UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

SEGMENTATION AND CLASSIFICATION OF HUMAN SPERM HEADS TOWARDS MORPHOLOGICAL SPERM ANALYSIS

TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS
MENCIÓN COMPUTACIÓN

VIOLETA NOEMÍ CHANG CAMACHO

PROFESOR GUÍA:
NANCY HITSCHFELD KAHLER

PROFESOR CO-GUÍA:
STEFFEN HÄRTEL GRUNDLER

PROFESOR CO-GUÍA:
LAURENT HEUTTE

MIEMBROS DE LA COMISIÓN:
MARIA CECILIA RIVARA ZUÑIGA
BENJAMIN BUSTOS CARDENAS
ALCEU DE SOUZA BRITTO JUNIOR

Este trabajo ha sido parcialmente financiado por CONICYT.

SANTIAGO DE CHILE
2015

Resumen

La infertilidad es un problema clínico que afecta hasta a 15% de parejas en edad reproductiva, con implicancias tanto emocionales como fisiológicas. Un análisis de semen es el primer paso en la evaluación de una pareja infértil. El énfasis en identificar no sólo cabezas normales de espermatozoides sino también categorías de cabezas anormales puede tener una significativa utilidad clínica al decidir por un tratamiento de fertilidad. Esta tesis propone una nueva metodología para detectar, segmentar, caracterizar y clasificar cabezas de espermatozoides humanos, con el objetivo de facilitar el posterior análisis morfológico, para diagnósticos de fertilidad, toxicología reproductiva, investigación básica o estudios de salud pública.

En la primera parte de este tesis, se ha tratado la detección y segmentación de cabezas de espermatozoides humanos. En este sentido, se propone un gold-standard para segmentación de espermatozoides construido con la cooperación de un experto referente en el área, para comparar métodos para detección y segmentación de espermatozoides. Además, se ha desarrollado un framework para la detección y segmentación de componentes de cabezas de espermatozoides humanos (incluyendo acrosoma y núcleo) que usa tres espacios de color además de técnicas de clustering y análisis estadístico del histograma. La evaluación experimental muestra que el método propuesto mejora el desempeño del estado del arte. Los resultados logran 98% de detección correcta a expensas de un número menor de falsos positivos, comparado con el estado del arte. Así mismo, los resultados de segmentación de cabeza, acrosoma y núcleo muestran más de 80% de solapamiento comparado con las máscaras de segmentación manual del gold-standard.

En la segunda parte de esta tesis, el enfoque estuvo en la caracterización y clasificación de cabezas de espermatozoides humanos. Así, se introduce un gold-standard para clasificación de cabezas de espermatozoides humanos, construido con la colaboración de tres expertos referentes en área, y de acuerdo al criterio de la OMS. Además, se ha formulado un nuevo descriptor para cabezas de espermatozoides que, combinado con otros descriptores basados en forma, permite discriminar entre cabezas de espermatozoides normales y anormales, identificando cuatro tipos de cabezas anormales. También se propone un esquema de clasificación, que permite categorizar las cabezas de espermatozoides en 5 clases diferentes, según la OMS. La evaluación experimental muestra que el esquema propuesto tiene mejor desempeño que distintos clasificadores monolíticos, así como varios esquemas de clasificación en cascada que fueron diseñados en el contexto de esta investigación. Los resultados muestran más de 70% de clasificación correcta usando un dataset de total concordancia entre expertos del área.

Abstract

Infertility is a problem that affects up to 15% of couples worldwide with emotional and physiological implications. Emphasis on identifying not only normal sperm heads, but also categories of abnormal, may have significant clinical utility when deciding on an infertility treatment. This thesis proposes a new methodology for accurately detecting, segmenting, characterizing and classifying human sperm heads which facilitates a further morphological analysis for fertility diagnoses, reproductive toxicology, basic research or public health studies.

In the first part of this thesis, we focused on detection and segmentation of human sperm heads. In this sense, we propose a segmentation gold-standard built with the cooperation of a referent expert in the field to compare methods for detecting and segmenting sperm cells. We also present a two-stage framework for detection and segmentation of human sperm head characteristics (including acrosome and nucleus) that uses three different color spaces. The first stage detects regions of interest (ROIs) that define sperm heads, using $k-$means. Candidate heads are then refined using mathematical morphology. In the second stage, we work on each ROI to accurately segment the sperm head as well as nucleus and acrosome, using clustering and histogram statistical analysis techniques. Our experimental evaluation shows that our proposed method outperforms the state-of-the-art. Our results achieve above 98% in the sperm head detection process at the expense of having significantly fewer false positives obtained by the state-of-the-art method. Our results also show an accurate head, acrosome and nucleus segmentation achieving over 80% overlapping against the hand-segmented gold-standard.

In the second part of this thesis, we focused on characterizing and classifying human sperm heads. Thus, we introduce a classification gold-standard for evaluating morphological sperm analysis methods, built with the active collaboration of three referent experts in the field. We present a new descriptor for human sperm heads that combined with different shape descriptors allows for discrimination among normal and abnormal sperm heads; separating four different types of abnormal heads. We also present a two-stage classification scheme for classifying human sperm heads using a combined classifier approach together with an ensembled feature selection technique. Our experimental evaluation shows that our proposed scheme outperforms a number of monolithic classifiers, as well as different cascade classification schemes designed in the context of this research. Our results achieved more than 70% of classification accuracy on a dataset with total agreement among domain experts.

To my encouraging husband José and to my lovely daughter Adri.

Acknowledgements

I have been very privileged to have Nancy Hitschfeld as my advisor. I appreciate all her contributions of time, ideas and funding to make my PhD experience productive and stimulating. Her support in the PhD pursuit, specially during tough times, was very encouraging for me. I am also thankful for the excellent example she has provided me as a successful computer science woman and professor.

I thank my co-advisor Steffen Härtel for welcoming me in his lab and the continuous support over the years. I am grateful for his guidance and the opportunities he has afforded me. He has taught me how good research is done.

I would like to express my sincere gratitude to my co-advisor Laurent Heutte for the continuous support of my PhD research, for his time, patience, motivation, enthusiasm, and immense knowledge. I have been stimulated by his constant flow of good ideas.

I also thank all members of SCIAN-Lab for their support and help in different stages of my PhD work: Alejandra García, Susana Vargas, Jorge Jara, Víctor Castañeda and Erika Labbe. I would like to thank Luis Sarabia for his patient while teaching me about morphological sperm analysis.

I would like to thank Caroline Petitjean, José Saavedra, Luiz Oliveira, Alceu Britto, Alessandro Koerich and Simon Bernard for their encouragement, and insightful comments in the last stage of my thesis.

Last but not the least, I would like to thank my family for all their love and encouragement. Thanks to my parents Violeta and Ricardo who supported me in all my pursuits. Thanks to my loving, supportive, encouraging, and patient husband José whose faithful support during all my PhD pursuit is very appreciated. Thanks to my lovely daughter Adri who kindly understood and supported my work and gave me the motivation to finish this thesis.

Contents

Resumen i
Abstract ii
Acknowledgements iv
Table of Contents v
List of Figures viii
List of Tables x

1 Introduction **1**
1.1 Thesis Impact 3
1.2 Thesis Contribution 3
1.3 Thesis Methodology 4
1.4 Thesis Deliverables 6
1.5 Thesis Outline 6

2 Background **8**
2.1 Biological Concepts 8
2.1.1 Sperm Structure 8
2.1.2 Spermiogram 10
2.1.3 Sperm Defects 11
2.1.4 Morphological Sperm Analysis 12
2.2 Image Processing 12
2.2.1 Pixel Connectivity 13
2.2.2 Mathematical Morphology 14
2.2.3 Anisotropic Diffusion 15
2.2.4 Active Contours 16
2.2.5 Color Spaces 17
2.3 Pattern Recognition 18
2.3.1 Shape-based Feature Extraction 18
2.3.2 Pattern Classification 19
2.3.3 Base Classifiers 20
2.3.4 Combination of Classifiers 22
2.3.5 Combination Rules 23
2.3.6 Clustering Methods 23
2.3.7 Dimensionality Reduction Techniques 24
2.4 Evaluation Metrics 25
2.4.1 Evaluation Metrics for Detection 25
2.4.2 Evaluation Metrics for Segmentation 26

2.4.3 Evaluation Metrics for Classification . . . 26
2.5 Summary . . . 27

3 Related Work **29**
3.1 Detection and Segmentation . . . 30
3.1.1 Human Sperm Cells . . . 30
3.1.2 Other Eukaryotic Cells . . . 31
3.2 Characterization and Classification . . . 31
3.2.1 Sperm Cells . . . 32
3.2.2 Grain Kernels . . . 33
3.3 Commercial Applications for Computer-Aided Sperm Analysis . . . 36
3.3.1 IVOS - Integrated Visual Optic System . . . 36
3.3.2 SCA - Sperm Class Analyzer . . . 37
3.4 Summary . . . 37

4 Gold-Standard Generation **38**
4.1 Sample Preparation . . . 38
4.2 Image Acquisition . . . 39
4.3 Source of Sperm Smears . . . 39
4.4 Segmentation Gold-Standard . . . 39
4.5 Classification Gold-Standard . . . 39
4.6 Summary . . . 45

5 Sperm Head Segmentation **46**
5.1 Detection of Sperm Head . . . 46
5.2 Segmentation of Sperm Head . . . 48
5.3 Segmentation of Sperm Nucleus and Acrosome . . . 50
5.4 Detection and Segmentation Algorithm . . . 51
5.5 Experimental Results . . . 51
5.5.1 Parameter Optimization . . . 51
5.5.2 Sperm Detection . . . 52
5.5.3 Sperm Head Segmentation . . . 53
5.6 Discussion . . . 55
5.7 Summary . . . 59

6 Sperm Head Characterization **60**
6.1 Shape Representation . . . 60
6.2 Feature Extraction . . . 61
6.2.1 Global Shape-based Measures . . . 61
6.2.2 Shape-based Descriptors . . . 64
6.2.3 Proposed Morphological Descriptor . . . 67
6.3 Feature Selection . . . 68
6.4 Summary . . . 70

7 Sperm Head Classification **71**
7.1 Classification Scheme . . . 72
7.1.1 Stage 1: Separation of amorphous sperm heads . . . 72
7.1.2 Stage 2: Verification of potential classes . . . 74

7.2 Experimental Results . . . 74
7.2.1 Dataset . . . 74
7.2.2 Stage 1: Separation of amorphous sperm heads . . . 75
7.2.3 Stage 2: Verification of potential classes . . . 77
7.2.4 Complete Classification Scheme . . . 80
7.3 Discussion . . . 86
7.4 Summary . . . 87

8 Conclusions and Future Work **88**
8.1 Future Work . . . 89

Bibliography **91**

A Sperm Segmentation Method by Carrillo et al. **106**
A.1 Method Description . . . 106
A.1.1 Detection and extraction of sperm cells . . . 106
A.1.2 Segmentation of sperm cells . . . 107
A.2 Algorithms . . . 109
A.3 Implementation Details . . . 109

B Discarded Classification Schemes **111**
B.1 Monolithic multi-class classifiers . . . 111
B.2 First cascade of two-class classifiers . . . 113
B.3 Second cascade of two-class classifiers . . . 114
B.4 Third cascade of two-class classifiers . . . 115
B.5 Justification on choosing *the* classification scheme . . . 116

List of Figures

1.1 Slight variations among sperm classes 2

2.1 Morphology of the normal human sperm 9
2.2 Sperm cells with vacuoles in their acrosomal region. 9
2.3 Human sperm abnormalities 11
2.4 Neighbors of point P 13
2.5 Monolithic vs. combined classifier approach 22

4.1 Segmentation gold-standard 40
4.2 Classification gold-standard 42
4.3 Inter-expert agreement 43
4.4 Partial and total inter-expert agreement 43
4.5 Inter-expert variability in five-class classification 44
4.6 Inter-expert variability in two-class classification 44

5.1 Detection of sperm heads 47
5.2 Segmentation of sperm heads 48
5.3 Segmentation of nucleus and acrosome 50
5.4 ROC curves for sperm head detection 52
5.5 Results of head, acrosome and nucleus segmentation 54
5.6 Dice coefficient for head, acrosome and nucleus 55
5.7 Probability density function for Dice coefficient 56
5.8 Hausdorff distance for head, acrosome and nucleus 57
5.9 Non-detected sperm cells 58

6.1 Shape representation of sperm heads 61
6.2 Quadrant fitness 68
6.3 Bilateral symmetry 69

7.1 High-level architecture of the proposed classification approach 72
7.2 First stage of the classification scheme 73
7.3 Second stage of the classification scheme 75
7.4 Impact of testing datasets 82
7.5 Comparison of accuracy per class using DS3 achieved with the proposed classification scheme and different monolithic classifiers 83
7.6 Comparison of accuracy per class using DST achieved with the proposed classification scheme and different monolithic classifiers 83

7.7 Inter-expert and automatic classification variability in five-class classification 84
7.8 Inter-expert and automatic classification variability in two-class classification 85

B.1 Comparison of accuracy per class using DS3 with different classification schemes 118
B.2 Comparison of accuracy per class using DST with different classification schemes 118

List of Tables

2.1 Metrics for quality of segmentation 26

4.1 Sample source details 41
4.2 Inter-expert agreement 42
4.3 Inter-expert variability 43

5.1 Variation of parameters for our proposed method 52
5.2 Detection accuracy 53

6.1 Summary of shape-based measures included in the proposed morphological descriptor 68
6.2 Summary of used shape-based descriptors 69

7.1 Dataset partition 76
7.2 Ranking of descriptor combinations for stage 1 76
7.3 Results of the first stage of the classification scheme 77
7.4 Selected descriptors for four verifiers in stage 2 78
7.5 Results of verifier 1 : normal vs. amorphous 78
7.6 Results of verifier 2 : tapered vs. amorphous 79
7.7 Results of verifier 3 : pyriform vs. amorphous 79
7.8 Results of verifier 4 : small vs. amorphous 79
7.9 Results of the whole classification scheme for five-class classification 81
7.10 Results of the whole classification scheme for two-class classification 82
7.11 Results of applying dimensionality reduction techniques using MLE intrinsic dimension 85
7.12 Results of applying dimensionality reduction techniques using the best number of dimensions 86

A.1 Variation of parameters for Carrillo's method 110

B.1 Results of monolithic multi-class classifiers using testing dataset DS3 112
B.2 Results of monolithic multi-class classifiers using testing dataset DST 112
B.3 Results of first cascade of two-class classifiers using testing dataset DS3 . . . 113
B.4 Results of first cascade of two-class classifiers using testing dataset DST . . . 114
B.5 Results of second cascade of two-class classifiers using testing dataset DS3 . 115
B.6 Results of second cascade of two-class classifiers using testing dataset DST . 115
B.7 Results of third cascade of two-class classifiers using testing dataset DS3 . . 116
B.8 Results of third cascade of two-class classifiers using testing dataset DST . . 117

Chapter 1

Introduction

Infertility is a problem that affects up to 15% of couples worldwide [WHO, 2010a]. This condition has emotional and physiological implications including stress, depression or sexual dysfunction [Domar et al., 1992]. A semen analysis according to standard criteria [WHO, 2010b], is the first step in the evaluation of the male factor and sets the basis for all posterior steps for medical treatment of the couple [Nafisi et al., 2005]. A typical spermiogram considers concentration, motility, vitality, and/or the fragmentation of the spermatic DNA. In addition, the morphology of the sperm cells is considered as an important parameter to elucidate the potential fertility of a sample [Katz et al., 1986]. As a result of morphological semen analysis, all the sperm cells in the semen sample are classified as *normal* or *abnormal* [Auger, 2010]. Many studies have demonstrated the close relationship between fertility and morphologically normal sperm [Moench and Holt, 1931, MacLeod and Gold, 1951, Kruger et al., 1986, Enginsu et al., 1991, Kobayashi et al., 1991].

The morphology of human sperm is considered to be a clinical tool dedicated to the fertility prognosis and serves, mainly, for making decisions regarding the options of assisted reproduction technologies (ART). Furthermore, in addition to a rigorous application of existing guidelines and respect to high laboratory standards, emphasis on identifying the categories of abnormal sperm heads may have significant clinical utility when deciding on an infertility treatment. For example, a significant increase in percentage of sperm with elongated heads may be due to stress caused by a male urogenital genital tract infection [Menkveld et al., 2011]. Also for clinical decisions, morphological details of sperm considered abnormal, such as diagnosing a genetic anomaly, can expedite the process of deriving directly to *in-vitro* fertilization/intracytoplasmic sperm injection (IVF/ICSI) [Menkveld et al., 2011]. There is also clinical significance regarding the shape of sperm heads, and each class is associated with different genetic and environmental factors that impact clinical decisions pertaining to an infertility treatment. Therefore, a complete analysis of not only normal sperm but also abnormal sperm turns out to be critical in this context. Specifically, a complete classification of head shapes becomes very important. However, the classification of abnormal sperm morphology is a difficult task since the spectrum of possible malformations is considerably wide [Kruger et al., 1993].

There is evidence from previous decades that the quantification of abnormalities is a

challenging task. In 1966, a comparative study in 47 laboratories dedicated to human sperm morphological analysis showed that the traditional method of performing the analysis was *personality oriented*, as well as *subjective, qualitative, non repeatable and difficult to teach to students and technicians* [Katz et al., 1986, Freund, 1966]. Despite of the fact that the classification rules for morphological semen analysis have been simplified [WHO, 2010b], the visual analysis of sperm morphology still presents a substantial challenge concerning reproducibility and objectivity (see Figure 1.1), while inter and intra observer still presents a well known problem [Barroso et al., 1999, Auger et al., 2000, Soler et al., 2003, Cipak et al., 2009, Rivera-Montes et al., 2013]. There are many authors revealing a lack of standardization of the methods used in laboratories in many countries [Walczak-Jedrzejowska et al., 2013, Rivera-Montes et al., 2013]. A sophisticated computational analysis might help overcome these problems.

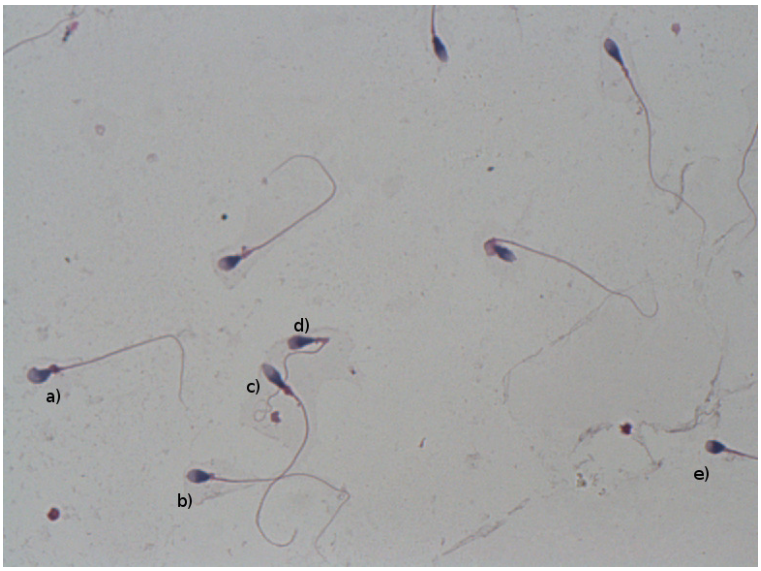


Figure 1.1: **Slight variations among sperm classes.** Representative bright field image from a semen sample, containing five different sperm classes: **a)** amorphous, **b)** normal, **c)** tapered, **d)** pyriform, and **e)** small (Image size: 277×144 pixels $\approx 58 \times 30$ μm).

Overall, the evaluation of cellular and sub-cellular regions (size of the sperm head, tail length, residual cytoplasm area, etc.) and pattern recognition (multiple heads or tails, absent tail, coiled tail, etc.) are required for categorizing defects according to normal and abnormal sperm definitions in visual sperm classification under the microscope [Auger, 2010]. An alternative to replacing the poor visual ability to assess the size and shape of sperm is to analyze the sperm morphology with the help of a computer [Auger, 2010].

Currently, there are Computer-Aided Sperm Analysis (CASA) systems. They were primarily developed to measure sperm concentration, the percentage of motile sperm and some details of sperm movement. To obtain useful information from the visual assessment of sperm morphology, it is crucial some kind of standardization of methods and variables to be analyzed [Wang et al., 1991, Davis and Gravance, 1993, Lacquet et al., 1996, Coetzee et al., 1999, Auger, 2010]. Although there are some commercial applications for sperm morphology assessment, none of these study abnormal sperm in depth, which has been proven to have

a significant impact in research. In addition, these kinds of systems are sold as *black boxes* without any possibility of modification for research purposes, and, of course, there are no publications related to the algorithms implemented in these systems.

