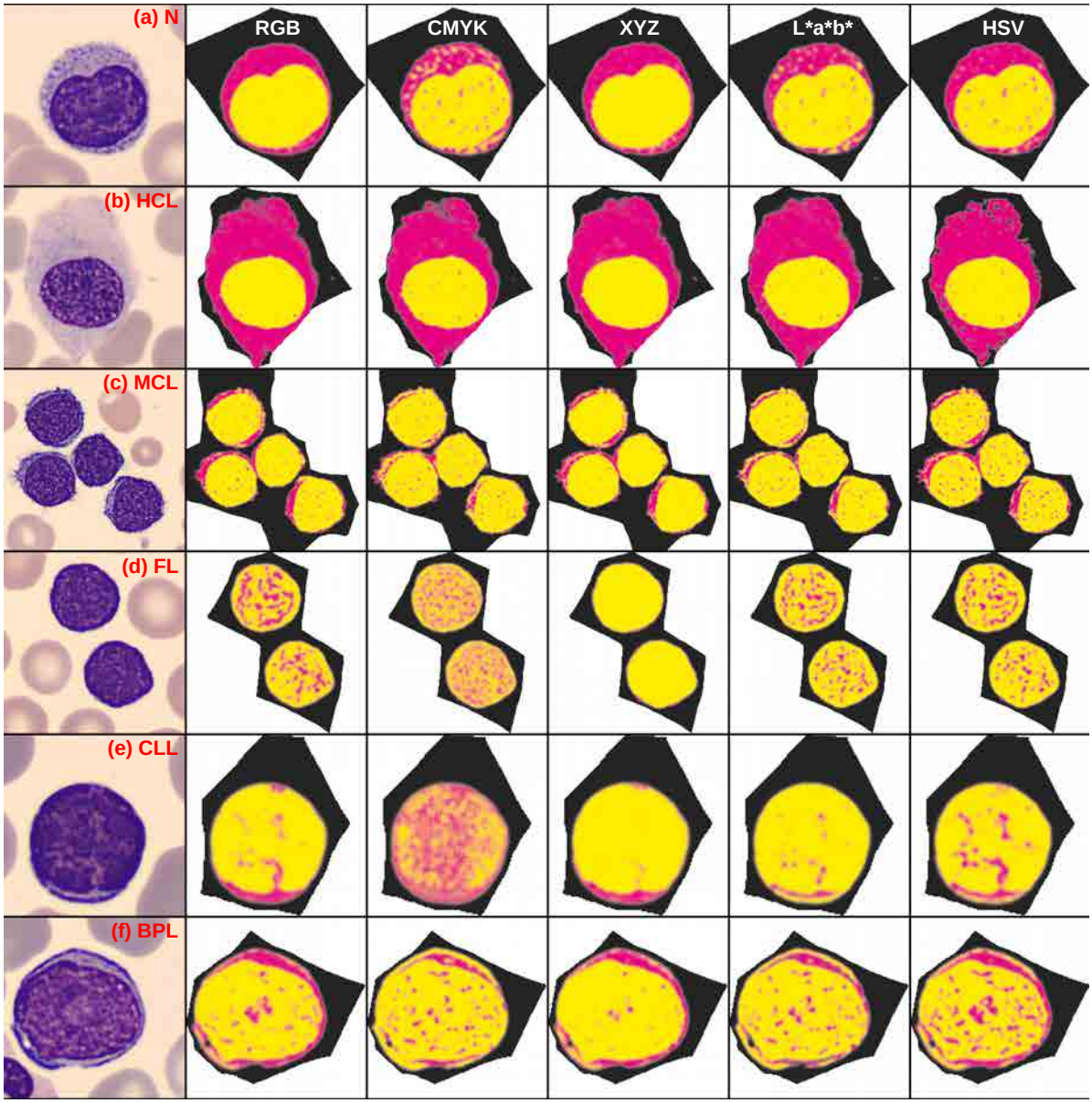
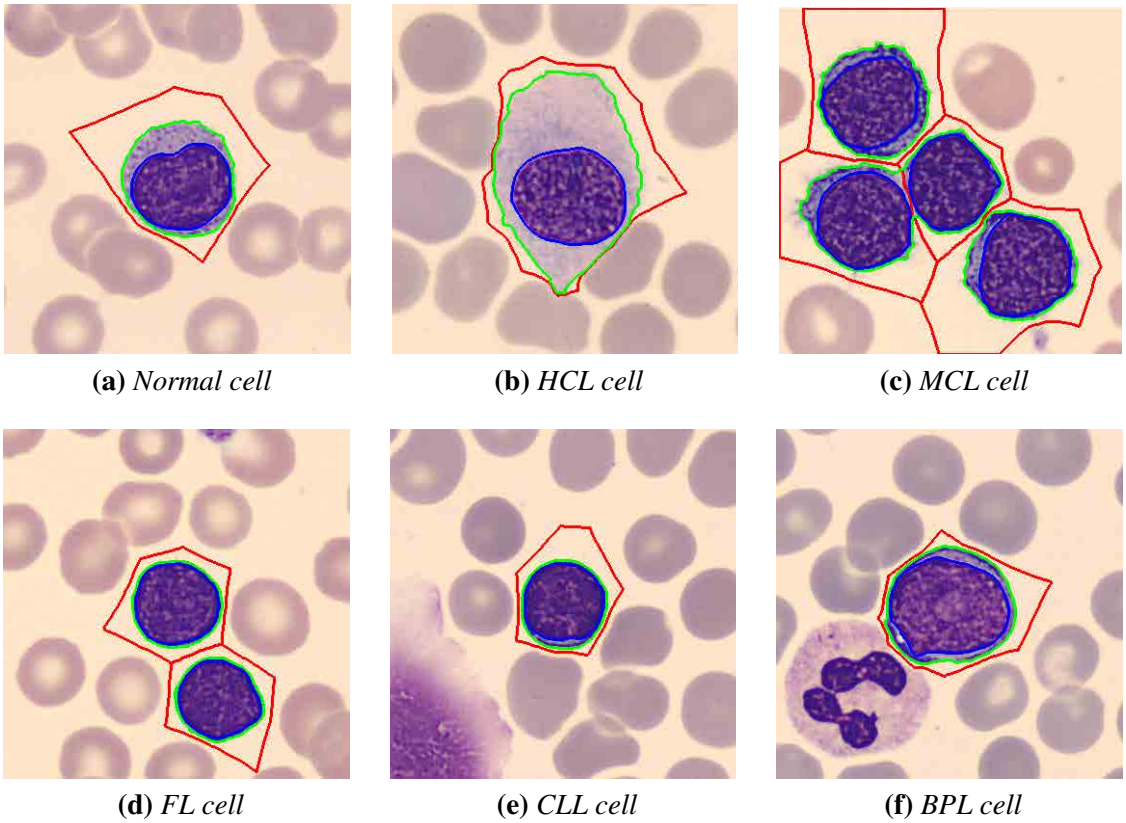


FIGURE 4.8: *sKFCM clustering experiments of six cell-type images (rows) for five color spaces. The parameters used for the algorithm were:* $m = 2; p = 0; q = 2; \sigma = 150;$ *and* $n = 3$ *clusters.*

last result became fundamental for the general methodology to segment the nucleus of the lymphoid cells.

4.5.4 Completed segmentation results

The segmentations of the cells in Figures 4.6-4.8 are shown in Figure 4.9. These results present an excellent performance for the six types of lymphoid cell: separating the whole cell even



(a) *Normal cell* **(b)** *HCL cell* **(c)** *MCL cell*

(d) *FL cell* **(e)** *CLL cell* **(f)** *BPL cell*

FIGURE 4.9: *Results of completed segmentation for the cell in Figures 4.6-4.8*

when RBC are touching it (Figure 4.9b), simultaneously segmenting multiple cells (Figure 4.9d) even when they are touching each other (Figure 4.9c), isolating shadows (Figure 4.9e), or not considering other types of WBCs (Figure 4.9f).

It is important to remark that several works have segmented only two regions of the cell (nucleus and cytoplasm) [39, 53, 59, 67]. In Chapter 3, the WT on the green color component was applied to segment three regions: the nucleus, the cell (the cytoplasm by difference), and the peripheral zone around the cell. This last region was especially significant because it allowed to evaluate the external profile of the cytoplasm. This chapter has implemented the segmentation of the same three regions, but with a higher efficiency and robustness, using the color information of the image instead of only one gray level.

4.5.5 Efficiency of the segmentation methodology

A total of 3394 individual PB lymphoid cells distributed between six types were segmented to measure the efficiency of the developed segmentation methodology. The efficiency was defined as the number of correctly segmented cell images divided by the total number of

TABLE 4.1: *Segmentation efficiency for various types of lymphoid cells*

Type	#Cells	Correct	Efficiency
N	287	260	90.59%
HCL	571	530	92.82%
MCL	827	732	88.51%
FL	582	551	94.67%
CLL	904	863	95.46%
BPL	223	214	95.96%
Total	*3394*	*3150*	*92.81%*

images (for each type and for the whole set). A segmentation was considered correct when the three regions were well-segmented: the nucleus, the cell and the peripheral zone around the cell. Then, if one of the segmented regions was wrong, the whole process would be seen as incorrect. The confirmation of the segmentation results during the experiments was done by expert cytologists. Table 4.1 presents the number of cells, the correctly segmented cell images and the segmentation efficiency for each type of lymphoid cell. Almost all cell types were segmented with an efficiency above 90%, except for the MCL cells which have a variable size, irregular nuclear profile and the chromatin may have different textural variations being condensed, lax or immature. But, although MCL cells is the most complicated cell type to segment, the 88.51% of efficiency achieved is quite good.

4.5.6 sKFCM clustering of cell images from different sources

All the results presented above were obtained from images acquired from the Cellavision DM96 system, following a standard protocol (e.g. the automated staining), to take advantage of pathological information obtained during several years in the clinical laboratory. However, the extension of the segmentation methodology to other images acquired by different equipment may allow the application of the DIP framework and the automatic detection of lymphoid neoplasms to simpler clinical laboratories. Accordingly, results of the application of the sKFCM technique over images from two different microscopes are presented in this section.

Figure 4.10 shows the clustering of the Y and C components of the CMYK color space for two cell images of patients with acute myeloid leukemia (AML) and AML with nucleophosmin gene mutations (NPM-AML). They were acquired with a pixel resolution of 640 $\times$ 480 from an Olympus BX50 microscope equipped with a compact 3CCD RGB color camera module SONY XC-003. Figure 4.11 also shows the Y-K clustering for two cell images of

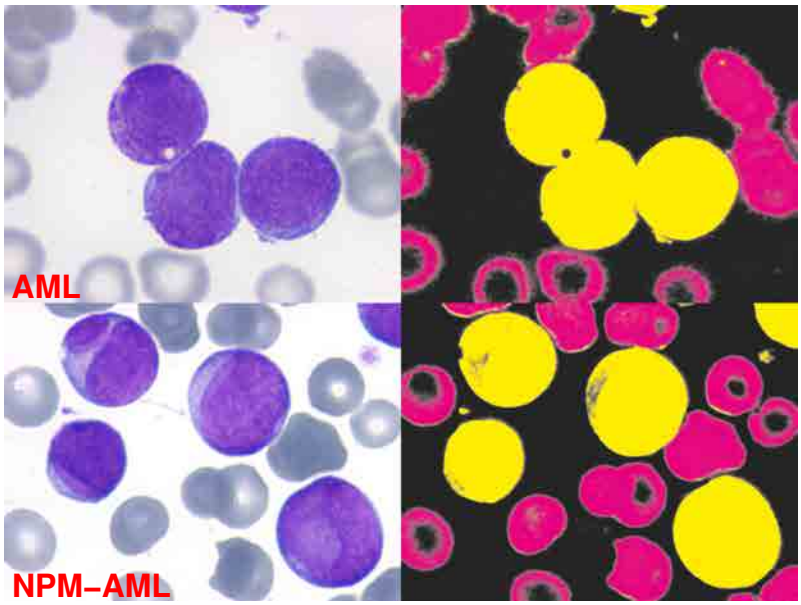


FIGURE 4.10: *Y-K sKFCM clustering of two cell images from patients with AML and NPM-AML. They are obtained using an Olympus BX50 microscope equipped with a compact 3CCD RGB color camera module SONY XC-003. The PB smear is stained with MGG. The parameters used for the algorithm were:* $m = 2; p = 0; q = 2; \sigma = 150;$ *and* $n = 3$ *clusters.*

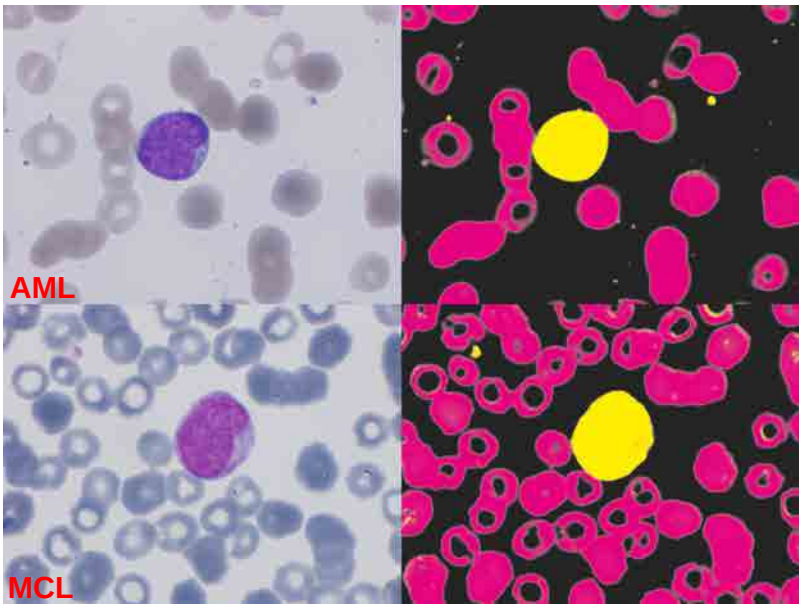


FIGURE 4.11: *Y-K sKFCM clustering of cell images from patients with AML and MCL, obtained from an Olympus BX43 microscope equipped with an Olympus DP73 camera. The PB smear is stained with MGG. The parameters used for the algorithm were:* $m = 2; p = 0; q = 2; \sigma = 150;$ *and* $n = 3$ *clusters.*

patients with AML and MCL. They were obtained by the Olympus BX43 equipped with an Olympus DP73 camera, the images having an original resolution of 4800×3600 pixels. It can be seen that all four color membership images show an excellent representation of each region, separating the lymphoid cell, the RBCs and the background. Therefore, this procedure can be the starting point to generalize the methodology to images obtained with different acquisition systems.

4.6 Discussion

Several works have studied the color clustering segmentation problem of blood cells. Most of them have been applied to the segmentation of normal WBC [24, 34, 50], other works have segmented immature WBC (from bone marrow to study leukemia) [52, 53], while neoplastic lymphoid cell has been scarcely studied [49]. The proposed color clustering method developed in this thesis is used mainly to segment atypical lymphoid cells.

Respect to Fuzzy clustering for blood cell segmentation, the original FCM has been applied for color clustering segmentation of WBC using the L^*a^*b color space and thresholding techniques [24]. The work in this chapter has combined the Kernel method for FCM clustering [95] with the spatial method proposed in [98], thus resulting into a novel segmentation method. While the papers [95, 98] apply the fuzzy clustering segmentation to Magnetic Resonance images, the sKFCM clustering method developed in this thesis has been extended to color images, increasing its applicability.

The sKFCM algorithm can be used for both the complete image or for a cropped section (it is not necessarily a rectangular region). Particularly, the sKFCM color clustering was very useful to separate different regions of PB cell images stained with MGG stain, providing excellent results for most color spaces, especially with the Y and K color components of the CMYK space. Furthermore, tuning the parameters that control the relative importance of the membership and spatial functions allowed to improve the membership images that represent the region of the cell image. For instance, the clustering results with $p = 0$ and $q = 2$ were much better than just using KFCM. Moreover, the color clustering by sKFCM of a cropped image produced more detailed membership images related to the ROI because the information was reduced and approximately limited to them, as it happened with the nucleus segmentation of the lymphoid cell image by using the XYZ color space clustering. Therefore, the sKFCM was used two times in a kind of cascade mode: firstly to crop the lymphoid cell from the remaining regions, and secondly to distinguish between the regions of the nucleus, the cytoplasm and the background.

Since atypical lymphoid cells[3] present more subtle differences compared to normal WBCs, their characterization is more complex, causing the segmentation algorithm has to be more specific and precise. This chapter is focused on the segmentation of different types of atypical lymphoid cells (and normal lymphocytes).

The general problem of segmenting atypical lymphoid cells has seldom been studied. Angulo et al. [59] uses the green and saturation components with thresholding and mathematical morphology tools (Watershed transformation and some morphological operators) to segment the nucleus and cytoplasm of lymphocytes, while Yang et al. [47] develops a segmentation methodology based on active contours (gradient vector flow) and robust estimation for the nucleus and cytoplasm of lymphoid cells from FL, MCL and CLL. The segmentation method proposed in this chapter, through clustering color and Watershed transformation, segments not only the nucleus and cytoplasm, but also the external region of the cell, which can supply relevant information in some types of cells such as HCL. In addition, the efficiency of the segmentation is tested on normal lymphocytes and five types of neoplastic lymphoid cells (HCL, FL, MCL, CLL and PLB).

4.7 Conclusion

A robust segmentation methodology has been developed, which separates three regions: the nucleus, the cytoplasm and the peripheral zone around the cell. The peripheral zone is particularly important to characterize some pathologies which present cells with hairy-like cytoplasmic projections. The segmentation methodology has been able to process different types of lymphoid cell images, which have varied morphologic characteristics, achieving a total efficiency of 92.81%. Thus, the segmentation method could be extended to other types of WBCs such as blasts, reactive lymphocytes, plasma cells and other neoplastic lymphoid cells. On the other hand, promising clustering results indicate that the methodology could be expanded to other images obtained from different acquisition systems.

Since the sKFCM clustering algorithm can work with color spaces, the potential applications are not only limited to PB cells images but to any kind of color images to segment different regions with color and spatial similarities.

In the remaining of this thesis, the developed segmentation methodology is used to complete the further steps of feature extraction/selection and automatic classification of atypical lymphoid cells.

[3] Atypical lymphoid cells include reactive lymphocytes and neoplastic lymphoid cells

Chapter 5

A Methodology for Automatic Recognition of Neoplastic Lymphoid Cell Images from Peripheral Blood

Based upon: S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, Automatic recognition of atypical Lymphoid cells from peripheral blood by digital image analysis, American Journal of Clinical Pathology, vol. 143, pp. 168-176, 2015. doi: 10.1309/AJCP78IFSTOGZZJN

Abstract

The objective of this chapter was the development of a methodology for the automatic recognition of different types of neoplastic lymphoid cells. In the methodology development, a training set (TS) of 1500 lymphoid cell images from peripheral blood was used. Clustering of color components and Watershed Transformation were use to segment the images. A number of 113 features were extracted for lymphocyte recognition by Linear Discriminant Analysis (LDA) with a 10-fold cross validation over the TS. Then, a new validation set (VS) of 150 images was used, performing two steps: 1) tuning the LDA classifier using the TS; and 2) classifying the VS in the different lymphoid cell types. The segmentation algorithm was very effective in separating cytoplasm, nucleus and peripheral zone around the cell. From them, descriptive features were extracted and used to recognize the different lymphoid cells. The accuracy for the classification in the TS was 98.07%. The precision, sensitivity and specificity values were above 99.7%, 97.5% and 98.6% respectively. The accuracy of the classification in the VS was 85.33%.

5.1 Introduction

Detecting lymphoma and leukemia cells timely to provide patients with an adequate treatment is decisive for their prognosis. Frequently, the blood smear provides the primary or the only evidence for a specific diagnosis, remaining an important diagnostic tool even in the age of molecular analysis [9, 79]. In the World Health Organization (WHO) classification, neoplastic cell morphology, along with immunophenotype and genetic changes, remains essential in defining lymphoid neoplasms [6]. Morphologic distinction between various lymphoid cell types requires experience and skill and, moreover, objective values do not exist to define cytological variables. Chronic Lymphocytic Leukemia (CLL) cells are typically small lymphocytes with clumped chromatin and scant cytoplasm. CLL cells are larger than normal lymphocytes (N) and they have abundant weakly basophilic cytoplasm with irregular *hairy* margins. However, subtle differences on morphologic characteristics exhibited by some lymphoma and leukemia cells leads to a significant number of false negatives in the routine

screening. Moreover, additional studies are expensive and time-consuming. This is why having an automated screening imaging system for decision support could reduce the cost and morbidity of the patients.

Some available equipments for Digital Image Processing (DIP) are able to pre-classify cells in different categories by applying neural networks, extracting a large number of measurements and parameters that describe the most significant cell morphologic characteristics [19]. These systems, when integrated in the daily workload, represent an interesting technological advance since they are able to pre-classify most of the normal blood cells in peripheral blood (PB) [17, 102].

Neoplasatic lymphoid cells are the most difficult pathological cells to classify using morphology features only [80]. Few studies about automatic recognition of different neoplastic lymphoid cells with satisfactory results have been published. In most of the previous studies, the lymphoid cell classification has been addressed with pattern recognition systems to separate the cells into categories [20, 59, 70, 74]. Nevertheless, the image processing techniques used in these works are not useful for the current digital images, since the present acquisition technology is based on charge-coupled device sensors [20].

Among the difficulties to overcoming the automation of the lymphoid cell classification process from their morphologic characteristics, the most relevant are: 1) to accurately solve the cell segmentation problem, which means separating the cells of interest from the whole image; 2) to obtain descriptive features from the cells of interest, which allow high accuracy during the classification step; and 3) to train a classifier to distinguish accurately among the different lymphoid subtypes [54, 103]. This chapter presents a methodology for lymphocyte recognition to allow the automatic classification of normal and four neoplastic lymphoid cells circulating in Peripheral Blood (PB) mature B cell lymphoid neoplasms.

5.2 Material and methods

This study was carried out in two stages: 1) the methodology development, and 2) the methodology validation. In the first one, a set of 1500 lymphoid cell images was used, which is referred to as training set. In the second one, a new independent set of 150 cell images was used, which is named validation set. Figure 5.1 shows the scheme of the complete methodology.

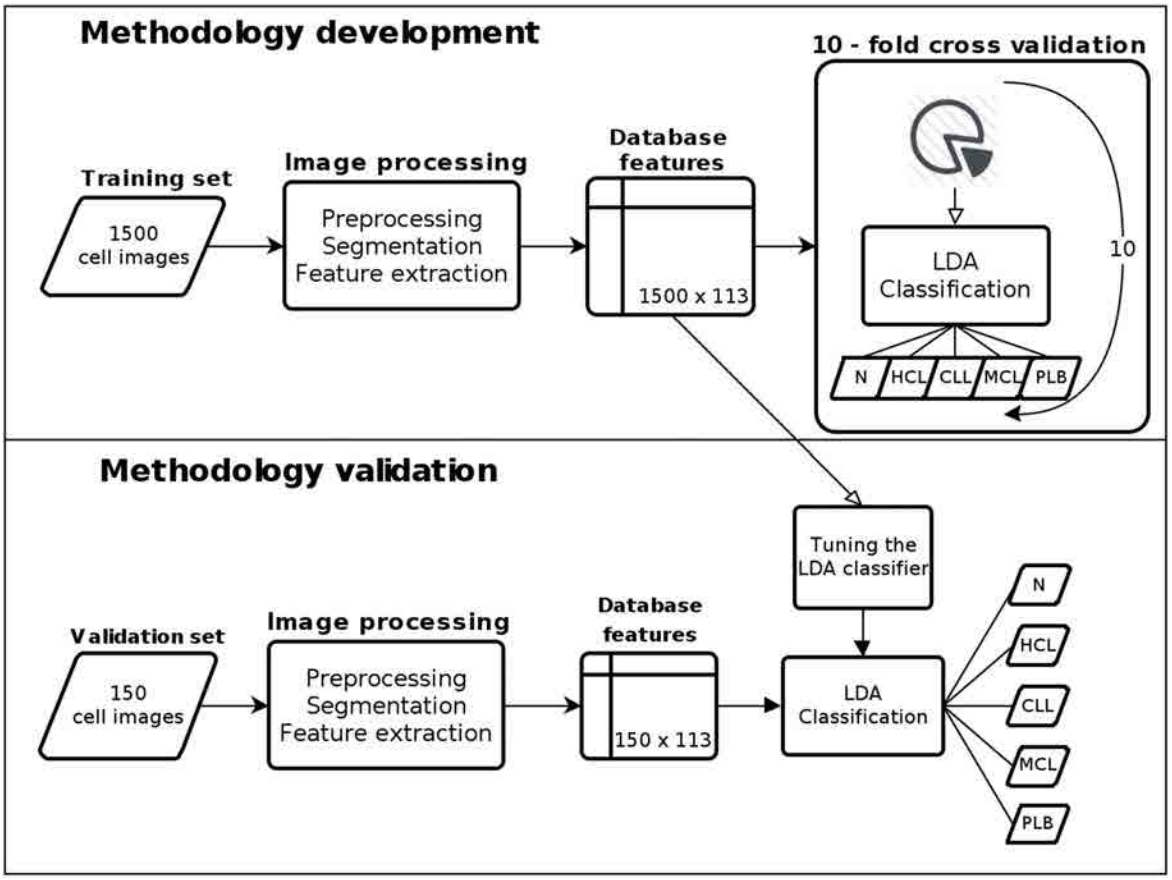


FIGURE 5.1: *The complete methodology has two stages: 1) the methodology development, and 2) the methodology validation. In the first one, a training set of 1500 lymphoid cell images is processed to obtain a database of features, which is used in a 10-fold cross validation with LDA classification to calculate some statistical measures. In the second one, a new independent validation set of 150 cell images was processed following the preceding steps. Then, the validation consists in two steps: a) tuning the classifier using the previous training set; and b) classifying the new validation set in the different lymphoid cell types.*

5.2.1 Methodology development

The methodology to achieve the automatic recognition of lymphoid cells was performed through the following steps:

5.2.1.1 Blood sample preparation and digital image acquisition

Samples from normal donors and patients with CLL, HCL and Mantle Cell Lymphoma (MCL) were included in this study. The diagnoses were established by clinical and morphologic findings as well as characteristic immunophenotype of the lymphoid cells. Specifically, CLL cells had the phenotype CD5+, CD19+, CD23+, CD25+, weak CD20+, CD10-, FMC7- and

dim surface immunoglobulin (sIg) expression. All the patients with HCL had lymphoid cells with the phenotype CD11c+, CD25+, FMC7+, CD103+ and CD123+. Patients with MCL showed lymphoid cells with the phenotype CD5+, FMC7+, CD43+, CD10- and BCL6-. B-prolymphocytes (BPL) images were obtained from transformed CLL.