This PhD thesis set out to develop, implement and calibrate a novel methodology to accurately detect, segment, characterize and classify sperm heads in the context of morphological sperm analysis with results within the variability among those of referent experts in the field, focusing on a depth analysis of abnormal sperm heads for fertility diagnosis, prognosis, reproductive toxicology, basic research or public health studies.

1.1 Thesis Impact

The results presented in this thesis will have a direct impact on social and public health. From the view point of education and training of new specialists, the classification gold-standard generated in this research is very relevant. There is no collection of semen smear images obtained from Chilean laboratories and manually labeled by referent domain experts. Counting with variability analysis and detailed labels from each expert, it will be possible to find out cases with difficult diagnoses that may require a deep review while training future andrologists.

From the view point of medical diagnosis, this research will have significant impact. As a result of this research, a pipeline for classifying sperm heads trained with experience from different domain experts was generated. The implementation of this pipeline is expected to increase the accuracy and efficiency of medical diagnosis. In this sense, laboratories without experts in morphological sperm analysis in their permanent staff will have the opportunity to remotely access this supporting tool.

Finally, it is worth mentioning that this work will have a significant impact in the image processing community, because it proposes a new approach for segmenting and describing biological microscopic structures under certain characteristics and conditions. Furthermore, this approach could be slightly modified to be applied to another kind of cells, different from sperm, facing the very challenging problem of cell segmentation in image processing research.

1.2 Thesis Contribution

In detail, this PhD thesis has the following contributions:

- We introduced a gold-standard[1] for sperm parts segmentation [Chang et al., 2014]. This dataset was built with the active cooperation of a referent expert in the field and contains twenty images with more than two hundred sperm cells plus hand-segmented masks. This gold-standard has been used to evaluate and compare our detection and segmentation results with the only reproducible detection and segmentation method

[1] Available in *http://morfologia.cedai.cl/public/*

that has been published in the past and therefore represents our state-of-the-art reference method.

- We developed a framework for detecting and segmenting human sperm heads that outperformed the state-of-the-art [Chang et al., 2014], since a reliable detection and segmentation presents the first step for all posterior classification algorithms. This fully automatic approach is based on a clustering method for detecting sperm heads, combining different color spaces as well as on image processing techniques specially adapted for this application. Another contribution is the proposal of a novel algorithm to determine which direction the sperm head points. This is a very important issue for posterior stages in the quest for an accurate morphological analysis.
- We introduced a gold-standard for sperm head classification [Chang et al., 2015] according to World Health Organization (WHO) criteria [WHO, 2010b]. This gold-standard was built with the collaboration of three referent domain experts. The combination of the three experts labels was performed using majority voting. We analyzed inter-expert variability.
- We proposed a morphological descriptor, based on single shape-based measures, for extracting features from segmented sperm heads, as well as a descriptor combining approach to take advantage of different descriptors proposed in the literature.
- We developed and validated a pipeline for sperm head classification, according to WHO criteria [WHO, 2010b]. In this sense, we proposed a classification two-stage scheme that permits classifying sperm heads among five different classes (one class for normal sperm heads and four classes for abnormal sperm heads) combining an ensembled strategy for feature selection and a cascade approach with SVM multi-classifiers. We conducted experiments with results within the variability among those of domain experts.

1.3 Thesis Methodology

This PhD thesis set out to develop, implement and calibrate a novel methodology to detect, segment, characterize and classify sperm heads in the context of morphological sperm analysis, focusing on a depth analysis of abnormal sperm heads for fertility diagnosis, prognosis, reproductive toxicology, basic research or public health studies (See Algorithm 1). This research followed the next procedure:

1. As an initial step in this research, a gold-standard[2] for sperm parts segmentation was built with the active cooperation of a referent expert in the field (Section 4.4). This dataset consists of 210 valid sperm cells. For each sperm cell, there are five hand-segmented masks: head, acrosome, nucleus, mid-piece and tail. Thus, the output of this step is a segmentation gold-standard to validate and evaluate sperm detection and segmentation methods.
2. Since a reliable detection and segmentation presents the first step for all posterior classification algorithms, a framework for detecting and segmenting human sperm head, acrosome and nucleus was developed (Chapter 5). The first stage was aimed to detect

[2]Available in *http://morfologia.cedai.cl/public/*

sperm heads (Section 5.1, Algorithm 2), while the second stage was devoted to segment sperm head (Section 5.2, Algorithm 3), acrosome and nucleus (Section 5.2, Algorithm 5) from detected sperm heads in previous stage (Section 5.2). The main idea beyond our framework was a fusion of color space combinations, $k-$means approach and mathematical morphology operations. In addition, our method proposed an ellipse fitting based algorithm to identify the head front direction (Section 5.2, Algorithm 4). The only reproducible detection and segmentation method that has been published in the past [Carrillo et al., 2007b, Carrillo et al., 2007a] has been implemented and represents the state-of-the-art reference method for this part of the thesis (Appendix A). The segmentation gold-standard has been used to evaluate and compare our detection and segmentation results with the state-of-the-art method. Thus, the contribution of this step to the main goal of this thesis are the sperm detection algorithm and the sperm head segmentation algorithm.

3. Having the segmented heads, the next step was the characterization and classification of sperm heads according to WHO criteria. In this sense, a gold-standard for sperm head classification was built with the collaboration of three referent domain experts, according to WHO criteria (Section 4.5). This dataset contains $1,854$ sperm heads that were manually classified by each expert in one of the following classes: normal, tapered, pyriform, small or amorphous. Thus, the output of this step is a classification gold-standard to validate and evaluate automatic classification methods.
4. Next step in this research was characterization sperm heads. To this end, one of the first decisions is about how would be the segmented sperm head representation. It was decided to use a continuous representation of the curve that defined the sperm head, using anisotropic diffusion and active contours (Section 6.1). We use a number of shape descriptor proposed in the literature and formulate a morphological descriptor, based on single shape-based measures, for extracting features from segmented sperm heads (Section 6.2.3), as well as a descriptor combining approach to take advantage of different descriptors proposed in the literature (Section 6.3). The classification gold-standard was used for making decisions about: 1) which shape-based descriptors to combine, 2) which shape-based measures to include in the proposed morphological descriptor, and 3) which topology of feature selection and combination to use. Thus, the output of this step is a new morphological descriptor for sperm heads and a descriptor combining approach.
5. Finally, the next step was focused on designing a pipeline for sperm head classification, according to WHO criteria. A number of monolithic classifiers were evaluated using the classification gold-standard (Appendix B). Also, different cascade classification schemes were designed and evaluated using the classification gold-standard (Appendix B). After evaluating different classification schemes, a two-stage classification scheme was designed (Section 7.1). The proposed classification scheme permits classifying sperm heads among five different classes (one class for normal sperm heads and four classes for abnormal sperm heads) combining an ensembled strategy for feature selection and a cascade approach with SVM multi-classifiers (algorithm 7). This classification scheme was trained using a subset of the classification gold-standard and validated using another subset of the same gold-standard. The training step returned the trained combined classifiers for each stage of the scheme (Tables 7.3, 7.5, 7.6, 7.7, and 7.8). The validation step returned the selected descriptors used for each combined classifier (Tables

7.2 and 7.4). Thus, the contribution of this step to the main goal of this thesis is the sperm head classification scheme, along with the trained combined classifiers and selected descriptors for each stage of the classification scheme.

1.4 Thesis Deliverables

The deliverables of this PhD thesis are:

- A sperm head segmentation gold-standard (Section 4.4)
- A sperm head classification gold-standard (Section 4.5)
- A pipeline for detection, segmentation, characterization and classification of human sperm heads (Algorithm 1) that encompasses all the algorithms designed in this thesis
- A published article in CMPB *Gold-standard and improved framework for sperm head segmentation* [Chang et al., 2014]
- A submitted article in CMPB *Gold-standard for computer-assisted morphological sperm analysis* [Chang et al., 2015]
- A manuscript in preparation for CMPB *Framework for automatic classification of human sperm heads*

Algorithm 1 Pipeline for detecting, segmenting, characterizing and classifying sperm heads developed in this thesis.

imRgb: original image containing several sperm cells
svm1: combination of three trained SVMs of Stage 1 {Section 7.1.1}
v1: trained verifier for normal and amorphous {Section 7.1.2}
v2: trained verifier for tapered and amorphous {Section 7.1.2}
v3: trained verifier for pyriform and amorphous {Section 7.1.2}
v4: trained verifier for small and amorphous {Section 7.1.2}
imgDetection ← detectionSpermHeads(imRgb) {Algorithm 2 (Section 5.1)}
sperms ← labelRegions(imgDetection)
for each sperm i **do**
  spermRgb ← imgDetection(sperm i)
  spermSegmented ← segmentationSpermHead(spermRgb) {Algorithm 3 (Section 5.2)}
  spermAni ← anisotropicDiffusion(spermRgb)
  contour ← activeContours(spermAni,spermSegmented)
  [MorphoD,FourierD,GeomD,ZernikeD,ConvD,EllipD] ← calculateDescriptors(contour) {Sections 6.2.3 and 6.2.2}
  label(i) ← testingClassificationScheme(MorphoD,FourierD,GeomD,ZernikeD,ConvD,EllipD, svm1,v1,v2,v3,v4) {Algorithm 7 (Section 7.2.4)}
end for
return label

1.5 Thesis Outline

Chapter 2 is devoted to detailing background knowledge about related areas that are critical for the best understanding of the whole thesis document. We first describe biological issues

about components and features of human sperm cells as well as a variety of sperm defects according to WHO criteria and the spermiogram as the result of a sperm smear evaluation, introducing the reader to the world of morphological sperm analysis. Second, some specific techniques from image processing are reviewed according to their promising utility in the final application of this work. Third, we present a brief description of clustering, shape-based feature extraction, classification and dimensionality reduction techniques that were used in the next chapters. Finally, we present a general description of metrics for evaluating final results of our proposed methods.

In Chapter 3 we present an up-to-date review of related research works. We will disscuss the most prominent approaches for detecting and segmenting eukaryotic cells. Regarding the small variability inter classes and the subjectivity at characterizing and classifying, a very related area is grain featuring and classification, including food and pollen grains. Therefore, we will review published research works that addressed the problem of characterizing and classifying sperm cells at first, and grain kernels, later.

Chapter 4 presents the design and construction of a segmentation gold-standard, regarding the opinion of one expert, as well as a classification gold-standard, regarding the opinion of three referent experts in the field. In this sense, we describe in detail the staining method, the features of the equipment we used to capture the images and specific details about images and manual methods in both of our gold-standards.

Chapter 5 presents a framework for detecting and segmenting human sperm heads as a reliable detection and segmentation presents the first step for all posterior classification algorithms. We will describe in detail our proposed approach and present experimental results on the introduced segmentation gold-standard. We include Appendix A to discuss the state-of-the-art segmentation method that we used as reference to compare our results.

In Chapter 6, we discuss common shape-based descriptors as well as describe a morphological descriptor designed particularly with this application in mind. Looking to get a feature vector as compact as possible, it is imperative to select the best combination of feature spaces. In this chapter we therefore explain the strategies available for selecting and combining feature spaces.

In Chapter 7 we propose a classification pipeline for sperm heads according to WHO criteria. We present a two-stage classification scheme as a cascade ensemble of SVMs. We explain in detail both stages of our classification scheme as well as present the results of applying each stage and the whole scheme to the classification gold-standard. The discarded, but evaluated, previous classification schemes that we designed are presented in Appendix B, as well as the justification on choosing *the* classificacion scheme.

Although we present a whole approach for detecting, segmenting, characterizing and classifying sperm heads in the context of morphological sperm analysis, still many extensions of this research deserve further consideration. Chapter 8 outlines directions for future research, summarizes this thesis and discusses its findings and conclusions.

Chapter 2

Background

From a computational point of view, a system for analyzing images requires the confluence of several methods from related areas such as image processing and pattern recognition. There are many methods for specific situations in each of these areas, but some are more promising for automated morphological analysis of sperm. Therefore, this chapter is devoted to detailing some background knowledge on related areas critical for better understanding the whole thesis.

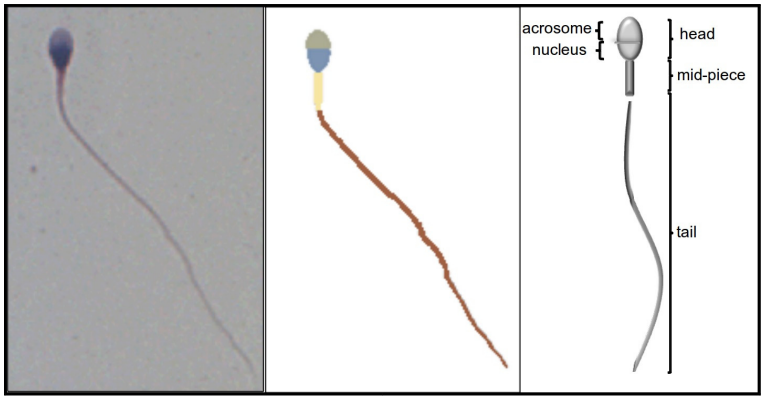
The first section 2.1 describes biological issues about components and features of human sperm cells as well as a variety of sperm defects acoording to WHO criteria and the result of a sperm smear evaluation: the spermiogram. This section also attempts to introduce the reader to the world of morphological sperm analysis. Some specific techniques from image processing are reviewed in Section 2.2, according to their promising utility in the final application of this work. In Section 2.3, the reader finds a brief review of pattern recognition issues like shape-based feature extraction, clustering, classification and dimensionality reduction techniques. Finally, a general description of metrics for evaluating final results is presented in Section 2.4.

2.1 Biological Concepts

2.1.1 Sperm Structure

A sperm cell consists of three major parts: head, mid-piece and tail (see Figure 2.1).

The human sperm has a head with a regular contoured ellipsoid form of $5\mu m$ long, $3\mu m$ wide and $1.5\mu m$ thick. Most of the head is occupied by the nucleus containing DNA with densely coiled chromatin fibers, surrounded anteriorly by the acrosome, which contains enzymes used for penetrating the female egg, comprising $40 - 70\%$ of the head area [Menkveld et al., 2001]. The acrosome region does not contain large vacuoles; however, it could have up to two small vacuoles which should not cover more than 20% of the sperm head (see Figure



a) b) c)

Figure 2.1: **Morphology of the normal human sperm. (a)** Representative bright field image of a normal human sperm (Image size: 277×144 pixels $\approx 58 \times 30$ μm). **(b)** Manually segmented ground-truth of the sperm: Head, acrosome, nucleus, mid-piece and tail of stained spermatozoa. **(c)** Schematic drawing of the principal components of a normal human sperm. Oval head: 5μm long and 3μm wide. Acrosome: $40-70\%$ of the head area. Mid-piece: 5μm. Tail: 55μm.

2.2). There should not be any vacuoles in the post-acrosomal region [WHO, 2010b].

Figure 2.2: Sperm cells with vacuoles in their acrosomal region.

The mid-piece of the sperm has a central filamentous core with many mitochondria spiraled around it, which provides the energy for sperm motion. The mid-piece is usually slender, regular and about the same length as the sperm head. The major axis of the sperm head should be aligned with the major axis of the mid-piece [WHO, 2010b].

Finally, the tail of the sperm is approximately $45\mu m$ long. The tail is made of protein fibers that contract on alternative sides, giving a characteristic wavelike movement that drives the sperm through the seminal fluid, which also supplies additional energy. The tail is usually thinner than mid-piece and has a uniform caliber along its length. If there is no sharp angle indicative of a flagellar break, the tail may be looped back on itself [WHO, 2010b].

2.1.2 Spermiogram

A spermiogram is a report of a microscopic semen smear analysis. This evaluation is the most important and initially asked laboratory test for the evaluation of men fertility potential.

A typical semen analysis measures, at least, the following parameters:

Concentration

Sperm concentration (count of sperm cells) is highly important for being able to fertilize the ovum. Sperm concentration is the number of sperm cells in one milliliter of semen. In a normal ejaculation, there are approximately $20E6/ml$ sperm cells. A sperm concentration lower than $15E6/ml$ might indicate a fertility problem.

Motility

Motility (mobility of sperm cells) is the percentage of moving sperm cells in one sample and indicates the direction and rate of their movement. Motility of sperm cells is also important for being able to fertilize the ovum because only moving sperms can go down the female genital tracts and fertilize the ovum. At least 50% of sperm cells should be able to move swiftly and travel in a straight line for an hour after ejaculation.

Vitality

Sperm vitality is a reflection of the proportion of live sperm cells determined by the evaluation of membrane integrity. Sperm vitality is especially important for samples with less than about 40% progressively motile sperm cells. Less than 58% of sperm cells with intact membrane might suggest a fertility problem.

Morphology

Another important factor for fertilization capacity is the form of the sperm cells and whether any structural deformations are present. It is not possible for a deformed sperm cell to fertilize the ovum. Morphology parameters indicate the form of the sperm cells and analyzes the size, form and appearance of cells. To evaluate this parameter, 200 sperm cells are analyzed and the deformed ones are noted. The greater number of abnormal sperm cells, the less the likelihood of fertility is. Abnormal forms include anomaly of head, neck and tail, and immature forms. At least 4% of sperm cells should present normal morphology.

2.1.3 Sperm Defects

There are sperm cells with different kinds of malformations in human semen samples. Depending on the types of anomalies, abnormal sperm generally have a lower fertilizing potential and may also have abnormal DNA. The following categories of defects should be noted in a reliable morphological analysis[1] [WHO, 2010b]:

1. Head defects: large or small, tapered, pyriform, round, amorphous, vacuolated (more than two vacuoles or $> 20\%$ of the head area occupied by unstained vacuolar areas), vacuoles in the post-acrosomal region, small or large acrosomal areas ($< 40\%$ or $> 70\%$ of the head area), double heads, or any combination.
2. Neck and mid-piece defects: asymmetrical insertion of the mid-piece into the head, thick or irregular, sharply bent, abnormally thin, or any combination.
3. Tail defects: short, multiple, broken, smooth hairpin bends, sharply angulated bends, of irregular width, coiled, or any combination.
4. Excess residual cytoplasm (ERC): this is associated with abnormal sperm produced from a defective spermatogenic process. Sperm characterized by large amounts of irregular stained cytoplasm, one third or more of the sperm head size, often associated with defective mid-pieces [Mortimer and Menkveld, 2001] are abnormal.

In Figure 2.3, schematic drawings of some abnormal forms of human sperm are shown, according to [WHO, 2010b].

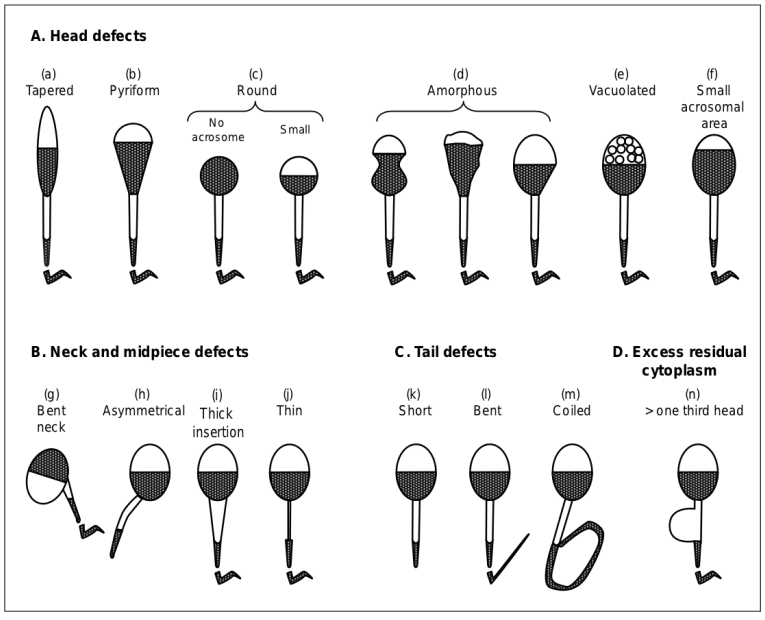


Figure 2.3: **Human sperm abnormalities.** Image reproduced exactly as appears in [WHO, 2010b].

[1]The next list has been taken exactly as appears in the cited reference

2.1.4 Morphological Sperm Analysis

Assessment of human sperm morphology is based on the microscopic analysis of stained smears. It could be done in the manual way (through visual observations) or by using image processing and analysis methods [Auger, 2010]. For any of these cases, an optimally stained smear is needed to provide sharp contrast for defining the sperm outline and cell details (see Figure 2.1a)). Recognition of these details also depends on the final magnification, for example 100× oil objective [Auger, 2010]. In the fifth edition of WHO laboratory manual [WHO, 2010b], pre-analytical procedures have been described.

In the manual method based on visual observations, each sperm cell is classified as normal or abnormal, and each anomaly encountered is sub classified using strictly defined criteria. By contrast, in the computer-assisted method, each sperm cell has different morphological features (mostly head parameters) that are measured [Auger, 2010].

The computer assisted sperm morphology assessment has been fueled by the inherent lack of objectivity in the evaluation of human sperm morphology, the difficulty in standardizing, implementing and controlling manual methods, and the high degree of variation within and between laboratories and technicians [Soler et al., 2003].

2.2 Image Processing

Modern digital image technology is widely used today. This has leaded research in areas aimed to manipulate multi-dimensional signals (images) with two different goals: image processing starts with one image and produces a modified version of that image, while image analysis is a process that transforms a digital image into something other than a digital image, such as a set of measurements [Shih, 2009].

In this sense, image processing is the study of any algorithm that takes an image as input and returns an image as output, while suppressing undesired distortions and highlighting some information for further analysis. Thus, image processing techniques are applied in the initial stage of any application that requires extracting some kind of information from images [Sonka et al., 2008]. For instance, getting edge information of an image is a very important task in many applications. In this case, the idea is to suppress undesired information about homogeneous-intensity regions and highlight information about intensity local-variations (edges). Other useful operations applied in the context of image processing include color balancing, noise reduction, exposure correction, and sharpening, among others [Szeliski, 2011].

The wide range of applications of image processing and analysis techniques has extended to almost all scientific and engineering fields: remote sensing, quality and security control, visual inspection, document imaging, biology, microscopy, and medical imaging, among others [Soille, 2010]. For example, some specific applications in biological and medical imaging [Soille, 2010] include: 1) the use of geostatistical and morphological tools for the analysis of confocal images of cells [Conan et al., 1992], 2) a methodology for segmenting skin biopsy

samples [Casas et al., 1994], and 3) a segmentation approach for connected fibroblasts on light micro-graphs [Metzler et al., 2000].

This section is aimed to briefly review image processing techniques that will be useful in the further development of our proposed algorithms. Even though these reviewed techniques do not seem related to each one, all of them have a significance impact in algorithms of Chapter 5. In this sense, we will review the basic notion of pixel connectivity, the basic mathematical morphology operations, a brief description of anisotropic diffusion filter, the basic notions of active contours, as well as the main features of some color spaces, useful for our detection and segmentation purposes.

2.2.1 Pixel Connectivity

Mathematical notion of connectivity plays an important role in a variety of problems related to image processing such as image filtering and image segmentation [Braga-Neto and Goutsias, 2003]. In general, pixel connectivity refers to a relation between two or more pixels on a region of interest (ROI) inside an image. The idea is that for two pixels to be connected, they need to meet certain criteria related to intensity values and spatial relationship.

Mathematically, connectivity could be define from two perspectives. The first one uses a topological framework [Dugundji, 1966] and it is useful for images defined over a continuous space. Whereas the second option uses a graph-theoretic framework [Diestel, 2010] useful for images defined over a discrete space. In this thesis we are interested in the notion of the neighborhood of a pixel in the image, according to digital topology.

In this sense, the digital plane Z is defined as a representative model of the Euclidian plane, consisting in the set of all points in the plane R^2 having integer coordinates [Eckhardt and Latecki, 2008]. In image processing, the digital plane is taken as a mathematical model of digital binary images [Eckhardt and Latecki, 2008]. In this application one usually has a given set, namely the set S of white points in the image and the complement $C(S)$, which is the set of black points. Given a point $P = (m, n) \in Z^2$, the 8−neighbors of P are all points with integer coordinates (k, l) such that

$$max(|m - k|, |n - l|) \leq 1$$

	column		
row	$n-1$	n	$n+1$
$m+1$	$N_3(P)$	$N_2(P)$	$N_1(P)$
m	$N_4(P)$	P	$N_0(P)$
$m-1$	$N_5(P)$	$N_6(P)$	$N_7(P)$

Figure 2.4: **Neighbors of point** P**.** The 4−neighbors of P are $N_0(P), N_2(P), N_4(P)$ and $N_6(P)$. The 8−neighbors of P are $N_{\mathrm{i}}(P), \mathrm{i} = 0 \ldots 7$.

Figure 2.4 shows the way in which the 8−neighbors of P are numbered [Eckhardt and Latecki, 2008]. Neighbors with even numbers are the direct or 4−neighbors of P, those with odd numbers are the in-direct neighbors. The 8−neighborhood of P is the set of all 8−neighbors of P (excluding P), while the 4−neighborhood of P is the set of all 4−neighbors of P (excluding P).

2.2.2 Mathematical Morphology

Mathematical morphology is a set theory approach, developed by Serra and Matheron [Serra, 1986], that provides lossy and nonlinear filters for image processing and pattern analysis. The basic principle of morphological operators is to simplify the image data while preserving their essential shape features and eliminating irrelevancies [Serra, 1986]. With morphological operators it is possible to extract shape characteristics such as edges, holes, and corners by operating with structuring elements of varied sizes and shapes [Shih, 2009]. In fact, most morphological filters are designed using a priori knowledge about the geometric properties and shape of ROIs in the image [Soille, 2010].

Mathematical morphology uses tools of non-linear algebra and operates with point sets, their connectivity and shape [Sonka et al., 2008]. A binary image can be treated as a 2D point set, where points belonging to ROIs in the image represent a set S (these points are pixels with value equal to one). Points of the complement set $C(S)$ correspond to the background with pixel values equal to zero. A morphological transformation Ψ is defined as the relation of the image with another small point set B called structuring element. To apply the morphological transformation $\Psi(S)$ to the image S means that the structuring element B is moved systematically across the entire image [Sonka et al., 2008]. It is important to note that the structuring element can be placed at any pixel in an image. The position (i, j) of a reference pixel defines where the structuring element has been placed $B(\mathrm{i}, j)$. The structuring element could be symmetric or asymmetric.

In this thesis, we are interested in morphological operations such as erosion, dilation, opening and closing on binary images using symmetric structuring elements. The primary morphological operations are dilation and erosion. The *erosion* of S by B is defined to be the set of all pixel locations for which B placed at that pixel is contained within S. This is denoted $S \ominus B$ and may be written as

$$S \ominus B = (\mathrm{i}, j) : B_{(\mathrm{i},j)} \subset S$$

Erosion is used to simplify the structure of an object, smoothing edges and it might decompose complicated objects into several simpler ones [Sonka et al., 2008].

A complementary operation to that of erosion is *dilation*. It is defined simply as the erosion of the complement of a set. The dilation of a set S by a structuring element B, denoted $S \oplus B$, is defined by

$$S \oplus B = C(C(S) \ominus B)$$

Dilation is used to fill small holes and narrow gulfs in objects [Sonka et al., 2008].

Erosion and dilation are not inverse transformations, it means that if an image is eroded and then dilated, the original image is not re-obtained. Instead, the result is a simplified and less detailed version of the original image [Sonka et al., 2008]. Erosion followed by dilation creates a morphological transformation called *opening*. The opening of an image S by the structuring element B is denoted by $S \circ B$ and is defined as

$$S \circ B = (S \ominus B) \oplus B$$

Dilation followed by erosion is called *closing*. The closing of an image S by the structuring element B is denoted by $S \bullet B$ and is defined as

$$S \bullet B = (S \oplus B) \ominus B$$

2.2.3 Anisotropic Diffusion

Anisotropic diffusion is a technique aiming at reducing image noise that encourages intraregion smoothing while inhibiting interregion smoothing introduced by Perona and Malik [Perona and Malik, 1990]. The main idea beyond this filter is a transformation with two features: non-linearity and scale-space. In this sense, anisotropic diffusion creates a scale space in which an image generates a parameterized family of successively more and more blurred images based on a diffusion process where each resulting image is a combination between the original image and a filter that depends on the local content of the original image.

According to Perona and Malik [Perona and Malik, 1990], the anisotropic difussion is defined as an evolving process as following:

$$\frac{\partial I}{\partial t} = \mathrm{div}(c(x,y,t)\nabla I) = \nabla c \cdot \nabla I + c(c,y,t)\Delta I,$$

where Δ is the Laplacian operator and ∇ denotes the gradient operator. In addition, $c(x,y,t)$ represents the diffusion coefficient which depends on the point (x,y) and the time variable t. The idea behind using this coefficient is to control the strength of the smoothing, so that the diffusivity is reduced (low diffusion coefficient) at those locations which have a larger likelihood to be edges, while non-edge regions must be linked to a higher coefficient [Weickert, 1998]. To this end, $c(\cdot)$ is chosen as a function of the image gradient. In the original work, Perona and Malik proposed the following two functions:

$$c(||\nabla I||) = \mathrm{e}^{-(||\nabla I||/\kappa)^2} \qquad c(||\nabla I||) = \frac{1}{1 + \left(\frac{||\nabla I||}{\kappa}\right)^2}$$

where κ is fixed to control the smoothing strength around edge regions. The evolution with respect to t is related to iterate the diffusion process T times. This allows us to get a scale-space representation of the diffusion and the final diffused image is obtained at time T. In a discrete implementation, it is needed a parameter λ for controlling the speed of diffusion [Perona and Malik, 1990].

2.2.4 Active Contours

Active contours are methods to locate object boundary curves in images. Active contours approach is more robust than conventional segmentation techniques when dealing with noisy images [Caselles V. et al., 1993]. The name of *active contours* is due to the fact that these boundary detectors iteratively move their final position forward according to the image forces [Blake and Isard, 1998].

One kind of these methods, introduced originally as *snakes* [Kass M. et al., 1988], is a two-dimensional spline curve that evolves towards image features such as strong edges according to a energy-minimizing equation [Szeliski, 2011]. However, active contours need an initial curve that will finally evolve to the desired shape.

Formally, an active contour is defined as a parametric curve $v(s) = x(s) + y(s)$, $s \in [0, 1]$ that moves on the image domain $D \in R$ until minimizing the energy equation [Kass M. et al., 1988, Giraldi and Oliveira, 2000]. The minimized energy equation can be written as:

$$\int E^*_{snake} = \int_0^1 E_{snake}(v(s))\mathrm{d}s$$

The snake energy can be interpreted as

$$\int_0^1 E_{snake} = \int_0^1 E_{int}(v(s))\mathrm{d}s + \int_0^1 E_{imagen}(v(s))\mathrm{d}s + \int_0^1 E_{ext}(v(s))\mathrm{d}s$$

where E_{int} represents the internal energy of the curve, E_{imagen} represents the image force, and E_{ext} the external force that acts on the snake. The image force attracts the snake to lines and edges.

The internal force of a snake is defined as:

$$E_{int} = \alpha(s)\left|\frac{\mathrm{d}v}{\mathrm{d}s}\right|^2 + \beta(s)\left|\frac{\mathrm{d}^2v}{\mathrm{d}^2s}\right|^2$$

where the first order derivative $\frac{\mathrm{d}v}{\mathrm{d}s}$ makes the curve elastic. The second order derivative $\frac{\mathrm{d}^2v}{\mathrm{d}^2s}$ makes the curve rigid. The values of α and β control the behavior of the snake: when it should be elastic or keep rigid.

The external force E_{ext} tries to put the snake on the edges. Regarding a gray-level image $I(\mathrm{i}, j)$, the external energy can be defined as:

$$E_{ext}(\mathrm{i}, j) = -|\nabla I(\mathrm{i}, j)|^2$$

$$E_{ext}(\mathrm{i}, j) = -|\nabla(G_\sigma(\mathrm{i}, j) * I(\mathrm{i}, j))|^2$$

where G is a Gaussian filter and σ stands for the standard deviation that controls the smoothing effect.

An expansion ballooning force can be added to the equation in order to move each point along its normal [Cohen and Cohen, 1993]. This is an interesting alternative because regular snakes tend to shrink [Szeliski, 2011].

2.2.5 Color Spaces

A color space is a method by which color can be specified, created and visualized. Several color spaces are used in practice, and these spaces can be transformed into each other. The components of each color space are impacted in different ways by an image processing technique, especially by segmentation techniques, and it is a good idea to take advantage of this. In the following, we will review three different color spaces that will be used in algorithms of Chapter 5.

RGB

The RGB color space has its origin in color television. RGB is one of most widely used color space for processing and storing the digital image data. The value of a particular color is expressed as a vector of three elements (intensities of three primary colors R-red, G-green and B-blue). This model is non-linear with visual perception. It is device dependent and specification of colors is semi-intuitive [Ford and Roberts, 1998]. There is a high correlation between the three components and chrominance and luminance components are mixed [Kaur and Kranthi, 2012].

L*a*b*

The L*a*b* color space is a color-opponent space with dimension L for luminance and a and b for the color-opponent dimensions. Unlike the RGB color space, L*a*b color space has perceptual uniformity [Vezhnevets et al., 2003], and its L component closely matches human perception of lightness. However, the perceptual uniformity is obtained at the expense of heavy computational transformations, because the computation of the luminance L and the chroma ab is obtained through a non-linear mapping of compressed coordinates [Kaur and Kranthi, 2012]. One of the most important attributes of the L*a*b* model is device independence.

YCbCr

YCbCr color space has been defined in response to increasing demands for digital algorithms in handling video information and has become a widely used model in digital video [Kaur and Kranthi, 2012]. In the YCbCr color space, luminance information is stored as a single luma component Y, and chrominance information is stored as two color-difference components (Cb and Cr). In this sense, Cb is the difference between blue and luma component and Cr is the difference between red and luma component. In contrast to RGB, the YCbCr color space is luminance independent [Kaur and Kranthi, 2012].

2.3 Pattern Recognition

2.3.1 Shape-based Feature Extraction

Shape-based description is one of the basic steps in the ROI classification process, and consists in quantifying some of the ROI properties, known as features. For instance, a ROI can be described in terms of its area, its perimeter, and its number of holes, among other features [Costa and Cesar, 2009].

To semantically describe the content of an image, ROI shape-based description has been proven to be much more effective than other descriptions, such as those based on texture or color [Persoon and Fu, 1977]. However, when invariance with respect to a number of possible transformations such as scaling, shifting, and rotation is required, the construction of shape-based descriptors is more complicated [Belongie et al., 2002].

Shape-based descriptors are categorized as contour-based shape descriptors and region-based shape descriptors. The first kind of shape-based descriptors, the contour-based ones, exploit information just at the boundary points focusing on contour features, which is critical for human perception of shapes [Zhang and Lu, 2004]. Whereas, region-based shape descriptors are useful at describing non-connected and disjoint ROIs because they combine information across an entire object, so they can capture the interior content of a ROI [Zhang and Lu, 2004]. Fourier descriptor is an example of a contour-based descriptor, while geometric moments is an example of a region-based descriptor [Amanatiadis et al., 2011]. In the following we briefly describe the shape-based descriptors used in this thesis, while a detailed formulation can be found in Chapter 6.

Contour-based Shape Descriptors

Convexity Measure Convexity in image processing has been studied for many years. There are many proposals of convexity measures that have been applied in object classification and recognition [Zunić and Rosin, 2004,Rosin and Mumford, 2006]. The most naive and common approach is the one based on the convex hull of the ROI boundary. Either the ratio of areas or perimeters can be used [Rosin, 2009]. Also, there are different convexity measures that are invariant with respect to translation, rotation and scaling transformations and range through the interval $(0, 1]$, taking the value of 1 only for convex shapes [Rahtu et al., 2006, Rosin and Mumford, 2006, Zunić and Rosin, 2004].

Fourier Descriptor Fourier descriptor has been successfully applied to many shape representation applications because of its simple derivation, simple normalization and its robustness to noise [Zhang and Lu, 2002]. The Fourier descriptor is obtained by applying a Fourier transform on a shape signature. The shape signature is an one-dimensional function that is derived from the ROI boundary coordinates. With Fourier descriptor, global features are captured by the first few low frequency terms, while higher frequency terms capture the finer features of the ROI [Zhang and Lu, 2004]. The Fourier descriptor overcomes the noise sensitivity in the shape signature representations.

Ellipticity Measure Several ellipticity measures already exist in the literature [Aktas and Žunić, 2011,Rosin, 2003,Sonka et al., 2008]. All of them use their own way of evaluating how the ROI considered differs from an ellipse. These measures assign a highest possible ellipticity to all the ellipses (including circles) [Aktas and Žunić, 2013]. Aktas et. al. [Aktas and Žunić, 2013] presented a family of shape ellipticity measures that evaluates how much a given ellipse S differs from an ellipse, ranges over $(0, 1]$, is invariant under similarity transformations and picks the value 1 if and only if S is an ellipse.

Region-based Shape Descriptors

Geometric Moments Image moments are considered as an example of region-based shape descriptors. Image moments are global descriptors characterized by representing a ROI in a compact way avoiding the effect of noise [Nixon and Aguado, 2008]. In addition, image moments describe the rate of change within a local area. Formally, descriptors based on image moments interpret a normalized gray-level image function as a probability density of a two-dimensional random variable. The properties of this random variable can be described using statistical characteristics or moments [Sonka et al., 2008]. Hu introduced the use of moments for shape description [Hu, 1962] and showed that information is preserved when using moments for describing ROIs.

Zernike Moments Zernike-moment descriptor (ZMD) is one of the best shape-based descriptors among the existing shape-based descriptors, since its proposal [Teague, 1980]. Many researchers report promising results of ZMD [Kim and Kim, 2000,Zhang and Lu, 2001]. Zernike moments allow independent moment invariants to be constructed to an arbitrarily high order using the theory of orthogonal polynomials. ZMD is obtained by using all the pixel information within a ROI, because it does not assume boundary information. The theory of Zernike moments is similar to that of the Fourier transform, to expand a signal into a series of orthogonal basis [Zhang and Lu, 2004]. The precision of shape representation depends on the number of moments truncated from the series expansion.

2.3.2 Pattern Classification

Pattern classification is a scientific area whose purpose is to classify objects (patterns) in certain classes or categories by a priori knowledge or statistical information extracted from the patterns. Generally, the data to be classified defines sets of observations which correspond to points in a multidimensional space [Zheng and Xue, 2009].

There has been extensive interest in applying pattern classification methods for biomedical applications including microcalcification detection [El-Naqa et al., 2002], fetal lung maturity analysis [Bhanu Prakash et al., 2002], and polyp detection [Gokturk et al., 2001].

The main difficulty in pattern classification is the variability problem that appears when objects belonging to the same class undergo variations due to illumination changes, geometric

distortions, and occlusion, among others.

A complete pattern recognition system consists of (1) a sensor that perceives observations, (2) a pre-processing module that performs any operation that contributes in defining a compact representation of the pattern, (3) a feature extraction module, and (4) a classification module [Zheng and Xue, 2009]. In the case of visual patterns, step 1 is carried out by any photosensitive device and step 2 is carried out by image processing operations.

Classification is a process in which a classifier assigns a class label to an input. Following the proposal of Jain et al. [Jain et al., 2000], a classifier can be designed following one of the next approaches:

- Similarity Approach: Similar patterns should be assigned to the same class. The *K-nearest neighbor classifier* and the *nearest mean classifier* belong to this category.
- Probability Approach: The optimal *Bayes decision rule* assigns a pattern to the class with the maximum posterior probability.
- Construction decision boundaries: This category of classifiers constructs decision boundaries by optimizing certain criterion. The *Support Vector Machine* (SVM) belongs to this class.
- Classifier Combination: The idea is to combine weak classifiers to produce a strong classifier. *Adaboost* [Viola and Jones, 2002] is the most representative classifier in this category.

2.3.3 Base Classifiers

$K-$Nearest Neighbor Classifier

The $k-$ nearest neighbor (kNN) classifier is a nonparametric learning algorithm [Vege, 2012] and one of the simplest techniques for machine learning [Duda et al., 2000, Dasarathy, 1991].

Nonparametric classification is often associated with the notion of a representative element from a class. The class label assigned to an input feature vector is based on the similarity of this vector to one or more representative elements. Typically, similarity is defined in a geometrical sense, that is, based on a certain distance [Kuncheva, 2004]. Euclidean distance can be used to compute distance between vectors [Vege, 2012].

In this sense, a data sample in kNN is classified on the basis of a selected number of k nearest neighbors. To classifiy an input feature vector, the k nearest neighbors are retrieved together with their class labels. The input feature vector is labeled with the most represented class label among the k nearest neighbors [Kuncheva, 2004].

Naive-Bayes Classifier

A naive-Bayes classifier is a simple probabilistic learning method based on Bayes theorem where every feature is assumed to be class-conditionally independent [Duin, 1996].

The Bayes classification rule requires the estimation of the probability density functions, based on the available training set of size N. In order to guarantee good estimates, N must be large enough. But, this turns in a difficult scenario when the dimension of the feature space increases. In this sense, if in an one-dimensional feature space N could be regarded as large enough, then in a D-dimensional space N^D points would be necessary [Theodoridis and Koutroumbas, 2009]. Assuming that individual features are statistically independent, the naive-Bayes classifier uses only DN points to estimate a D one-dimensional probability density functions for each of the classes [Theodoridis and Koutroumbas, 2009].