Blood samples were obtained from the routine workload of the Core Laboratory of the Hospital Clínic of Barcelona. Venous blood was collected into tubes containing K3EDTA as anticoagulant. Samples were analyzed by a cell counter Advia 2120 (Siemens Healthcare Diagnosis, Deerfield, USA) and PB films were automatically stained with May-Grünwald-Giemsa in the SP1000i (Sysmex, Japan, Kobe) within 4 hours of blood collection.

Individual lymphoid cell images from PB had a resolution of 363 x 360 pixels and they were obtained by the CellaVision DM96 system (Lund, Sweden). The quality of the smears and cell morphology was assessed by hematologists prior to the image study.

A training set of 1500 lymphoid cell images from PB films was selected to evaluate the accuracy of the developed methodology. They were distributed as follows: 181 normal lymphocytes from healthy patients, 301 HCL, 401 MCL and 617 from patients with CLL. This group was divided into 542 CLL clumped chromatin typical lymphocyte images and 75 BPL images.

5.2.1.2 Clustering color segmentation

Through the segmentation step, lymphoid cells were separated from other objects in the image [26, 35, 81]. In this chapter, the segmentation methodology developed in Chapter 4 was applied to segment the cells, obtaining three regions for each cell: (1) cell region, (2) nucleus region, and (3) peripheral zone around the cell. The cytoplasm region was obtained by the difference between the regions of the whole cell and the nucleus.

5.2.1.3 Feature extraction

The objective of this step is to obtain quantitative information about the objects in the image under analysis. In the present study, 113 features were used related respectively to: geometry (10), color and texture (102), and cytoplasm external profile (1). Table 5.1 shows the distinct categories of the features used in this work. They are summarized below.

Geometric features They were calculated for each individual cell as described in Chapter 3. These features are quantitative geometric interpretations of morphologic characteristics such as the size and the shape of both nucleus and cell, and the nucleus-cytoplasm ratio.

TABLE 5.1: *Set of 113 lymphoid cell features that were extracted in the methodology described in this chapter.*

Feature Description	Quantitative Feature	
Cytoplasmic profile feature (1)	Estimation of the Hairy projections	
Geometric features (10)	Cell area	Cell diameter
	Cell conic eccentricity	Cell perimeter
	Nucleus area	Nucleus diameter
	Nucleus conic eccentricity	Nucleus perimeter
	Nucleus/cytoplasm ratio	
	Nucleus eccentricity respect to the cytoplasm	
Color and texture features (102)*	**Cytoplasm**	**Nucleus**
First-order statistical features (12 x 3)	Mean	Mean
	Standard Deviation	Standard Deviation
	Skewness	Skewness
	Kurtosis	Kurtosis
	Energy	Energy
	Entropy	Entropy
Second-order statistical features (14 x 3)	Contrast	Contrast
	Homogeneity	Homogeneity
	Correlation	Correlation
	Energy	Energy
	Entropy	Entropy
	Variance	Variance
	Difference Variance	Difference Variance
Granulometric features (8 x 3) (Based on granulometric curve)	Mean	Mean
	Standard Deviation	Standard Deviation
	Skewness	Skewness
	Kurtosis	Kurtosis

* Color and texture features were applied to the nucleus and cytoplasm for each component of the L*a*b* color space.

Color and texture features First-order statistical features (based on the histogram of each color component) [82], second-order statistical features [83], and granulometric features [84] were applied in this chapter, and also used in [103]. In addition, the novelty in the present chapter was to apply them over the nucleus and the cytoplasm for each component of the *L*a*b** color space to extract new features.

Cytoplasmic profile feature As it was introduced in Chapter 3, this feature was extracted to characterize the cytoplasm profile. It estimates the projections of the cytoplasm using the peripheral region around the segmented cell. Cytoplasmic profile feature is obtained by using threshold segmentation to the green component and counting the pixels of this region.

All features were stored in a numerical data matrix, which was used as the input data for the classification step.

5.2.1.4 Feature analysis

The objective of this step is to determine the most relevant features for each cell type, which was achieved within the context of the information theory feature selection using the so called Conditional Mutual Information (CMI) Criteria [104, 105].

Principal Component Analysis (PCA) was used as a tool to visualize all the features. PCA is a technique that is commonly used to reduce the dimensionality of a big dataset by its transformation into a new set of principal components linearly uncorrelated, searching the causes of variability and sorting the components by their importance [106].

5.2.1.5 Classification

The aim of this step was to obtain the automatic recognition of normal and different types of neoplastic lymphoid cells from PB using the supervised classification method called Linear Discriminant Analysis (LDA) [107].

To assess the efficiency of the proposed method, a 10-fold cross validation technique was performed over the training set of 1500 lymphoid cells. This technique randomly divides the data set into 10 equal size subsets. A single subset is used as the testing data, while the remaining data are used for training. Then, the process is repeated 10 times. Finally, a confusion matrix was obtained to calculate some overall statistical measures.

5.2.2 Methodology validation

In order to validate the methodology as illustrated in Figure 5.1, an independent validation set of 150 new lymphoid cell images was distributed as follows: 34 normal lymphocytes, 19 HCL, 37 MCL, 30 LLC and 30 BPL. These images were acquired, segmented and their features were extracted precisely through the same steps previously used for the training set. Next, the validation of the methodology consisted in performing the LDA technique in two steps: 1) tuning the classifier using the previous training set; and 2) classifying the new validation set in the different lymphoid cell types.

5.3 Results

The developed segmentation algorithm was very effective in separating three different regions of the cell image: cytoplasm, nucleus and peripheral zone around the cell. This procedure

is an essential part of the methodology to ensure the success of the final classification step. Figure 5.2 shows different examples of the segmentation step obtained in different normal and neoplastic lymphoid cell images from PB.

For the classification, the 113 features mentioned in the Section 5.2 and listed in Table 5.1 were used. In addition, for each cell type the ten most ranked features by relevance and interdependence were identified using information theory feature selection [104, 105]. The two most important features were among the geometric ones: (1) cell perimeter and (2) nucleus-cytoplasm ratio. Regarding the remaining eight, they were color and texture features: (3) standard deviation of the granulometric curve of the L component of the nucleus, (4) mean of the b component of the nucleus, (5) standard deviation of the b component of the cytoplasm, (6) entropy (first order statistical) of the a component of the nucleus, (7) correlation of the L component of the nucleus, (8) homogeneity of the a component of the nucleus, (9) kurtosis of the L component of the nucleus, and (10) eccentricity of the cell.

Figure 5.3 shows the first and second principal components derived from the whole set of 113 features obtained by PCA dimension reduction. It can be observed that the different subtypes of lymphoid cells presented a different position according to these principal components. LLC cells showed a pattern more similar to normal lymphocytes but very different with respect to HCL cells and BPL.

5.3.1 Methodology performance evaluation

The 1500 cell images of the training set were classified using LDA with 10 fold cross-validation. Table 5.2 shows these results (confusion matrix) where the rows represent the true diagnosis supplied by physicians and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. Appendix A gives details on the performance classification parameters based on the confusion matrix. Every row was normalized with respect to the total number of cells of its respective type to represent the percentages with

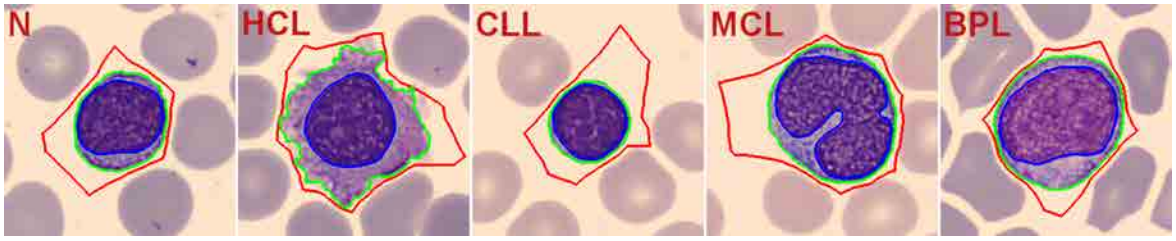


FIGURE 5.2: *Segmentation results obtained in some images of lymphoid cells from peripheral blood: Normal lymphocyte (N), Hairy cell leukemia (HCL), Chronic lymphocytic leukemia (CLL), Mantle Cell Lymphoma (MCL) and B-prolymphocyte (BPL). The outer line delimits the cell from the external area; the middle line corresponds to the cytoplasm margin and the inner line to the nucleus perimeter.*

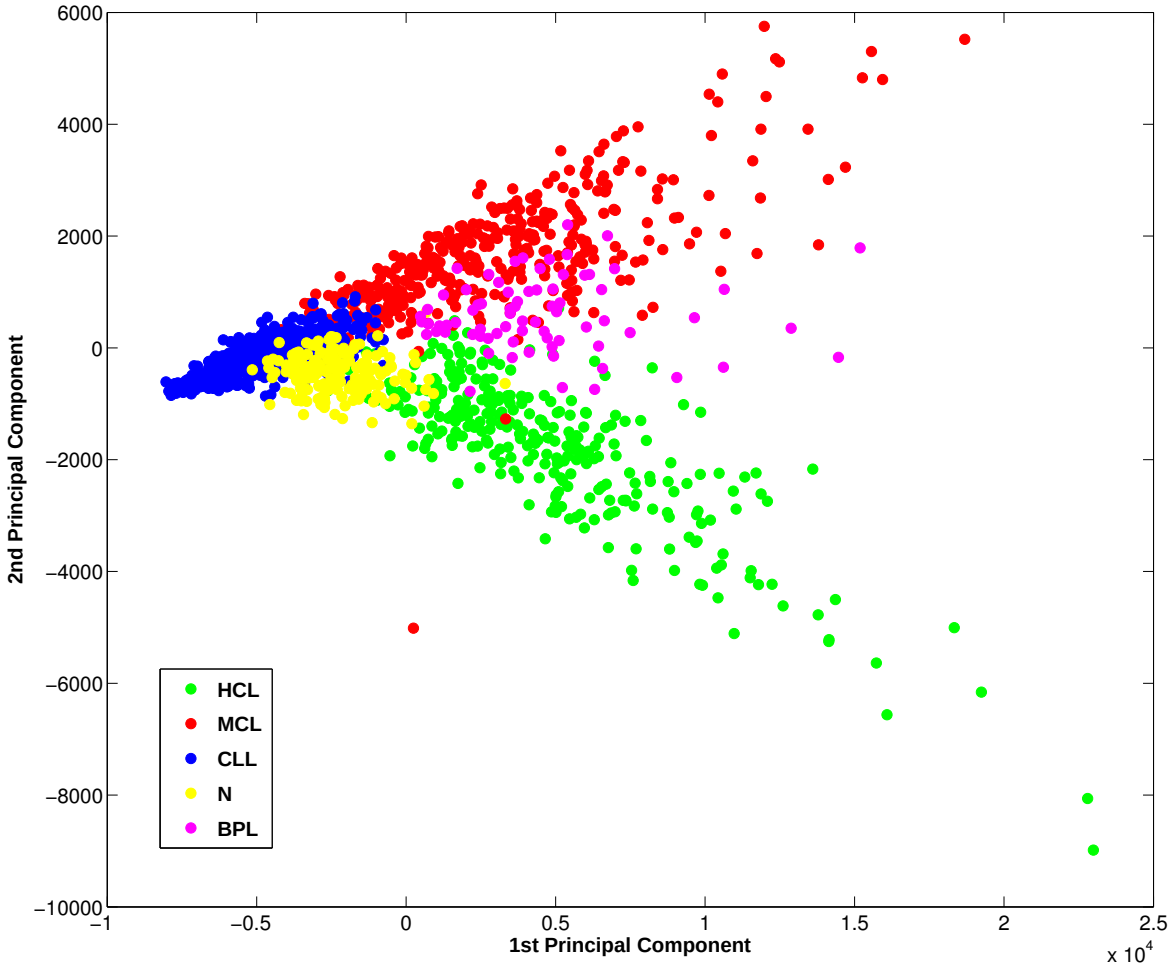


FIGURE 5.3: *First and second principal components of all set of features obtained by Principal Components Analysis (PCA) showing that the different subtypes of lymphoid cells presented a different position taking into account these principal components of the set of features. Chronic lymphocytic leukemia (CLL) cells showed a similar pattern with respect to normal lymphocytes (N) but very different from Hairy cell leukemia (HCL) cells and B-prolymphocytes (BPL).*

TABLE 5.2: *Confusion Matrix of the LDA classification and 10-fold cross-validation for the training set.*

		Predicted*				
		N	**HCL**	**CLL**	**MCL**	**BPL**
True	**N**	**99.45**	0.00	0.00	0.00	0.55
	HCL	1.66	**97.67**	0.00	0.00	0.67
	CLL	0.92	0.00	**98.71**	0.19	0.18
	MCL	1.25	0.00	0.00	**97.50**	1.25
	BPL	4.00	1.33	0.00	0.00	**94.67**

* The rows represent the true diagnosis and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. The values are in percentage. Accuracy = 98.07 % and standard deviation = 0.80.

respect to the true diagnosis. The five-class classification accuracy was 98.07% and also its standard deviation (STD) was calculated to measure the variability between folds, which was 0.80%. Diagonal values are the true positive rates for each cell subtype, showing values of 99.45% for normal lymphoid cells, 97.67% for HCL, 98.71% for CLL, 97.5% for MCL and 94.67% for BPL.

Figure 5.4 shows some statistical measurements of the five-type classification. Most of the precision values are above 99.7% with their respective STDs lower than 1%, except for N and BPL cells, which present values above 88.8% and STDs lower than 9.6%. Most of the sensitivity values are above 97.5% with their STD lower than 2%, while PLB value is 94.7% with a STD of 6.9%. All the specificity values are above 98.6% with their STDs lower than 0.9%.

Figure 5.5 shows some images corresponding to the different lymphoid subtypes classification results using the last fold of the first experiment. Each row corresponds to a different lymphoid cell type (N, HCL, CLL, MCL and BPL).

5.3.2 Validation of the methodology

Table 5.3 shows the classification results (confusion matrix) in which the five-class classification accuracy is 85.33%. True positive rates for each cell subtype are: 94.12% for normal lymphoid cells, 94.74% for HCL, 80% for CLL, 89.19% for MCL and 70% for BPL.

5.4 Discussion

Morphologic examination of PB cells is the first analytical step in the hematological diagnosis and it is a truly useful aid for the indication of further necessary tests. Since neoplastic

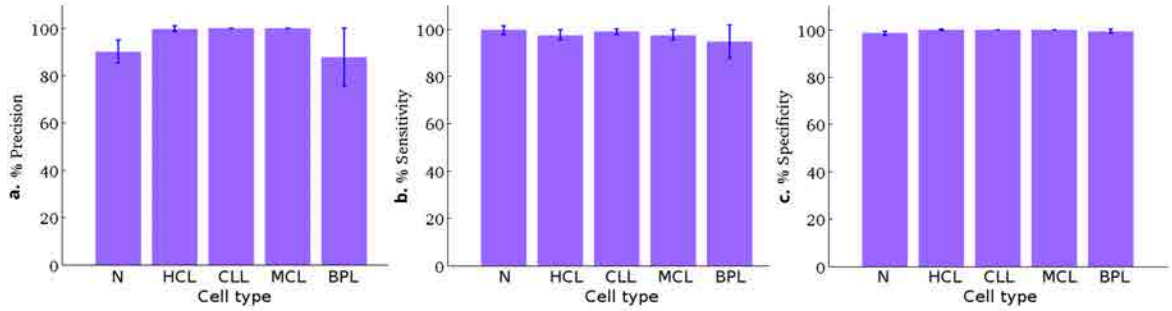


FIGURE 5.4: *Precision (a), sensitivity (b) and specificity (c) of the 10-fold cross validation with Linear Discriminant Analysis (LDA) classification of the training set. Standard deviations are represented by the lines on the top of the bars.*

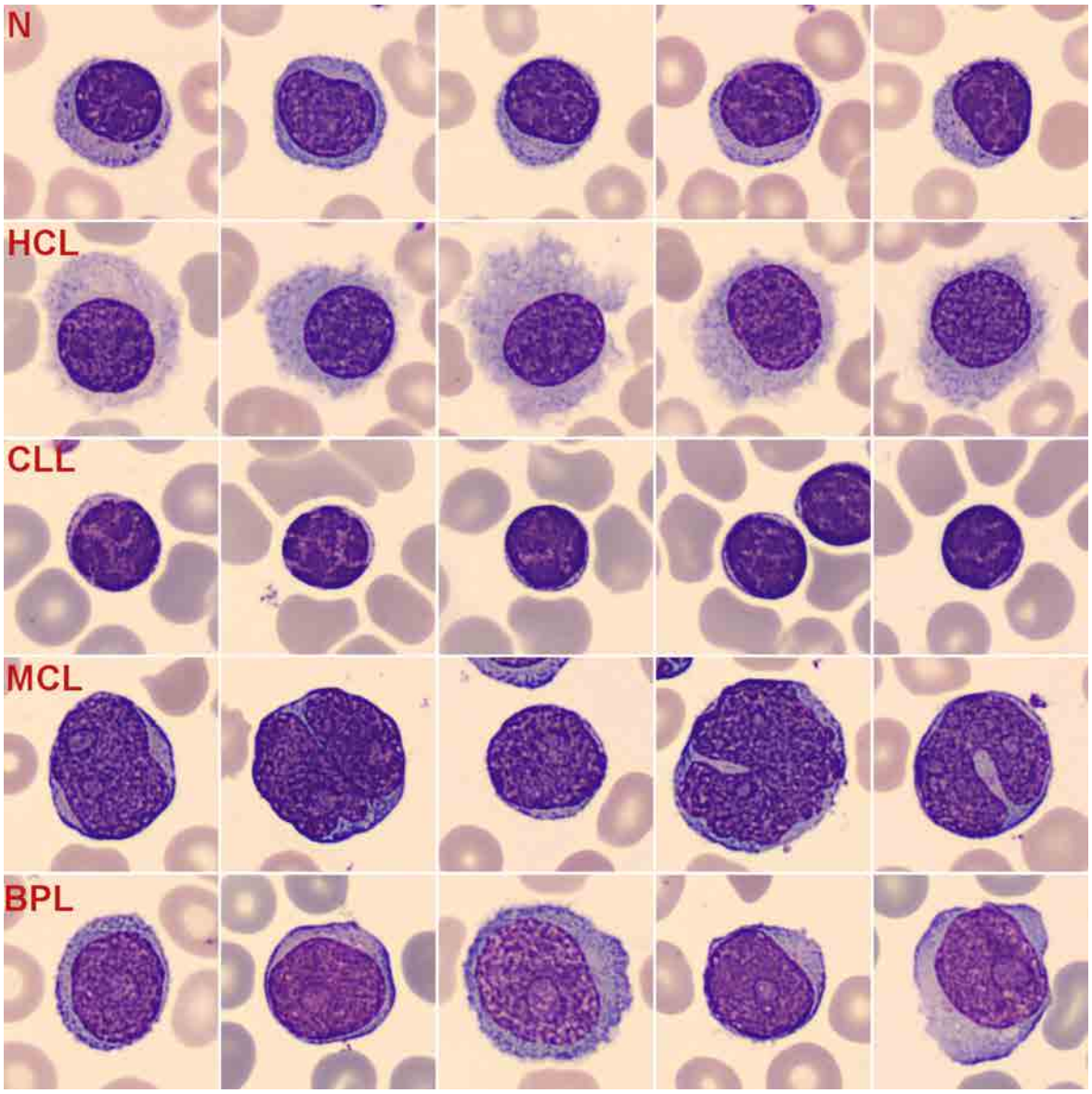


FIGURE 5.5: *Images corresponding to the different lymphoid subtypes classification results using the last fold of our first experiment. Each row corresponds to a different lymphoid cell type: normal lymphocytes (N), hairy cell leukemia (HCL), chronic lymphocytic leukemia cells (CLL), Mantle Cell Lymphoma (MCL) and B-prolymphocytes (BPL).*

lymphoid cells are the most difficult ones to be classified using only morphologic features [80], the major goal of the work in this chapter is to design a methodology combining segmentation, feature extraction and classification algorithms that can be useful as a diagnosis support tool.

Cell morphology is susceptible to variations in slide making and staining process. With the aim of decreasing that variation, the images were obtained by a standard and reproducible method using automatic staining and the Cellavision DM96 analyzer. This equipment can

TABLE 5.3: *Confusion Matrix of the LDA classification for the validation set.*

		Predicted*				
		N	**HCL**	**CLL**	**MCL**	**BPL**
True	**N**	**94.12**	0.00	0.00	5.88	0.00
	HCL	0.00	**94.74**	0.00	0.00	5.26
	CLL	16.67	0.00	**80.00**	3.33	0.00
	MCL	2.70	0.00	0.00	**89.19**	8.11
	BPL	0.00	0.00	0.00	30.00	**70.00**

* The rows represent the true diagnosis and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. The values are in percentage. Accuracy = 85.33 %

scan the slides identifying different types of WBC by artificial neural networks [17, 19], but it is not able to separate the diverse neoplastic lymphoid cells circulating in PB [18].

Normal and four neoplastic (CLL, HCL, MCL and BPL) lymphoid cells were selected due to their characteristic morphology and the high number of them that were collected from daily workload in the clinical laboratory.

With respect to the segmentation method, some authors have only segmented two regions of the cell (nucleus and cytoplasm) [39, 53, 59, 67]. In Chapter 3, watershed transformation on the green color component to segment three regions (nucleus, cytoplasm and the peripheral zone around the cell) was used. In that study, the peripheral region of the cell was especially significant because it allowed us to evaluate the external profile of the cytoplasm. In this chapter, the full extended segmentation procedure of Chapter 4 has been used to segment the three regions of each lymphoid cell, which uses the color information of the image instead of only one gray level. This has resulted to be very useful to extract the appropriate description about the cell.

Regarding the feature extraction step, Scotti calculated only a few geometric features over the cell, nucleus and cytoplasm [39]. Other authors calculated geometric and statistical features of the nucleus [53, 67]. Angulo et al. [59] used extensively the granulometry over the luminance component to compute features for the lymphocyte identification. In this chapter, the texture features have been extended to multiple color components obtaining a total of 113 features, combining geometric, statistical and granulometric features to analyze the texture of the nucleus and the cytoplasm. This extension has permitted to find a better color and texture description of these regions. Nuclear characteristics are important features in morphologic diagnosis. The nuclear staining pattern reflects chromatin organizations and, in addition, CLL cells typically contain clumped chromatin [73]. Not only the cytoplasm profile feature described by the first time in [103] has been calculated, but also its texture, which has been scarcely analyzed in the literature.

Furthermore, it has been demonstrated that the size of the lymphoid cell (perimeter) and the nucleus-cytoplasm ratio are the two main features, which comprise the most information regarding the type of cell. Eight relevant features have also been found, six of them containing information of the nuclear texture, one related to cytoplasm and one relative to the cell shape. All above morphologic features are used in the conventional differential analysis of the cells under the microscope.

Moreover, the whole set of features has been represented using their two principal components. From the PCA dimension reduction, HCL cells were the subtype most separated from the rest of cells, which corresponds to the differences in the morphology of these cells with respect to the others. LLC cells were located very close to the normal lymphocytes and a portion of the MCL cells, which is in accordance with the subtle morphologic differences between them. Figure 5.3 shows that the group of PLB cells overlaps the group of MCL cells, which is according to the morphologic similarities that they share. The separation observed from the PCA plot is an indicator that both segmentation and feature extraction steps have been satisfactorily performed.

Concerning the classification process, Angulo et al. [59] classified the morphologic features of the lymphoid cells in categories using decision trees, but that work was not completed with further studies toward the specific discrimination among different groups of similar diagnosis. In addition, Yang et al. [75] classified five types of neoplastic blood cells but they included precursor lymphoid neoplasms and acute myeloid leukemia blast cells. One of the main practical contributions of the present work is that the proposed methodology succeeded in classifying five types of lymphoid cells (normal and mature B-cell neoplasms).

The accuracy of the methodology has been truly satisfactory, with very low standard deviation. Precision, sensitivity and specificity values have presented excellent values, while they were lower for N and BPL cells because the number of the images for these two types of cells was smaller and therefore less representative. Moreover, the methodology has been satisfactorily validated through the classification of an independent validation set using a group of cells from new patients.

5.5 Conclusion

In summary, the methodology presented in this chapter started with the design of a segmentation procedure to obtain regions of interest in lymphoid cells; from them, descriptive features were extracted and a combination of these characteristics was used in algorithms to classify normal and four different types of neoplastic lymphoid cells.

The addition of more color and texture features and exploring other classification methods to the described methodology will allow classifying other types of neoplastic lymphoid cells, which is important in view of a potential use as a diagnosis support tool in clinical practice. With this aim, the next chapter presents further developments related to feature extraction/selection and classification procedures.

Chapter 6

Feature Extraction and Classification of Neoplastic Lymphoid Cells from Peripheral Blood Digital Images

Based upon: S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, A methodology for automatic recognition of neoplastic lymphoid cell images from peripheral blood, manuscript in preparation.

Abstract

Morphologic analysis of peripheral blood cells is the first analytical step in the hematological diagnosis, which is useful to indicate further laboratory tests. The objective of this chapter was to develop a feature extraction and automatic classification methodology of neoplastic lymphoid cells, since they are the most difficult to be recognized by only qualitative morphologic features. The feature set implemented in this work consists of geometric features and color-texture features (statistical, wavelet and granulometric features) from various color spaces. This full feature set was extracted from 1834 cell images of six different types of lymphoid cells to be used in several classification experiments. The best experiment configuration was obtained for 20 features selected by theoretic feature selection and classified by support vector machines with radial basis function kernel, achieving a lymphoid cell classification accuracy of 97.93%. The feature extraction and classification methodology developed in this chapter will be integrated into the automatic classification system of neoplastic lymphoid cells developed in this thesis.

6.1 Introduction

As it has been pointed out in previous chapters, morphologic analysis over different types of neoplastic lymphoid cells is difficult and requires skills developed through experience [79]. In addition, cytological variables are very difficult to define objectively, resulting in doubts on the recognition of pathologies due to small variations of the morphologic features [80]. Then, the mathematical characterization of these parameters can lead to objective values improving the identification (manual or automatic) of neoplastic lymphoid cells. The characteristics of the cell can be obtained by quantitative measures calculated in the feature extraction step of the digital image processing (DIP) framework. They can represent morphologic qualitative features usually employed by the hematologist [3] or can be just abstract quantitative parameters [35].

Feature extraction/selection and classification techniques have been investigated in different papers (see Chapter 2, Sections 2.6.4 and 2.6.5) with the purpose of automatically

recognize normal white blood cells (WBCs), neoplastic lymphoid cells and leukemia blood cells.

Neoplastic lymphoid cells are the most difficult pathological cells to classify (manually and automatically) using only morphologic features [80]. A problem still open is to achieve a successful automatic discrimination among a relevant number of different neoplastic lymphoid cells in the context of the currently known B cell neoplasms [6]. Up to our knowledge, the literature has reported automatic classification tools able to recognize only a limited number of neoplastic lymphoid cells [21, 59, 70, 74, 75]. In previous Chapters 3 and 5 in this thesis, a method for lymphocyte recognition to allow the automatic classification of normal and up to four types of neoplastic lymphoid cells circulating in PB in mature B cell neoplasms is presented. This chapter extends the classification to include a new group: Follicular Lymphoma (FL). This extension involves significant methodological improvements in the feature extraction/selection and in the classification methodology. The present chapter presents the formulation in detail, including extensive classification experimental results to evaluate the merits of different alternative designs. Chapter 7 will present a strategy for automatic recognition of PB neoplastic lymphoid cells implementing and evaluating the methodology in a scenario where it could be useful in clinical practice.

The remainder of this chapter is organized as follows. Section 6.2 reviews the segmentation method used to separate the regions of interest (ROI). Section 6.3 explains the geometric features, which include the geometric features and the Elliptical Fourier Descriptors (EFD). Section 6.4 describes the color-texture features, introducing: the first and second order statistical features, the wavelet features and the granulometric features; all of them applied on different color spaces. Section 6.5 introduces the normalization methods applied over the features. Section 6.6 reviews the information theoretic feature selection used to reduce the feature set, considering the redundancy and the relevance. Support Vector Machines (SVM) classifiers with several kernels are briefly explained in Section 6.7. In Section 6.8 various classification experiments are shown to analyze the features and to find the best possible configuration of a lymphoid cell classification system. Finally, conclusions are presented in Section 6.9.

6.2 Regions of interest obtained by color segmentation

Three ROI were obtained for each individual cell image using the color clustering segmentation methodology developed in Chapter 4: the entire cell, the nucleus and the peripheral zone around the cell. These three regions are originally binary masks, from which geometric features can be calculated directly or they can define the corresponding color regions to extract

color-texture features. The cytoplasm region was obtained by the difference between the entire cell and the nucleus regions.

6.3 Geometric features

This kind of features aims to measure basic morphologic features such as size and shape. The geometric features extracted in this work are described below.

6.3.1 Geometric-size features

These features measure quantities related to the size of the cell, the nucleus and the cytoplasm. They also include a feature related to the relative orientation and a simple feature about the shape.

Area The area was calculated by counting the number of pixels in the region.

Diameter It was calculated as the equivalent diameter of a circle with the same area as the ROI, i.e. $diameter = \sqrt{4Area/\pi}$.

Perimeter The perimeter was obtained as the border length of the ROI, i.e the sum of the distances between adjacent pixels on the boundary.

Conic eccentricity It was measured as the conic eccentricity of an equivalent ellipse that describes the ROI. An eccentricity equal to 1 represents a straight line, whereas if it is 0 represents a circle.

Compactness It was calculated as the relation between the squared perimeter and the area, $Compactness = perimeter^2/area$.

Nucleus eccentricity respect to the cytoplasm This feature was calculated as the distance between the cell center and the nucleus center.

Nucleus-cytoplasm ratio The nucleus-cytoplasm ratio was obtained as the relation between the nucleus and cytoplasm areas, using the following expression:

$$NCratio = \frac{area_{nucleus}}{area_{cell} - area_{nucleus}}$$

Cytoplasmic profile feature This feature estimates the projections of the cytoplasm using the peripheral region around the segmented cell. It was obtained by using threshold segmentation to the green component and counting the pixels of this region (See Chapter 3 for more details).

Area, diameter, perimeter, conic eccentricity and compactness were calculated for both the nucleus and the cell regions.

6.3.2 Elliptical Fourier descriptors

Fourier descriptors represent a closed contour using a limited number of coefficients, which are derived from the Fourier series of the border. However, these descriptors are not invariant in translation, scale and rotation. The elliptical Fourier descriptors, proposed by Kuhl and Giardina [108], are a contour representation taking the Fourier series individually over each coordinate of each point $[x(t), y(t)]$ seen as a parametric function, where t is the time required to reach the point. If the "velocity" is constant, this parameter can be considered as the arc length of the chain code or any piecewise linear representation of a contour. Then, the EFD for the nth harmonic of a contour with K points are:

$$\begin{aligned}
a_n &= \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\triangle x_p}{\triangle t_p}\left[\cos\frac{2n\pi t_p}{T}-\cos\frac{2n\pi t_{p-1}}{T}\right]\\
b_n &= \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\triangle x_p}{\triangle t_p}\left[\sin\frac{2n\pi t_p}{T}-\sin\frac{2n\pi t_{p-1}}{T}\right]\\
c_n &= \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\triangle y_p}{\triangle t_p}\left[\cos\frac{2n\pi t_p}{T}-\cos\frac{2n\pi t_{p-1}}{T}\right]\\
d_n &= \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\triangle y_p}{\triangle t_p}\left[\sin\frac{2n\pi t_p}{T}-\sin\frac{2n\pi t_{p-1}}{T}\right]
\end{aligned} \tag{6.1}$$

where $t_p = \sum_{i=1}^{p}\triangle t_i$, $\triangle t_p = \sqrt{\triangle x_p^2 + \triangle y_p^2}$, $\triangle x_p = x_p - x_{p-1}$, $\triangle y_p = y_p - y_{p-1}$, and $T = t_K$ is the period trough all points of the closed contour (or the total arc length of the closed contour).

The EFD geometric interpretation as an addition of rotating phasors, which are linked by ellipses corresponding to each harmonic, allows to make the descriptors to be invariant in terms of: (1) the starting (and arbitrary) point of the contour, (2) spatial rotations and, (3) size variations (translation invariance is obtained eliminating the DC[2] components of the Fourier

[2]DC component is really the first component, at zero frequency.

series). This normalization procedure is based on whether the locus of the first harmonic phasor is elliptical or circular.

In this work, 8 harmonics for the nucleus contour and 8 harmonics for the cell contour were calculated, obtaining 64 EFD. The elliptical Fourier descriptors were not made size-invariant in this work since the size is important in the cell morphologic analysis.

6.4 Color-texture features

These features allow to extract information about the color and texture of the image. The texture was characterized by using statistical features, wavelet statistical features (which included first and second order statistical features) and granulometric features. The color information was described by applying the texture features on different color components of several color spaces. Each of these features was applied to the nucleus and the cytoplasm.

6.4.1 First order statistical features

These features are based on the histogram of a grayscale digital image. The histogram is a discrete function that shows the number of pixels *H(i)* on the image having the pixel intensity value i (frequencies) [35]. Since the digital image is a discretized matrix, its pixels can take values between the range $[0, L-1]$ (L is not necessarily the number of values that can take the pixels, but the number of containers under which the frequencies of the histogram are calculated). The histogram can also be interpreted as a probability density of occurrence of the intensity values if the frequencies are divided by the number of pixels in the image (or the number of containers):

$$p(i) = H(i)/L, \qquad i = 0, 1, ..., L-1 \tag{6.2}$$

From the simple statistical information about the image supplied by the histogram, several first order statistical features can be obtained [82]. Table 6.1 shows the six features used in this work. The mean is the average of the whole intensity values of the image (or the ROI), the standard deviation measures the dispersion of the intensity values around the mean, the skewness describes the symmetry of the histogram around the mean; the kurtosis measures the flatness, the energy the uniformity, and the entropy[3] describes the variability of the histogram.

[3]There are two entropies, one for the first order and another for the second order statistical features.

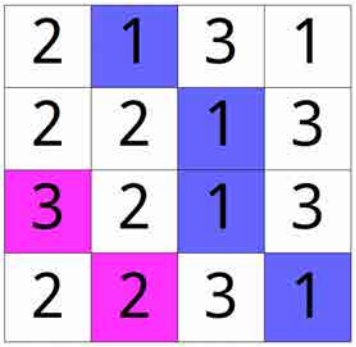
Feature	Expression
Mean	$\mu = \sum_{i=0}^{L-1} ip(i)$
Standard deviation	$\sigma = \sum_{i=0}^{L-1} (i-\mu)^2 p(i)$
Skewness	$\mu_3 = \sigma^{-3} \sum_{i=0}^{L-1} (i-\mu)^3 p(i)$
Kurtosis	$\mu_4 = \sigma^{-4} \sum_{i=0}^{L-1} (i-\mu)^4 p(i) - 3$
Energy (uniformity)	$E_1 = \sum_{i=0}^{L-1} [p(i)]^2$
Entropy 1	$H_1 = \sum_{i=0}^{L-1} p(i) \log_2 (p(i))$

TABLE 6.1: *First order statistical features*

6.4.2 Second order statistical features

Haralick et al. [83] defines the second order statistical features, which provide more information about the texture of the regions to be analyzed. These parameters are defined based on the gray level co-occurrence matrix (GLCM) of a digital image, which represents the joint probability $P(i, j)$ that a *pair of pixels* have intensity values of i and j, respectively, at a distance d in a particular direction θ. This probability can be calculated as the frequency count of occurrences (second order histogram) divided by the total number of neighbouring pixels [82]. Thus, the co-occurrence matrix considers not only the information about the intensity values, but also the position of the pixels with similar intensities. Figure 6.1 shows an example for an simple image of size 4×4 pixels with three levels of intensity (Figure 6.1a). Its co-occurrence matrix, for $d = 1$ and $\theta = 135°$ (northwest), has a size of 3×3 because the intensity levels in the image are *1*, *2* and *3* (Figure 6.1b). The element in *(1, 1)* position of the GLCM indicates that the level *1* is adjacent to the level *1* in the northwest direction 2 times. The element in *(2, 3)* relates that the level *2* is adjacent to the level *3* in the same direction, appearing only once.

Considering that the grayscale image has N levels of gray and its pixels can take values between the range of $[1, N]$ (it can be obtained by normalization or by reducing the number of intensities and quantifying a few levels), the second order statistical features are defined as follows [109, 110].



(a) *A simple image*

2	1	0
0	2	1
1	1	1

(b) *Co-occurrence matrix*

FIGURE 6.1: *Example of a gray level co-occurrence matrix with $d = 1$ and $\theta = 135°$ for a simple image*

Angular second moment, energy or uniformity

$$s_1 = \sum_{i=1}^{N} \sum_{j=1}^{N} [P(i,j)]^2 \tag{6.3}$$

Contrast or inertia

The contrast, inertia or texture contrast can be defined in two ways, completely equivalent:

Contrast calculated directly This is the most simple way to calculate the texture contrast. It uses the level intensities and the co-occurrence matrix.

$$s_2 = \sum_{i=1}^{N} \sum_{j=1}^{N} (i-j)^2 P(i,j) \tag{6.4}$$

Contrast calculated by pair-differences Haralick et al. [83] defines the difference or the pair-differences:

$$P_{x-y}(k) = \sum_{\substack{i \\ |i-j|=k}} \sum_{j} P(i,j) \tag{6.5}$$

where k=$\{0, 1, ..., N-1.\}$ Then, the texture contrast can also be defined:

$$s_2 = \sum_{k=0}^{N-1} P_{x-y}(k) \tag{6.6}$$

Alternative contrast Another way to estimate the texture contrast is using the following formula:

$$s_2 = \sum_{i=1}^{N}\sum_{j=1}^{N} |i-j|P(i,j) \tag{6.7}$$

Marginal probabilities

The marginal probabilities are defined as follows:

$$P_x(i) = \sum_{j=1}^{N} P(i,j) \qquad P_y(j) = \sum_{i=1}^{N} P(i,j) \tag{6.8}$$

The *means and standard deviations* of the marginal probabilities are:

$$\mu_x = \sum_{i=1}^{N} iP_x(i) = \sum_{i=1}^{N}\sum_{j=1}^{N} iP(i,j) \qquad \mu_y = \sum_{j=1}^{N} jP_y(j) = \sum_{i=1}^{N}\sum_{j=1}^{N} jP(i,j) \tag{6.9}$$

$$\sigma_x = \sqrt{\sum_{i=1}^{N}(i-\mu_x)^2 P_x(i)} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}(i-\mu_x)^2 P(i,j)} \tag{6.10}$$

$$\sigma_y = \sqrt{\sum_{j=1}^{N}(j-\mu_y)^2 P_y(j)} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}(j-\mu_y)^2 P(i,j)} \tag{6.11}$$

P_x and P_y are types of histogram of an image and μ_x and μ_y are their respective means.

Correlation or texture correlation

The original paper [83] defines the following formula:

$$s_3 = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{ijP(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{6.12}$$

More recent authors have defined the correlation with another representation [35, 109]:

$$s_3 = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{(i-\mu_x)(j-\mu_y)P(i,j)}{\sigma_x\sigma_y} \tag{6.13}$$

Variance

$$s_4 = \sigma_x^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(i-\mu_x)^2 P(i,j) \tag{6.14}$$

Homogeneity or inverse difference moment

$$s_5 = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{P(i,j)}{1+(i-j)^2} \tag{6.15}$$

The *texture homogeneity* is another version of the above term [35], producing similar results:

$$s_5 = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{P(i,j)}{1+|i-j|} \tag{6.16}$$

Sum average or mean of pair - sums

The pair-sums is defined as

$$P_{x+y}(k) = \sum_{\substack{i}}\sum_{\substack{j\\(i+j)=k}} P(i,j) \tag{6.17}$$

where $k = \{2,3,...,2N\}$. Then, the sum average or mean of pair-sums is

$$s_6 = \sum_{k=2}^{2N} kP_{x+y}(k) \tag{6.18}$$

Sum variance or variance of pair-sums

$$s_7 = \sum_{k=2}^{2N}(k-f_6)^2 P_{x+y}(k) \tag{6.19}$$

Sum entropy or entropy of pair-sums

$$s_8 = -\sum_{k=2}^{2N} P_{x+y}(k)\log_2\left[P_{x+y}(k)\right] \tag{6.20}$$

Entropy 2

$$s_9 = \sum_{i=1}^{N}\sum_{j=1}^{N} P(i,j)\log_2\left(P(i,j)\right) \tag{6.21}$$

The logarithm for entropy usually has base 2, but sometimes is used with base 10 or also with natural base.

Difference variance or variance of pair-differences

$$s_{10} = \sum_{k=0}^{N-1} (k - \mu_{x-y})^2 P_{x-y}(k) \tag{6.22}$$

where the mean of pair of differences is: $\mu_{x-y} = \sum_{k=0}^{N-1} k P_{x-y}(k)$.

Difference entropy or entropy of pair-differences

$$s_{11} = -\sum_{k=0}^{N-1} P_{x-y}(k) \log_2 [P_{x-y}(k)] \tag{6.23}$$

Information measure of correlation 1

$$s_{12} = \frac{HXY - HXY_1}{\max(HX, HY)} \tag{6.24}$$

where,

- $HXY = s_9 =$ Entropy 2 (Equation 6.21)
- $HXY_1 = -\sum_{i=1}^{N} \sum_{j=1}^{N} P(i,j) \log [P_x(i) P_y(j)]$
- $HX = -\sum_{i=1}^{N} P_x(i) \log (P_x(i))$
- $HY = -\sum_{j=1}^{N} P_y(j) \log (P_y(j))$

Information measure of correlation 2

$$s_{13} = \sqrt{\left|1 - e^{-2(HXY_2 - HXY)}\right|} \tag{6.25}$$

where $HXY_2 = -\sum_{i=1}^{N} \sum_{j=1}^{N} P_x(i) P_y(j) \log [P_x(i) P_y(j)]$.

Maximal correlation coefficient

$$s_{14} = (Second\ largest\ eigenvalue\ of\ Q)^{1/2} \tag{6.26}$$

where

$$Q(i,j) = \sum_{k=1}^{N} \frac{P(i,k) P(k,j)}{P_x(i) P_y(j)} \tag{6.27}$$

Maximum probability

$$s_{15} = \max_{i,j}(P(i,j)) \tag{6.28}$$

There are two extra features that were not defined in the original paper but they are commonly used to describe texture [111]:

Cluster shade

$$s_{16} = \sum_{i=1}^{N}\sum_{j=1}^{N}(i-\mu_x+j-\mu_y)^3 P(i,j) \tag{6.29}$$

Cluster prominence

$$s_{17} = \sum_{i=1}^{N}\sum_{j=1}^{N}(i-\mu_x+j-\mu_y)^4 P(i,j) \tag{6.30}$$

In this thesis the GLCM was calculated with $d = 1$. Every feature was computed for $\theta = \{0°, 45°, 90°, 135°\}$ and then the average of the four values was taken as the feature value, to give rotational invariance.

6.4.3 Wavelet statistical features

Wavelet transform is widely used in signal processing due to its ability to decompose the signal in trends (lower resolution) and fluctuations (local changes). The analysis of the fluctuations of a digital image allows to characterize the texture, separating the details. The dyadic continuous wavelet transform of a 1D signal $f(x)$ is defined as:

$$\langle f(x), \psi_{a,b}(x)\rangle = \int_{-\infty}^{\infty} f(x)\frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)dx \tag{6.31}$$

where $a = 2^{-s}$ and $\psi_{a,b}$ is the mother wavelet translated by b and dilated by a.

Mallat [112] develops a wavelet decomposition and reconstruction algorithm using quadrature mirror filters (QMF) to implement the discrete wavelet transform (DWT). This algorithm computes the wavelet coefficients using a lowpass wavelet filter and a highpass wavelet filter, implementing the DWT without directly using wavelet functions [82]. Figure 6.2 shows the QMF scheme for a first-level decomposition and reconstruction of a signal f. In the decomposition stage the filters h (highpass) and g (lowpass) are used with downsampling by 2, while in the reconstruction stage the filters h' and g' are used with upsampling by 2. The approximation coefficients a_1 and the detail coefficients d_1 are about half the size of the

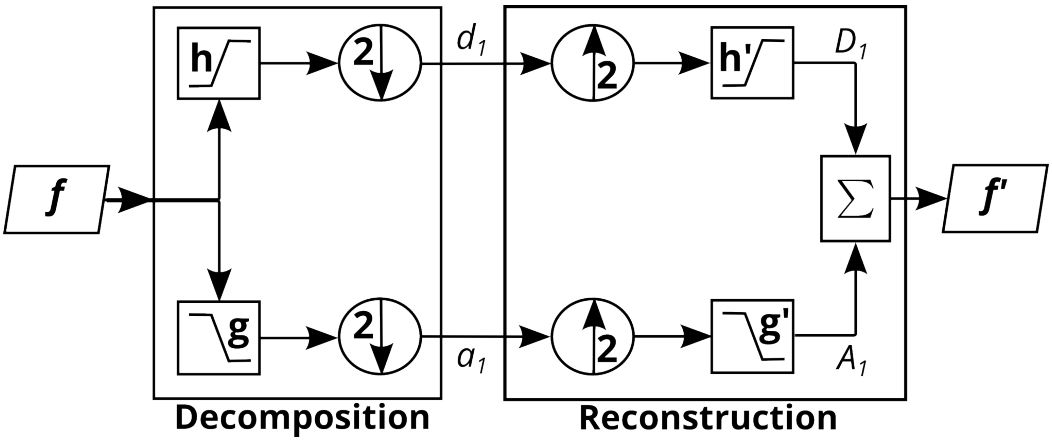


FIGURE 6.2: *First level wavelet decomposition and reconstruction for a 1D signal f using QMF.*

original signal f. Otherwise, the lower frequency sub-band signal A_1 and the high frequency sub-band signal D_1 (i.e. the reconstructions of the coefficients) are about the same size of the original signal f. The second level wavelet decomposition consists in applying the first level decomposition over the first approximation coefficients a_1, resulting in the second level detail and approximation coefficients d_2 and a_2, respectively. Figure 6.3 illustrates the complete second level decomposition and reconstruction methodology by the QMF. At the end of this method, the sub-bands A_2 (second approximation), D_2 (second detail) and D_1 (first detail) reconstruct the signal. It is important to remark that the coefficients d_2 and a_2 are about a quarter the size of the original signal, while the reconstruction of the coefficients D_2 and A_2 (also D_1 and A_1) are about the same size of the original signal.