Decision Trees

The decision tree classifier is a multistage nonlinear classifier, in which classes are sequentially rejected until we reach a finally accepted class [Theodoridis and Koutroumbas, 2009]. To this end, the feature space is sequentially split into unique regions, corresponding to the classes [Theodoridis and Koutroumbas, 2009].

Upon the arrival of a feature vector, the searching of the region to which the feature vector will be assigned is achieved via a sequence of decisions along a path of nodes of a tree [Theodoridis and Koutroumbas, 2009]. The first decision is made at the root of the tree by asking a question with a small number of different answers. Depending of the answer, a branch is selected and the child node is visited. Another decision is made at this node, and so on, until a leaf is reached. The leaf contains a single class label, which is assigned to the feature vector being classified [Kuncheva, 2004]. To this end, the sequence of decisions is applied to individual features, and the questions to be answered are of the form *is feature* $x_{\mathrm{i}} \leq \alpha$*?*, where α is a threshold value [Theodoridis and Koutroumbas, 2009].

Decision tree classifiers are intuitive because the decision process can be traced as a sequence of simple decisions [Kuncheva, 2004]. In addition, tree structures can capture a knowledge base in a hierarchical arrangement [Kuncheva, 2004]. However, this hierarchical nature of the tree classifiers yields their high variance. A small change in the training dataset might result in a very different tree. In this way, an error that occurs in an early stage in the tree propagates all the way down to the leaves below it [Theodoridis and Koutroumbas, 2009].

Support Vector Machine Classifier

Support Vector Machine (SVM) is a machine learning tool that is widely used for pattern classification, which implements non-linear mapping of input vectors to high dimensional feature space, while performing minimization of error function by an iterative training algorithm

to construct its optimal separation hyper-plane [Ramakrishnan and Emary, 2010].

According to SVM, optimal classification of a separable two-class problem is achieved by maximizing the width of the empty area (margin) between the two classes [Sonka et al., 2008]. The margin width is defined as the distance between the discrimination hypersurface in $n-$dimensional feature space and the closest training patterns: these are called support vectors. In this sense, the support vectors specify the discrimination function. In addition, SVMs need a kernel function that maps the data in the input space to feature space where they are linearly separable. There are many kernel mapping functions available; however, the Radial Basis Function (RBF) kernel is the most commonly used [Ukil, 2007].

2.3.4 Combination of Classifiers

Methods for combining classifiers have been widely explored over recent years. Two main goals have fueled this field: a) to reduce the error rate in classification tasks in opposite to monolithic classifiers, and b) to make the performance of the whole system more robust against the difficulties that each individual classifier may have on each particular dataset [Dietterich, 2000]. Figure 2.5 shows the monolithic and the combined classifier paradigms.

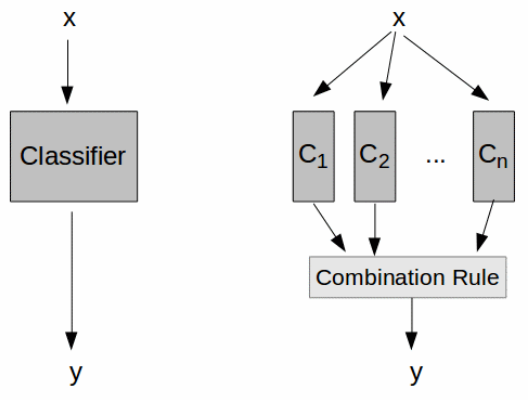


Figure 2.5: **Monolithic vs. combined classifier approach**. We show a schematic explanation of the difference between a monolithic classifier **(a)** and a combined classifier **(b)**. x stands for input feature vector, y stands for output class label and *Classifier* and C_{i} $(\mathrm{i} = 1 \ldots n)$ stand for any base classifier.

Combining classifier methods can be grouped acoording to the level at which they operate: at feature or at decision level. Combining classifiers at feature level has the advantage that the features from two or more sets at the same time can potentially provide additional information about the classes. If combination methods operate at decision level, they combine outputs of the classifiers using a combination rule with the main advantage of lower complexity compared to the first group [Tulyakov et al., 2008].

A combination of classifiers can be constructed either in a parallel or serial manner. The selection of a proper topology depends on the type of problem at hand. Some tasks can be

divided into subtasks, and consequently, each subtask can be implementd using a different classifier. However, the most common use of combination of classifiers involves the development of diverse classifiers which solve the same task and the combination of them in a parallel manner [Ranawana and Palade, 2006].

2.3.5 Combination Rules

Once the individual classifiers have been designed and implemented, the next most important task involves the combination of the individual results obtained through individual classifiers [Kittler et al., 1998]. In the following, we list common combination rules that were used to obtain the results presented in Chapter 7.

Plurality

This rule selects the relevant class by polling all the classifiers to see which class is the most popular. The class with the highest vote is selected.

Majority

Similar to plurality, except for the fact that there is the constrain about the selected class. It should be the most voted class, concentrating simple majority of all individual classifiers.

Unanimity

In the unanimity combination rule, a decision is made for some class if all individual classifiers agree on that class; otherwise the combiner refuses to decide.

Maximum

The classifier output with the highest value (confidence) is chosen as the output of the overall combiner. A slight variation consists of setting a threshold value and only considers classifier outputs that overcome that threshold.

2.3.6 Clustering Methods

The purpose of clustering methods is to detect similar subgroups among a large collection of data points and to assign those data points to the clusters [Nisbet et al., 2009]. These methods are very suitable for color image segmentation, where the data points are represented by the

color intensities of pixels and the subgroups are represented by the objects in the image. Here, we review the most common algorithm used for clustering in segmentation approaches.

$K-$means

The classic $k-$means algorithm was introduced by Hartigan [Hartigan, 1975] and is based on a simple idea. Suppose we have a data set $x_1, \ldots, x_n$ consisting of n data points from a multidimensional Euclidean space. The main goal is to partition the data set into a given number k of clusters, assigning each data point to one of those clusters so that the mean across clusters are as different from each other as possible [Nisbet et al., 2009]. A cluster is defined as a group of data points whose inter-point distances are small compared with the distances to points outside the cluster [Bishop, 2006].

The $k-$means algorithm is based on the use of squared Euclidean distance as the measure of dissimilarity between a data point and the center of a cluster [Bishop, 2006].

2.3.7 Dimensionality Reduction Techniques

Dimensionality reduction is an important task related to compression and coding, statistics, with latent variables, as well as machine learning and sampling theory. In essence, the goal is to change the representation of data sets, originally in a form involving a large number of variables, into a low-dimensional description using only a small number of free parameters. The new representation should describe the data in a faithful manner, by preserving some quantities of interest such as local mutual distances [Coifman and Lafon, 2006]. In the following, we will review four dimensionality reduction techniques that will be referenced in the discussion of the results of Chapter 7.

Principal Component Analysis

Principal Component Analysis (PCA) is a technique that is widely used for dimensionality reduction [Jolliffe, 2002]. PCA is a linear transform that represents the data in a new coordinate system in which basis vectors follow modes of greatest variance in the data: it is the optimal linear transformation which divides an observed space into orthogonal subspaces with the largest variance [Sonka et al., 2008]. PCA involves evaluating the mean and the covariance matrix of the $D-$dimensional data set and then finding the M eigenvectors corresponding to the M largest eigenvalues [Bishop, 2006]. This represents a compression of the data set, because for each data point we replaced the $D-$dimensional vector with a $M-$dimensional vector.

Multidimensional Scaling

Multidimensional scaling (MDS) is a linear dimensionality reduction technique used to project data onto a lower-dimensional space [Cox and Cox, 2008]. It finds a low-dimensional projection of the data looking to preserve, as closely as possible, the pairwise distances between data points, and involves finding the eigenvectors of the distance matrix. In case where the distances are Euclidian, it gives equivalent results to PCA [Bishop, 2006].

Kernel Principal Component Analysis

Kernel Principal Component Analysis (Kernel PCA) is a nonlinear generalization of PCA [Schölkopf et al., 1998]. The main idea is that a dataset in the original data space is projected by a nonlinear transformation ϕ into a feature space. The function ϕ maps the original dataset into a larger feature space by creating nonlinear combinations of the dataset. By performing PCA in the feature space, the principal components are obtained [Bishop, 2006].

Diffusion Maps

Diffusion maps is a nonlinear dimensionality reduction technique introduced by Coifman and Lafon [Coifman and Lafon, 2006]. It finds a family of embeddings of a data set into a lower-dimensional Euclidean space whose coordinates can be obtained from the eigenvectors and eigenvalues of a diffusion operator on the data. The diffusion distance between probability distributions centered at points in the embedded space is equal to the Euclidean distance between those points.

2.4 Evaluation Metrics

Due to the fact that this PhD thesis is aimed to develop and calibrate a pipeline for detecting, segmenting, characterizing and classifying human sperm heads, we need a number of evaluation metrics for each stage of the proposed pipeline. In this sense, in this section we list the evaluation metrics that we decided to use for detection, segmentation and classification of human sperm heads. There are no specific evaluation metrics for the characterization stage as this stage was validated along with the classification scheme.

2.4.1 Evaluation Metrics for Detection

Given a detection gold-standard, we have four possible scenarios to evaluate a detection algorithm: true positives (TPs), the number of correctly detected objects, true negatives (TNs), the number of correctly detected non objects, false positives (FPs), the number of

wrongly detected non objects, and false negatives (FNs), the number of wrongly non-detected objects.

To evaluate our detection algorithm (introduced in Chapter 5), we use *true-positive rate* (TPR), *false-positive rate* (FPR) and *precision* as follows.

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$
$$precision = \frac{TP}{TP + FP}$$

2.4.2 Evaluation Metrics for Segmentation

Given a segmentation gold-standard, we can use different metrics to evaluate the quality of a segmentation method. In general, the idea is that the automatic segmentation S has to be compared to the manually segmented image (gold-standard) G, by computing some evaluation metrics. These metrics can be based on spatial overlap measures (e.g., Dice coefficient [Dice, 1945]) and on distance measures (e.g., Hausdorff distance [Alt and Guibas, 2000]). Table 2.1 summarizes the evaluation metrics for segmentation results used in this thesis.

Metric	Dice Coefficient	Hausdorff Distance
Expression	$D = \frac{2\|S \cap G\|}{\|S\|+\|G\|}$	$H(S,G) = max(h(S,G), h(G,S))$ $h(S,G) = \max_{s \in S}(\min_{g \in G} \mathrm{d}(s,g))$
Range	[0,1]	[0,1]
Interpretation	0: no spatial overlap between S and G 1: complete overlap	0: maximum agreement between perimeters of S and G 1: maximum disagreement between perimeters

Table 2.1: Metrics for quality of segmentation

2.4.3 Evaluation Metrics for Classification

Taking advantage of having a classification gold-standard, the idea is to evaluate the performance of a classification method against that gold-standard. Furthermore, it is possible to compare the proposed method with each of the individual experts who collaborated in generating the gold-standard. In this thesis, we used four metrics to evaluate the two different stages of our classification scheme as well as the whole classification method (see Chapter 7 for details).

Confusion matrix

A confusion matrix is used to describe the performance of a classification method on a gold-standard [Kohavi and Provost, 1998]. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The entries in the confusion matrix are related to: true positives (TPs), the number of correctly classified objects; true negatives (TNs), the number of correctly rejected objects; false positives (FPs), the number of wrongly classified objects; false negatives (FNs), the number of wrongly rejected objects. Here, we present a confusion matrix for a two-class classification method that can be extended for a multi-class classifier:

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

Accuracy rate

The accuracy rate indicates the proportion of the total number of predictions made by the classification method that were correct compared against the gold-standard. It is defined in terms of confusion matrix entries as:

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

True positive rate

True positive rate, also called sensitivity and recall, measures the proportion of positive cases that were correctly identified as such. It is defined in terms of confusion matrix entries as:

$$TPR = \frac{TP}{TP + FN}$$

2.5 Summary

In this chapter, we have discussed relevant topics required to understand the subsequent chapters. The discussed topics come from a variety of research areas. First, we have discussed biological concepts related to the structure and morphological deffects of human sperms (Section 2.1). These concepts will be useful at understanding Chapters 4 to 7. Second, with respect to the image processing notions (Section 2.2) discussed above, it is important to realize that their main application will be focused in Chapter 5. For the sperm detection and segmentation algorithms, it will be relevant topics about pixel connectivity, mathematical morphology, color spaces and clustering methods ($k-$means). The anisotropic diffusion filter and the active contour approach will be useful to understand the characterization stage of our

methodology (Chapter 6). Third, the topics related to pattern recognition (Section 2.3) will be required mainly in Chapters 6 and 7 as well as Appendix B. For understanding Chapter 7, the reader will find helpful the sections that discussed combination of classifiers, combination rules and dimensionality reduction techniques. While the discussion about base classifiers is referenced in the Appendix B.

Chapter 3

Related Work

This chapter presents an up-to-date review of related research. For easier reading, this chapter is organized according to the main stages of this PhD thesis.

First, Section 3.1 discusses the most prominent approaches for detecting and segmenting eukaryotic cells and separating sperm cells (see Section 3.1.1) from other eukaryotic cells (see Section 3.1.2). This section pays special attention to methods for working specifically with images from bright-field microscopy. Section 3.2 reviews published research works that addressed the problem of characterizing and classifying sperm cells at first, and grain kernels, later.

There are not many published research papers for sperm cell computer characterization and classification that encompasses a full morphological analysis of sperm (see Section 3.2.1). There is only one proposed approach for characterizing and classifying human sperm cells as normal and abnormal sperm. Besides, to the best of our knowledge, there is no published framework for identification of human sperm cell abnormalities. On the other hand, there are a number of commercial applications for morphological analysis of sperm. Therefore, the characteristics of the three major commercial applications are also described.

Regarding the small variability inter classes and the visual features of objects to characterize and classify, a very related area is grain featuring and classification, including food and pollen grains (see Section 3.2.2). In this area, the problematic situation is the same as in sperm classification: the manual method is a hard-laborius task with subjective results. Thus, the development of automatic classification methods in all these research fields would increase the overall classification rate.

3.1 Detection and Segmentation

3.1.1 Human Sperm Cells

Few publications are concerned with the detection and segmentation problems associated with these cells. The following lists the most representative approaches with published experimental results.

A two-stage method for segmentation of sperm heads and mid-pieces was presented by Carrillo et al. looking for an objective analysis of human sperm morphology [Carrillo et al., 2007b,Carrillo et al., 2007a]. At the first stage, the objects obtained by thresholding using the Otsu method [Otsu, 1979] are classified through histogram analysis. Some particles are then removed according to their size. After that, each sperm cell detected (head and mid-piece) is enclosed in a bounding box. Next, each sperm cell is extracted from the original RGB color image. At the second stage, the authors proposed to segment the head and mid-piece by applying a nth-fusion method to the enhanced image. The nth-fusion method is based on nth-level thresholding of an image followed by intersection with n special growing masks, constructed using prior object morphological models. The proposed method achieved 89.5% of correct segmentation of sperm heads. To the best of our knowledge, this proposal is the state-of-the-art reproducible method using images from bright-field microscopy, and other research works compare their results with it [Bijar and Mikaeili, 2011, Bijar et al., 2012].

Gonzalez et al. presented a method to segment images of live and dead spermatozoa in positive phase contrast images [González-Castro et al., 2009]. The authors addressed the sperm head segmentation by means of applying an intelligent thresholding. The threshold changed its value when the binary image obtained did not fulfill some surface and eccentricity conditions. Then, using the same automatic criteria, the badly segmented images were processed using the watershed transform. Experimental results showed almost 91% of correct segmentation of sperm cells.

Bijar et al. proposed a method for segmentation of sperm acrosome, nucleus, mid-piece and identification of tails [Bijar et al., 2012]. The segmentation step is performed by means of a Bayesian classifier which uses entropy based expectation maximization and a Markov random field. For sperm tail identification, the authors proposed to use a structural similarity index and local entropy techniques. The paper presented results that outperformed those of Carrillo et al. achieving 94.3% of correct segmentation of sperm heads, however the experimental framework is so weak that it makes the validation of those results very difficult.

Shojaedini and Heydari presented an approach for automatic segmentation of human sperm [Shojaedini and Heydari, 2014]. The authors proposed a sperm detection technique based on minimization of the information distance between the original and processed images using co-occurrence matrix. The final segmentation stage consisted of an algorithm based on the watershed transform using candidates from the previous stage. The proposed framework achieved 96% of correct segmentation of sperm cells in experiments conducted on microscopic videos captured from human semen.

3.1.2 Other Eukaryotic Cells

There are two types of cells: prokaryotic and eukaryotic. Prokaryotic cells are simpler cells like bacteria cells, while eukaryotic cells include plant and animal cells [Karp, 2013]. Numerous image processing techniques were used in different approches for detection and segmentation of various kinds of eukaryotic cells different from sperm cells.

Korzynska et al. presented a semi-automatic method for segmentation of cells from bright-field microscopy [Korzynska et al., 2007]. The algorithm consisted of a combination of the use of texture threshold in a window sliding approach and of Prewitt edge detection method. It was tested with neuthophils and lymphocytes and compared against common segmentation methods as watershed and active contours. The main drawback is that the method required much human interaction to provide competitive results. No specific segmentation accuracy rate was reported.

Shah proposed an automatic cell image segmentation method using a shape-classification model [Shah, 2008]. The main idea was to use k-means approach coupled with cluster merging, based on a fitness function using proximity graphs to obtain cell localization. A joint segmentation-classification approach incorporating an ellipse as a shape model was used to detect the final cell contour based on the level-set method. The method was evaluated with cytological images for thyroid lesions, reaching 92% of segmentation accuracy.

Bradbury introduced a combined method for segmentation of bright-field eukaryotic cell images for cancer studies [Bradbury, 2009]. The proposed approach involved the combination of multiphase active contour and watershed method. There were three stages in the algorithm: 1) to use level-set method to select the seeds for a watershed technique, 2) after applying a watershed technique, a binary image with two regions (cell and background) is obtained, and 3) to use a simple level-set method on this binary image to extract the boundary of the cell. No specific segmentation accuracy rate was reported.

The proposed method of Ali et al. was a complete approach for cell detection and boundary and nucleus segmentation [Ali et al., 2012]. The authors used thresholding based on deconfused images for cell detection. While the boundary segmentation was achieved using local orientations as well as level-set based active contour. They proposed to use a texture model to segment the nucleus of detected cells. The evaluation of this algorithm was done using several types of carcinoma cells (mammary and cervical among others) achieving 80% of detection precision and 75% of segmentation accuracy compared to their own gold-standard obtained from experts.

3.2 Characterization and Classification

In this section, relevant research works related to characterization and classification of sperm heads are discussed. A few research studies were published about characterization of animal sperm cells. In case of classification sperm cells, there is only one proposed approach for classifying human sperm cells as normal and abnormal sperm. The other research stud-

ies discussed in this section are related not to morphology sperm analysis, but to vitality sperm analysis as they proposed membrane integrity assessment techniques. We also discuss research studies from a very related area: grain featuring and classification. In this case, the problematic situation is the same as in sperm classification: the manual method is a hard-laborius task with subjective results.

3.2.1 Sperm Cells

A number of image analysis and machine learning methods have been suggested for description and classification of sperm cells, from human and animal semen.

Beletti et al. described the morphology of animal sperm heads from a computational point of view [Beletti et al., 2005a]. Sperm morphology was quantified in terms of the following morphological features: head area, perimeter, width, length, aspect ratio, ellipticity, shape factor, width of sperm basis, the three first Fourier values, side and anterior-posterior symmetry and hydrodynamic coefficient.

Beletti et al. characterized the animal sperm head shape using a multiscale curvature estimation [Beletti et al., 2005b]. This work described how a spectral approach to derivative estimation was useful for extracting relevant features for sperm head morphology assessment. Using the Fourier transform, the proposed method calculated three morphological features: width of sperm head, implantation symmetry and bending energy of the frontal portion of the head. Experiments with real data from several animal species are shown.

Severa et al. developed a framework to characterize stallion sperm heads and evaluate the intrinsic shape variability [Severa et al., 2010]. Sperm head shape characteristics including aspect ratio, position of the center of gravity, curvature and degree of roundness were assessed and analysed using Fourier descriptors and inverse Fourier transformation.

Abbiramy and Tamilarasi evaluated the accuracy of neural networks for human sperm classification, discriminating normal from abnormal sperm cells [Abbiramy and Tamilarasi, 2011]. The feature vector proposed by the authors includes first order statistics, textural features (gray level co-ocurrence matrix) and morphological features (head area, perimeter width and length, eccentricity and orientation, among others). For classification purposes, the authors evaluated three neural networks techniques: feed forward, radial basis and Elman back propagation. The experiments were done using images taken from WHO laboratory manual [WHO, 2010b], and showed that the radial basis network produced the highest classification accuracy of 60%, 75% and 70% when trained with statistical features, combined features (statistical, textural and morphological) and morphological features, respectively.