The two-dimensional DWT over a grayscale digital image is implemented by the decomposition using the QMF method separately along its rows and columns [112]. Figure 6.4a illustrates the QMF scheme for the first level decomposition of a two dimensional image,

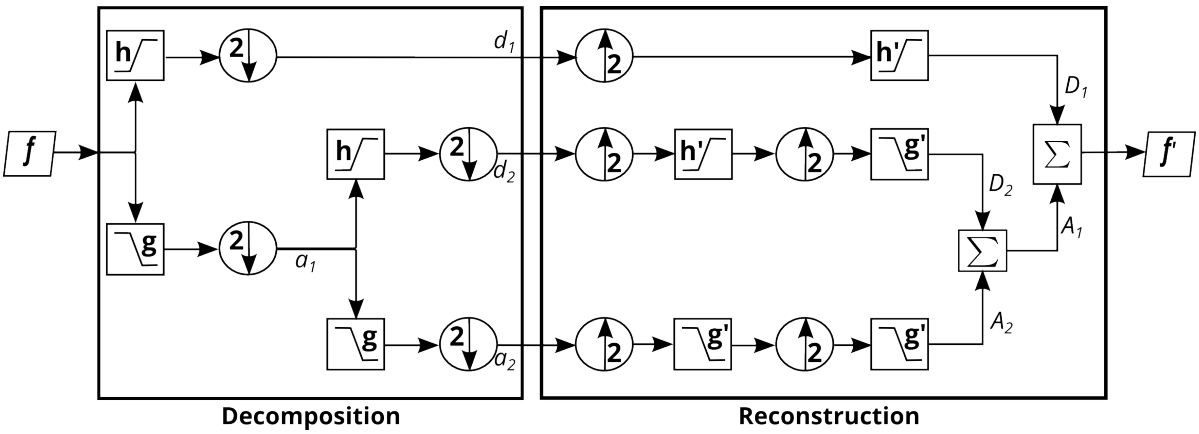
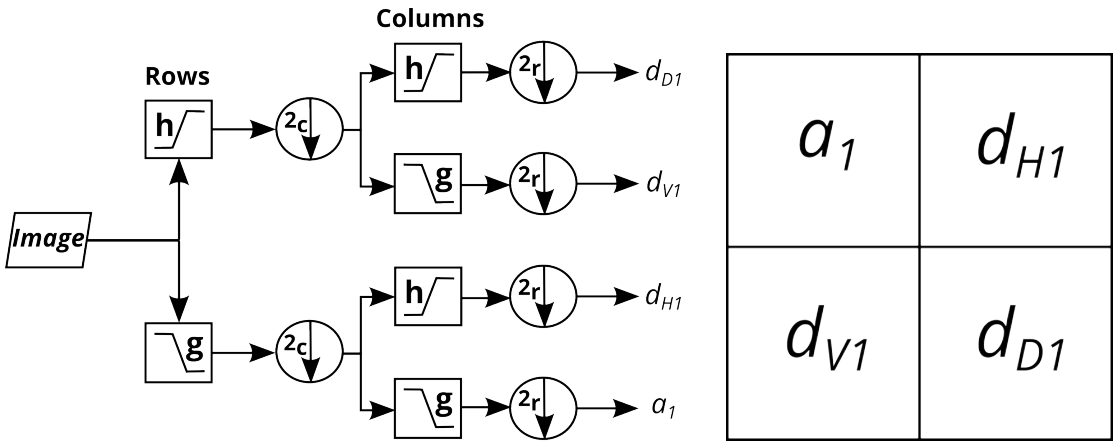


FIGURE 6.3: *Second level wavelet decomposition and reconstruction for a 1D signal f using QMF.*



(a) *QMF image scheme* **(b)** *Square representation*

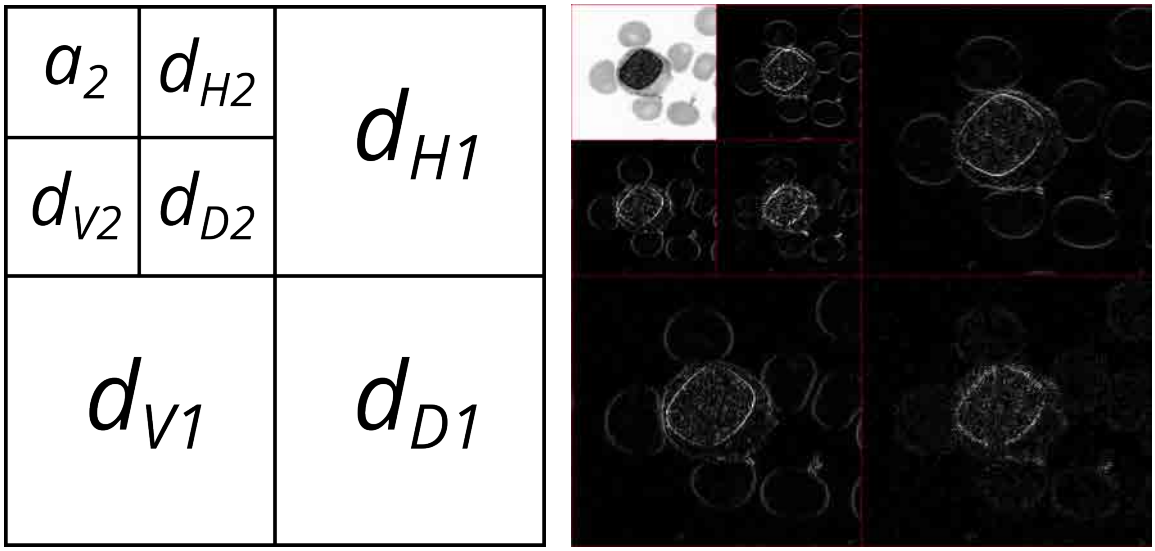
FIGURE 6.4: *First level wavelet decomposition for a two dimensional digital image.*

which consists basically of two stages: the filtering trough the rows followed by downsampling (by 2) of the columns, and the subsequent filtering through the columns followed by downsampling of the rows. Then, four wavelet coefficients are produced: the approximation coefficient a_1 (lowpass→lowpass), the horizontal detail coefficient d_{H1} (lowpass→highpass), the vertical detail coefficient d_{V1} (highpass→lowpass) and the diagonal detail coefficient d_{D1} (highpass→highpass). All the coefficients are about half the size of the original image as shown in the representation in Figure 6.4b. The second level wavelet decomposition of a two dimensional image is obtained in the same way as for the one dimensional case, i.e. by reapplying the QMF method of Figure 6.4a over the approximation coefficient, resulting in 4 new coefficients. Figure 6.5a shows the square representation of the second-level wavelet decomposition of an grayscale image and Figure 6.5b illustrates an example of this decomposition for a grayscale image of a reactive lymphocyte.

In this thesis, the reconstruction of detail coefficients (high frequency sub-band images) of the second level wavelet decomposition and the two-dimensional image were used to calculate some texture features [113–115]. That is, the first and second statistical features were extracted in the ROI over each color component and its corresponding six details sub-band of the wavelet decomposition.

6.4.4 Granulometric features

Granulometry measures the particle size distribution of an image by mathematical morphology operations, such as dilation, erosion, opening and closing. Let $f(x,y) \in R$ be a grayscale image and $b(x,y) \in R$ a structuring element, for $(x,y) \in Z^2$, the following operations are defined as:



(a) *Square representation* **(b)** *Decomposition of a reactive lymphocyte*

FIGURE 6.5: *Representation and example of the second level wavelet decomposition of a two dimensional image.*

Dilation

$$\delta(f) = (f \oplus b)(u,v) = \max_{(x,y)\in D_b} \{f(u-x, v-u) + b(x,y)\} \tag{6.32}$$

Erosion

$$\varepsilon(f) = (f \ominus b)(u,v) = \min_{(x,y)\in D_b} \{f(u+x, v+y) - b(x,y)\} \tag{6.33}$$

Opening

$$\gamma(f) = f \circ b = (f \ominus b) \oplus b \tag{6.34}$$

Closing

$$\varphi(f) = f \bullet b = (f \oplus b) \ominus b \tag{6.35}$$

where D_b is the domain of the structuring element b [35]. Dilation is equivalent to a maximum filter (for a flat structuring element), reducing or eliminating dark details depending on how their values and shapes are related to the structuring element. Erosion is equivalent to a minimum filter (flat structuring element), reducing the bright details in the image that are smaller in area than the structuring element. Opening removes small bright details in the image while leaving the rest of the image relatively unchanged including the larger bright details. Closing suppress small dark regions, whilst leaving relatively unchanged the global gray levels and large dark details [116].

Formally a granulometry is defined as a family of transformations $\Psi = (\psi_\lambda), \lambda \geq 0$ that satisfies the following conditions [117, 118]:

1. Identity transformation: $\psi_0(f) = f$.
2. ψ_λ is increasing.
3. ψ_λ is anti-extensive: $\psi_\lambda(f) \leq f, \forall \lambda \geq 0$.
4. Absorption: $\psi_\lambda \psi_\mu = \psi_\mu \psi_\lambda = \psi_{\max(\lambda,\mu)}, \forall \lambda \geq 0, \forall \mu \geq 0$.

Then, a granulometry of openings (or simply granulometry) is defined by: $\Gamma = (\gamma_\lambda); \gamma_\lambda = f \circ \lambda b, \lambda \geq 0$, while a granulometry of closings (or anti-granulometry) is defined by the family: $\Phi = (\varphi_\lambda); \varphi_\lambda = f \bullet \lambda b, \lambda \geq 0$. Thus, if λb are structuring elements with increasing size, the four above properties are satisfied. In this thesis a disk-shaped structuring element with an increasing radius was always used.

The granulometry analysis is performed through a size measurement $m(\gamma_\lambda(f))$ for each step, which is the area (intensity sum of the pixels) of each opening. Then, the granulometric curve (pattern spectrum) for a family of openings $\Gamma = (\gamma_n); n \geq 0$ is defined by the following transformation:

$$PS_\Gamma(f,n) = \frac{m(\gamma_n(f)) - m(\gamma_{n+1}(f))}{m(f)} \tag{6.36}$$

Therefore, for each size (*nth* iteration) corresponds a particular measure of the bright structures in the image obtained by the brightness loss between two successive openings. PS_Γ is a size-histogram in which a high value at a specific size indicates the presence of many bright structures with similar size in the image. In a similar way, the anti-granulometric curve for a family of closings $\Psi = (\psi_n), n \geq 0$ is defined as

$$PS_\Phi(f,-n) = \frac{m(\varphi_n(f)) - m(\varphi_{n-1}(f))}{m(f)} \tag{6.37}$$

which is used to describe dark structures in the image.

The joint between the granulometric curve for openings associated with "positive" sizes and the anti-granulometric curve for closings associated with "negative" sizes results in the completed *granulometric curve* [118], which is defined as follows:

$$\{-n, 0, n\} \rightarrow PS(f,n) = \{PS_\Phi(f,-n), 0, PS_\Gamma(f,n)\}; n \geq 1 \tag{6.38}$$

Although dilation and erosion do not satisfy the absorption condition, it is quite useful to define a pseudo-granulometry for a family of erosions $\varepsilon = (\varepsilon_n) = f \ominus nb$ and an anti-pseudo-

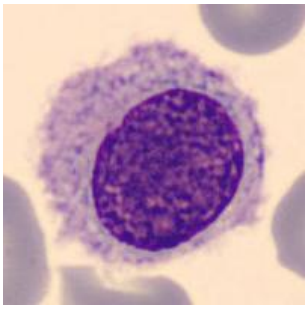
granulometry for a family of dilations $\Delta = (\delta_n) = f \oplus nb$ with $n \geq 0$, such that the pseudo-pattern spectra for erosion and dilation are, respectively:

$$\begin{aligned} PPS_{\varepsilon}(f,n) &= \frac{m(\varepsilon_n(f)) - m(\varepsilon_{n+1}(f))}{m(f)} \\ PPS_{\Delta}(f,-n) &= \frac{m(\delta_n(f)) - m(\delta_{n-1}(f))}{m(f)} \end{aligned} \tag{6.39}$$

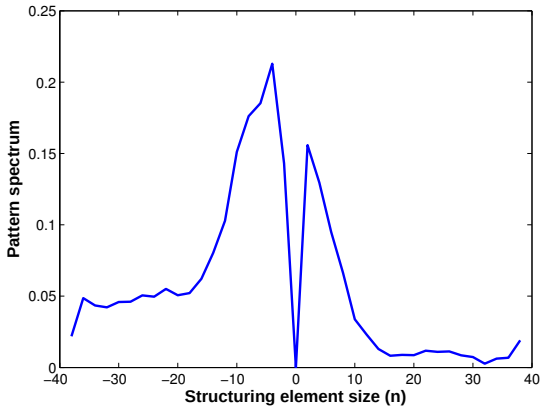
In the same way as it was shown above, the complete *pseudo-granulometric curve* is defined as follows [118]:

$$\{-n,0,n\} \rightarrow PPS(f,n) = \{PPS_{\Delta}(f,-n),0,PS_{\varepsilon}(f,n)\}; n \geq 1 \tag{6.40}$$

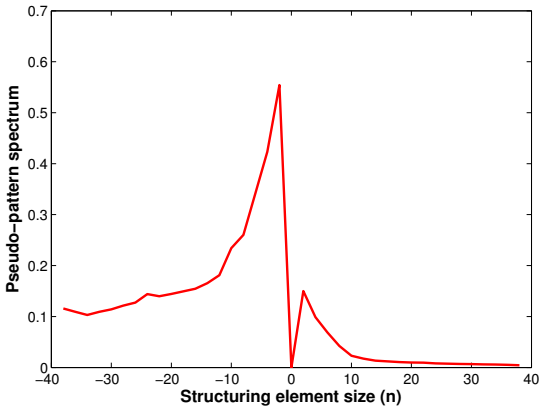
Figure 6.6 shows an example of the granulometric curve and the pseudo-granulometric curve for a lymphoid cell of a patient with Hairy Cell Leukemia. The left part of the curves (negative sizes) represents the size density of dark granules, while the right part of the curves



(a) *Hairy cell*



(b) *Granulometric curve*



(c) *Pseudo-granulometric curve*

FIGURE 6.6: *Granulometric and pseudo-granulometric curves for the nucleus of a Hairy cell (leukemia).*

represents the size density of bright granules.

In this thesis, a flat disk-shaped structuring element with its specified (and increasing) radius (n) was always used to calculate the granulometric curves.

The mean, standard deviation, skewness and kurtosis (described in Section 6.4.1) were calculated over the granulometric and pseudo-granulometric curves. These eight parameters constitute the *granulometric features* used in this work.

6.4.5 Summary of the color-texture features

Figure 6.7 shows the scheme of the application of the color-texture features on four color components. The 23 statistical features consist of 6 first order statistical features and 17 second order statistical features. The 138 wavelet statistical features are constituted by the application of the statistical features over the six detail sub-bands images of the second wavelet decomposition. And the 8 granulometric features are calculated over the granulometric (4) and pseudo-granulometric (4) curves. Then, a total of 169 features are extracted for the two ROIs: the nucleus and the cytoplasm. Thus, 338 color-texture features are extracted for each color component.

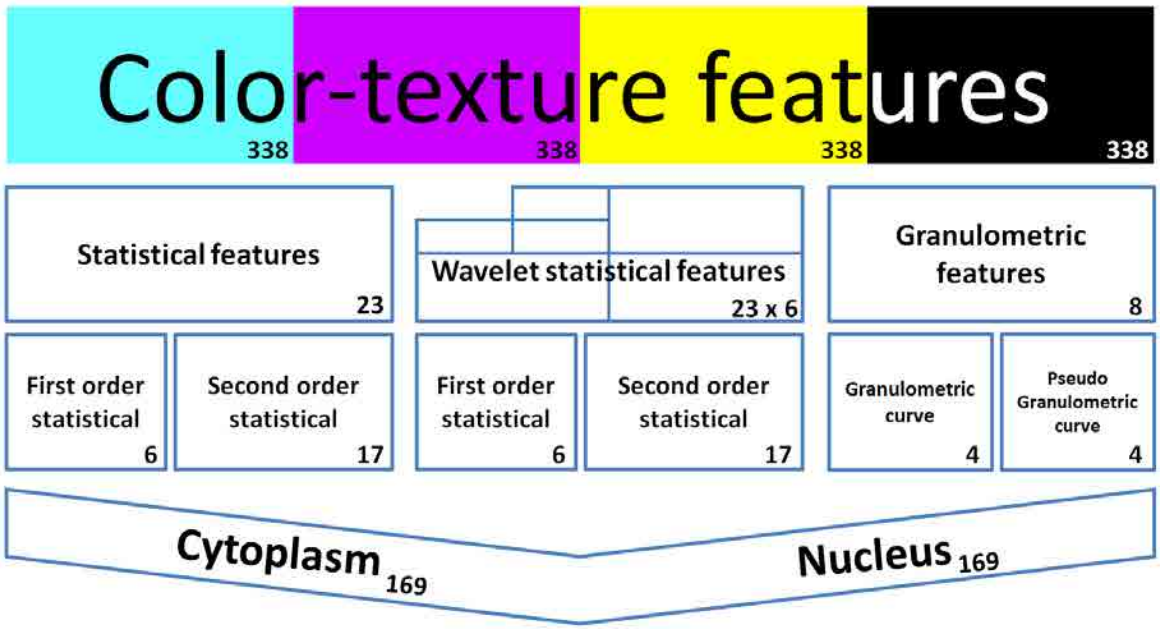


FIGURE 6.7: *Scheme of the application of the color-texture features for various color components.*

6.5 Feature normalization

In any process which requires exploring data, it is important to perform a preprocessing of them. One of the most basic methods is the normalization, which is especially useful when the features have different units and scales. There are two widely known procedures to make a linear scaling of the data between the interval [*0, 1*] (it is also common [*-1, 1*]). Let x be a particular feature value, and let $x_{min}$and x_{max} be the respective minimum and maximum values of that feature. Then the *max-min* scaling is defined as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6.41}$$

where x' is the normalized feature value. The other procedure is the standardization, in which all the feature values are centered such that the mean is *0* and standard deviation is equal to *1*. This can be done by the following definition:

$$x' = \frac{x - \mu}{\sigma} \tag{6.42}$$

where μ is the mean and σ the standard deviation of the feature values.

If there are atypical feature values (outliers), the above normalization procedures scale linearly all the data, considering these extreme values. Furthermore, the standardization method assumes that the data have a Gaussian distribution, which is not necessarily true. A solution to these problems is to use a nonlinear normalization as the *softmax* scaling. This procedure presents a linear part (which can be controlled by a parameter) and the space allocated to the outliers can be controlled by the uncertainty level of the sample. The expression for the *softmax* normalization is:

$$x' = \frac{1}{1 + \exp\left(\frac{x-\mu}{r\sigma}\right)} \tag{6.43}$$

where r is the parameter that controls the size of the linear response.

Throughout this thesis, it has been used mainly the *softmax* scaling (for the classification process) and the standardization (for the feature selection process).

6.6 Information theoretic feature selection

When many features are extracted, the complexity of the problem description becomes high, making difficult to build a good classification system. Feature selection defines a topic commonly used in data mining to select the more significant features. Feature selection allows

to improve the classification performance, makes faster and more profitable classifiers and provides a better understanding of the data processing [119]. Usually, there are two types of feature selection techniques: those that depend on the classification and those that are independent of the classification. The former evaluates the feature subsets through the accuracy of a particular classifier (wrapper method), or it takes advantage of the framework of a particular classifier (embedded method). The second considers filter methods, which select the features by defining a scoring with a specific criterion (estimated classification accuracy), being faster and easier to understand than the classifier-dependent methods [105]. The feature selection procedure used in this thesis lies within this second group.

Roughly, the filter methods select the features by an usefulness score that maximizes the relevance and minimizes the redundancy of the features, evaluating their utility of inclusion in a specific set. This filtering process can be performed through *forward selection step* (a feature is added according to the maximum score) or *backward elimination step* (a feature is removed according to the minimum score). Brown et al. [105] presents a single unifying framework for feature selection using information theoretic that includes about twenty years of heuristic scoring criteria research, by the optimization of *the conditional likelihood of the class label given the features*.

To define the framework of information theory, it is necessary to review some basic concepts. The entropy of a random variable X (of the feature x) is defined as follows:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

where $p(x)$ is the probability of the particular event $X = x$. This entropy measures the uncertainty of the distribution of X. On the other hand, the entropy of the feature can be conditioned on other events, e.g. on the presence of the random variable of the class label Y. Thus the conditional entropy of X given Y is defined in the form

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y)$$

With the two above concepts, the mutual information between X and Y is defined as follows:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

The mutual information conditioned on another event or conditional information is,

$$I(X;Y|Z) = H(X|Z) - H(X|YZ) = \sum_{z\in Z} p(z) \sum_{x\in X} \sum_{y\in Y} p(xy|z) \log \frac{p(xy|z)}{p(x|z)p(y|z)}$$

where Z is another random variable (another feature, the label class, etc.).

Through the above definitions, Table 6.2 shows four criteria of the information theoretic feature selection framework. It is important to remark that one of the more general criteria, *Conditional Mutual Information* (CMI), can be expressed by three terms: the first corresponds to the relevance (mutual information between the unselected feature and the class label), the second is related to the redundancy (mutual information between the unselected feature and the selected features), and the third corresponds to the conditional redundancy (mutual information between the unselected feature and the selected features given the class label). Then, most of the criteria, especially those that combine these terms linearly, make a balance between these three parts, showing that it is not only important to include features that increase the relevance and reduce redundancies, but also to include some correlated features [105].

In this thesis, the criteria MRMR, JMI and CMIM were used to analyze and select the features, with the purpose of developing the lymphoid cell recognition methodology.

TABLE 6.2: *Summary of the framework for information theoretic feature selection*

Criterion	Score formula
Conditional Mutual Information (CMI) [105]	$I_{cmi}(X_k) = I(X_k;Y\|S) = I(X_k;Y) - I(X_k;S) + I(X_k;S\|Y)$
Minimum-Redundance Maximum-Relevance (MRMR) [120]	$J_{mrmr}(X_k) = I(X_k;Y) - \frac{1}{\|S\|}\sum_{X_j\in S} I(X_k;X_j)$
Joint Mutual Information (JMI) [121]	$J_{jmi}(X_k) = \sum_{X_j\in S} I(X_kX_j;Y)$
Conditional Mutual Info Maximization (CMIM) [122]	$J_{cmim}(X_k) = \min_{X_j\in S} \left[I(X_k;Y\|X_j)\right]$
S: current set of selected features; Y: class label; X: random variable of the feature x; $X_k \notin S$; $X_j \in S$ X_kX_j is a joint variable between the candidate X_k with each previous selected feature.	

6.7 Support Vector Machines classification

SVM defines a well known and widely used classification (and regression) technique, which aims to produce a model using training data sets that can predict the corresponding class labels of validation data sets. Given a training set of N pairs of feature vectors and their corresponding labels $(\mathbf{x}_i, y_i)$, with $i = 1,...,N$, where $\mathbf{x}_i \in R^D$ and $y_i \in \{-1,1\}$, the original

SVM approach obtains the linear discriminant $f(\mathbf{x_i}) = \langle \mathbf{w}, \Phi(\mathbf{x_i}) \rangle + b$ with maximum margin in the transformed feature space (higher-dimensional) by the mapping function $\Phi : R^D \rightarrow R^H$, through solving the following optimization problem [90, 123]:

$$\begin{aligned} & \min_{\mathbf{w},b,\xi} \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & y_i \left(\langle \mathbf{w}, \Phi(\mathbf{x_i}) \rangle + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \mathbf{w} \in \mathbf{R}^H, b \in \mathbf{R} \end{aligned} \tag{6.44}$$

where ξ is the vector of *slack* variables, $\mathbf{w}$ is the weight coefficients vector, C is the penalty parameter (to balance the classification error and the margins), and b is the bias term of the separating hyperplane. Since $\mathbf{w}$ can have a very high dimension, the optimization problem is solved indirectly by the Lagrangian dual function obtaining the following approach:

$$\begin{aligned} & \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) - \sum_{i=1}^{N} \alpha_i \\ \text{subject to} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned} \tag{6.45}$$

where the inner product $K{:}\mathbf{R}^D \times \mathbf{R}^D$ is the *kernel function* defined by $K\left(\mathbf{x_i}, \mathbf{x}_j\right) = \left\langle \Phi(\mathbf{x_i}), \Phi(\mathbf{x_j}) \right\rangle$, and α_i are the dual variables corresponding to the constraints. The solution of the last problem is $\mathbf{w}{=}\sum_{i=1}^{N} \alpha_i y_i \Phi(\mathbf{x_i})$, therefore the discriminant function becomes $f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K\left(\mathbf{x_i}, \mathbf{x}\right) + b$.

The most common and profitable Kernels used in the literature are the following functions:

Linear $K\left(\mathbf{x_i}, \mathbf{x}_j\right) = \left\langle \mathbf{x_i}, \mathbf{x}_j \right\rangle$

Polynomial $K\left(\mathbf{x_i}, \mathbf{x}_j\right) = \left(\gamma \left\langle \mathbf{x_i}, \mathbf{x}_j \right\rangle + r\right)^d, \gamma > 0$

Radial Basis Function (RBF) $K\left(\mathbf{x_i}, \mathbf{x}_j\right) = \exp\left(-\gamma \left\| \mathbf{x_i} - \mathbf{x}_j \right\|^2\right), \gamma > 0$; with $\left\| \mathbf{x}^2 \right\| = \langle \mathbf{x}, \mathbf{x} \rangle$

Sigmoid $K\left(\mathbf{x_i}, \mathbf{x}_j\right) = \tanh\left(\gamma \left\langle \mathbf{x_i}, \mathbf{x}_j \right\rangle + r\right)$

where γ, r, and d are parameters of the kernel function.

SVM was originally proposed as a binary classifier but it can be extended to multiclass problems using the construction of different binary classifiers to recognize a particular label from the remaining (one-against-all), or for every pair of classes (one-against-one) [124].

In this work, SVM were mainly applied with Linear and RBF kernel functions to classify the cell features, implementing the multi-classification through the one-against-one technique. All the procedures were implemented using the library package LIBSVM [92].

6.8 Experimental Results

In this chapter, several experiments were done to analyze the features (feature selection) and evaluate the performance of the different classifiers, mainly SVM classifiers with different kernels and parameters. Up to six color spaces were used to extract color-texture features for their subsequent analysis and classification, separately and in combination with the geometric features. The implemented color spaces were: RGB[4] (Red, Green and Blue), CM$\mathbb{Y}$K[5] (Cyan, Magenta, Yellow and Black), XYZ (or CIE 1931 XYZ, where Y is the luminance, X and Z are the chromaticities), *L*a*b** (or CIELAB 1976, where *L** is the luminance and, *a** and *b** are the chromaticities), *L*u*v* (or CIELUV 1976 where *L** is the luminance and, *u** and *v** represent the chromaticities) and HSV (Hue, Saturation and Value) [35, 125]. Thus, the number of computed features was dependent on the color spaces and the color components that were selected in each experiment.

6.8.1 Blood sample preparation and digital image acquisition

The blood images used in the experiments were samples from normal donors and patients with Chronic Lymphocytic Leukemia (CLL), Hairy Cell Leukemia (HCL), Mantle Cell Lymphoma (MCL) and Follicular Lymphoma (FL), where FL cells were the new type of neoplastic lymphoid cell included in this chapter with respect to the previous Chapter 5. The diagnoses were established by clinical and morphologic findings as well as characteristic immunophenotype of the lymphoid cells. Specifically, CLL cells had the phenotype CD5+, CD19+, CD23+, CD25+, weak CD20+, CD10-, FMC7- and dim surface immunoglobulin (sIg) expression. All the patients with HCL had lymphoid cells with the phenotype CD11c+, CD25+, FMC7+, CD103+ and CD123+. Patients with MCL showed lymphoid cells with the phenotype CD5+, FMC7+, CD43+, CD10- and BCL6-. Follicular lymphoma cells showed B-cell associated antigens (CD19, CD20, CD22, CD79a) BCL2+, BCL6+, CD10+, CD5- and CD43-. B-prolymphocytes (BPL) images were obtained from transformed CLL. Blood samples were obtained from the routine workload of the Core Laboratory of the Hospital Clínic of Barcelona. Venous blood was collected into tubes containing K3EDTA as anticoagulant. Samples were analyzed by a cell counter Advia 2120 (Siemens Healthcare Diagnosis, Deerfield, USA) and PB films were automatically stained with May Grünwald-Giemsa in the SP1000i (Sysmex, Japan, Kobe) within 4 hours of blood collection.

Individual lymphoid cell images from PB had a resolution of 363 x 360 pixels and they were obtained by the CellaVision DM96 system (Lund, Sweden). The quality of the smears

[4] Standard RGB

[5] The blackboard style of $\mathbb{Y}$ is to avoid confusion with the second component of the XYZ color space

was assessed by cytologists prior to the image study. A training set of 1834 lymphoid cell images from PB films was selected by the cytologist to evaluate the accuracy of the proposed methodology, which were distributed as follows: 181 normal lymphocyte images from healthy patients (N), 301 HCL, 401 MCL, 334 FL and 617 from patients with CLL. This group was divided into 542 CLL clumped chromatin typical lymphocyte images and 75 BPL images.

6.8.2 Feature analysis and classification experiments

The first classification experiment was made using the full set of features, which combines 77 geometric features and 6422 color-texture features calculated over the six color spaces (RGB, CM$\mathbb{Y}$K, XYZ, L*a*b*, L*u*v, and HSV). Due to the high number of features, a SVM classifier with only a linear kernel ($C = 10$) was applied. Table 6.3 shows the results of this experiment through the confusion matrix of a 10-fold cross validation classification procedure. Appendix A gives details on the performance classification parameters based on the confusion matrix. The average classification accuracy for the 10 folds was 95.96%. Table 6.3 also shows that almost all true positives rates (diagonal of the confusion matrix) exceed 90%, except for BPL type because of its low number of samples. Although these results indicate a great performance, the large amount of information makes the classification a very complex and time-consuming process.Therefore, an additional feature selection step is quite useful to reduce the information (considering the redundancy and relevance), thereby optimizing the process.

Table 6.4 presents the accuracies of several classifications using SVM with Linear ($C = 10$) and RBF ($C = 10, \gamma = 0.5$) kernels for 20 features selected from the full set (geometric and color-texture features of the six color spaces), by using information theoretic with three different criteria: CMIM, MRMR and JMI. These results show that the introduction of the

TABLE 6.3: *10-fold cross validation confusion matrix of a SVM classification using the full set of features (6499).*

		Predicted*					
		N	**HCL**	**CLL**	**FL**	**MCL**	**BPL**
True	**N**	92.27	2.21	1.10	0.00	3.87	0.55
	HCL	1.00	97.34	0.33	0.00	0.66	0.66
	CLL	0.74	0.00	98.34	0.55	0.00	0.37
	FL	0.60	0.00	1.20	95.51	2.69	0.00
	MCL	1.25	0.25	0.00	1.75	96.26	0.50
	BPL	1.33	4.00	2.67	0.00	9.33	82.67

* The rows represent the true diagnosis and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. Every row was normalized in relation to the total number of cells of its respective type to obtain the percentages respect to the true diagnosis. Linear kernel, C = 10; Classification accuracy = 95.96%.

TABLE 6.4: *Lymphoid cell classification accuracy (Acc) using SVM for 20 selected features from the full set.*

Criterion	Kernel	Acc (%)
CMIM	Linear	97.11
	RBF	97.49
MRMR	Linear	95.97
	RBF	96.46
JMI	Linear	96.07
	RBF	96.46

The color-texture features were extracted from six color spaces: RGB, CM$\mathbb{Y}$K, XYZ, *L*a*b**, *L*u*v*, and HSV. SVM classifiers: linear kernel, $C = 10$; RBF kernel, $C = 10, \gamma = 0.5$.

feature selection step improves the accuracy in all cases, especially using the criterion CMIM and the SVM classifier with RBF kernel, which obtains a 97.49% of accuracy utilizing only 20 features.

Previously, the cell-type classification using the full set of features was made, but it is also important to analyze separately the specific weight of each feature category to improve and optimize the DIP process. Thus, several experiments according to the different types of features are presented below, separating the geometric-size features, the EFD, the color-texture features extracted from each color component, and some combinations between them (always on the same color space).

Table 6.5 shows the accuracy of cell-type classification resulting from applying SVM with Linear ($C = 10$) and RBF ($C = 10, \gamma = 0.5$) kernels over three categories of features excluding color-texture features: (1) only geometric-size features (Geo-s), (2) only EFD and (3) the combination between Geo-s and EFD (Geo). The first case uses the full set of features, i.e. no feature selection (NFS) was made, while the second and third cases use a set of 20 features selected by using information theoretic (CMIM, MRMR and JMI criteria). The best result of this experiment is the SVM classification with RBF kernel of the geometric-size features with an accuracy of 88.39%, while the classification accuracy of the EFD is below 80% and the accuracy for the classification of the full geometric features reaches a value up to 88%. Consequently, the EFD have very little impact on the description of the lymphoid cells studied in this chapter.

Table 6.6 presents the accuracy results in the lymphoid cell classification using SVM with Linear ($C = 10$) and RBF ($C = 10, \gamma = 0.5$) kernels, for only the full set of 6422 color-texture features (no feature selection, NFS), as well for the case of 20 features selected from this full set (calculated over all the color spaces: RGB, CM$\mathbb{Y}$K, XYZ, *L*a*b**, *L*u*v*, and HSV), by using three information theoretic selection criteria (CMIM, MRMR and JMI). All the

TABLE 6.5: *Lymphoid cell classification accuracy (Acc) using SVM for 20 features selected from three feature categories.*

Case	#Features	Criterion	SVM kernel	Acc (%)
Geo-s	13	NFS	Linear	87.24
			RBF	88.39
EFD	64	NFS	Linear	75.63
			RBF	74.1
		CMIM	Linear	74.65
			RBF	76.06
		MRMR	Linear	73.83
			RBF	76.44
		JMI	Linear	73.72
			RBF	76.5
Geo	77	NFS	Linear	84.35
			RBF	83.1
		CMIM	Linear	87.19
			RBF	86.37
		MRMR	Linear	88
			RBF	87.95
		JMI	Linear	87.73
			RBF	87.24

SVM classifiers: linear kernel, $C = 10$; RBF kernel, $C = 10, \gamma = 0.5$. Geo-s, geometric-size features; EFD, elliptical Fourier features; Geo, combination between Geo-s and EFD; NFS, no feature selection.

TABLE 6.6: *Lymphoid cell classification accuracy (Acc) using SVM for the full set of color texture features and for 20 features selected from this full set.*

Criterion	Kernel	Acc (%)
NFS	Linear	95.64
CMIM	Linear	94.71
	RBF	95.58
MRMR	Linear	94.66
	RBF	95.15
JMI	Linear	94.44
	RBF	95.36

The color-texture features were extracted from six color spaces: RGB, CM$\mathbb{Y}$K, XYZ, *L*a*b**, *L*u*v*, and HSV. SVM classifiers: linear kernel, $C = 10$; RBF kernel, $C = 10, \gamma = 0.5$. NFS, no feature selection.

accuracies of the color-texture classification exceed 94%, showing the importance of these color-texture features in the description of the lymphoid cells. The best result is for the full color-texture feature set, but the SVM-RBF classification of the features selected using the criterion CMIM is quite close, with the extra advantage of reducing computation times. This is the main reason why the following experiments use only this criterion in the feature selection step and the SVM with the RBF kernel in the classification process.

With the purpose of further investigating and analyzing the contribution of the color to the description of the lymphoid cells, several experiments were done by calculating the corresponding features on each color component and using the SVM classifier (RBF kernel, $C = 10, \gamma = 0.5$) and the feature selection step (CMIM criterion, 20 features) with various combinations of feature sets, depending on the color space. Subsequently, the same process was done but adding the geometric features to the combined feature sets. Table 6.7 presents the lymphoid cell classification accuracy results of the experiments described above. The first column represents the color components that were used to calculate the features, the second one is the cell classification accuracy using only the color-texture features, and the third one is the accuracy of the combination between the geometric and the specific color-texture features. By comparing columns 2 and 3, it is observed that the addition of the geometric features improves the classification up to 7% in terms of accuracies. The best classification result was obtained with the CM$\mathbb{Y}$K color space both for only the color-texture features and the combination of them with the geometric features. Moreover, three combinations of color-texture features (and geometric ones) achieved a classification accuracy greater than 97%: CM$\mathbb{Y}$K, *L*a*b* and CK. As the above results show, the color-texture features are very important to describe the types of lymphoid cells analyzed in this chapter.

From this point, the experiments are focused on the set that presented better results: 20 features selected from the set of geometric and the color-texture features calculated over the CM$\mathbb{Y}$K color space (1429 features). They are selected by information theory with CMIM criterion (best performance feature set). The selected features of this configuration are shown in Table 6.8. The two most important features are the cell perimeter and the nucleus-cytoplasm ratio, however the color-texture features (granulometric and statistical features) are also presented, specially those calculated on the Cyan and Black color components.

In order to build the best possible classifier, several lymphoid cell classification experiments were made using SVM with four different kernels, varying their parameters over the feature set with the conditions mentioned above. Table 6.9 shows a comparison of the corresponding results using linear, RBF, polynomial and sigmoid kernels. In these experiments, most classification accuracies were above 97%, but the best result was the SVM-RBF with $C = 5, \gamma = 0.8$, obtaining a 97.93% of accuracy. In addition, a comparison between

TABLE 6.7: *Lymphoid cell classification accuracy using SVM for a set of 20 features selected by using information theoretic (CMIM) from various features sets depending on the color components.*

Color Components	Accuracy (%)	
	Color	Geo+Color
RGB	94.49	96.13
CM$\mathbb{Y}$K	95.417	97.71
XYZ	92.58	95.91
*L*a*b**	95.2	97.11
*L*u*v**	94.22	96.78
HSV	95.31	96.13
R	92.42	96.29
G	93.07	95.41
B	90.24	93.4
C	92.09	96.13
M	92.09	95.04
$\mathbb{Y}$	88.33	94.22
K	92.42	96.46
X	91.98	95.42
Y	90.78	94.77
Z	91.82	94.71
*L**	92.15	95.15
*a**	88.39	92.91
*b**	93.78	96.62
*l**	90.78	94.77
*u**	87.73	94.82
*v**	91.6	95.53
H	89.15	94.71
S	92.31	95.31
V	92.64	95.42
RG	94.11	96.29
CK	94.66	97.33
XZ	92.97	96.4
*L*b**	94.33	96.78
*l*v**	93.95	95.8
SV	94.77	96.73

SVM classifier with RBF kernel, $C = 10, \gamma = 0.5$; "Color component" indicates which components were used to calculate the color-texture features; Color, classification accuracy utilizing only the color-texture set; Geo+color, classification accuracy utilizing the geometric and the color-texture features.

TABLE 6.8: *20 features selected by information theoretic (CMIM criterion) from the geometric and CM𝕐K color-texture feature set.*

1. Perimeter – cell	11. IMC2 – K – nucleus
2. Nucleus – cytoplasm ratio	12. Mean – 𝕐 – cytoplasm
3. IMC1 – C – nucleus	13. Kurtosis – K – nucleus
4. Standard deviation – H1 – M – nucleus	14. Homogeneity – C – cytoplasm
5. IMC2 – C – nucleus	15. Standard deviation – PGC – K – nucleus
6. Diameter – nucleus	16. Entropy (2) – K – nucleus
7. Mean – PGC – C – nucleus	17. Standard deviation – GC – M – cytoplasm
8. Mean – C – nucleus	18. Standard deviation – M – nucleus
9. Standard deviation – C – nucleus	19. Skewness – PGC – 𝕐 – cytoplasm
10. IMC2 – D2 – K – cytoplasm	20. Standard deviation – K – cytoplasm

C, cyan; M, Magenta; 𝕐, yellow; K, black; IMC, information measure of correlation (1 and 2); GC, granulometric curve; PGC, pseudo-granulometric curve. H1, V1, D1, H2, V2, and D2 are the details sub-images corresponding to the one and second levels of the wavelet decomposition. The application of each feature over the sub-images, color components, curves and regions goes from left to right.

TABLE 6.9: *Comparative lymphoid cell classification accuracies (Acc) using SVM with several kernels of 20 features selected by information theoretic (CMIM criterion) from a set of geometric and color-texture features of the CM𝕐K color space.* C *is the penalty parameter and* γ ,d *and* r *are the different parameters corresponding to the particular kernel.*

Kernel	***C***	γ	***d***	***r***	**Acc (%)**
Linear	1				97.55
	10				97.27
	100				97.00
	1000				96.78
	5				97.33
RBF	10	0.5			97.71
	10	0.05			97.71
	100	0.05			97.71
	1000	0.05			97.65
	5	0.9			97.87
	5	0.8			97.93
Polynomial	10	0.5	2	-1	97.76
	10	0.5	2	-0.95	97.82
	10	0.5	3	0	97.49
	10	0.5	4	0	97.44
	1	3	1	0	97.65
Sigmoid	10	0.05		0	96.78
	100	0.05		0	97.44

the classification techniques SVM (RBF kernel), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN) and Naive Bayes classifier (NB), was done on the same feature set previously used. The respective best classification accuracies are shown in Table 6.10. The SVM-RBF classifier (it was also presented in Table 6.9) presents the best performance in the cell classification. However, the other classifiers also show good results (above of 92% of accuracy), thus indicating the high quality of the cell description.

Table 6.11 shows the confusion matrix for the best experimental result, i.e. the SVM-RBF classifier ($C = 5, \gamma = 0.8$) over the best performance feature set. It can be observed that this confusion matrix is superior in all aspects to that shown in Table 6.3.

Finally, Figure 6.8 illustrates the accuracies of the SVM with a RBF kernel ($C = 5, \gamma = 0.8$) respect to the number of selected features from the CM$\mathbb{Y}$K space, by using information theoretic with CMIM criterion (best performance in above experiments). This figure helps to explain why the number of 20 features was chosen as the most appropriate in the feature selection step. Indeed, beyond this threshold, the accuracy remains approximately constant with the increasing of the number of features.

TABLE 6.10: *Lymphoid cell classification accuracy using various classifiers over 20 features selected by information theoretic (CMIM criterion) from a set of geometric and color-texture features of the CM$\mathbb{Y}$K color space.*

Classifier	Parameters/note	Accuracy (%)
SVM	RBF kernel, $C = 5, \gamma = 0.8$	97.93
LDA		95.58
k-NN	Euclidean distance, 3 neighbors	95.96
NB	Gaussian distribution	92.47

SVM, Support Vector Machines; LDA, Linear Discriminant Analysis; kNN, k-Nearest Neighbors; NB, Naive Bayes classifier.

TABLE 6.11: *10-fold cross validation confusion matrix of the best experiment result.*

		Predicted*					
		N	**HCL**	**CLL**	**FL**	**MCL**	**BPL**
True	**N**	98.34	1.66	0.00	0.00	0.00	0.00
	HCL	0.66	98.34	0.00	0.00	0.33	0.66
	CLL	0.18	0.00	99.45	0.18	0.00	0.18
	FL	0.30	0.00	1.50	96.41	1.80	0.00
	MCL	0.75	0.00	0.00	1.50	97.26	0.50
	BPL	0.00	0.00	1.33	0.00	4.00	94.67

* The rows represent the true diagnosis and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. The data set consists of 20 color-texture features from the CM$\mathbb{Y}$K space selected by information theoretic (CMIM criterion). SVM classifier with RBF kernel ($C = 5, \gamma = 0.8$). Classification accuracy = 97.93%.

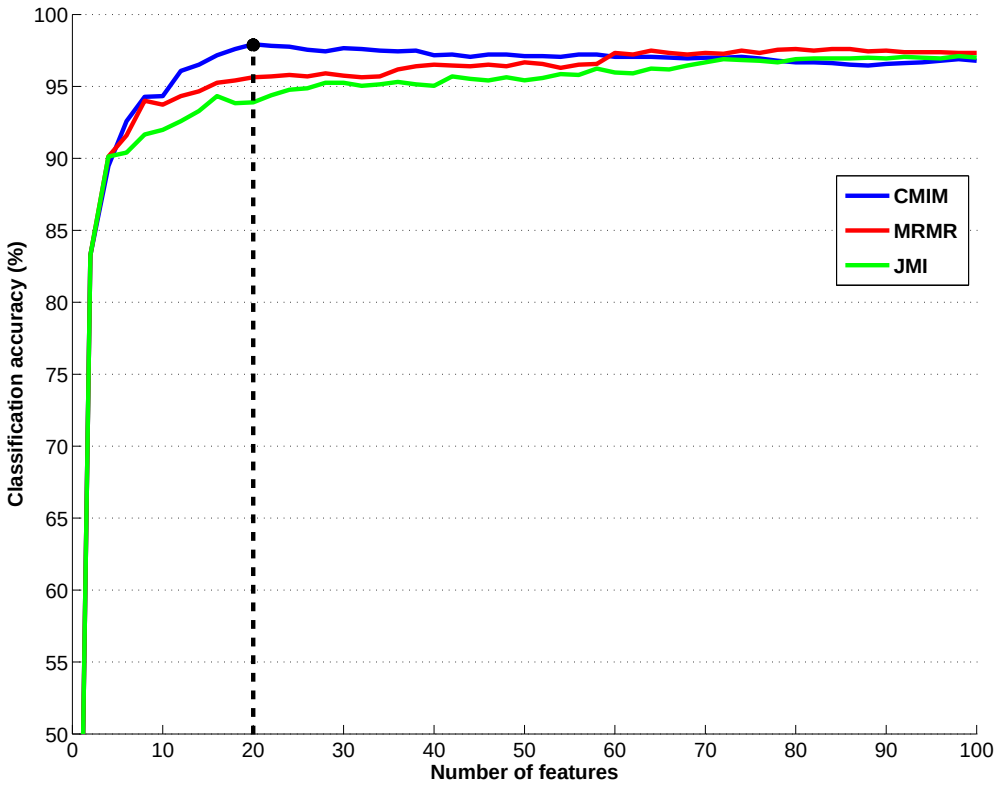


FIGURE 6.8: *Variation of the lymphoid cell classification (SVM-RBF $C = 5, \gamma = 0.8$) accuracy respect to the number of features selected by information theoretic using three different criteria.*

6.9 Discussion

The work in this chapter is an evolution of the methodology presented in Chapter 5, resulting in important differences and extensions of the methodology in the feature extraction and classification stages.

In the methodology developed in this chapter, the geometric features include the compactness and the EFD. The second order statistical features are extended to the full set of Haralick [83] and two more features [111]. Besides, in this chapter the statistical features are also applied over the detail sub-band images of the second wavelet decomposition, achieving a more elaborated description of the texture. The granulometric features in Chapter 5 are calculated based only on the granulometric curve, while in this chapter they are also calculated not only on the granulometric curve, but also on the pseudo-granulometric curve. In Chapter 5 the color-texture features are applied directly over the *L*a*b** color space. However, in this chapter, they are applied up to six color spaces, producing up to 6499 features.

The feature selection procedure in Chapter 5 is used only to analyze the behavior and importance of the full set of features. On the other hand, in this chapter, the feature selection step is used and included within the pattern recognition methodology.

In this chapter, various classification experiments are done, showing that the SVM classifier works better than the LDA classifier, which is used in Chapter 5. The number of types of lymphoid cells automatically recognized has been increased, including the Follicular Lymphoma cells (respect to Chapter 5), showing promising results in the respective classifications.

The full set of extracted features provides an excellent cell description obtaining a successful classification. However, higher number of features makes the feature extraction a more complex process (particularly with new cells), with longer run time and more difficult interpretations. In this work, three strategies were adopted to optimize the DIP methodology of the lymphoid cells: (1) separation of the features according to their type and color space, (2) information theoretic feature selection, and (3) implementation of various classifiers.

Individually, the geometric-size features reach a good cell classification performance, while the EFD have the worst performance of the experiments. That is, the "size" is more important than the "shape" in the DIP of the lymphoid cells studied in this work.

In spite of the differences are not very large, most of the experiments presented better cell classification accuracies with the information theoretic criteria CMIM, particularly those that used the color and texture information. Furthermore, the SVM with RBF kernel showed the best cell classification performance (with a reduced feature set). This explains why the combination between feature selection with CMIM and SVM-RBF produced the best classification results.

The separated color experiments presented excellent classification performance, showing the importance and the specific weight of the color-texture features in the lymphoid cell description. Besides, the features extracted from CM$\mathbb{Y}$K, *L*a*b** and CK presented the best accuracy results (in combination with the geometric features).

The best lymphoid cell classification performance through all experiments was obtained with 20 features selected (by CMIM criteria) from the set formed by the geometric and the color-texture features from the CM$\mathbb{Y}$K color space. This reduced set included geometric-size features, statistical features, wavelet statistical features and granulometric features, mostly extracted from the C and K color components.

Although the color-texture features have individually a better classification accuracy than the geometric features, when they are combined, the most important features are the geometric ones, particularly the cell perimeter and the nucleus cytoplasm ratio. These two were the most important features in all experiments that involved some combination with the geometric-size features.

The slight differences between the classification accuracy of the different SVM configurations (variations of the kernel and the parameters) and other classifiers, show a certain independence of the classification process respect to the classifier, i.e. the lymphoid cell feature extraction and selection has been successful.

6.10 Conclusion

This chapter has focused on three main issues (feature extraction, feature selection and classification), which are consecutively linked in a cascade form to obtain a final recognition procedure for the lymphoid cells under study.

The successful cell classification results show that the extracted full feature set has been able to describe size, shape, texture and color of the different types of lymphoid cells, involving the calculation of geometric and color-texture features. First and second order statistical features have been calculated for each color component. The wavelet statistical features have been implemented by the application of the first and second order statistical features over the six detail sub-bands images of the second wavelet decomposition of each color component. The granulometric features have been calculated on the granulometric and pseudo-granulometric curves of each color component. All the above features describe the texture and color of the lymphoid cells, and with the geometric features (size features and EFD) constitute the full dataset with almost 6500 features. This complete feature extraction process, which joins several texture concepts, is a novel methodology in the description framework of different types of neoplastic lymphoid cells.

This chapter has presented automatic classifications of different types of lymphoid cells without considering the correlation between cells of the same patient. In the next chapter, the cells are separated in a training and a validation set where individual patients are not previously included. The purpose is to validate the methodology in a horizon where the overall system could be useful in clinical practice.

Chapter 7

A System for Automatic Identification of Atypical Lymphoid Cells from Peripheral Blood Cells Images

Based upon: S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, A new strategy for automatic identification of atypical lymphoid cell images from peripheral blood, International Journal of Laboratory Hematology, 2015, submitted for publication.