From a different perspective, a number of publications are concerned with the membrane integrity validation. Sanchez et al. proposed several methods for the classification of boar spermatozoa heads based on their intracellular intensity distribution and using a least square classifier [Sánchez and Petkov, 2009]. Experiments were conducted on real images classified by veterinarian experts as intact or damaged, and showed a missclassification error of 20% when classifying sperm cells as intact or damaged.

Alegre et al. developed a method for classification of the acrosome integrity of boar spermatozoa images [Alegre et al., 2012]. In this scenario, the main goal is to discriminate live from dead sperm cells. The proposed approach aimed to characterize the acrosomes by means of textural and moment-based features. The head texture was described using several Haralick descriptors computed after applying a discrete wavelet transform. In case of moment-based features, the authors proposed to use Hu, Legendre and Zernike moments. For classification purposes, kNN classifier and multilayer perceptron neural network were used as classifiers. With the combination of textural features and multilayer perceptron, this approach achieved 94.9% as correct classification of sperm cells discrimating live from dead cells.

Gonzalez et al. used a curvelet-based descriptor to classify acrosoma integrity [González-Castro et al., 2012]. The classification stage used a multilayer perceptron neural network. Experiments were conducted to evaluate the proposed methods against other texture descriptors based on the wavelet transform, as well as moment-based descriptors. The results showed that textural descriptors performed better than other evaluated descriptors, and the proposed descriptor achieved a correct classification in 97% of cases, trying to discriminate live from dead sperm cells.

A method for assessing the acrosome state of boar sperm heads was proposed in [Alegre et al., 2013]. The authors introduced a n-contour descriptor that includes local texture features (local maximum gradient values, local mean gray level values and local standard deviation). In the classification evaluation, three different classifiers were used: Relevance learning vector quantization, class-conditional mean and kNN. The experimental results showed that gradient magnitude data offered the best rate with an overall test error of 1%, while using 11NN at classifying sperm cells according to their vitality.

A recent proposal of Garcia et al. used an early fusion of texture and contour descriptors for acrosome integrity assessment [García-Olalla et al., 2015]. With this purpose, three descriptors have been considered, and combined in a whole feature vector: 1) thirteen Haralick features on the co-occurrence matrix of the original image and on the images obtained in the first Haar discrete wavelet transform decomposition, 2) a rotation invariant and uniform LBP (local binary patterns), and 3) Fourier shape descriptors for the contour of the sperm head. The classification was performed with a SVM backed by a least-square training algoritm and a linear kernel. The proposed approach provided 99.2% of correct classification of sperm cells in two classes (live and dead), outperforming the previous approaches.

3.2.2 Grain Kernels

Food Grains

Various research works have been published addressing characterization and classification of food grains, especially corn and cereal grains. An approach for classifying rice seeds variety was introduced by Liu et al. [Liu et al., 2005]. The main goal was to classify rice seeds in six classes. This approach proposed the use of seven color features and fourteen morphological

features in combination with discriminant analysis, achieving almost 85% of average correct classification in six classes.

Anami et al. proposed an automatic method for grain classification in six classes, including wheat, corn and redgram, among others [Anami et al., 2005]. The classification was performed based on color and morphological features. Using neural networks approach, the proposed method achieved up to 90% as classification accuracy.

Chen et al. developed a corn variety identification approach [Chen et al., 2010]. Several morphological (13), geometric (17) and color features (28) were extracted. After ranking features using stepwise discriminant analysis, the optimum sets of features for two models were created individually. A two-stage classifier combining distance discriminant and a back propagation neural network was built for identification. Experiments showed that the average classification accuracy for five corn varieties (including white, yellow and mixed corn) was up to 94%.

Kaur and Sinngh presented a method for classifying and grading rice kernels [Kaur and Singh, 2013]. The main idea was to use a multi-class SVM based approach for grading rice kernels in one of the following classes: Premium, Grade A, Grade B and Grade C. For characterizing the rice kernels, the authors used ten geometric features and reported 86% of correct classification.

Mebatsion et al. developed a method for classification of cereal grains using morphological and color features [Mebatsion et al., 2013]. The authors used chain code representation for grain boundary and extracted elliptic Fourier coefficients from them. The morphological features included symmetrical Fourier index, aspect ratio, major diameter and roundness. Using these features, a least-square classifier achieved around 98% of correct classification, discriminating among four classes: barley, oat, rye and wheat.

Pollen Grains

Over the past decade, there has been increasing research interest in the automatic classification of pollen grains. Accurate identification of pollen types is a relevant topic in different scenarios. For instance, quantifying the concentration of airborne pollen may help people suffering from allergic reactions to adopt adequate treatment strategies [Marcos et al., 2015]. In apiculture, pollen classification is required to identify nectar sources, which determine the quality of the product and enable the authentication of its origin [Kaya et al., 2013]. Currently, pollen identification is based on visual inspection of microscopy images. It is a time-consuming and costly procedure since a trained expert must manually classify each pollen grain [Mitsumoto et al., 2009]. The combination of image processing techniques with machine learning approaches achieves very promising results.

Treloar et al. proposed an automatic pollen grain classification method [Treloar et al., 2004]. In this work, the authors used morphological, geometrical and texture features for the charaterization stage. These features were used with a multi-variable statistical classifier for recognizing twelve species of pollen reporting a 95% classification accuracy.

Zhang et al. presented a method for pollen classification that used neural networks [Zhang et al., 2004]. The descriptor was obtained by means of combining texture features computed by the Gabor transform with geometrical features computed by invariant moments. The whole approach showed a correct classification in 97% of test cases while classifying five species of pollen grains.

Rodriguez-Damian et al. proposed a method for detecting and classifying pollen grains [Rodríguez-Damián et al., 2006]. The authors used shape and texture features as well as conventional first-order and second-order statistical features computed from Haralick's co-ocurrence analysis to describe grains of pollen. These descriptors were combined with SVM, neural networks and minimum distance classifiers. Using texture features, the proposed method achieved 88% of correct classification, while using shape features reached 80% of correct classification evaluated on the same dataset.

Allen et al. developed an automatic system for localization and classification of three species of pollen grains [Allen et al., 2008]. The authors used geometric and texture features to characterize grains of pollen. The chosen classifier was neural networks, achieving 90% of correct classification.

Punyasena et al. developed a supervised, layered, instance-based classification system for pollen grains [Punyasena et al., 2012]. The features used to describe pollen grains include shape, size and texture. For the classification stage, the authors proposed to use a kNN classifier with leave-one-out bias optimization. The method was evaluated on two datasets (with two and three types of pollen grains, respectively) achieving a similar rate around 93% of correct classification in both cases.

Chica proposed a method for authenticating pollen grains in bright-field microscopic images [Chica, 2012]. The author proposed to combine one-class classification techniques to reject unkown pollen grain objects, after a comparison of five models: Gaussian estimator, support vector data description, and three variants of the $k-$NN. The features used to describe pollen grains included shape, color and textural features as well as the features for describing the exine of the pollen grain. The overall accuracy of classification reached 92%.

Kaya et al. presented a method for pollen classification in honey [Kaya et al., 2013]. GLCM (grey level co-occurrence matrix) texture features and neural networks were used for the identification of pollen grains in honey. GLCM has been calculated in four different angles and offsets for the pollen of the plant and the honey samples. The method was evaluated using ten different species of pollen grains achieving a classification accuracy of 88%.

Redondo et al. developed a novel method for automatic pollen classification in bright-field microscopy [Redondo et al., 2015]. The authors used morphological, textural and statistical features for charaterizing pollen grains. In addition, they proposed a new descriptor to measure contour profile. For classification purposes, the use of random forest, Fisher classifier and SVM was evaluated. The results showed that combining descriptors could improve overall classification accuracy up to 99%.

3.3 Commercial Applications for Computer-Aided Sperm Analysis

In the context of commercial applications, companies in the interface of research and development offer computer assisted systems for semen analysis, usually referred to as CASA (Computer Aided/Assisted Sperm Analysis). CASA technology was developed in the late 1980s for analyzing sperm movement characteristics or kinematics and has been highly successful in enabling this field of research [Davis et al., 1992,Kruger et al., 1993]. CASA has also been used with great success for measuring semen characteristics such as sperm concentration and proportions of progressive motility in many animal species, including wide application in domesticated animal production laboratories and reproductive toxicology. However, the same is not yet for human clinical laboratories due to biological and technical limitations [Mortimer et al., 2015].

Despite many unresolved problems, the state-of-the-art in CASA has progressed greatly in the last decade so that several modern CASA systems are no longer *black boxes* but embody additional (and often user-adjustable) operational parameter settings that address many of the problems that constrained earlier systems [Mortimer et al., 2015]. Furthermore, CASA products offer solutions for sperm concentration, motility, morphology, vitality and fragmentation with different degrees of accuracy. Specifically, in the case of sperm morphology assessment, the repeatability and validity of CASA systems could be higher than any subjective morphological evaluation, but it is required that the quality of the preparation, choice and quality of fixing, thickness of the preparation, choice of dyes, type of light and adjustment of optics are carefully chosen [Wang et al., 1991,Davis and Gravance, 1993,Lacquet et al., 1996,Coetzee et al., 1999,Auger, 2010].

Common CASA systems that offer morphological analysis tools of human sperm are *IVOS* - Integrated Visual Optic System (Hamilton-Thorne Biosciences, Beverly, MA, USA) [1] and *SCA* - Sperm Class Analyzer (Microptic Automatic Diagnostic Systems SL, Barcelona, Spain) [2].

3.3.1 IVOS - Integrated Visual Optic System

Hamilton-Thorne IVOS-II features improved integrated phase contrast optics. It offers operational settings such as frame rate, number of images to be captured, interactive/automated illumination and focus control to optimize tail detection, as well as user-adjustable settings to control more aspects of cell detection and static/ non-progressive/progressive motility differentiation. Options for sperm morphology include the research-orientated *Metrix* system [3] that learns from user interaction for improving classification of sperm head morphology (normal and abnormal), reporting sperm head parameters such as elongation and area. However, this system can not analyze either the sperm mid-piece or tail morphology, and hence

[1] *http://www.origio.com/en/products/all-products/ivos-sperm-analyzer*
[2] *http://www.micropticsl.com/products/sperm-class-analyzer-casa-system/*
[3] *http://www.hamiltonthorne.com/index.php/products/sperm-analyzer-options/metrix-morphology*

only provides a classification of *normal* sperm cells based on sperm head morphology [Mortimer et al., 2015]. This CASA system has been used for research purposes in a number of published works [Douglas-Hamilton, 1995,Coetzee et al., 1999,Atkins et al., 2003,Phetudomsinsuk et al., 2008].

3.3.2 SCA - Sperm Class Analyzer

The Sperm Class Analyzer (SCA) comprises four modules (concentration and motility, morphology, vitality, and fragmentation) that are fully integrated and associated with an extensive database that will allow results imported and exported in many of the formats used by hospitals. With respect to sperm morphology, SCA needs to be better elaborated in humans and different animal species by including the entire cell, especially tail characteristics, and also using polychromatic stains. SCA already employs adaptive thresholding to visualize the entire spermatozoon and then measure all its components, at least in ram spermatozoa [Mortimer et al., 2015]. SCA has been used in several research studies [Soler et al., 2003, Hidalgo et al., 2008, Hidalgo and Dorado, 2009, Maree and van der Horst, 2013, Dearing et al., 2014].

3.4 Summary

In this chapter we have reviewed relevant research and commercial works related to the main topic of this thesis. In this sense, we have reviewed four segmentation approaches for human sperm cells (Section 3.1.1), two of which used images from bright-field microscopy achieving from 90% to 96% of correct segmentation. We also reviewed four segmentation methods tested with other eukaryotic cells different from sperm cells (Section 3.1.2) with correct segmentation range between 75% to 92%. For the case of characterization and classification related work, our review was focused on sperm cells and grain kernels, due to the similar problematic issues in both research areas. In the sperm cells scenario (Section 3.2.1), we have reviewed three proposals for characterizing animal sperm cells using morphological parameters. We discussed only one research work related to human sperm classification regarding two classes: normal and abnormal, achieving from 60% to 75% of correct classification depending on which kind of features are used. The remaining five publications commented in this section are related to vitality assessment, by means of membrane integrity validation, achieving from from 80% to 99% of correct classification. In this case, the main goal is to discriminate between live and dead sperm cells (or intact and damaged membrane). With respect to grain kernels, we presented an up-to-date review of research works focused on food grains (mainly rice and corn kernel classification with correct classification range between 85% and 98%) as well as pollen grains achieving up to 99% of correct classification in different numbers of species of pollen. Computer aided sperm analysis technology (Section 3.3) has enormous potential as a research tool, in reproductive toxicology, in animal production, and for human clinical analyses (sperm concentration and motility). However, urgent improvements in CASA technology need to address sperm morphology analysis.

Chapter 4

Gold-Standard Generation

A ground-truth represents the absolute truth for a certain application. For instance, in cancer detection from medical images, a suspicious region is malignant or benign. The absolute truth (whether it is cancer or not) can be obtained from biopsies. The biopsy results form the ground-truth for those medical images [Raykar et al., 2009]. Unfortunately, for many applications, especially in biomedicine, such as morphological analisis of sperm cells, it is impossible to count with a ground-truth because of the subjectivity of the task [Yan et al., 2010]. A valid alternative consists in asking many experts in the field for their opinion about punctual cases, in order to generate a gold-standard [Fuchs and Buhmann, 2011].

This chapter presents the design and construction of a segmentation gold-standard, regarding the opinion of one expert, as well as a classification gold-standard, regarding the opinion of three referent experts in the field.

In this sense, we describe in detail the staining method, the features of the equipment we used to capture the images and specific details about images and manual methods in our both gold-standards.

This represents a significant contribution of our work, because at present there is no segmentation neither a classification public gold-standard, so the few existing methods cannot be properly evaluated.

4.1 Sample Preparation

Sperm samples were stained with a modified Hematoxylin/Eosin assay, in order to distinguish different parts of sperm morphology (Figure 2.1). First, the sperm smear was fixed with ethanol 70% and immersed in Harris' Hematoxylin for ten seconds for nuclear staining. Slides were washed with tap water for ten minutes to remove residual staining. Later, slides were immersed in 1% Eosin for two minutes to stain the acrosome, mid-piece and tail in a pink-orange color. Finally, the sample was washed with distilled water for one minute. Samples were then air dried and fixed. This staining procedure allows samples to be used for more

than one year.

4.2 Image Acquisition

Digital images were acquired using an optical microscope (Axiostar Plus, Carl Zeiss Inc, Wetzlar, Germany), a 63x objective (oil, NA 1.4) with an adapter of 0.63x and a digital camera (scA780-54gc, Basler AG, Ahrensburg, Germany).

4.3 Source of Sperm Smears

We obtained the sperm smears from:

- Laboratory of Spermiogram, Program of Anatomy and Developmental Biology (ICBM), Faculty of Medicine, University of Chile, Santiago, Chile
- Maternal Child Research Institute (IDIMI), San Borja Arriarán Hospital, Santiago, Chile
- Reproductive Medicine Unit, Clínica Las Condes, Santiago, Chile.

In both cases, we obtained sperm smears from males between 28 and 35 years old, with children of less than one year old.

4.4 Segmentation Gold-Standard

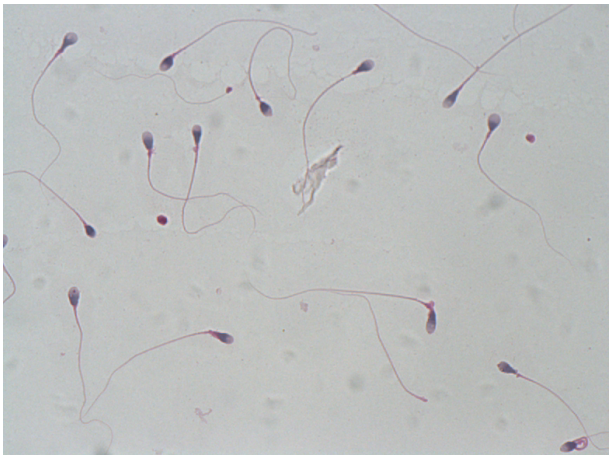
The segmentation gold-standard consists of 20 images with 264 sperm cells, where 210 are valid sperm cells (not at the border of image, without containing noise, etc). Each image has 780×580 pixels (see Figure 4.4a). All images were obtained only from one patient.

For each of these images, hand-made segmentation masks have been generated under supervision of a referent expert in the field. The manual segmentation masks include the whole sperm cell, head, mid-piece, and tail (see Figure 4.4b) as well as acrosome and nucleus (see Figure 4.4c).

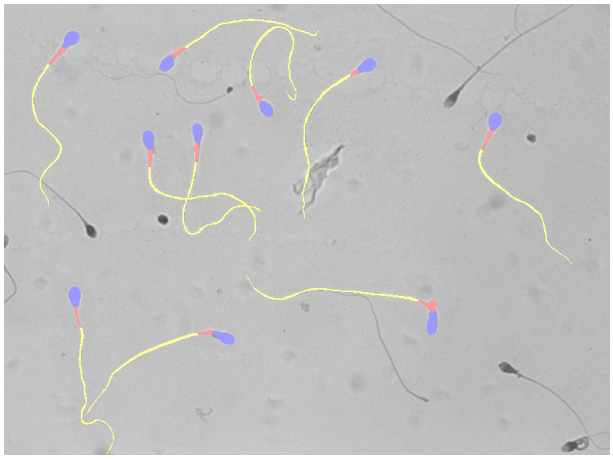
4.5 Classification Gold-Standard

For the classification gold-standard, we collected smears from four patients (named p1, p3, p4 and p5). For each patient, we had one or two smears (pl1 and pl2), with a total of six smears, as is detailed in Table 4.1.

a)



b)



c)

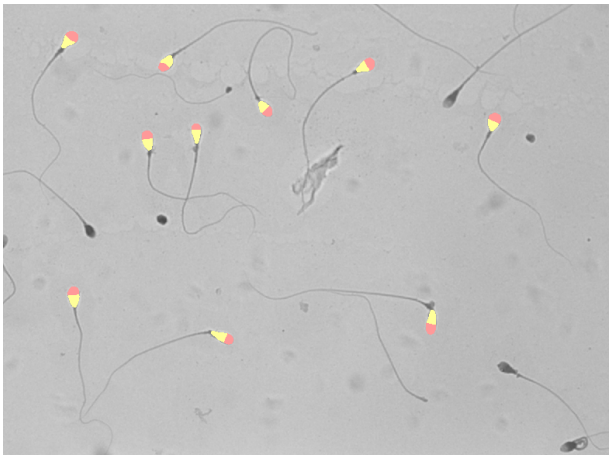


Figure 4.1: **Segmentation gold-standard.** **(a)** Bright field image: 780×580 pixels $\approx 163 \times 121$ μm. **(b)** Overlay with masks from manual segmentation for head (blue), mid-piece (red) and tail (yellow). **(c)** Overlay with masks from manual segmentation for acrosome (red) and nucleus (yellow).

Patient	Smear	Sperm cells
p1	pl1	399
	pl2	319
p3	pl1	213
p4	pl1	253
	pl2	157
p5	pl1	187

Table 4.1: Sample source details

We obtained 1,872 sperm head images that could be classified according to 11 head defects as WHO stablishes [WHO, 2010b]. We decided to build a classification gold-standard of human sperm head, considering only the five most representative classes among the obtained smears. Finally, the gold-standard contains 1,854 observable and evaluable sperm cells whose class is one of the following: *normal, tapered, pyriform, small*, and *amorphous*. In Figure 4.2, representative sperm cells from each class are showed. The manual classification process has been performed independently, per patient/smear, by three referent experts with vast experience in morphological sperm analysis.

A very important aspect in the analysis of the gold-standard is the discussion on the inter-expert agreement distribution. As this gold-standard was built with the cooperation of three experts, there are three different agreement scenarios: one (basis set), two experts (partial agreement - PA), or three experts agree on the same label for a given sperm head (total agreement - TA). The first set contains 1,854 sperm head labels, but a sperm head can be classified into three different classes by the three different experts. The second set contains 1,132 sperm heads, meaning that there are 1,132 sperm heads with partial agreement and without overlapping. The third set contains only 384 sperm heads, with total agreement between the three expert technicians.