Abstract

The objective of the work presented in this chapter was to go forward in the development of a system capable of the successful automatic discrimination of a significant number of atypical lymphoid cell images in view of practical diagnosis purposes. A training set of 3617 lymphoid cell images from peripheral blood from 70 patients was used. They were segmented using clustering of color components and watershed transformation. From the regions of interest of each image, 6499 features were extracted and selected to choose the most significant for lymphocyte recognition. A 10-fold cross validated classification using Support Vector Machine was done. The complete classification system was tested through processing lymphoid cell images from 21 new patients individually. The number of features obtained in the extraction step was reduced to the 95 most relevant, which were used for the lymphoid recognition. The accuracy for the cell classification in the training set was 92.3%. In the validation for each patient, the system was able to successfully recognize most of the cells corresponding to the true diagnosis. The cell classification accuracy of the whole validation process was 85.2%. The strategy was able to achieve high precision in the automatic recognition of 7 different types of lymphoid cells. The performed patient-based validation results are important in view of the use of this system as a support tool for initial B lymphoid neoplasm detection in peripheral blood.

7.1 Introduction

Atypical lymphoid cells are the most difficult pathological cells to classify using only morphology features [80]. Since morphologic evaluation is primarily focused on the cytological features, quantitative measurements or descriptors of individual lymphoid cells by imaging analysis may contribute to define morphologic features of malignant lymphocytes.

If one aims to advance in the objective that an automatic classification could become an useful practical diagnosis support tool in the near future, a problem still open is to achieve a successful discrimination among a relevant number of different atypical lymphoid cells in the context of the currently known B cell neoplasms [6]. Up to the authors knowledge, the

literature has reported classification tools able to recognize only a limited number of atypical lymphoid cells [21, 59, 70, 74, 75]. In Chapter 5, normal and four types of neoplastic lymphoid cells were classified: Chronic Lymphocytic Leukemia (CLL), Hairy Cell Leukemia (HCL), Mantle Cell Lymphoma (MCL) and B-prolymphocytes (BPL).

In previous Chapters 3 and 5 a method for lymphocyte recognition to allow the automatic classification of normal and several types of neoplastic lymphoid cells circulating in PB in mature B cell neoplasms was presented. This method was based on solving three problems: 1) cell segmentation to separate the regions of interest from the overall image; 2) feature extraction for a relevant quantitative description of morphologic characteristics; and 3) trained classification to automatically identify the different lymphoid subtypes. In Chapter 6 a detailed methodology for automatic recognition of neoplastic lymphoid cells, including a new group (Follicular Lymphoma FL), was described. This development involved significant improvements in the feature extraction/selection and classification steps. The present chapter extends the work of Chapter 6, with the following three main contributions: 1) it uses a significantly higher number of cell images, 2) it includes a new type of atypical lymphoid cells (Reactive Lymphocytes RL), and 3) it evaluates the effectiveness of the completed classification system in such a way that a number of lymphoid cell images from individual patients (previously not included) are processed by the system achieving the recognition of the different types of cells. This validation is important in view of practical diagnosis purposes, where the system would operate to automatically recognize the atypical lymphoid cells of each new single patient.

The remainder of this chapter is organized as follows. Section 7.2 describes the development of the methodology through various digital image processing steps and the validation of the classification system. Section 7.3 shows the most important selected features, the evaluation of the performance of the methodology and specifically the validation of the methodology as a support diagnostic tool. Finally, Sections 7.4 and 7.5 provide discussions and conclusions.

7.2 Material and methods

The work presented in this chapter was developed in two stages: 1) system development, resulting in an operative classification system; and 2) validation of the system through tests in which cell images from individual patients were classified for diagnosis purposes. In the first stage, a set of 3617 lymphoid cell images from 70 patients was used, which is referred to as training set. In the second one, a number of 910 new images obtained from 21 patients, previously selected by the cytologist, were used to evaluate the system. The set of lymphoid cells from each independent patient was named as validation set. Figure 7.1 shows the scheme of the two-stage process. The bottom part of this scheme can be visualized as a diagnostic

support tool whose inputs are the cell images from the patient and the output is the cell classification in the different types of normal or atypical lymphoid cells.

All the algorithms for the development and validation of the classification system have been implemented using the scientific and high-level language MATLAB®.

7.2.1 System development

The introduction of new types of atypical lymphoid cell implies more complexity to the classification problem, which led to modify the developed methodology with improvements in almost all the steps of the digital image processing, specifically the feature extraction and classification process. The optimized methodology to carry out the automatic classification of 7 different types of lymphoid cells was done through the following steps: 1) blood sample preparation and digital image acquisition; 2) clustering color segmentation and Watershed transformation; 3) feature extraction; 4) feature selection; and 5) classification. The details of the methodology have been described in Chapters 4 and 6.

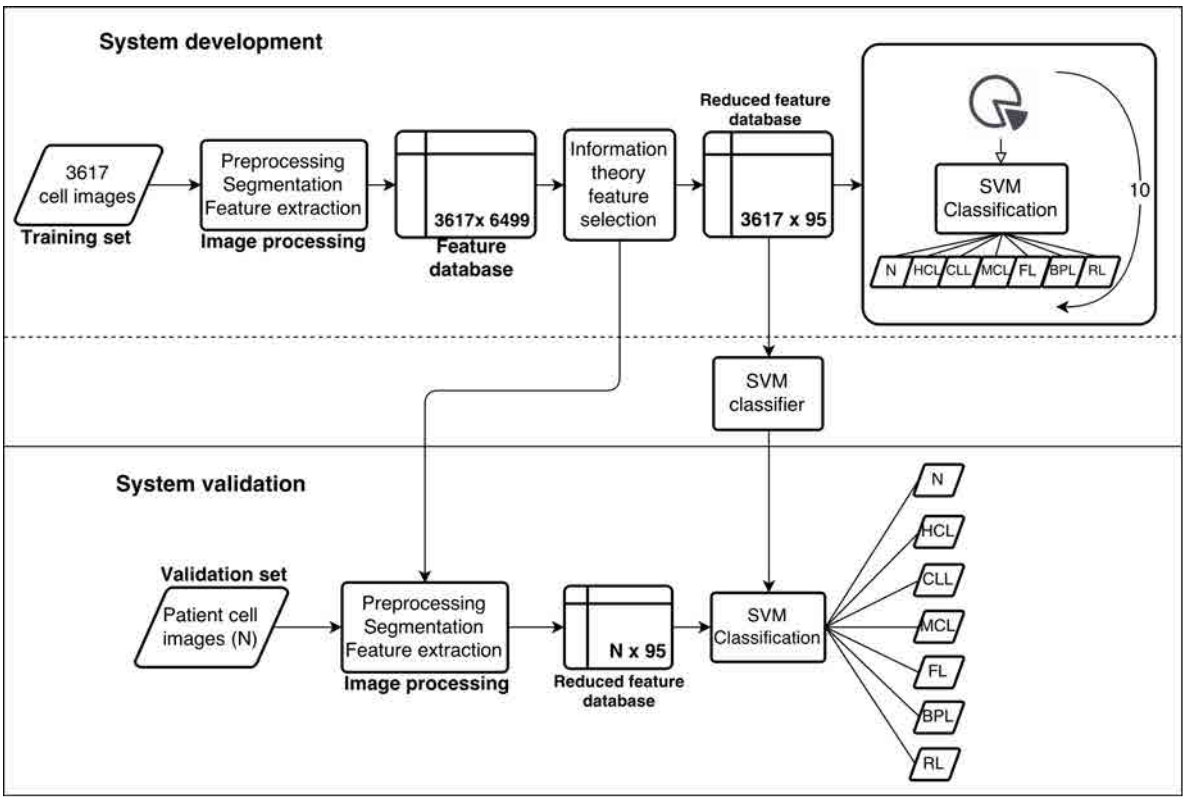


FIGURE 7.1: *The whole process has two stages: 1) the system development (digital image processing is applied over the training set), and 2) the system validation (the methodology is applied over lymphoid cells of individual patients).*

7.2.1.1 Blood sample preparation and digital image acquisition

Samples from normal donors and patients with CLL, HCL, MCL and FL were included in this study. The diagnoses were established by clinical and morphologic findings as well as characteristic immunophenotype of the lymphoid cells. Specifically, CLL cells had the phenotype CD5+, CD19+, CD23+, CD25+, weak CD20+, CD10-, FMC7- and dim surface immunoglobulin (sIg) expression. All the patients with HCL had lymphoid cells with the phenotype CD11c+, CD25+, FMC7+, CD103+ and CD123+. Patients with MCL showed lymphoid cells with the phenotype CD5+, FMC7+, CD43+, CD10- and BCL6-. Follicular lymphoma cells showed B-cell associated antigens (CD19, CD20, CD22, CD79a) BCL2+, BCL6+, CD10+, CD5- and CD43-. BPL images were obtained from transformed CLL. The reactive lymphocyte images were obtained from patients with the diagnoses of infectious mononucleosis. Blood samples were obtained from the routine workload of the Core Laboratory of the Hospital Clínic of Barcelona. Venous blood was collected into tubes containing K3EDTA as anticoagulant. Samples were analyzed by a cell counter Advia 2120 (Siemens Healthcare Diagnosis, Deerfield, USA) and PB films were automatically stained with May Grünwald-Giemsa in the SP1000i (Sysmex, Japan, Kobe) within 4 hours of blood collection.

Individual lymphoid cell images from PB had a resolution of 363 x 360 pixels and they were obtained by the CellaVision DM96 system (Lund, Sweden). The quality of the smears was assessed by cytologists prior to the image study. A training set of 3617 lymphoid cell images from PB films was selected by the cytologist to evaluate the accuracy of the proposed methodology, which were distributed as follows: 320 normal lymphocyte images from healthy patients (N), 408 RL, 529 HCL, 732 MCL, 551 FL and 1077 from patients with CLL. This group was divided into 863 CLL clumped chromatin typical lymphocyte images and 214 BPL images. For the validation step cell images from a total of 21 different patients previously selected by the cytologist were used.

7.2.1.2 Clustering color segmentation and Watershed transformation

As described in Chapter 4, lymphoid cells were segmented from other objects in the image using sKFCM clustering technique over the XYZ and CMYK color spaces and the Watershed Transformation (WT). Thus, three regions were obtained: cell, nucleus and peripheral zone around the cell.

7.2.1.3 Feature extraction

The feature extraction methods described in Chapter 6 are used in this work. A summary of the implementation of the feature extraction step is presented here.

Geometric features These features are numerical interpretations of morphologic attributes such as size, shape, nucleus-cytoplasm ratio, etc. A total of 77 geometric features were calculated: 13 geometric-size features and 32 Elliptic Fourier Descriptors (EFD) for each region of interest. They also included a cytoplasmic profile feature, which estimates the external projections of the cytoplasm [103].

Color and texture features Three different methods to characterize the color and texture were used, obtaining the following three types of features for each color component from six color spaces (RGB, CM$\mathbb{Y}$K, XYZ, *L*a*b**, *L*u*v*, and HSV): (1) statistical features; (2) wavelet statistical features; and (3) granulometric features. Each of these features was applied to the nucleus and the cytoplasm.

1. **Statistical Features**

   In Chapter 5, 6 first order and 7 second order statistical features were used [83]. As in Chapter 6, from 7 to 15 second order statistical features were extended and 2 more features were also added: cluster shade and cluster prominence [111]. These 23 features were calculated over each color component of the image.

2. **Wavelet statistical features**

   A novelty respect to earlier works (Chapters 3 and 5) is that the above 23 statistical features were applied not only over the color components of the original image but also over 6 sub-images derived from a two level wavelet decomposition [113–115] for each color component. As derived in Chapter 6, this procedure consists in the application of the discrete wavelet transform (DWT) over an image (in this case a color component). It decomposed the image into in 4 sub-images: an approximation of the image and three highlighted versions of the horizontal (H1), vertical (V1) and diagonal (D1) details. This process was repeated in a second level decomposition over the first approximation image and four more sub-images were obtained: A2, H2, V2 and D2. These 6 detail sub-images were used to obtain the 23 x 6 wavelet statistical features.

3. **Granulometric features**

   As in Chapter 6, 8 granulometric features were extracted: 4 of them were calculated on the granulometric curve (it uses successive operations of opening and closing) and the remaining 4 were calculated on the pseudo-granulometric curve, which uses successive applications of the mathematical morphology operations dilation and erosion [84].

In summary, 23 statistical features for the original image, 23 x 6 wavelet statistical features and 8 granulometric features were obtained. These 169 features were calculated for each color

component of the six color spaces. All of them were applied for the nucleus and the cytoplasm regions. All features were stored in a numerical data matrix, which was used as the input data for the feature selection step.

7.2.1.4 Feature selection

Due to the large number of cell features extracted, it was necessary to apply feature selection to reduce their interdependence, their redundancy and to make the classification process more feasible. The purpose of this step was to determine the most significant features for the further classification step. In the present work, as it was studied in Chapter 6, the information theoretic feature selection using the so-called Conditional Mutual Info Maximization (CMIM) criteria was used [105, 122].

7.2.1.5 Classification

The objective of this step was to obtain the automatic recognition of normal and reactive lymphocytes, and five types of neoplastic lymphoid cells from PB. Accordingly, the most relevant features from the selection were used as inputs to the supervised learning classifier based on Support Vector Machines (SVM) using a radial basis function kernel [91, 92] as it was developed and tested in Chapter 6. The classification performance was evaluated by the application of the 10-fold cross validation technique over the training set of 3617 lymphoid cell images. This technique randomly divides the data set into 10 equal size subsets. A single subset is used as the testing data, while the remaining data are used for training. Then, the process is repeated 10 times and a confusion matrix is calculated to get significant overall statistical measures.

7.2.2 System validation

After the methodology development stage, a *classification system* was assembled by tuning the SVM classifier with the use of the set of the reduced selected features. The aim of this part of the work was to evaluate the effectiveness of the cell classification system for diagnosis purposes. A total of 910 images from 21 patients were selected by the cytologist, and distributed as follows: 69 normal lymphocytes, 175 RL, 84 HCL, 93 MCL, 236 CLL, 136 FL and 117 BPL. These images were processed following the steps described above but extracting only the previously selected features (see Figure 7.1, lower part). Then, this new feature database was processed by the tuned SVM classification system, producing the recognition of the different types of lymphoid cells for each patient. This procedure was repeated for each patient, thereby completing the system validation. It is important to remark

that the cell images in the validation set of the patients were not used before in the training of the overall classifier.

7.3 Results

7.3.1 Feature selection

As it was explained before, feature selection was used to reduce the redundancy of the variables and the complexity of the classification. A total of 6499 features were obtained in the previous extraction step, 77 were geometric and 6422 were color and texture features. Through several tests, the feature selection step allowed improving the classification process reducing the number of features from 6499 to 95, which were the only inputs for the classification step. The first twenty features of that set are listed in order of relevance in Table 7.1. A total of 6 features were geometric and the remaining 89 were color and texture features. However, the three most important features obtained in the selection process were geometric: the nucleus - cytoplasm ratio, the perimeter of the nucleus and diameter of the cell.

7.3.2 Methodology performance evaluation

The 3617 cell images included in the training set were classified into 7 types using SVM with 10-fold cross validation. A confusion matrix is shown in Table 7.2, which evaluates the performance of the classification of these images. Rows denote the true diagnosis and columns represent the predicted diagnosis supplied by the classification process for each type

TABLE 7.1: *The 20 most relevant features (from a total of 95) were obtained using the Conditional Mutual Info Maximization criteria of the information theoretic feature selection step.*

1. Nucleus – cytoplasm ratio	11. Correlation – G – nucleus
2. Perimeter – nucleus	12. Mean – PGC – C – nucleus
3. Diameter – cell	13. Mean – X – nucleus
4. Mean – b* – nucleus	14. Entropy 1 – a* – nucleus
5. IMC1 – M – nucleus	15. Entropy 1 – K – nucleus
6. Cytoplasmic profile feature	16. Mean – PGC – M – nucleus
7. Mean – u - nucleus	17. Skewness – GC – v – nucleus
8. Standard deviation – PGC – B – cytoplasm	18. Skewness – PGC – X – nucleus
9. Mean – C – nucleus	19. Mean – PGC – b* – nucleus
10. Mean – PGC – S – nucleus	20. Correlation – H – nucleus

Color spaces involved: L*a*b*, **CMYK**, L**uv**, R**GB**, **HSV**, **XYZ**. Abbreviations: IMC, information measure of correlation (1 and 2); PGC, pseudo-granulometric curve; GC: granulometric curve. The application of each feature over the sub-images, color components, curves and regions goes from left to right.

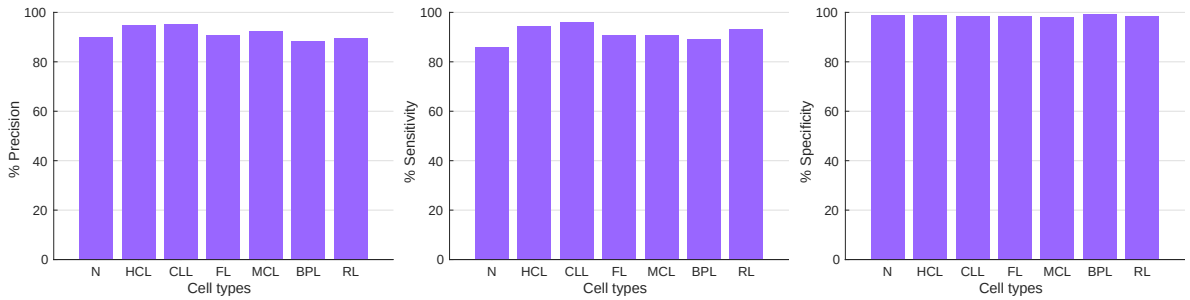
TABLE 7.2: *Confusion Matrix of the support vector machines classification and 10-fold cross-validation for the training set.*

		Predicted*						
		N	**HCL**	**CLL**	**FL**	**MCL**	**BPL**	**RL**
True	**N**	**85.94**	1.56	6.88	0.94	1.25	0.94	2.50
	HCL	1.70	**94.33**	0.57	0.00	0.57	0.57	2.27
	CLL	1.85	0.35	**96.18**	0.93	0.23	0.23	0.23
	FL	0.36	0.36	2.54	**90.93**	5.63	0.00	0.18
	MCL	0.27	0.27	0.41	5.19	**90.71**	1.78	1.37
	BPL	0.00	0.93	0.00	0.47	3.74	**89.25**	5.61
	RL	0.49	3.43	0.25	0.00	1.72	0.98	**93.14**

* The rows represent the true diagnosis and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. The values are in percentage. Accuracy = 92.34 % and standard deviation = 0.92%.

of lymphoid cell. Every row was normalized in relation to the total number of cells of its respective type to obtain the percentages respect to the true diagnosis. Appendix A gives details on the performance classification parameters based on the confusion matrix. The overall 7-type classification accuracy was 92.34%. Its standard deviation (STD) was also computed (which was 0.92%) to measure the variability between folds. Diagonal values are percentage values of the true positive rates for each cell subtype: 85.94% for normal lymphoid cells, 94.33% for HCL, 96.18% for CLL, 90.93% for FL, 90.71% for MCL, 89.29% for BPL and 93.14% for RL (see Table 7.2).

Figure 7.2 shows three statistical measurements of the 7-type classification obtained with the automatic classification methodology described in this chapter. The precision values were above 88.43% for all neoplastic lymphoid cell subtypes and, normal and reactive lymphocytes. The sensitivity values were above 89.25% except for the normal cell subtype, in which sensitivity value was 85.94%. Finally, all the specificity values were above 98.09%.



(a) *Precision* **(b)** *Sensitivity* **(c)** *Specificity*

FIGURE 7.2: *Precision (a), sensitivity (b) and specificity (b) values (in percentages) of the different lymphoid cell subtypes obtained in the system development.*

7.3.3 Validation of the methodology

The classification system was tested by individually processing lymphoid cell images from 21 new patients. Normal lymphoid cell images were selected from 5 healthy subjects, reactive lymphocytes from 6 patients with the diagnoses of infectious mononucleosis, and neoplastic B cells from the remaining 10 patients (2 HCL, 2 MLC, 2 FL, 2 CLL and BPL cells were selected from other 2 CLL patients). As shown in Table 7.3, in the validation for each patient, the system was able to successfully recognize most of the cells corresponding to the true diagnosis: 92% to 100% for RL, 97% to 100% for HCL cells, 54% to 66% for MCL cells, 63% to 95% for FL cells, 80% to 82% for CLL cells, 85% to 100% for normal lymphocytes and 68% to 97% for BPL cells. The cell classification accuracy of the whole validation process was 85.17%.

Figure 7.3 shows some images used in the validation step, which correspond to different

TABLE 7.3: *Percentage of the classification results obtained in the validation stage using lymphoid cells corresponding to individual patients.*

Patient	*True diagnosis*	*Cell images*	*Predicted**						
			HCL%	MCL%	FL%	CLL%	BPL%	N%	RL%
1	RL	10	0	0	0	0	0	0	**100**
2	RL	53	2	0	0	0	0	0	**98**
3	RL	29	0	0	0	0	0	0	**100**
4	RL	26	0	0	0	0	4	4	**92**
5	RL	44	2	0	0	0	0	0	**98**
6	RL	13	0	0	0	0	0	0	**100**
7	HCL	52	**100**	0	0	0	0	0	0
8	HCL	32	**97**	0	0	0	0	3	0
9	MCL	13	0	**54**	31	0	0	15	0
10	MCL	80	0	**66**	31	1	1	0	0
11	FL	63	0	37	**63**	0	0	0	0
12	FL	73	0	1	**95**	4	0	0	0
13	CLL	165	0	4	15	**80**	1	1	0
14	CLL	71	4	4	0	**82**	1	7	1
15	N	16	0	0	0	0	0	**100**	0
16	N	14	0	0	0	7	0	**93**	0
17	N	11	0	9	0	0	0	**91**	0
18	N	13	0	0	8	0	0	**85**	8
19	N	15	0	0	0	7	0	**93**	0
20	BPL	64	0	2	0	0	**97**	2	0
21	BPL	53	0	23	0	0	**68**	6	4

* The values of the cell types are in percentage. Global cell classification accuracy = 85.17%

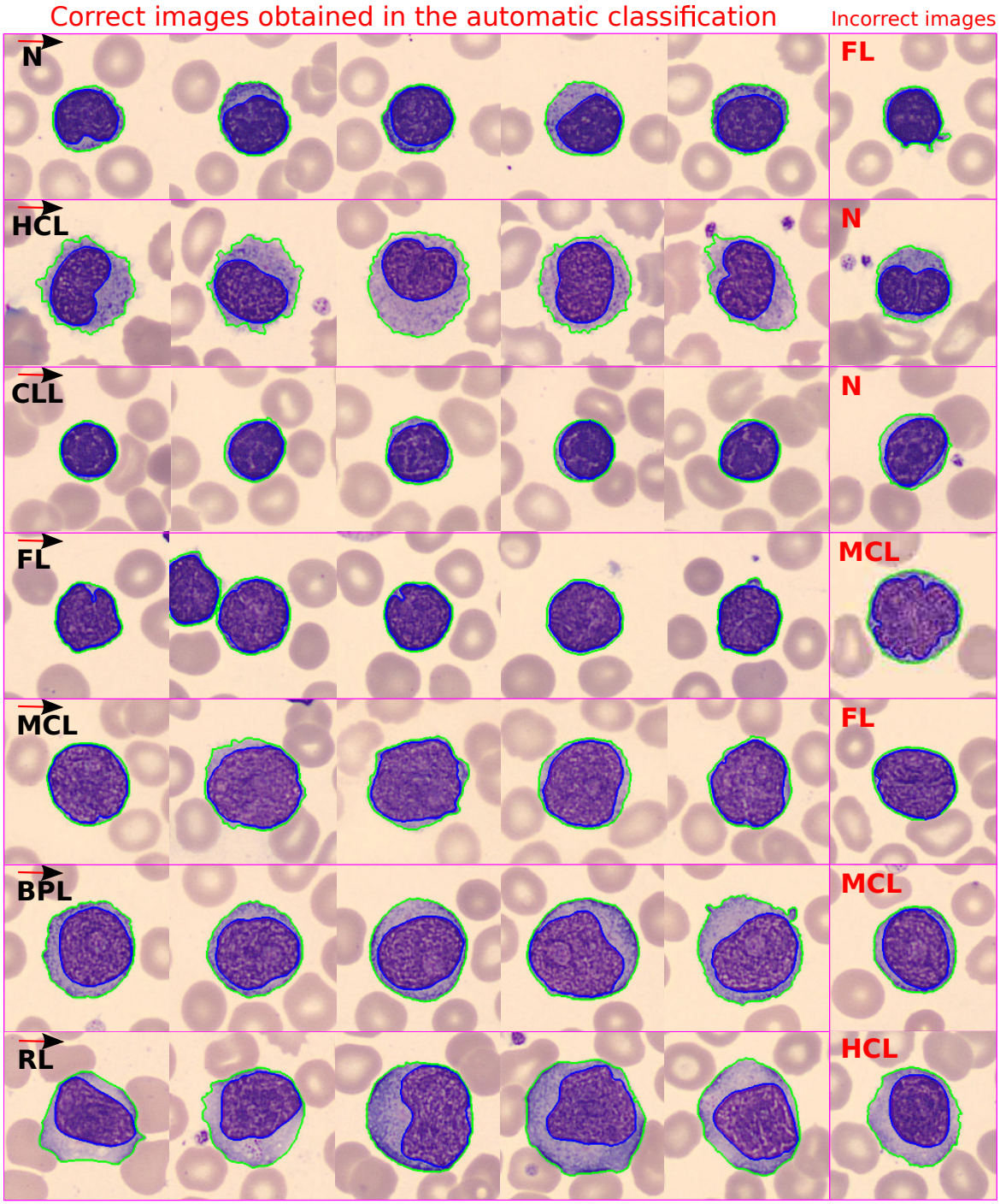


FIGURE 7.3: *Examples of lymphoid cell images of individual patients obtained after the automatic classification process in the validation stage. Correct classified images are shown in the first 5 columns, while in the last column images incorrectly classified are included (magnifications:* $\times 1000$ *and stain: May-Grünwald-Giemsa).*

lymphoid cell types obtained in the classification process. Each row corresponds to cell images from an individual patient (N, HCL, CLL, FL, MCL, BPL and RL). The first five columns show correct images obtained after the automatic classification, while the last column shows images that were not correctly classified.

7.4 Discussion

Morphologic analysis of PB cells is the first analytical step in the hematological diagnosis, being very useful for the indication of further laboratory tests. Since atypical lymphoid cells are the most difficult to be recognized by only morphologic features [80], the major goal of the current work was to use the methodology introduced in Chapter 6, including reactive lymphocytes and validating the performance of the classification system using a number of lymphoid cell images from individual patients.

In this chapter, normal, reactive lymphocytes and five types of neoplastic (HCL, CLL, FL, MCL and BPL) lymphoid cells were selected considering their characteristic morphology and the high number collected from daily workload in the clinical laboratory. In the present work, the segmentation method described in Chapter 4 was implemented, using the color information of the image to separate the main regions for each lymphoid cell (nucleus, cytoplasm and the peripheral zone around the cell). This procedure has been very useful to extract information about the cell.

Geometric features over the cell, the nucleus and the cytoplasm were used previously by Scotti [39]. In other papers, geometric and statistical features were also used for the analysis of the shape and the nucleus texture [53, 67]. Granulometry on the luminance color component of the $L^*a^*b^*$ space to calculate texture features was also used by Angulo et al [59] for the recognition of lymphoid cells. In this work, the number of morphologic characteristics considered in Chapter 5 were extended from 113 to 6499 features. New geometric features were used to describe the shape of the cell and the whole set of second order statistical features were expanded. In addition, the application of the wavelet decomposition and the use of the pseudo granulometric curve made possible to characterize with more detail the texture of both the nucleus and cytoplasm. This feature extraction strategy is novel in the automatic recognition of the atypical lymphoid cells.

A number of 95 important features were obtained after their selection, and it is interesting to remark that the nucleus-cytoplasm ratio, the perimeter of the nucleus and the diameter of the cell were the three more relevant features. This finding is according to the more relevant size-morphologic characteristics that the cytologist uses to differentiate among lymphoid cells.

Concerning the classification procedure, Ushizima et al [72] investigated the use of SVM classifiers to recognize five different types of normal leukocytes but only one subtype of neoplastic lymphoid cell (CLL cells). Moreover, Angulo et al [59] classified the morphologic features of the lymphoid cells in categories using decision trees, but that work was not completed with further studies toward the specific discrimination among different groups of similar diagnosis. In addition, the methodology proposed by Yang et al [75] was able to recognize five types of neoplastic blood cells, but the images selected in this study were precursor lymphoid cells and myeloid blast cells from acute leukemia patients. In this work, the developed methodology has been extended to automatically recognize 7 types of lymphoid cells, obtaining very satisfactory performance measures. This high number of different types of lymphoid cell classification, to our knowledge, has not been previously published in the literature.

7.5 Conclusion

In order to validate the methodology, a group of new patients were selected, not included in the training set. A good accuracy was obtained in the classification results of the different lymphoid cells using the images of these new patients with higher percentages of lymphoid cell subtypes recognized. These results in the validation step of the classification method are encouraging toward the idea that the system could be useful for diagnosis purposes in the future.

In summary, an overall system for the automatic classification of different types of atypical lymphoid cells has been assembled with the combination of the robust segmentation method, the best selected feature set and the best classifier, all of them developed in this work. The classification algorithm recognized normal and reactive lymphocytes, and five different types of neoplastic B lymphoid cells. The main contribution of this work, combining medical, engineering and mathematical backgrounds, is the development of a complete methodology that could allow in the next future to design a practical diagnosis support tool. The classification of the different lymphoid cells considering a group of new patients was made not only to validate the methodology but also in a horizon where this methodology could be useful in clinical practice.

Chapter 8

Conclusions and contributions. Future perspectives

This thesis presents a methodology for the automatic classification of peripheral blood (PB) lymphoid cells in atypical lymphoid cells through several steps of digital image processing and pattern recognition. The thesis has grown through the evolution of various works, starting with a discrimination between normal lymphocytes and two types of neoplastic lymphoid cells, and ending with the design of a system for the automatic recognition of normal and reactive lymphocytes, and five types of neoplastic lymphoid cells. All this work has involved the development of a robust lymphoid cell segmentation, a complete cell description by feature extraction/selection and a successful classification using support vector machines (SVM).

The conclusions derived along this research are summarized in Section 8.1. The main contributions of this thesis are highlighted in Section 8.2. Finally, some future perspectives open from this work are outlined in Section 8.3.

8.1 Conclusions

- The segmentation method proposed in this thesis combines the novel fuzzy clustering method spatial Kernel Fuzzy C-means (sKFCM) on different color components and watershed transformation, to segment not only the nucleus and cytoplasm, but also the external region of the cell, which can supply relevant information in some types of cells such as HCL. In addition, the efficiency of the segmentation is tested on normal and reactive lymphocytes, and five types of neoplastic lymphoid cells (HCL, FL, MCL, CLL and PLB).

- A novel cytoplasmic profile feature has been proposed based on a simple thresholding of the peripheral zone around the cell. This feature has been important for the HCL cells recognition, since they show a soft, blue-gray cytoplasm with hair-like cytoplasmic projection. On the other hand, this feature could be used for the detection of another neoplastic lymphoid cells with cytoplasmic villous, such as the splenic marginal zone lymphoma.

- The methodology uses mainly two steps to segment the regions of interest: (1) cropping the lymphoid cell from the remaining regions, and (2) separating the regions of the nucleus, the cytoplasm and the background. For the cropping of the lymphoid cell from the remaining regions, it has been found that Y and K are the best combination of color components using sKFCM clustering. For the separation of the regions of interest, it has been found that XYZ is the best color space using the sKFCM clustering.

- The procedure has achieved a high level of efficiency in the segmentation of the different types of lymphoid cell images included in this thesis, which exhibit a variety of morphologic features.

- The extracted full feature set has been able to describe size, shape, texture and color of different types of lymphoid cells included in this work, involving the calculation of almost 6500 geometric and color-texture features. The further successful cell classification results show that this extracted feature set provides excellent cell description.

- In general, the feature selection is a necessary step to simplify the complexity of the process, reducing the run time of the algorithms associated to the high number of descriptors. In this thesis many classification experiments have been carried out with different combinations of color components and/or geometric features, applying information theoretic feature selection tools.

- The experiments using only geometric features have shown good performance, while features extracted only from combinations of color components have shown better classification performance. On the other hand, when both types of features are combined, the classification performance improves. Particularly, the best results have been obtained using the following color sets: CM$\mathbb{Y}$K, *L*a*b** ,CK and the combination of all six color spaces studied.

- It has been found that 95 features were the most relevant and less redundant for the classification step. Each selected set included geometric and the following color-texture features: statistical, wavelet statistical and granulometric features. The three most

important selected features for the automatic lymphoid cell recognition are the nucleus-cytoplasm ratio, the perimeter of the nucleus and the diameter of the cell.

- In this work several classifiers have been implemented to automatically recognize different types of lymphoid cells. The best classification results have been achieved using support vector machines (SVM) with radial basis function (RBF) kernel.

- The information theoretic feature selection procedure developed in this thesis is independent of the classification process. This explains that, although the SVM classifier has shown the best results, the differences with respect to the use of other classifiers have been slight.

- An overall system for the automatic classification of different types of normal and atypical lymphoid cells has been assembled with the combination of the robust segmentation method, the best selected feature set and the best classifier, all of them developed in this work.

- The methodology for the automatic lymphoid cell recognition has shown excellent accuracies in the system development stage.

- In the development of the methodology, the information theoretic feature selection criterion CMIM (conditional mutual info maximization) applied over the color-texture set extracted from the six color spaces: RGB, CM$\mathbb{Y}$K, XYZ, *L*a*b*, L*u*v*, and HSV, and the geometric features has been the best configuration of feature extraction/selection according to the results of the classification of the lymphoid cells included in the study group.

- In the validation of the methodology, the feature selection with the criterion CMIM applied over the color-texture and the geometric features has been found as the best configuration of feature extraction/selection according to the results of the lymphoid cell classification from individual patients.

- The classification system using SVM classifiers has been designed and experimentally assessed in two stages: (1) development, and (2) validation. In the first stage, the digital image processing (DIP) steps are developed to build the classifier using a training set of image cells collected from a study group of subjects with normal or reactive lymphocytes and patients with different B-cell neoplasms. In the second stage, the DIP steps are implemented considering the previous stage with a validation set of lymphoid cell images collected from individual patients not included in the previous study group.

The classification of the different lymphoid cells considering new patients was made not only to validate the methodology but also in a horizon where this methodology could be useful in clinical practice. The developed methodology has been extended to automatically recognize 7 types of lymphoid cells, obtaining very satisfactory performance measures. The methodology developed, combining medical, engineering and mathematical backgrounds, could allow to designing a practical diagnosis support tool in the next future.

8.2 Main contributions of this thesis

The most significant novel contributions with respect to the available state of the art are the following:

- The clustering segmentation method developed in this thesis combines the fuzzy C-means technique with the kernel method and it also takes in account the spatial information. This combination results in a novel spatial kernel fuzzy C-means clustering for color segmentation of digital images.
- A robust segmentation methodology has been developed using mainly color clustering and Watershed transformation, which is able to separate three regions of interest: cell, nucleus and peripheral zone around the cell. This peripheral zone is particularly important to discriminate lymphoid cells with hairy-like cytoplasmic projections, which are present in specific B-cell neoplasms such as hairy cell leukemia. To the best of our knowledge, the segmentation of this peripheral zone of the cell has not been previously published in the literature.
- A novel cytoplasmic profile feature has been proposed based on a simple thresholding of the peripheral zone around the cell. This feature has shown to be one the most relevant for the recognition of the atypical lymphoid cells.
- In particular, the cytoplasmic profile feature has been crucial for the Hairy Cell Leukemia (HCL) automatic recognition, since in peripheral blood the lymphoid cells show a soft, blue-gray cytoplasm with hair-like cytoplasmic projection. The automatic recognition of hairy cells has been scarcely studied in the literature.
- An innovative feature extraction method has been proposed in this thesis. This method computes geometric features (size and shape) and color-texture features. The novelty lies in that the color-texture information is obtained through the extraction of statistical, wavelet statistical and granulometric features on different color components.

- In addition, a feature selection procedure using information theory has been applied to identify and select the best features that provide useful information about the characterization of the lymphoid cells. This feature extraction/selection strategy is novel in the automatic recognition of the atypical lymphoid cells.

- The novel segmentation and feature extraction/selection have been completed with a support vector machine classifier. This particular pattern recognition strategy has resulted in a complete methodology to automatically recognize normal and reactive lymphocytes, and five types of neoplastic lymphoid cells circulating in peripheral blood in different mature B-cell neoplasms with a very satisfactory classification performance. The classification of this high number of different types of lymphoid cells, to the best of our knowledge, has not been previously published in the literature.

8.3 Future perspectives

The results obtained in this thesis open some issues that could be further developed. Some of them are outlined below:

- The methodology developed could be extended to the classification of target cells from other hematological diseases, such as precursor B/T cells lymphoid neoplasms, or other mature B/T cells neoplasms not included in the present work. In other words, the future goal is to identify any abnormal lymphoid cells in a peripheral blood sample.

- It would be interesting to extend the methodology to the automatic classification of abnormal cells in body fluids in different hematological or non hematological neoplasms.

- Since the full set of features extracted and selected in this thesis has reached a very good characterization of lymphoid cells, it could be interesting to apply them over other types of abnormal hematopoietic cells to detect myelodysplastic morphologic changes of some hematological diseases.

- It would be interesting to apply the methodology developed to asses their effectiveness in the classification of abnormal red blood cells circulating in peripheral blood in different types of anemias.

- Since the sKFCM clustering algorithm can work with color spaces, the potential applications are not only limited to cell images obtained from peripheral blood but to other medical application areas, which require segment images with different regions with color and spatial similarities.

- Although the sKFCM segmentation experiments were implemented with three clusters, the algorithm can be used with more clusters, increasing the details in the regions or the number or regions obtained.
- The peripheral zone around the cell has been explored only to calculate an approximated measure of the hairy-like cytoplasmic projections. It would be interesting to extract more features from this region to analyze their contribution to the characterization of the cell.
- The new applications proposed above will require, as has occurred in the development of this thesis, further extensions in the DIP steps.

8.4 Publications derived from this thesis

8.4.1 Conferences and Communications

1. S. Alférez, A. Merino, L.E. Mujica, and J. Rodellar,. Morphological Analysis Using Digital Image Processing in Lymphoid Neoplasias. Abstracts of the XXIV International Symposium on Technical Innovations in Laboratory Hematology. International Journal of Laboratory Hematology, Issue s1, vol 33, pp. 53, May 2011, Wiley. ISNN: 1751-5521
2. A. Merino, S. Alférez, L.E. Mujica, J. Rodellar. Análisis Morfológico de las Células Linfoides Atípicas Mediante Técnicas de Segmentación: Aplicaciones al Diagnóstico. Revista del Laboratorio Clínico, Volumen 4, Especial Congreso Noviembre 2011, 0911, pp. 425-426, ISSN: 1888-4008.
3. S. Alférez, A. Merino, M. Ruiz, L.E. Mujica, J. Rodellar. Atypical Lymphoid Cells Detection and Classification using Mathematical Morphology and Fuzzy Clustering on Digital Blood Image Analysis. Abstracts of the XXV International Symposium on Technological Innovations in Laboratory Hematology, International Journal of Laboratory Hematology, Issue s1, pp. 75, June 2012, Wiley. Online ISSN:1751-553X
4. S. Alférez, A. Merino, M. Ruiz, L.E. Mujica, J. Rodellar. Detección y clasificación de células linfoides B atípicas mediante morfología matemática y agrupamiento Fuzzy en imágenes de sangre periférica. Libro de Comunicaciones del VI Congreso Nacional de Laboratorio Clínico, 2012, pp. 431, ISBN: 978-84-695-5381-7

5. S. Alférez, A. Merino, M. Ruiz, L.E. Mujica, J. Rodellar. Atypical lymphoid cells detection and classification on digital blood image analysis. Abstracts of the XXVI International Symposium on Technological Innovations in Laboratory Hematology. International Journal of Laboratory Hematology, Issue s1, pp. 100, June 2013, Wiley. doi: 10.1111/ijlh.12105

6. S. Alférez, A. Merino, L. Bigorra, L.E. Mujica, M. Ruiz, and J. Rodellar. Detección y clasificación de células linfoides atípicas mediante procesamiento digital de imágenes de sangre periférica. Libro de Comunicaciones del VII Congreso Nacional del Laboratorio Clínico, 2013, pp. 466, ISBN 10: 84-695-8699-8.

7. S. Alférez, A. Merino, L. Bigorra, M. Ruiz, L.E. Mujica, and J. Rodellar. A new strategy for automatic identification of atypical lymphoid cells from peripheral blood cell images. Invited Presentation Abstracts of the XXVII International Symposium on Technological Innovations in Laboratory Hematology. International Journal of Laboratory Hematology, Issue s1, vol 36, pp. 7, June 2014, Wiley. doi: 10.1111/ ijlh.12265

8. S. Alférez, A. Merino, L. Bigorra, L.E. Mujica, M. Ruiz, and J. Rodellar. Identificación automática de céulas linfoides atípicas mediante análisis de imágenes digitales de sangre periférica. LVI Congreso Nacional de la Sociedad Española de Hematología y Hemoterapía, 2014, Madrid.

9. S. Alférez, A. Merino, L. Bigorra, L.E. Mujica, M. Ruiz, and J. Rodellar. Method for automatic recognition of neoplastic lymphoid cells using peripheral blood cell images. Plenary session of the XXVIII International Symposium on Technological Innovations in Laboratory Hematology. May 2015, Chicago.

8.4.2 Awards

- Young Investigator Award. Method for automatic recognition of neoplastic lymphoid cells using peripheral blood cell images. Plenary session of the XXVIII International Symposium on Technological Innovations in Laboratory Hematology. May 2015, Chicago.

8.4.3 Journals

1. S. Alférez, A. Merino, L.E. Mujica, M. Ruiz, L. Bigorra, and J. Rodellar, Automatic classification of atypical lymphoid B cells using digital blood image processing, Inter-

national Journal of Laboratory Hematology, vol. 36, no. 4, pp. 472-80, Aug. 2014. doi: 10.1111/ijlh.12175

2. S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, Automatic recognition of atypical Lymphoid cells from peripheral blood by digital image analysis, American Journal of Clinical Pathology, vol. 143, pp. 168-176, 2015. doi: 10.1309/AJCP78IFSTOGZZJN

3. S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, A new strategy for automatic identification of atypical lymphoid cell images from peripheral blood, International Journal of Laboratory Hematology, 2015, *submitted for publication*.

4. S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, Color clustering segmentation of lymphoid cell images using spatial kernel fuzzy c-means, *manuscript in preparation*.

5. S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, A methodology for automatic recognition of neoplastic lymphoid cell images from peripheral blood, *manuscript in preparation*.

8.4.4 Patent

- S. Alférez, A. Merino, J. Rodellar, L.E. Mujica, M. Ruiz. Spanish Patent ES 2428215 A1. Priority date 09/05/2013. Método implementado por ordenador para el reconocimiento y clasificación de células sanguíneas anormales y programas informáticos para llevar a cabo el método. International patent classification G06K 9/00 (2006.01). International patent extension in process by Patent Cooperation Treaty (PCT).

References

[1] B. Ciesla, *Hematology in Practice*. FA Davis, 2011.

[2] B. J. Bain, D. M. Clark, I. A. Lampert, and B. S. Wilkins, *Bone Marrow Pathology*. John Wiley & Sons, 2008.

[3] A. Merino, *Manual de Citología de Sangre Periférica*. Grupo Acción Médica, 2005.

[4] K. Kottke-Marchant and B. H. Davis, *Laboratory Hematology Practice*. Oxford, UK: Wiley-Blackwell, Apr. 2012.

[5] D. Longo, *Harrison's Hematology and Oncology*. McGraw Hill Professional, 2010.

[6] S. H. Swerdlow, E. Campo, N. L. Harris, E. S. Jaffe, S. A. Pileri, H. Stein, J. Thiele, and J. W. Vardiman, *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. IARC Press, 2008.

[7] V. Kumar, A. K. Abbas, N. Fausto, and J. C. Aster, *Robbins and Cotran Pathologic Basis of Disease, Professional Edition: Expert Consult-Online*. Elsevier Health Sciences, 2009.

[8] G. J. Sinclair, "Blood film review by biomedical scientists," *British Journal of Biomedical Science*, vol. 62, pp. 77–80, Jan. 2005.

[9] B. J. Bain, "Diagnosis from the blood smear," *The New England Journal of Medicine*, vol. 353, pp. 498–507, Aug. 2005.

[10] Sysmex, "Medica EasyCell." https://www.sysmex.com/us/en/Products/Hematology /CellImageAnalysis/Pages/Medica- EasyCell.aspx. Accessed August 15, 2014.

[11] C. Münzenmayer, T. Schlarb, D. Steckhan, E. Haß lmeyer, T. Bergen, S. Aschenbrenner, T. Wittenberg, C. Weigand, and T. Zerfaß, "HemaCAM - A computer assisted microscopy system for hematology," in *Microelectronic Systems*, pp. 233–242, Springer, 2011.

[12] Fraunhofer Institute for Integrated Circuits, "HemaCAM." http://www.iis.fraunhofer.de/en/bf/med/mbv/hemacam.html. Accessed August 15, 2014.

[13] Bloodhound and Roche Website, "cobas m 511." http://www.roche-rdh.com. Accessed August 15, 2014.

[14] "Cellavision Website." http://www.cellavision.com. Accessed August 15, 2014.

[15] A. Kratz, H.-I. Bengtsson, J. E. Casey, J. M. Keefe, G. H. Beatrice, D. Y. Grzybek, K. B. Lewandrowski, and E. M. Van Cott, "Performance evaluation of the CellaVision DM96 system: WBC differentials by automated digital image analysis supported by an artificial neural network," *American Journal of Clinical Pathology*, vol. 124, pp. 770–781, Nov. 2005.

[16] E. Cornet, J.-P. Perol, and X. Troussard, "Performance evaluation and relevance of the CellaVision DM96 system in routine analysis and in patients with malignant hematological diseases," *International Journal of Laboratory Hematology*, vol. 30, pp. 536–42, Dec. 2008.

[17] C. Briggs, I. Longair, M. Slavik, K. Thwaite, R. Mills, V. Thavaraja, A. Foster, D. Romanin, and S. J. Machin, "Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system.," *International Journal of Laboratory Hematology*, vol. 31, pp. 48–60, Feb. 2009.

[18] A. Merino, R. Brugués, R. García, M. Kinder, F. Torres, and G. Escolar, "Estudio comparativo de la morfología de sangre periférica analizada mediante el microscopio y el CellaVision DM96 en enfermedades hematológicas y no hematológicas," *Revista del Laboratorio Clínico*, vol. 4, pp. 3–14, Jan. 2011.

[19] H. Ceelie, R. B. Dinkelaar, and W. van Gelder, "Examination of peripheral blood films using automated microscopy; evaluation of Diffmaster Octavia and Cellavision DM96," *Journal of Clinical Pathology*, vol. 60, pp. 72–9, Jan. 2007.

[20] M. Bergmann, H. Heyn, H. K. Müller-Hermelink, H. Harms, and H. M. Aus, "Automated recognition of cell images in high grade malignant lymphoma and reactive follicular hyperplasia," *Analytical Cellular Pathology*, vol. 2, no. 2, pp. 83–95, 1990.

[21] D. 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*, vol. 4, pp. 265–273, Dec. 2000.

[22] J. Juan, F. Sigaux, and G. Flandrin, "Automated classification of lymphoid cells," *Analytical and Quantitative Cytology*, vol. 7, no. 1, pp. 38–46, 1985.

[23] R. S. Riley, J. M. Ben-Ezra, D. Massey, and J. Cousar, "The virtual blood film," *Clinics in Laboratory Medicine*, vol. 22, pp. 317–45, Mar. 2002.

[24] F. Scotti, "Robust segmentation and measurements techniques of white cells in blood microscope images," *2006 IEEE Instrumentation and Measurement Technology Conference Proceedings*, pp. 43–48, Dec. 2006.

[25] T. Würflinger, J. Stockhausen, D. Meyer-Ebrecht, and A. Böcking, "Automatic coregistration, segmentation and classification for multimodal cytopathology," *Studies in Health Technology and Informatics*, vol. 95, pp. 218–23, Jan. 2003.

[26] F. Sadeghian, Z. Seman, A. R. Ramli, B. H. Abdul Kahar, and M.-I. Saripan, "A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing," *Biological Procedures Online*, vol. 11, pp. 196–206, June 2009.

[27] V. Piuri and F. Scotti, "Morphological classification of blood leucocytes by microscope images," *2004 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 2004. CIMSA.*, no. July, pp. 103–108, 2004.

[28] R. Ravindraiah, D. M. G. Prasad, and M. V. Srinu, "Qualitative evaluation of enhancement methods for analysis of acute leukemia images," *International Journal of Engineering Science and Technology (IJEST)*, pp. 6447–6453, 2011.

[29] A. N. Aimi Salihah, M. Y. Mashor, N. H. Harun, and H. Rosline, "Colour image enhancement techniques for acute leukaemia blood cell morphological features," *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*, pp. 3677–3682, 2010.

[30] S. Eom, S. Kim, V. Shin, and B. Ahn, "Leukocyte segmentation in blood smear images using region-based active contours," *Advanced Concepts for Intelligent Vision ...*, pp. 867–876, 2006.