Table 4.2 shows the number of sperm cells per class for each agreement scenario. Considering the manual classification agreement by at least one, two, or three experts, the class *Amorphous* is the largest class in all cases, concentrating over 68% in the total agreement scenario (see Figure 4.3). The class *Tapered* is the second largest class, slightly decreasing assignment percentage as agreement among experts increases and concentrating around the fifth part of the samples. The *Pyriform* and *Small* classes decrease their assignment percentage while increasing agreement among experts (from 188 to 7 and from 152 to 11 sperm cells, respectively). It is important to note that in the case of class *Pyriform*, there is less than 2% of the samples with total agreement of experts (only 7 sperm cells). The only class that maintains its assignment percentage is the class *Normal*, regardless if one considers the label agreement of at least one, two, or three experts, and consisting of less than 10% of the samples, but ranging from 175 to only 35 sperm cells with total agreement among experts.

The underlying complexity in the sperm head classification task can be studied by eval-

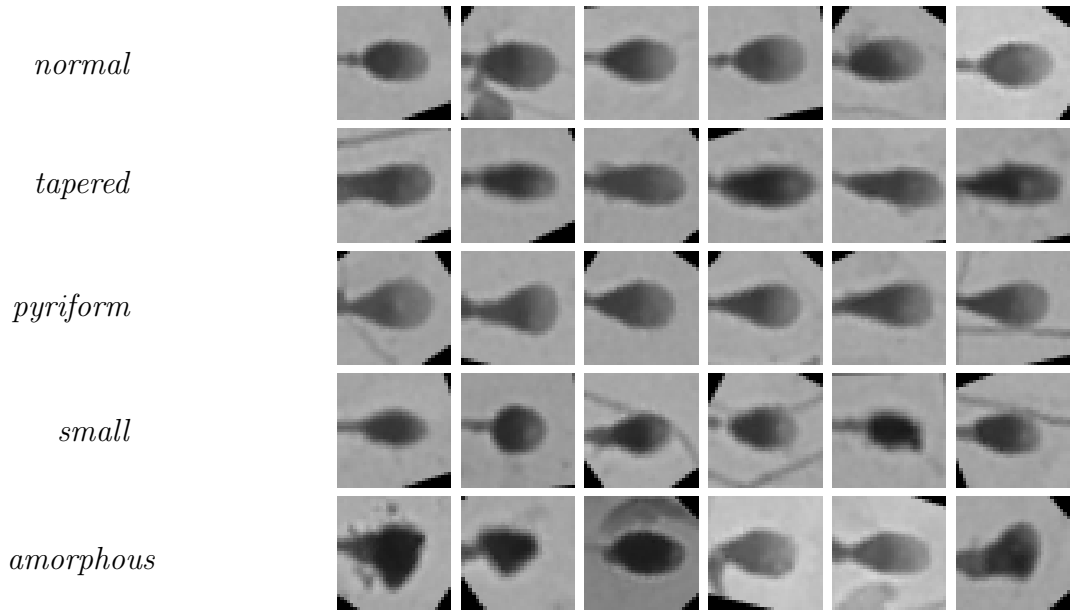


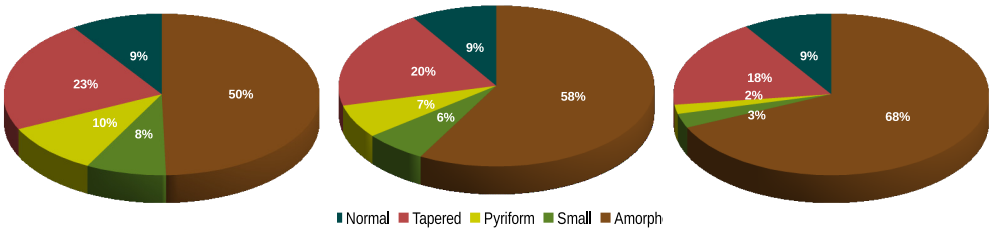
Figure 4.2: **Classification gold-standard.** Representative examples of normal, tapered, pyriform, small and amorphous sperm cells (Image size: 35×35 pixels $\approx 7 \times 7$ μm).

Agreement among experts	Normal	Tapered	Pyriform	Small	Amorphous
At least one (Basis set)	175	420	188	152	919
Partial agreement (PA)	100	228	76	72	656
Total agreement (TA)	35	69	7	11	262

Table 4.2: Inter-expert agreement

uating the degree of agreement between different experts. Figure 4.4 shows the inter-expert agreement per class. Considering the number of heads that have been classified into a certain class by at least one expert as the size of the basis set, figure 4.4 shows the percentage of partial agreement and total agreement. There are some classes in which it is really difficult to reach an agreement, for example, class *Pyriform*. There is only 40% of sperm heads that were classified as *Pyriform* by at least one expert that reach partial agreement and less than 2% that reach total agreement. It seems to be the most difficult class to get agreement among experts, even seems to be the most morphologically defined class by theory. While the most morphologically ambiguous class in theory, *Amorphous*, turns out to be the class that has the greatest rate of agreement among experts. For this class, this trend is the same for whether partial agreement as in total agreement analysis.

To demonstrate the subjectivity of morphological analysis and dependence of the specialist who performs it, Figure 4.5 shows inter-expert variability per class. *Pyriform* and *Small* are the most defined classes according to their morphological features, and both show a high degree of agreement between two out of the three experts, while the discrepancy with the third expert is really significant. In the case of classes *Normal* and *Tapered*, a high degree of agreement is reached between two technicians, whereas the discrepancy with the remaining expert is very high in the case of class *Tapered*. Class *Amorphous* shows a high degree of discrepancy among all experts. In general, the inter-expert variability analysis shows 60%



a) b) c)

Figure 4.3: **Inter-expert agreement.** (a) Manual classification by at least one expert assigning a class label *Amorphous* covers almost 50% with similar presence of classes *Normal*, *Pyriform* and *Small* (around 10%). (b) For partial agreement, the class *Amorphous* is the biggest class (almost 60%), while classes *Tapered*, *Small* and *Pyriform* sightly decrease. (c) For total agreement, the class *Amorphous* covers almost 70%, while *Pyriform* covers almost 2%. The only class that maintains its assignment percentage is the class *Normal*, and maybe *Tapered*, regardless if one considers the label agreement of at least one, two, or three experts.

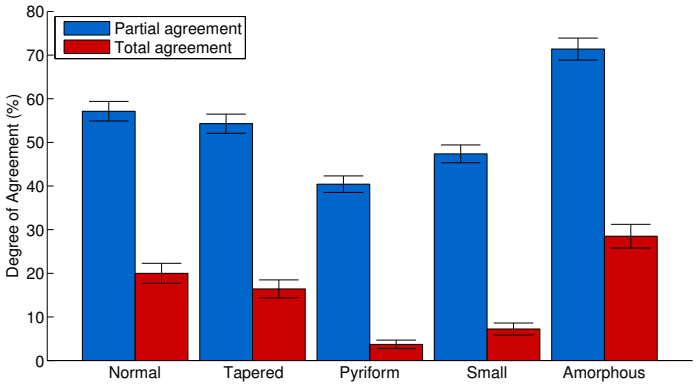


Figure 4.4: **Partial and total inter-expert agreement**. For each class, we show the percentage ± SE of partial and total agreement among experts normalized by the size of the basis set.

of pairwise expert agreement. We calculate Fleiss' Kappa coefficient [Fleiss, 1971] as a way to measure the inter-expert agreement, and it shows and a fair degree of agreement with a coefficient of 0.36 ($\alpha = 0.05$). Furthermore, assuming that a semen analysis would show the percentage of sperm heads in each of the five selected classes, Table 4.3 shows how the spermiogram would be quantified by each expert considered in this study, working with the partial agreement data set.

	Normal %	Tapered %	Pyriform %	Small %	Amorphous %	Other %
Expert1	6.2 ± 0.7	32.4 ± 1.7	11.3 ± 1.0	9.5 ± 0.9	39.6 ± 1.9	1.1 ± 0.3
Expert2	10.1 ± 1.0	16.3 ± 1.2	11.0 ± 1.0	9.0 ± 0.9	52.6 ± 2.1	1.1 ± 0.3
Expert3	11.1 ± 1.0	14.6 ± 1.1	1.7 ± 0.4	2.3 ± 0.4	70.1 ± 2.5	0.2 ± 0.1

Table 4.3: Inter-expert variability (including standard error)

However, a more promising situation is observed while regarding a rough classification

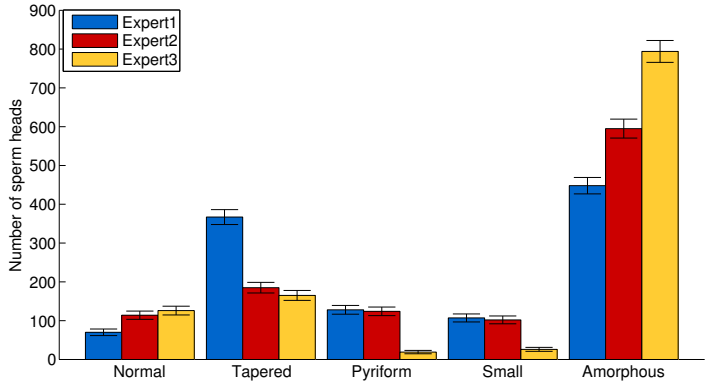


Figure 4.5: **Inter-expert variability in five-class classification**. For research purposes, the expert classifies sperm heads in a number of classes (five in our case: *normal*, *tapered*, *pyriform*, *small* and *amorphous*). We show the number of sperm heads ± SE that belongs to each class according to each of the three experts. The expert manual classification shows a fair agreement among experts with Fleiss' Kappa coefficient of 0.36 ($\alpha = 0.05$).

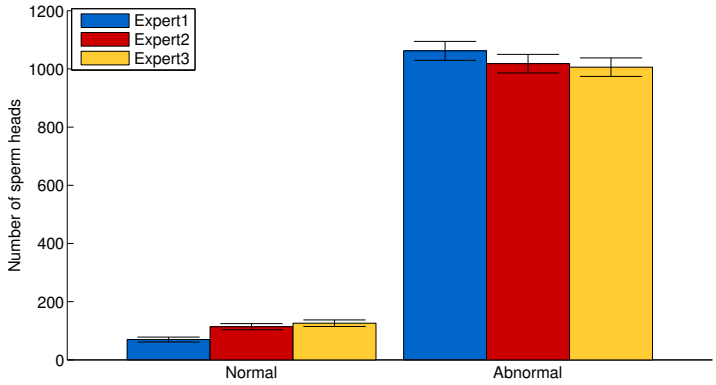


Figure 4.6: **Inter-expert variability in two-class classification**. For normal clinical purposes, the expert only classifies between *normal* and *abnormal*. We show the number of sperm heads ± SE that are considered as normal or abnormal by each of the three experts. The expert manual classification shows a substantial agreement among experts with Fleiss' Kappa coefficient of 0.61 ($\alpha = 0.05$).

(two-class classification problem), used for clinical purposes. Instead of considering five clases, we only regard two categories of sperm heads as normal and abnormal, and the classification rates have a substantial degree of agreement among three experts achieving 0.61 of Fleiss' Kappa coefficient [Fleiss, 1971] ($\alpha = 0.05$), as shown in Figure 4.6. In general, in the two-class classification problem, the inter-expert variability analysis shows almost 94% of pairwise expert agreement.

4.6 Summary

In this chapter we have introduced two gold-standards: a segmentation gold-standard and a classification gold-standard. Both of our gold-standards where built from images captured with the same equipment and following the same tinction protocol using Hematoxylin/Eosin. Another common feature for both gold-standards is the source of the sperm smears. With respect to the segmentation gold-standard, it was built with the cooperation of one domain expert and consists of 20 images with 264 sperm cells. For each sperm cell, there are five hand-segmented masks: head, acrosome, nucleus, mid-piece and tail. With respect to the classification gold-standard, it was built with the cooperation of three domain experts and consists of $1,854$ sperm head images. Each sperm head was manually classified by each expert in one of the following classes: normal, tapered, pyriform, small or amorphous.

Chapter 5

Sperm Head Segmentation

This chapter presents an improved framework for detecting and segmenting human sperm heads as a reliable detection and segmentation presents the first step for all posterior classification algorithms. This fully automatic approach is based on a clustering method as well as on image processing techniques especially adapted for this application. In addition, we propose to combine different color spaces, instead of using only RGB color space. The proposed framework consists of two stages. In the first, our goal is to identify the ROIs of sperm heads (Figure 5.1a). In the second stage, we work on each ROI to accurately segment the sperm head as well as the nucleus and acrosome.

Our main contribution is the application of a clustering algorithm for detecting sperm heads and combining different color spaces. Another contribution is the proposal of a novel algorithm to determine which direction the sperm head points. This is a very important issue for posterior stages in the quest for an accurate morphological analysis.

Our experimental evaluation shows that our proposed method outperforms the state-of-the-art and is supported by the results of different evaluation metrics. Our results achieve notable improvement, yielding above 98% in the sperm head detection process at the expense of having significantly fewer false positives obtained by the state-of-the-art method. Our results also show an accurate head, acrosome and nucleus segmentation achieving over 80% overlapping against the hand-segmented gold-standard. Our method achieves a higher Dice coefficient, lower Hausdorff distance and less dispersion with respect to the results achieved by the state-of-the-art method.

5.1 Detection of Sperm Head

First, we transform the RGB color space to L*a*b*. We choose RGB and L*a*b* after experimental evaluation of the impact of different color spaces such as RGB, L*a*b*, YCbCr, and YQM (Algorithm 1, step 2). We evaluate Hausdorff distance and Dice coefficient values for each color space combination against hand-segmented masks. We then apply the *k-means* clustering algorithm looking for separation of the sperm cells from the background

(Figure 5.1b). We separate the pixels belonging to sperm cells (heads, mid-piece and/or residual cytoplasm, and tails) in one cluster, and the pixels belonging other structures and background in a second cluster (Algorithm 2, step 3). The resulting image contains, in the cluster of a smaller area, the ROIs that we need for the second stage (Algorithm 2, step 4).

Considering that our detection and segmentation algorithm aims for an accurate morphological analysis, there are some conditions that we must meet. Therefore, these regions need to be refined. This refinement includes eliminating sperm cells which touch the border of the image (Algorithm 2, step 5).

In order to eliminate most of the pixels that are not part of sperm heads, we use a binary morphology-based idea. We propose a convolution process with a disk-shaped kernel of size r and unitary weight (Algorithm 2, step 6). After convolution, we remove all pixels with a resulting value below a threshold $sumV$. We refer to this procedure as *eraseTails* (Figure 5.1c).

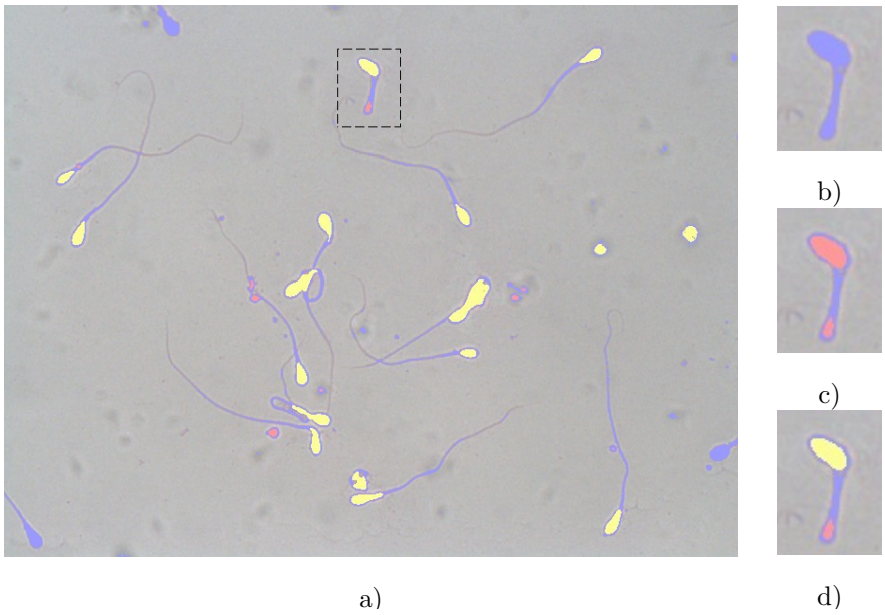


Figure 5.1: **Detection of sperm heads.** **(a)** Original image in RGB color space with resulting ROIs marked on it. **(b)** Blue color represents ROIs after applying *k-means* in RGB and L*a*b* color spaces. **(c)** Red color represents ROIs after erasing tails and sperm cells at border. **(d)** Yellow color represents ROIs after erasing by size. Yellow pixels constitute the final ROIs of this stage. Image size: 780×580 pixels $\approx 164 \times 122 \mu m$.

In the next section, we present our approach to segment sperm heads. After the ROI detection, we individualize each sperm head and work separately with each one. We segment the whole head and then process it to identify the regions of nucleus and acrosome. Afterward, we describe the algorithm that we propose for each step.

Algorithm 2 Detection of sperm heads

imRgb: original image
r: size of neighborhood for e*raseTails*
sumV: threshold value for sum inside neighborhood in e*raseTails*
1: imLab ← transformRGBtoLAB(imRgb)
2: data ← [imRgb(1) imRgb(2) imRgb(3) imLab(1) imLab(2) imLab(3)]
3: [cluster1,cluster2] ← kmeans(data,2)
4: cluster ← chooseMinorCluster(cluster1,cluster2)
5: noBorderImage ← eraseBorderSperms(cluster)
6: finalImage ← eraseTails(noBorderImage,r,sumV)
7: return finalImage

5.2 Segmentation of Sperm Head

For each individual sperm head (Figure 5.2a), we first refine the detected candidate head by means of applying morphological opening and discarding objects whose size are out of range between *minTs1* and *maxTs1* (Algorithm 3, step 2). We work with the color-opponent dimensions of L*a*b* color space and with the Cr component of the YCbCr color space (Algorithm 3, step 3). We then apply *k-means* only in the particular ROI (Figure 5.2b) to separate the darkest part of the head from the rest (that could be acrosome and residual cytoplasm). As this portion of the head is smaller than the real head (Figure 5.2c), we need to enlarge it, up to the region of interest, but without residual cytoplasm or mid-piece. Thus, it is important to determine the front direction of the head (Algorithm 3, step 7). We use *maxTc* as a maximum size threshold (Algorithm 3, step 6) to consider a set of pixels as a candidate head whose pointing direction is important.

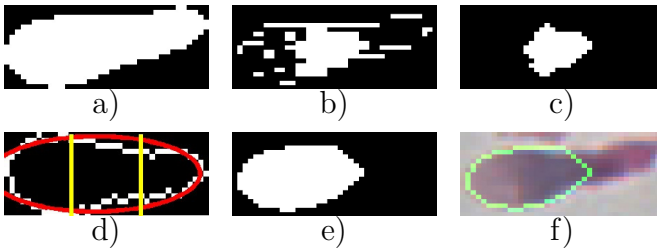


Figure 5.2: **Segmentation of sperm heads. (a)** Detection of sperm head (as returned from Algorithm 2). **(b)** ROI after applying *k-means* in L*a*b* and YCbCr color spaces. **(c)** smallHead detected as the smallest cluster (this constitutes the darkest part of the head). **(d)** Fitness of (a) with an ellipse (red) to indicate in which direction the head points. **(e)** smallHead detected in (c) grows according to growing mask up to (a). **(f)** Contour superposition of sperm head segmentation.

To determine the direction in which the head points, we propose a two-step method (Algorithm 4). First, we determine the orientation of the sperm head as the angle between the X axis and the major axis of the ROI, and using this angle, we rotate the head to a horizontal position in which the major axis of the fitted ellipse is parallel to the X axis (Algorithm 4, step 2). We then divide this major axis in three similar portions (Figure 5.2d) and calculate a fitness value [Yao et al., 2005] of the two extreme portions with the fitted ellipse (Algorithm 4, steps 10/11). The portion corresponding to the lowest fitness

value indicates the direction to where the head points. The pointing direction allows us to build a growing mask for segmenting the whole head, and not only the darkest part of the head (Algorithm 3, step 8). The next step consists in setting a growing mask according to head pointing direction and angle $\in [0, 2\pi)$ and apply it to the part of the head previously segmented (Algorithm 3, step 9). As a final refinement, we eliminate the small and big objects according to *minTs2* and *maxTs2* threshold values (Figure 5.2e).

Algorithm 3 Segmentation of one sperm head.

```
    white: binary image with a candidate sperm detected
    minTs1: minimum number of pixels of a sperm head before k-means
    maxTs1: maximum number of pixels of a sperm head before k-means
    maxTc: maximum number of pixels of a candidate head after kmeans
    minTs2: minimum number of pixels of a sperm head after growing
    maxTs2: maximum number of pixels of a sperm head after growing
 1: white ← opening(white)
 2: white ← eraseBySize(white,minTs1,maxTs1)
 3: data ← [imLab(2) imLab(3) imYCbCr(3)]
 4: [cluster1,cluster2] ← kmeans(data,2)
 5: smallHead ← chooseMinorCluster(cluster1,cluster2)
 6: smallHead ← eraseBySize(smallHead,0,maxTc)
 7: [angle,direction] ← getPointingDirection(white,smallHead)
 8: mask ← generateGrowingMask(white,angle,direction)
 9: head ← makeHeadBigger(white,mask,smallHead)
10: finalImage ← eraseBySize(head,minTs2,maxTs2)
11: return finalImage
```

Algorithm 4 Sperm head direction.

```
    white: binary image with a candidate sperm detected
 1: angle ← getOrientation(white)
 2: whiteRot ← rotateImage(white,angle)
 3: imgBorder ← getPerimeter(whiteRot)
 4: maximum ← number of columns of imgBorder
 5: x1 ← ceil(maximum/3)
 6: x2 ← 2*x1
 7: [xc,yc] ← getCentroid(whiteRot)
 8: majorAxis ← getMajorAxis(whiteRot)
 9: minorAxis ← getMinorAxis(whiteRot)
10: value1 ← fitnessFunction([xc,yc,majorAxis,minorAxis,0],imgBorder,1,x1)
11: value2 ← fitnessFunction([xc,yc,majorAxis,minorAxis,0],imgBorder,x2,maximum)
12: if value1 < value2 then
13:     direction ← 0
14: else
15:     direction ← 1
16: end if
17: return angle and direction
```

The parameters returned by Algorithm 4 can be used to generate eight different growing masks, according to all the possible orientations that a sperm head could present. The growing mask is created using angle and direction, as in the following.

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$Angle = 0 \qquad Angle = \frac{\pi}{2}$$
$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$
$$Angle = \pi \qquad Angle = \frac{3\pi}{2}$$

5.3 Segmentation of Sperm Nucleus and Acrosome

The segmented sperm head is used for the posterior segmentation of nucleus and acrosome. To this end, we performed a statistical analysis to determine the intensity characteristics of both head components, working on the red channel of RGB color space. We use the R channel of the RGB color space because R offers a better differentiation between sperm nucleus and acrosome, compared to the G or B channels. The Otsu threshold calculated in this region allows us to separate pixels of the nucleus from pixels of the acrosome (Algorithm 5, step 2).

Algorithm 5 Segmentation of nucleus and acrosome

white: binary image with a segmented sperm head
imRGB: rgb image with a segmented sperm head
r: size of neighborhood for e*raseTails*
sumV: threshold value for sum inside neighborhood in e*raseTails*

1: imRed ← imRGB(:,:,1)
2: threshold ← OtsuThreshold(imRed)
3: acrosome ← pixels in imRed whose value is > threshold
4: acrosome ← eraseTails(acrosome,r,sumV)
5: acrosome ← getBiggestROI(acrosome)
6: nucleus ← imageDifference(white,acrosome)
7: nucleus ← eraseTails(nucleus,r,sumV)
8: return acrosome and nucleus

The nucleus region is darker than the acrosome region because of a staining effect. Therefore, we pick the biggest ROI in the lighter region as the acrosome (Figure 5.3c). The segmented regions are obtained as a difference set operation between the pixels in the acrosome and pixels in the nucleus (Figure 5.3d). We use our proposed procedure *eraseTails* to smooth the resulting acrosome and nucleus.


a) b) c) d) e)

Figure 5.3: **Segmentation of nucleus and acrosome. (a)** Segmentation of sperm head in R channel (as returned by Algorithm 3). **(b)** Image after applying statistical-defined threshold (Algorithm 5, step 2). **(c)** Only the biggest ROI is kept and smoothed. **(d)** $ROI(a) - ROI(c)$, **(e)** Contour superposition of nucleus (green) and acrosome (red).

5.4 Detection and Segmentation Algorithm

In this section, we present an algorithm that encompass both of detection and segmentation steps (Algorithm 6), towards a posterior morphological sperm analysis.

Algorithm 6 Detection and segmentation of sperm heads, nucleus and acrosome.

```
    imRgb: original image containing several sperm cells
1:  imgDetection ← detectionSpermHeads(imRgb) {Algorithm 2}
2:  sperms ← labelRegions(imgDetection)
3:  for each sperm i do
4:      spermRgb ← imgDetection(sperm i)
5:      head(i) ← segmentationSpermHead(spermRgb) {Algorithm 3}
6:      [acrosome(i),nucleus(i)] ← segmentationNucleusAcrosome(head(i),spermRgb) {Algorithm 5}
7:  end for
8:  return head, acrosome, nucleus
```

5.5 Experimental Results

This section presents the results we achieved using the segmentation gold-standard (see Chapter 4), aiming to compare our method to previously published methods [Carrillo et al., 2007b, Carrillo et al., 2007a].

We implemented Carrillo's method [Carrillo et al., 2007b, Carrillo et al., 2007a], since it is not available as a source code by the authors to compare our detection and segmentation precision. A complete review of Carrillo's method can be found in Appendix A

We have conducted experiments for parameter estimation, and for comparison of our approach and Carrillo's method using Matlab[1]. The proposed method obtains a significant improvement respect to Carrillo's method as shown in following pages.

5.5.1 Parameter Optimization

In order to choose the best set of parameters, we have performed different experiments varying the values of seven parameters that are described as follows. The first two are referred to the procedure *eraseTails*, related to the neighborhood size (r) and the threshold value for the sum in the underlying neighborhood ($sumV$). This procedure is called from the Algorithms 2 and 4. The remaining five parameters are used in Algorithm 3 and are related to the allowed size of sperm heads at different stages of the segmenting process. The parameters *minTs1* and *maxTs1* are used for a refinement step before applying *kmeans*, while *maxTc* indicates the maximum size of sperm heads allowed to grow with a growing mask. The last two parameters, *minTs2* and *maxTs2* indicate size threshold values for the final segmented heads. In Table 5.1 the variation of parameters is shown. The final values were chosen according to the best tradeoff between true positive and false negative values, as well as the Dice coefficient.

[1]Matlab *R*2013*a* 8.1.0.604

Parameter name	Range	Best value
r	2 : 1 : 4	3
sumV	15 : 20 : 55	35
minTs1	70 : 10 : 90	80
maxTs1	800 : 100 : 1000	900
maxTc	1200 : 100 : 1500	1400
minTs2	30 : 10 : 50	40
maxTs2	350 : 50 : 450	400

Table 5.1: Variation of parameters for our proposed method

5.5.2 Sperm Detection

The performance of our detection results was determined by ROC curve. ROC curve takes into account the area under the curve (AUC) as a quality measure. The higher the AUC, the better the quality of a method. The ROC curve determines the cost in terms of false positives when a high correct detection is desired (see Figure 5.4). To create each point of the ROC curve, we calculated the percentage of correct detection (according our gold-standard) and the number of false positives for a given instance of the parameter values.

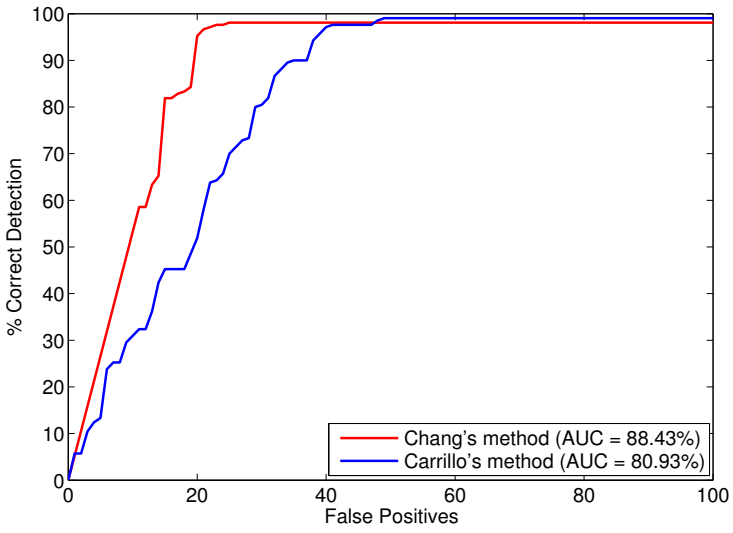


Figure 5.4: **ROC curves for sperm head detection.** Detection rate and number of false positives according to the results of our proposed method (red) and Carrillo's method (blue), versus hand-segmented masks.

Our approach achieves an AUC value of 0.88 while Carrillo's method accomplishes 0.81. In addition, our proposal achieves a correct detection rate over 97% at the expense of having only 23 false positives. To achieve a comparable result with a correct detection rate over

97%, Carrillo's method achieves 41 false positives. Carrillo's method reports a correct rate over 95% [Carrillo et al., 2007b, Bijar and Mikaeili, 2011]. According to our experiments, this rate is achieved with 39 false positives. In Table 5.2 the relationship between false positives and correct detection is shown for both evaluated methods.

	Our proposed method	Carrillo's method
Precision	$97,6\%$	$95,7\%$
True Positives (TP)	205	201
False Positives (FP)	23	39

Table 5.2: Detection accuracy

5.5.3 Sperm Head Segmentation

We performed different experiments to assess segmentation of sperm head, acrosome and nucleus. In each case, we calculate two quality measures for our results: the Dice Coefficient to assess the accuracy of our results with hand-segmented masks and the Hausdorff Distance (considering d as Euclidian distance) to assess the disagreement of segmentation against hand-segmented mask.

As mentioned before, we compare our segmentation results with the results obtained by implementation of Carrillo's method. We also calculate Dice Coefficient and Hausdorff Distance, using the same testing images to compare the results.

Figure 5.5 shows an image gallery with some segmentation results, considering head, acrosome and nucleus segmentation. For each segmentation (head, acrosome and nucleus), we present our best, average and worst result, in terms of Dice coefficient. For each image, we show the result of applying Carrillo's method to the same sperm head.

Figure 5.6 shows Dice coefficients for both segmentation results (our proposed method and Carrillo's method). The Dice coefficient assesses quality of sperm head segmentation as well as acrosome and nucleus segmentation, by means of measuring the overlap with ground-truth. We applied our proposed method to the testing images and calculated Dice's coefficients for each segmented sperm head. The same procedure was followed in the case of acrosome and nucleus. We then applied Carrillo's method to the same data set and calculated the Dice coefficient for each segmented sperm head, acrosome and nucleus. For every component (head, acrosome and nucleus), the Dice coefficients of our proposed method are significantly better than those achieved by Carrillo's method. Our average results have more than 80% of overlapping against hand-segmented masks, with average Dice coefficients of 0.88, 0.83 and 0.82 for head, acrosome and nucleus segmentation, respectively.

Figure 5.7 shows a graphical representation of the probability density function (PDF) for Dice coefficients, corresponding to acrosome, nucleus and sperm head segmentation, together. This is a comparison between the PDF corresponding to our results and the one corresponding to the results achieved by Carrillo's method. As we can observe, the distribution of values

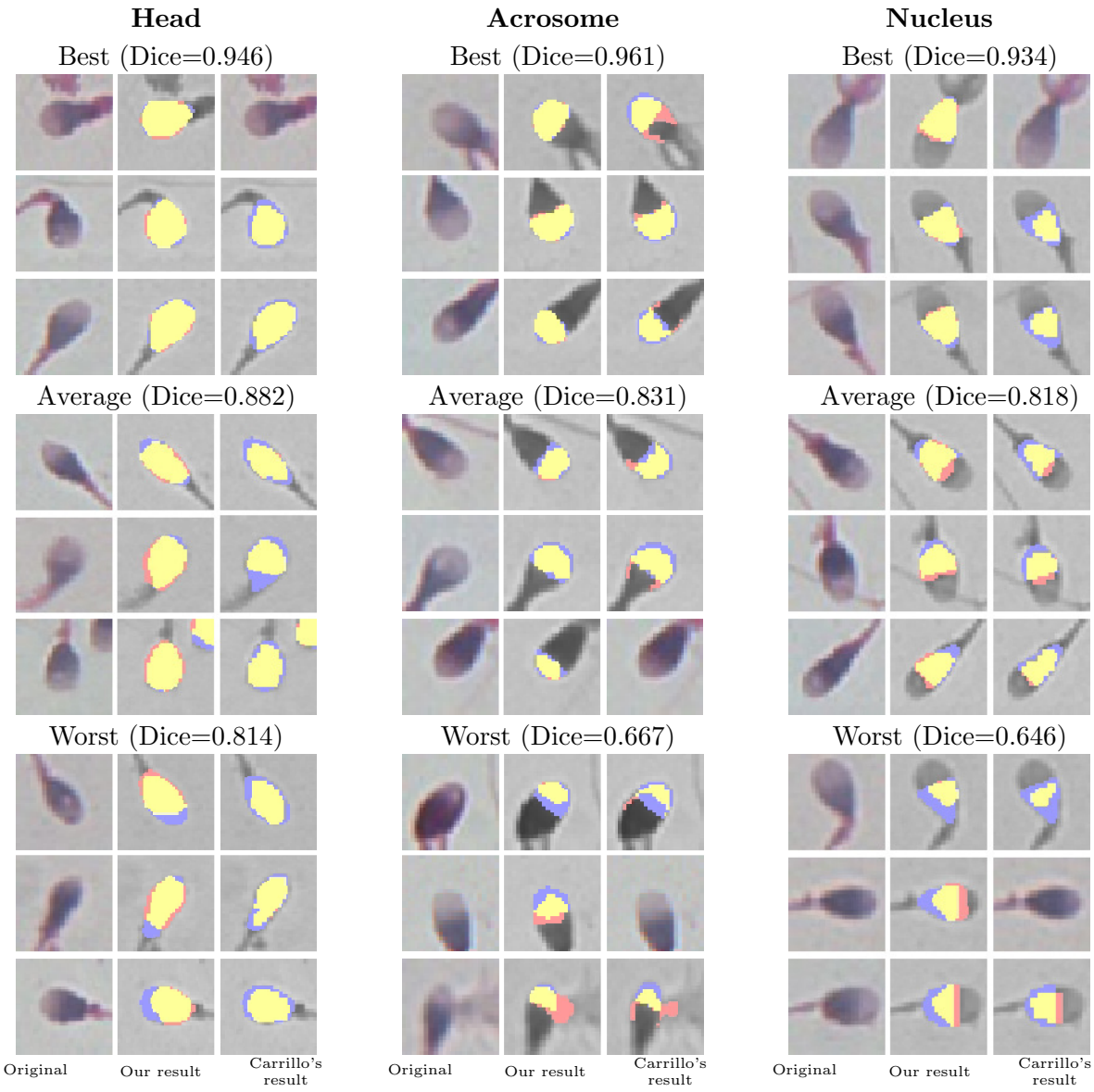


Figure 5.5: **Results of head, acrosome and nucleus segmentation.** The upper row shows representatives for best results, the middle row for average results, and the last row for worst results. For each part (head, acrosome and nucleus), we present the original (first column), our result (second column) and Carrillo's method result (third column). The blue color represents the gold-standard, red presents our proposed/Carrillo's method and yellow the overlap between gold-standard and our proposed/Carrillo's method.

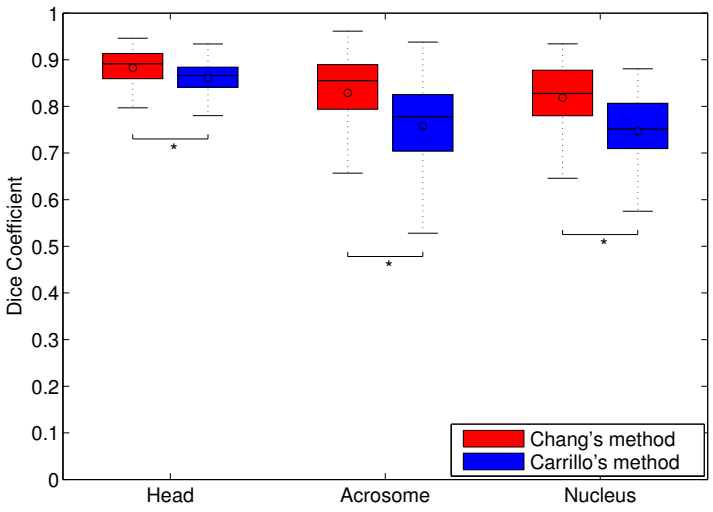


Figure 5.6: **Dice coefficient for head, acrosome and nucleus.** On each box, the edges are the $25th$ and $75th$ percentiles and the whiskers extend to the most extreme data points that are not outliers. For each box, we show the median value (horizontal line) and the sample mean ($\circ$). Statistically significant differences between our proposal (red) and Carrillo's method (blue) using Wilcoxon rank sum test are indicated ($*p < 0.05$).

for the Dice coefficient achieved by our method (with $\sigma = 0.06$) is shifted to higher Dice coefficients and provides smaller variance than that of Carrillo's method (with $\sigma = 0.08$).

Our average results have less than 25% of disagreement with segmentation gold-standard, with average Hausdorff distance values of 0.15, 0.20 and 0.24 for head, acrosome and nucleus segmentation, respectively. Additionally, we present a comparison of Hausdorff distance corresponding to our proposed framework and results of Carrillo's method, in Figure 5.8. As in the previous case, our aim is to assess quality of segmentation of sperm heads and sperm head parts (acrosome and nucleus), but now by means of measuring the disagreement with segmentation gold-standard. We followed the same procedure as in the Dice coefficient case. For every component (head, acrosome and nucleus), the Hausdorff distance values of our proposed framework are better than those achieved by Carrillo's method, because ours show a smaller distance (on average) between perimeters of segmentation gold-standard and segmentation results. Our average results have less than 25% of disagreement with segmentation gold-standard, with average Hausdorff distance values of 0.15, 0.20 and 0.24 for head, acrosome and nucleus segmentation, respectively.

5.6 Discussion

In this chapter we have presented a framework for sperm cell segmentation achieving significant improvement with respect to Carrillo's method. Our approach is different from the sperm segmentation approaches so far known (Chapter 3) in three different aspects:

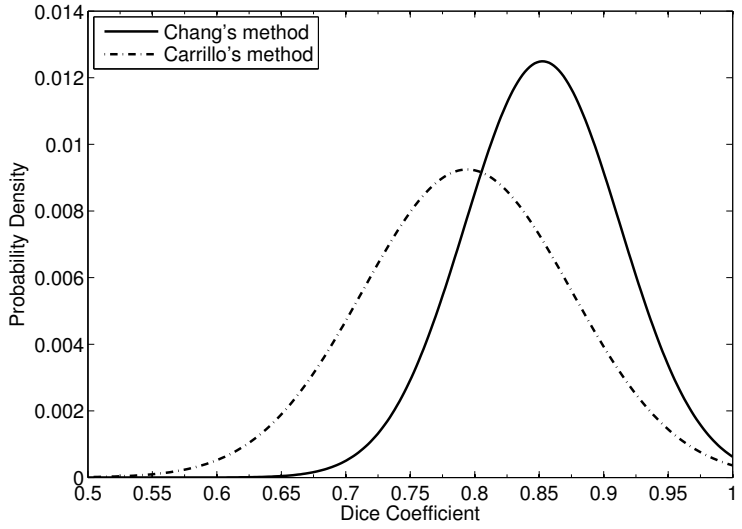


Figure 5.7: **Probability density function for Dice coefficient.** Acrosome, nucleus and sperm head segmentation results are showed together, considering the Dice coefficients achieved by our proposed method (continuous line, $\mu = 0.85 \pm 0.0026$, $s^2 = 0.0036$) and by Carrillo's method (dotted line, $\mu = 0.79 \pm 0.0035$, $s^2 = 0.0065$).