[31] D. M. Ushizima, R. T. Calado, and E. G. Rizzatti, "LNCS 4091 - Leukocyte detection using nucleus contour propagation," *Differential Equations*, pp. 389–396, 2006.

[32] J. Angulo and G. Flandrin, "Automated detection of working area of peripheral blood smears using mathematical morphology," *Analytical Cellular Pathology : the Journal of the European Society for Analytical Cellular Pathology*, vol. 25, pp. 37–49, Jan. 2003.

[33] M. Ghosh, D. Das, C. Chakraborty, and A. K. Ray, "Automated leukocyte recognition using fuzzy divergence," *Micron*, 2010.

[34] N. Sinha and A. Ramakrishnan, "Automation of differential blood count," *TENCON 2003. Conference on*, no. i, 2003.

[35] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice Hall, 3 ed., 2007.

[36] E. Gelsema, H. Bao, A. Smeulders, and H. Den Harink, "Application of the method of multiple thresholding to white blood cell classification," *Computers in Biology and Medicine*, vol. 18, no. 2, pp. 65–74, 1988.

[37] I. Cseke, "A fast segmentation scheme for white blood cell images," *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol. IV. Conference D: Architectures for Vision and Pattern Recognition,*, pp. 530–533, 1992.

[38] H. A. Madhloom, H.T., S.A. Kareem, H. Ariffin, A.A. Zaidan and B. Zaidan, "An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold," *Journal of Applied Sciences*, vol. 10, no. 11, pp. 959–966, 2010.

[39] F. Scotti, "Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images," in *CIMSA. 2005 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 2005.*, no. July, pp. 96–101, IEEE, 2005.

[40] N. Ritter and J. Cooper, "Segmentation and border identification of cells in images of peripheral blood smear slides," in *Proceedings of the Thirtieth Australasian Conference on Computer Science-Volume 62*, no. Rowan 1986, pp. 161–169, Australian Computer Society, Inc., 2007.

[41] A. Katz, *Image Analysis and Supervised Learning in the Automated Differentiation of White Blood Cells from Microscopic Images*. PhD thesis, 2000.

[42] K. Jiang, Q. Q.-M. Liao, and S. S.-Y. Dai, "A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering," in *Machine Learning and Cybernetics, 2003 International Conference on*, vol. 5, pp. 2820–2825, IEEE, Ieee, 2003.

[43] A. N. Aimi Salihah, M. Y. Mashor, N. H. Harun, A. A. Abdullah, and H. Rosline, "Improving colour image segmentation on acute myelogenous leukaemia images using contrast enhancement techniques," in *Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on*, no. December, pp. 246–251, IEEE, 2010.

[44] H. T. Madhloom, S. A. Kareem, and H. Ariffin, "An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images," *Journal of Medical Systems*, vol. 36, pp. 2149–58, Aug. 2012.

[45] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, S. Beksac, and M. Beksaç, "Automated contour detection in blood cell images by an efficient snake algorithm," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 47, no. 9, pp. 5839–5847, 2001.

[46] G. Ongun and U. Halici, "An automated differential blood count system," in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, no. 2, pp. 2583–2586, 2001.

[47] L. Yang, P. Meer, and D. J. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society*, vol. 9, pp. 475–86, Sept. 2005.

[48] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 7, pp. 359–69, Jan. 1998.

[49] D. Comaniciu, P. Meer, and D. J. Foran, "Image-guided decision support system for pathology," *Machine Vision and Applications*, vol. 11, pp. 213–224, Dec. 1999.

[50] H. Ramoser, V. Laurain, H. Bischof, and R. Ecker, "Leukocyte segmentation and classification in blood-smear images," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, vol. 4, pp. 3371–3374, IEEE, Jan. 2006.

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

[52] S. Mohapatra and D. Patra, "Automated leukemia detection using hausdorff dimension in blood microscopic images," in *Emerging Trends in Robotics and Communication Technologies (INTERACT), 2010 International Conference on*, pp. 64–68, IEEE, Ieee, Dec. 2010.

[53] S. Mohapatra, S. S. Samanta, D. Patra, and S. Satpathi, "Fuzzy based blood image segmentation for automated leukemia detection," *2011 International Conference on Devices and Communications (ICDeCom)*, pp. 1–5, Feb. 2011.

[54] J. Gonzalez, I. Olmos, L. Altamirano, B. A. Morales, C. Reta, M. C. Galindo, J. E. Alonso, and R. Lobato, "Leukemia identification from bone marrow cells images using a machine vision and data mining strategy," *Intelligent Data Analysis*, vol. 15, pp. 443–462, 2011.

[55] L. B. Dorini, R. Minetto, and N. J. Leite, "White blood cell segmentation using morphological operators and scale-space analysis," *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, pp. 294–304, Oct. 2007.

[56] J. Roerdink and A. Meijster, "The watershed transform: definitions, algorithms and parallelization strategies," *Fundamenta Informaticae*, vol. 41, pp. 1–40, 2000.

[57] S. Beucher, "The watershed transformation applied to image segmentation," *Scanning Microscopy-Supplement*, pp. 299–314, 1992.

[58] J. Angulo, G. Flandrin, M. Math, L. Central, and U. Necker, "Microscopic image analysis using mathematical morphology: Application to haematological cytology," *Science, Technology and Education of Microscopy: An overview*, pp. 304–312, 2003.

[59] J. Angulo, J. Klossa, and G. Flandrin, "Ontology-based lymphocyte population description using mathematical morphology on colour blood images," *Cellular and Molecular Biology*, vol. 52, no. 6, pp. 2–15, 2006.

[60] T. Markiewicz, S. Osowski, and B. Marianska, "White blood cell automatic counting system based on support vector machine," *Adaptive and Natural Computing Algorithms*, pp. 318–326, 2007.

[61] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, and S. Beksac, "Feature extraction and classification of blood cells for an automated differential blood count system," in *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, vol. 4, pp. 2461–2466, IEEE, Ieee, 2001.

[62] S. Sanei and T. K. M. Lee, "Cell recognition based on pca and bayesian classification," *Digital Signal Processing*, no. April, pp. 239–243, 2003.

[63] C. Pan, X. Yan, and C. Zheng, "Recognition of blood and bone marrow cells using kernel-based image retrieval," *IJCSNS International Journal of Computer Science and Network Security*, vol. 6, no. 10, pp. 29–35, 2006.

[64] P. Rodrigues, M. Ferreira, and J. Monteiro, "Segmentation and classification of leukocytes using neural networks: a generalization direction," *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, vol. 392, pp. 373–396, 2008.

[65] T. Markiewicz and S. Osowski, "Data mining techniques for feature selection in blood cell recognition," in *14th European Symposium on Artificial Neural Networks. Bruges: Proceedings-European Symposium on Artificial Neural Networks*, no. April, pp. 26–28, Citeseer, 2006.

[66] R. Siroic, S. Osowski, T. Markiewicz, and K. Siwek, "Support vector machine and genetic algorithm for eficient blood cell recognition," *2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007*, pp. 1–6, May 2007.

[67] H. T. Madhloom, S. A. Kareem, and H. Ariffin, "A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia," *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, pp. 330–335, Nov. 2012.

[68] A. Aimi, M. Mashor, and H. Rosline, "Classification of acute leukaemia cells using multilayer perceptron and simplified fuzzy ARTMAP neural networks," *The international Arab Journal of Information Technology*, vol. 10, no. 4, pp. 356–364, 2013.

[69] L. Benattar and G. Flandrin, "Morphometric and colorimetric analysis of peripheral blood smears lymphocytes in B-cell disorders: proposal for a scoring system," *Leukemia & Lymphoma*, vol. 42, pp. 29–40, June 2001.

[70] D. M. U. Sabino, L. Dafontouracosta, E. Gilrizzatti, M. Antoniozago, L. da Fontoura Costa, E. Gil Rizzatti, and M. Antonio Zago, "A texture approach to leukocyte recognition," *Real-Time Imaging*, vol. 10, pp. 205–216, Aug. 2004.

[71] D. Sabino, L. Costa, E. Rizzatti, and M. Zago, "Toward leukocyte recognition using morphometry, texture and color," *2004 2nd IEEE International Symposium on Biomedical Imaging: Macro to Nano (IEEE Cat No. 04EX821)*, vol. 2, pp. 121–124, 2004.

[72] D. Ushizima, A. Lorena, and A. de Carvalho, "Support vector machines applied to white blood cell recognition," in *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, p. 6 pp., IEEE, 2005.

[73] S. a. H. Jahanmehr, M. Rogers, J. Zheng, R. Lai, and C. Wang, "Quantitation of cytological parameters of malignant lymphocytes using computerized image analysis," *International Journal of Laboratory Hematology*, vol. 30, pp. 278–85, Aug. 2008.

[74] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of hematologic malignancies using texton signatures," *Pattern Analysis and Applications: PAA*, vol. 10, pp. 277–290, Oct. 2007.

[75] L. Yang, O. Tuzel, W. Chen, P. Meer, G. Salaru, L. a. Goodell, and D. J. Foran, "PathMiner: a web-based tool for computer-assisted diagnostics in pathology," *IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society*, vol. 13, pp. 291–9, May 2009.

[76] S. Mircic and N. Jorgovanovic, "Automatic classification of leukocytes," *Journal of Automatic Control*, vol. 16, no. 1, pp. 29–32, 2006.

[77] M. Colunga, O. Siordia, and S. Maybank, "Leukocyte Recognition using EM-algorithm," *MICAI 2009: Advances in Artificial Intelligence*, pp. 545–555, 2009.

[78] C. Reta, L. Robles, and J. Gonzalez, "Segmentation of bone marrow cell images for morphological classification of acute leukemia," *FLAIRS Conference*, no. Flairs, pp. 86–91, 2010.

[79] B. Houwen, "The differential cell count," *Laboratory Hematology*, vol. 7, no. 2, pp. 89–100, 2001.

[80] G. Gutiérrez, A. Merino, A. Domingo, J. M. Jou, and J. C. Reverter, "EQAS for peripheral blood morphology in Spain: a 6-year experience," *International Journal of Laboratory Hematology*, vol. 30, pp. 460–6, Dec. 2008.

[81] E. Alférez, A. Merino, L. Mujica, and J. Rodellar, "Morphological features using digital image processing in lymphoid neoplasias [Abstract]," *International Journal of Laboratory Hematology*, vol. 33, pp. 53–136, May 2011.

[82] A. Materka and M. Strzelecki, "Texture analysis methods - a review," 1998.

[83] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, Nov. 1973.

[84] J. Angulo, "A mathematical morphology approach to cell shape analysis," in *Progress in Industrial Mathematics at ECMI 2006* (L. L. Bonilla, M. Moscoso, G. Platero, and J. M. Vega, eds.), vol. 12 of *Mathematics in Industry*, pp. 2–6, Berlin, Heidelberg: Springer, 2008.

[85] N. Pal and J. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

[86] T. A. Summers and E. S. Jaffe, "Hairy cell leukemia diagnostic criteria and differential diagnosis," *Leukemia & Lymphoma*, vol. 52 Suppl 2, pp. 6–10, June 2011.

[87] T. J. Ross, *Fuzzy Logic with Engineering Applications*. John Wiley & Sons, 2009.

[88] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, 1981.

[89] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[90] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[91] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, 2008.

[92] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[93] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *Neural Networks, IEEE Transactions on*, vol. 12, no. 2, pp. 181–201, 2001.

[94] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge university press, 2000.

[95] D. Zhang and S. Chen, "Fuzzy clustering using kernel method," in *in The 2002 International Conference on Control and Automation, 2002. ICCA, 2002*, Citeseer, 2002.

[96] D.-Q. Zhang and S.-C. Chen, "A novel kernelized fuzzy C-means algorithm with application in medical image segmentation," *Artificial Intelligence in Medicine*, vol. 32, pp. 37–50, Sept. 2004.

[97] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 34, pp. 1907–1916, Aug. 2004.

[98] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics : the Official Journal of the Computerized Medical Imaging Society*, vol. 30, pp. 9–15, Jan. 2006.

[99] F. Meyer, "Topographic distance and watershed lines," *Signal Processing*, vol. 38, no. 1, pp. 113–125, 1994.

[100] "Image Segmentation and Mathematical Morphology." Centre of mathematical morphology, MINES ParisTech. http://cmm.ensmp.fr/˜beucher/wtshed.html. Accesed August, 2014.

[101] N. Nikolaou and N. Papamarkos, "Color reduction for complex document images," *International Journal of Imaging Systems and Technology*, vol. 19, no. 1, pp. 14–26, 2009.

[102] C. Briggs, N. Culp, B. Davis, G. D'Onofrio, G. Zini, and S. J. Machin, "ICSH guidelines for the evaluation of blood cell analysers including those used for differential leucocyte and reticulocyte counting," *International Journal of Laboratory Hematology*, Mar. 2014.

[103] S. Alférez, A. Merino, L. E. Mujica, M. Ruiz, L. Bigorra, and J. Rodellar, "Automatic classification of atypical lymphoid B cells using digital blood image processing," *International Journal of Laboratory Hematology*, vol. 36, pp. 472–80, Aug. 2014.

[104] H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li, "Conditional mutual information-based feature selection analyzing for synergy and redundancy," *ETRI Journal*, vol. 33, pp. 210–218, Apr. 2011.

[105] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 27–66, 2012.

[106] I. Jolliffe, *Principal Component Analysis*. Wiley Online Library, 2005.

[107] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, Sept. 1936.

[108] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," *Computer Graphics and Image Processing*, vol. 18, pp. 236–258, Mar. 1982.

[109] M. Bevk and I. Kononenko, "A statistical approach to texture description of medical images: a preliminary study," in *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)*, pp. 239–244, IEEE Computer. Soc, 2002.

[110] L. Tesar, D. Smutek, A. Shimizu, and H. Kobatake, "3D extension of Haralick texture features for medical image analysis," in *SPPR 2007 Proceedings of the Fourth Conference on IASTED International Conference*, no. 1, pp. 350–355, 2007.

[111] F. Albregtsen, "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," tech. rep., Image Processing Laboratory, Department of Informatics, University of Oslo, 1995.

[112] S. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 2091–2110, 1989.

[113] G. Van de Wouwer, P. Scheunders, and D. Van Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 8, pp. 592–8, Jan. 1999.

[114] A. Latif-Amet, A. Ertüzün, and A. Erçil, "An efficient method for texture defect detection: sub-band domain co-occurrence matrices," *Image and Vision Computing*, vol. 18, pp. 543–553, May 2000.

[115] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern Recognition Letters*, vol. 24, pp. 1513–1521, June 2003.

[116] C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. John Wiley & Sons, 2011.

[117] J. Serra, *Image Analysis and Mathematical Morphology*. Academic press, 1982.

[118] J. Angulo, *Morphologie mathématique et indexation d'images couleur: application à la microscopie en biomédecine*. PhD thesis, Mines ParisTech, 2003.

[119] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[120] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–38, Aug. 2005.

[121] H. Yang and J. Moody, "Feature selection based on joint mutual information," in *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pp. 22–25, 1999.

[122] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, Dec. 2004.

[123] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, ACM, 1992.

[124] C.-W. Hsu, C.-C. Chang, C.-J. Lin, and Others, "A Practical Guide to Support Vector Classification," 2003.

[125] W. Burger and M. J. Burge, *Digital Image Processing: an Algorithmic Introduction Using Java*. Springer, 2009.

[126] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2008.

[127] "A Tutorial on Clustering Algorithms." Dipartimento di Electronica, Informazione e Bioingenieria. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/. Accessed August 20, 2014.

[128] G. Kumar and P. K. Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems," *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, pp. 5–12, 2014.

[129] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[130] M. Natrella, "NIST/SEMATECH e-handbook of statistical methods," 2010.

[131] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

Appendix A

Performance classification parameters

In this thesis, the confusion matrix has been applied to evaluate the performance of the classification experiments. In the confusion matrix each column represents the predicted classes, while each row represents the true classes. For example, Table A.1 was obtained in Chapter 5 from a classification by Linear Discriminant Analysis (LDA) (using 10-fold cross validation) of five types of lymphoid cells (classes): normal lymphocytes (N), Hairy Cell Leukemia (HCL), Chronic Lymphocytic Leukemia (CLL), Mantle Cell Lymphoma (MCL) and B prolymphocytes (BPL). In this table, the true and confirmed diagnosis correspond to the rows, while the predicted lymphoid cell types correspond to the columns. For instance, 542 CLL cells were predicted as: 5 N, 0 HCL, 535, CLL, 1 MCL and 1 PBL. It is important to remark that the diagonal of the confusion matrix correspond to the true positives (classes predicted correctly).

In order to calculate some useful performance classification parameters based on the confusion matrix, some definitions for a *specific type of cell* (e.g. CLL) are described as follows.

TABLE A.1: *Confusion Matrix of the LDA classification and 10-fold cross-validation for the training set.*

		Predicted*				
		N	**HCL**	**CLL**	**MCL**	**BPL**
True	**N**	180	0	0	0	1
	HCL	5	294	0	0	2
	CLL	5	0	535	1	1
	MCL	5	0	0	391	5
	BPL	3	1	0	0	71

* The rows represent the true diagnosis and the columns the predicted diagnosis given by the classification algorithm for each type of lymphoid cell. Accuracy = 98.07 %.

Performance classification parameters

True positives (TP) Predicted cells that were correctly classified.

False positives (FP) Predicted cells that were incorrectly classified as CLL cells.

False negatives (FN) CLL cells that were incorrectly classified as other cell type.

True negatives (TN) All the remaining cells correctly classified as non CLL cells.

Following the example for the CLL cells, the TP correspond to the diagonal value of 535. The FP are obtained summing all the values of the CLL column except the TP, in this case 0. The FN are the sum of all values in the CLL row except the TP, which in this case is 7. Finally, the TN are calculated by the sum of all the values of the confusion matrix except the row and the column corresponding to the CLL type, i.e. 958.

From these basic definitions, several performance classification parameters can be computed:

Sensitivity or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$$

Specificity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{FP+TN}$$

Precision or positive predictive value (PPV)

$$PPV = \frac{TP}{TP+FP}$$

Accuracy (ACC)

$$ACC = \frac{\#Correct}{Total}$$

where *#Correct* is the number of correct classifications and *Total* is the complete number of cells under study. The accuracy can also be obtained from the confusion matrix by summing all the diagonal values and dividing it by the sum of all the values of the matrix.

Following the example of Table A.1, for the CLL type the above parameters in percentage are: *TPR = 99.45%, TNR = 100%, PPV = 100%* and *ACC = 98.07%*.

Appendix B

Technical glossary

10-fold cross validation This technique randomly divides the data set into 10 equal size subsets. A single subset is used as the testing data, while the remaining data are used for training. Then, the process is repeated 10 times. Finally, a confusion matrix was obtained to calculate some overall statistical measures.

Active contours Active contours, or snakes, are computer-generated curves that move within images to find object boundaries [48].

Binary mask It is the usual representation of a segmented region of the image, where the pixels have the value 1 if they are within the region and 0 if they are outside.

Classification It is the process of automatically assigning classes to objects based on the information extracted from them [126].

Clustering It is the process of organizing objects into groups (clusters) whose members are similar in some way. Then, a cluster collects objects which are similar between them and are dissimilar to the objects that belong to other clusters [127].

Color space It is a organized representation of colors following some standard. Generally, a color space follows a mathematical color model that specifies each color as a point on a coordinate system [35].

Conditional information $I(X;Y|Z) = H(X|Z) - H(X|YZ)$ It is the information still shared between X and Y after the value of a third variable, Z, is revealed [105].

Digital Image Processing (DIP) DIP consists of applying multiple algorithms to process digital images by a digital computer [35].

Discrete Wavelet Transform (DWT) While Fourier transform decomposes the signal into sines and cosines (functions localized in Fourier space), the wavelet transform uses functions that are localized in both the real and Fourier space. The DWT is an implementation of the wavelet transform using a discrete set of the wavelet scales and translations with some rules. Particularly, this transform decomposes the signal into mutually orthogonal set of wavelets [112, 113].

Distance transform For each pixel in a binary image, the distance transform assigns a number that is the distance between that pixel and the nearest nonzero pixel of the image [126].

Entropy (information) The entropy $H(X)$ quantifies the uncertainty present in the distribution of a random variable (see 6.6). If there is little uncertainty over the outcome, then the entropy is low. If all events are equally likely, that is, there is maximum uncertainty over the outcome, then the entropy is maximal [105].

Feature extraction The main objective of feature extraction is to obtain information of the objects of interest of the digital image. It is a form of dimensionality reduction, because the image is represented by a set of features (feature vector) [35, 128].

First order statistical features These features are based on the histogram of a grayscale digital image. From the simple statistical information about the image supplied by the histogram, several first order statistical features can be obtained, such as: mean, standard deviation, skewness, kurtosis, energy and entropy [82].

Fuzzy C-means (FCM) It is an unsupervised method that partitions the data into clusters, minimizing the distance between each data point in the cluster and its center, and maximizing the distance between cluster centers. However, under certain restrictions each data point can belong to several groups at the same time in a fuzzy way [87].

Granulometric curves The granulometric curve is a size-histogram in which a high value at a specific size indicates the presence of many bright structures (or dark structures) with similar size in the image [118].

Granulometry It is a field of the mathematical morphology that deals with determining the size distribution of particles in an image [35, 118, 126].

Gray Level Co-ocurrence Matrix (GLCM) represents the joint probability $P(i, j)$ that a *pair of pixels* have intensity values of i and j, respectively, at a distance d in a particular direction θ. This probability can be calculated as the frequency count of occurrences

(second order histogram) divided by the total number of neighbouring pixels. Thus, the co-occurrence matrix considers not only the information about the intensity values, but also the position of the pixels with similar intensities [82].

Histogram The histogram is a discrete function that shows the number of pixels *H(i)* on the image having the pixel intensity value *i* (frequencies) [35].

Image Gradient The gradient of a image is based on local derivatives of the image. Then, each pixel of the gradient image shows the intensity change of that point on the original image, in a particular direction [35, 125].

k-Nearest Neighbors (k-NN) It is a non parametric method used for classification (or regression). The input consist of the k closest training examples in the feature space. The output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor [129].

Kernel trick It consists in the operation of the kernel functions in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but simply calculating the inner products between the images of all pair of data in the feature space [91].

Kurtosis It is a measure of whether the data are peaked or flat relative to a normal distribution [130].

Linear Discriminant Analysis (LDA) LDA This method maximizes the ratio between-class variance to the within-class variance in any particular data set, ensuring maximum separability. This method can be used for classification or dimensionality reduction [129].

Mathematical morphology It is a set of mathematical tools for extracting useful image components to represent and describe regions of interest of the image, based on set theory [35, 118, 126].

Membership image In fuzzy logic, elements of a fuzzy set have varying degrees of membership in the set. They are mapped to an universe of membership values. If this sets are images they respective representation are membership images [87].

Mutual information $I(X;Y) = H(X) - H(X|Y)$ It is the amount of uncertainty in X which is removed by knowing Y, i.e. the amount of information that one variable provides about another [105].

Naive Bayes classifier The naive Bayes classifier is designed for use when predictors are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. It classifies data in two steps: 1) Using the training data, the method estimates the parameters of a probability distribution, assuming predictors are conditionally independent given the class. 2) For any unseen test data, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test data according the largest posterior probability [131].

Preprocessing Image preprocessing suppresses information that is not relevant to specific image processing tasks. Preprocessing consists of a set of image processing operations to enhance certain fine features in the data and to remove certain noise [126].

Principal Component Analysys (PCA) PCA is a technique that is commonly used to reduce the dimensionality of a big dataset by its transformation into a new set of principal components linearly uncorrelated, searching the causes of variability and sorting the components by their importance [106].

Region-based segmentation It segments an image into a number of regions, where for each pixel a criteria decides or estimates which class it belongs to [35].

Second order statistical features They provide more information than the first order statistical features. These features are based on a joint probability that a pair of pixels have a particular combination of a pair of values at a specific distance in a given direction [35, 82, 83].

Segmentation It consists of subdividing the images into its constituent regions or objects, selecting mainly the objects of interest [35].

Skewness It s a measure of symmetry, or lack of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point [130].

Support Vector Machines SVM defines a well known and widely used classification (and regression) technique, which aims to produce a model using training data sets that can predict the corresponding class labels of validation data sets. An SVM model is a representation of the examples as points in space, mapped so that the examples of the

separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on (see Section 6.7).

Texture Texture refers to properties that represent the surface or structure of an object, but it has not precise definition because to its wide variability [126].

Thresholding segmentation It is the simplest segmentation method. By a threshold value it can decide which pixels make the object of interest and which pixels are just the background of the image.

Watershed Transformation (WT) A grey-level image might be seen as a topographic relief. The concept of Watershed transformation is based on visualizing the maximum and minimum intensity values as peaks and basins. Then, water falling on this relief flows to reach a minimum. Intuitively, the watershed of a relief corresponds to the limits of the adjacent basins of the water regions [57, 100].