1. **Use of color space combinations.** Choices of color space have significant influences on the result of image segmentation. Cheng et al. [Cheng et al., 2001] compared several color spaces including RGB and L*a*b* for color image segmentation purposes, and they stated that the selection of a color space for image processing is image/application dependent. All the research works cited in Chapter 3 use RGB color space to segment sperm cells, including Carrillo's method. However, RGB is not suitable for color segmentation and analysis because of the high correlation among the R, G, B components [Pietikainen et al., 1996, Littmann and Ritter, 1997]. Besides, it is impossible to evaluate the similarity of two colors from their distance in RGB space because RGB space does not represent color differences in a uniform scale.
   We choose to work with a hybrid color space combining RGB, YCbCr and L*a*b* color spaces. Therefore, for the detection stage (Section 5.1) we used six redundant color features in RGB and L*a*b* color spaces. L*a*b* color space represents perceptual uniformity, and it is especially efficient in the measurement of small color difference, that can be calculated as the Euclidian distance between two color points [Cheng et al., 2001]. For us, it was important that L*a*b* space controls color and intensity information more independently and simply than RGB. Also, it has been shown that L*a*b* color space gives better results than others in color segmentation [Ohta et al., 1980]. We decided to keep RGB along with L*a*b* color space after intensive experimental evaluation, and regarding related work techniques. In addition, for segmentation stage (Section 5.2) we choose to combine L*a*b* and YCbCr color spaces. A chromatic component of YCbCr was introduced for two reasons: (a) the color difference of human perception can be directly expressed by a Euclidean distance in the color space, and (b) the intensity and chromatic components can be easily and independently controlled. In general, YCbCr color space has been extensively used for skin color segmentation [Hsu

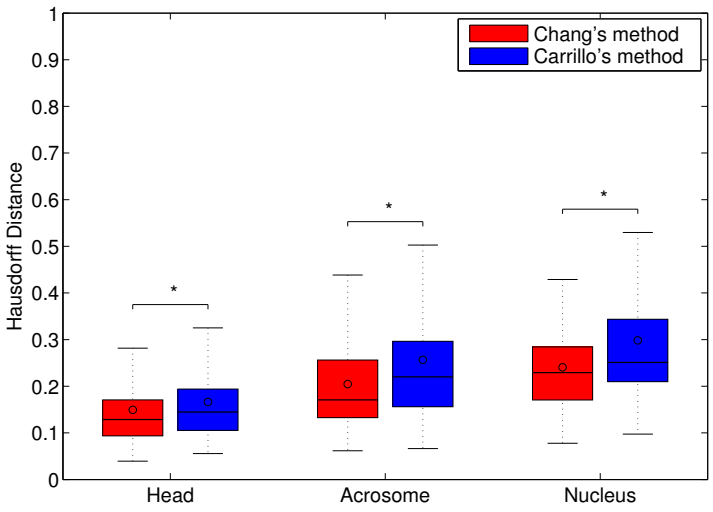


Figure 5.8: **Hausdorff distance for head, acrosome and nucleus.** On each box, the edges are the $25th$ and $75th$ percentiles and the whiskers extend to the most extreme data points that are not outliers. For each box, we show the median value (horizontal line) and the sample mean ($\circ$). Statistically significant differences between our proposal (red) and Carrillo's method (blue) using Wilcoxon rank sum test are indicated ($*p < 0.05$).

et al., 2002, Garcia and Tziritas, 1999].

2. **Use of clustering method.** All the existing color image segmentation approaches are ad hoc, because they are strongly application-dependent. Among several segmentation methods, clustering has been widely used for color image segmentation [Celenk, 1990, Hall et al., 1992, Jurio et al., 2009, Isa et al., 2009]. This is due to the fact that for color images, a color space is a natural feature space, and colors tend to form clusters in the color space. From the viewpoint of color clustering, the desired image is represented by color features which constitute a space possessing uniform characteristics such as the L*a*b* color space [Celenk, 1990]. We choose to apply *k-means* clustering in both stages of our proposal: detection and segmentation of sperm heads. As a traditional clustering algorithm, *k-means* is popular for its simplicity of implementation. it can also be adopted to solve the illumination variation problem. This, combined with YCbCr and L*a*b* color spaces, provide us an exceptional tool for illumination invariant segmentation approach, outperforming the state-of-the-art technique.
3. **Identification of sperm head direction.** One of the key contribution of our work is the proposal of a novel algorithm to determine which direction the sperm head points. This has not been considered before in any of the research works reviewed in Chapter 3. Proper identification of the head front direction could serve to accurately segment the sperm head and to discard mid-piece regions or residual cytoplasm areas that may have been included in the result of the detection. Related works [Park et al., 1997, Yi et al., 1998,Nafisi et al., 2005,Sánchez and Petkov, 2009], observe that authors proposed ellipse fitting for head morphological analysis and experimental evaluation, but they have not used it to refine the segmentation itself. Our proposal generates eight types of different growing masks, regarding all possible positions in which a sperm head may appear.

To our knowledge, none of the proposed approaches (including Carrillo's method) have taken into account the head front direction, however, it is a very important issue that could help in many other stages in the quest for an accurate morphological analysis.

Our results have shown that our approach, based on those aspects described above, outperforms the results achieved by Carrillo's method. In fact, we showed that our method achieves a higher Dice coefficient, lower Hausdorff distance and less dispersion with respect to the results achieved by Carrillo's method. This is clearly shown in the gallery presented in Figure 5.5. We believe that this outperforming occurs mainly because of the results of the first stage of our proposed method (Algorithm 2), combining RGB and L*a*b* color spaces. This stage result is extremely accurate at segmenting sperm heads. However, as part of the heads, we also segmented mid-piece areas, residual cytoplasm areas, and even tails. This pixel separation is probably due to the higher Euclidean distance between color pixels corresponding to the cells and background, incorporating L*a*b* color space. It is important to note that non-coiled tails are removed (though not all of their extension) by the procedure *eraseTails* (Figure 5.1). In addition, when including a YCbCr chromatic component for separating the sperm nucleus, our proposed approach removes mid-piece and residual cytoplasm areas, as well as rests of coils and coiled tails (Figure 5.5, first row of nucleus segmentation). This is a substantial difference against Carrillo's that discards such cells because they overcome size threshold. It should be noted that individualizing each sperm head for a more accurate segmentation makes it possible to work with spatially close cells and we obtain significantly better results than Carrillo's method (Figure 5.5, first row of head segmentation). We believe that this contributes by having an average Dice coefficient greater than the state-of-the-art method (Figure 5.7), observing that most of our Dice coefficients are greater than the average Dice coefficient obtained by the Carrillo's method.

There is the expectation that no detection method can achieve a detection rate of 100%. Our method is able to correctly detect up to 206 sperm cells (98%), while the method proposed by Carrillo et al. is able to correctly detect 208 sperm cells (99%) (Figure 5.4). There are specific situations in test images that are affecting our detection rate (Figure 5.9). We observe few sperm cells with an excessive residual cytoplasm area, whose front head direction is erroneously detected (because the cytoplasm area fits better with an ideal ellipse than a frontal region of its head). Then, in size validation of the candidate head, the head is removed by having a larger area than $maxTs2$ threshold (Section 5.5.1). Also, we observed a particular sperm cell overlapping an incomplete tail of another sperm cell. Thus, as they both are connected, the *eraseTails* procedure would remove both (overlapped head and incomplete tail) as if it were only one tail.

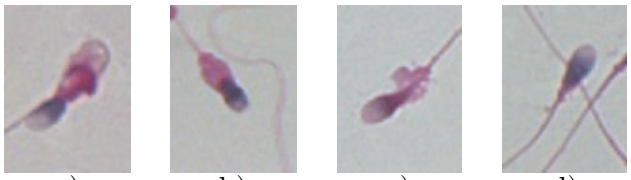


Figure 5.9: **Non-detected sperm cells. (a-c)** Sperm heads are not detected because of excessive residual cytoplasm area. **(d)** Sperm head is not detected because head overlaps incomplete tail of another sperm cell.

There is a tradeoff between correct detection rate and number of false positives. Our method provides 25 false positives with detection rate of 98%, while the method proposed by Carrillo et al. provides 49 false positives with detection rate of 99% (Figure 5.4). It is a very significant difference, in terms of false positives that we believe is due to three factors. First, we validate pixels of the object/background using the Euclidean distance to not include objects with a darker intensity than that of a sperm head, and combining RGB, L*a*b*, and YCbCr color spaces (Algorithms 2 and 3), rather than using only RGB space color, as Carrillo's method does. Second, our method removes objects according to their size and roundness, after applying *k-means* (Algorithm 3), while Carrillo's method only regards size validation. Therefore, Carrillo's method is more likely to report strangely shaped objects but with a size similar to a sperm head. Third, our method removes incomplete sperm cells which touch the border of the image (Algorithm 2). Carrillo's method does not take this into account, and reports heads of incomplete sperm cells as correctly detected heads. It is important to note that all the false positives reported by our approach are actually sperm heads, however they are not drawn in the gold-standard because they do not present a complete tail in the image.

5.7 Summary

In this chapter, we have presented a two-stage improved framework for detection and segmentation of human sperm head, acrosome and nucleus. The first stage was aimed to detect sperm heads, while the second stage was devoted to segment sperm head, acrosome and nucleus from detected sperm heads in previous stage. The main idea beyond our framework was a fusion of color space combinations, $k-$ means approach and mathematical morphology operations. In addition, our method proposed an ellipse fitting based algorithm to identify the head front direction. Experimental evaluation shows that our proposed framework outperforms the state-of-the-art, with a higher Dice coefficient, lower Hausdorff distance and less dispersion with respect to the results achieved by Carrillos method [Carrillo et al., 2007b, Carrillo et al., 2007a].

Chapter 6

Sperm Head Characterization

After properly segmenting the sperm head, the next task is to characterize each head, estimating features to differentiate samples of one class from samples of another class. In this work, special emphasis was given to a continuous representation of the curve defining the outline of a head rather than a discrete representation based on pixels.

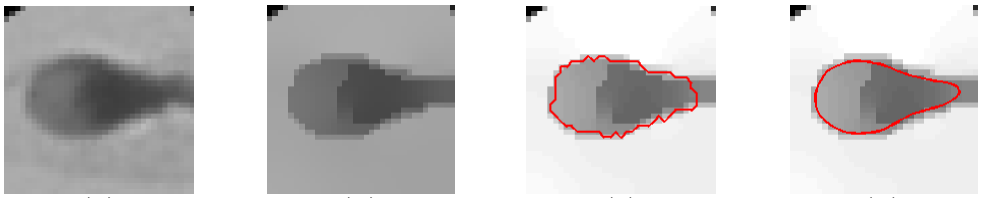
In order to have a reliable continuous representation, it is necessary to preprocess the image corresponding to each segmented head (regarded as a ROI in this thesis) according to the previous chapter (Section 6.1). For characterizing each one of the objects defined by ROIs, we used shape descriptors that can be found in the literature as well as one designed particularly with this application in mind (Section 6.2). An important goal is to get a feature vector as compact (low dimensionality) as possible, according to the objective of the next stage (classification of sperm heads). Therefore, it is imperative to select the best combination of feature spaces and Section 6.3 explains the strategies available to select feature spaces.

6.1 Shape Representation

The aim of this part of the work is to obtain a reliable representation of the closed curve constituting the contour of the head, given a 35×35 gray level image containing a sperm head (see Figure 6.1(a)). To do this, first we need to remove as much present noise as possible from the image. We propose using anisotropic diffusion (see Section 2.2.3) to preserve the border while the image is simplified, greatly reducing the noise in it (see Figure 6.1(b)). We apply 40 iterations, with $\lambda = 0.25$ and $\kappa = 7$.

From an image as presented in Figure 6.1(b)), now we need to generate a continuous representation of the curve that defines the sperm head. For this, we use active contours (see Section 2.2.4) regarding as initial curve the result of the segmentation approach presented in Chapter 5 (see Figure 6.1(c)). An example of how the outline for a given head would look after applying active contours, considering *elasticity* $\alpha = 0.1$, *rigidity* $\beta = 0.01$, *ballon force* $\gamma = 0.05$ and *external force weight* $\kappa = 0.42$ is shown in Figure 6.1(d). Finally, each sperm head is represented by a sequence B of n real coordinates $(p_{\mathrm{i}}, q_{\mathrm{i}})$ (and not in terms of

pixels), where $(p_1, q_1) = (p_n, q_n)$. The number of points that form each contour is variable and depends exclusively on the perimeter of the curve. In this sense, a contour with a perimeter around 35 will be represented by 41 points approximately, while another with a perimeter around 64 will need approximately 73 points for its representation.



(a) (b) (c) (d)

Figure 6.1: **Shape representation of sperm heads. (a)** Original gray level image (Image size: 35×35 pixels $\approx 7 \times 7$ μm). **(b)** Original image filtered using anisotropic diffusion. **(c)** Segmentation of sperm head (as returned from Algorithm 3) used as the initial curve for active contours. **(d)** Shape contour after applying active contours.

6.2 Feature Extraction

There are many shape-based descriptors in the literature, e.g. geometric moments [Hu, 1962], Fourier descriptors [Zahn and Roskies, 1972], and Zernike moments [Teague, 1980]. This section reviews different shape descriptors. These types of approaches have been proven to be effective in some applications, although a drawback is that their values are often not easily understable. In many applications it is preferable if the shape-based measures can be analysed by the domain experts, as this aids validation of the whole proposed scheme. This section presents a detailed review of common shape-based measures and descriptors. In addition, a morphological descriptor is proposed, combining and adapting some single shape-based measures.

6.2.1 Global Shape-based Measures

Global shape-based measures are a convenient way to describe ROIs. They are generally simple and efficient to extract, and provide an easy means for high level tasks such as classification. Although global shape-based measures have direct intuitive meaning, they can only discriminate shapes with large differences. Therefore, they are usually used as filters to eliminate false hits or combined with other shape-based measures and/or descriptors to discriminate shapes. In this section a selection of global shape-based measures for a ROI S are described. S is defined as the set of m points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ that are enclosed by curve B (see Section 6.1 for details).

Area

The simplest and most natural property of a ROI S is its area. The area of ROI S is given by m, regarding that a set of m points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ defines S.

Perimeter

The arc length of a spacially sampled curve (of contour) can be estimated suming up the Euclidian distance between each pair of adjacent points in the contour B of ROI S, i.e.:

$$\sum_{\mathrm{i}=0}^{n-1} \mathrm{d}\left(p_{\mathrm{i}} q_{\mathrm{i}}, p_{\mathrm{i}+1} q_{\mathrm{i}+1}\right)$$

Eccentricity

Eccentricity measures the aspect ratio of ROI S. Principal axes of S can be uniquely defined as the segments of lines that cross each other orthogonally in the centroid of S and represent the directions with zero cross-correlation [Peura and Iivarinen, 1997]. A common and convenient measure uses the central moments μ_{pq} (see Section 6.2.2 for a detailed explanation on geometric moments), and is computed as the ratio of the lengths of the axes of the image ellipse [Rosin, 2005]:

$$\frac{\sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{\mu_{20} + \mu_{02}} = \frac{majorAxis(S)}{minorAxis(S)}$$

Regularity

The common approach to compute regularity of ROI S [Maree et al., 2010] is to compute

$$\frac{\pi * majorAxis(S) * minorAxis(S)}{4 * area(S)}$$

Ellipticity

Sometimes a shape should be compared against a template. Clearly, a circle must be one of the simplest and most general possible choices. However, a natural choice to increase the degrees of freedom is to allow elongation, i.e., to fit an ellipse and measure the mapping mean-squared error. In this sense, ellipticity measures how close the contour B is to an ellipse with the same centroid, major axis, minor axis and orientation. A fitness value given a contour against a perfect ellipse is calculated as the shortest Manhattan distance between them [Saavedra and Bustos, 2014].

Circularity

The common approach to compute circularity (also called compactness, roundness and shape factor) is to compute

$$\frac{perimeter(S)^2}{area(S)}$$

This quantity measures how much a given ROI S differs from a perfect circular disk, which is understood as the most round shape [Rosin, 2005]. Low values correspond to circular shapes (4π in the limit).

Rectangularity

The standard approach to measuring rectangularity is to use the ratio of the area of ROI S to the area of its minimum bounding rectangle (MBR) [Rosin, 2003], i.e.

$$\frac{area(S)}{area(MBR(S))}$$

Maxima and minima curvature

The fact that not all points along a contour are equally important (in the sense of conveying information about the shape) motivates the analysis of dominant points such as those points where the curvature is either a positive maximum, or a negative minimum. The position along the curve of such points and the curvature values of them can be used as shape-based measures.

Convexity

Convexity of ROI S can be measured with respect to its convex hull [Zunić and Rosin, 2004, Rosin and Mumford, 2006]. If we denote the convex hull of ROI S by $CH(S)$ then the standard convexity measure is defined as:

$$\frac{area(S)}{area(CH(S))}$$

Symmetry

There are many approaches for detecting and quantifying symmetry [Atallah, 1985, Leou and Tsai, 1987, Masuda et al., 1993]. A simple scheme for measuring $n-$fold rotation symmetric is to measure the amount of overlap between S and the rotated versions of itself [Rosin,

2005]. This can be averaged over $n-1$ rotations by increments of $\frac{2\pi}{n}$ radians. If the result of rotating S by α is $T(S,\alpha)$ then the symmetry measure is:

$$\frac{1}{n * area(S)} \sum_{\mathrm{i}=1}^{n-1} area\left(S \cap T\left(S, \frac{2\pi \mathrm{i}}{n}\right)\right)$$

6.2.2 Shape-based Descriptors

Fourier Boundary Descriptor

Fourier descriptor has been successfully applied to many shape representation applications because of its simple derivation, simple normalization and its robustness to noise [Zhang and Lu, 2002]. Let B be a closed curve (boundary of ROI S) represented as the sequence of coordinates $(p_\mathrm{i}, q_\mathrm{i})$, for $\mathrm{i} = 0 \ldots n-1$. Each point in the curve can be treated as a complex number $r_\mathrm{i} = p_\mathrm{i} + jq_\mathrm{i}$. The Discrete Fourier Transform (DFT) of r_i is

$$a_u = \frac{1}{n} \sum_{\mathrm{i}=0}^{n-1} r_\mathrm{i} \mathrm{e}^{\left(-\frac{j2\pi ui}{n}\right)}$$

where $u = 0 \ldots n-1$. The coefficients a_u are called Fourier Descriptors (FDs), which represent the discrete contour of ROI S in the spectral domain [Zahn and Roskies, 1972]. The Fourier descriptor method uses a series of circles with different sizes and frequencies to build up a 2D plot of the boundary. Each coefficient is the frequency representation of a circle in the 2D plane [Sonka et al., 2008]. This descriptor contributes with 15 values to the whole feature vector.

Geometric Moment Descriptor

Moment based techniques are popular and useful in shape-based description. They were first applied as a way to describe images by Hu [Hu, 1962]. Hu proved that each image (expressed as a function) $f(x,y)$ has a one-to-one correspondence to a unique set of moments and vice versa [Vorobyov, 2011]. For a given ROI S, its geometric moment $m_{p,q}(S)$ is defined as

$$m_{p,q}(S) = \iint_S x^p y^q \,\mathrm{d}x\,\mathrm{d}y$$

where p and q are non-negative integers and $r = p + q$ is called the order of the moment.

Trivially, $m_{00}(S)$ equals the area of S. The position of a given ROI S is described by the ROI centroid $(x_c(S), y_c(S))$ defined as

$$x_c(S) = \frac{m_{10}}{m_{00}}, \quad y_c(S) = \frac{m_{01}}{m_{00}}$$

Since moments $m_{p,q}(S)$ are not translation invariant, it is suitable to consider the central moments $\bar{m}_{p,q}(S)$ defined as

$$\bar{m}_{p,q}(S) = \iint_S (x - x_c(S))^p \,(y - y_c(S))^q \,\mathrm{d}x\,\mathrm{d}y$$

A normalized moment $\mu_{p,q}(S)$ is also scale invariant and it is defined as

$$\mu_{p,q}(S) = \frac{\bar{m}_{p,q}(S)}{m_{0,0}(S)^{1+\frac{p+q}{2}}}$$

For the application of geometrical moments in recognition schemes, Hu introduced a set of 7 moment invariants using geometric moments, which are invariant under translation, scaling and rotation. These invariants are listed below:

$$\begin{aligned}
I_1 &= \mu_{20} + \mu_{02} \\
I_2 &= (\mu_{20} - \mu_{02})^2 + 4(\mu_{11}^2) \\
I_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\
I_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \\
I_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})((\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2) \\
&\quad +(3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})(3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2) \\
I_6 &= (\mu_{20} - \mu_{02})((\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2) \\
&\quad +4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\
I_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})((\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2) \\
&\quad +(\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})(3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2)
\end{aligned}$$

Therefore, this descriptor contributes with 7 values to the whole feature vector.

Zernike Moment Descriptor

Zernike-moment descriptor (ZMD) is one of the best shape-based descriptors as many researchers report promising results of ZMD [Kim and Kim, 2000, Zhang and Lu, 2001]. ZMD is obtained by using all the pixel information within a ROI, because it does not assume boundary information. ZMD is esentially a transform-based descriptor which is derived from 2D transform of shape in polar space [Zhang and Lu, 2002]. The complex Zernike moments are derived from Zernike polinomials [Teague, 1980],

$$V_{ab}(x, y) = V_{ab}(r\cos\theta, r\sin\theta) = R_{ab}(r)\mathrm{e}^{(jb\theta)}$$

and

$$R_{ab}(r) = \sum_{s=0}^{(a-|b|)/2} (-1)^s \frac{(a-s)!}{s!\left(\frac{a+|b|}{2} - s\right)!\left(\frac{a-|b|}{2} - s\right)!} r^{a-2s}$$

where r is the radius from (x, y) to the shape centroid, θ is the angle between r and x-axis, a and b are integers, and subject to $a - |b| =$even, $|b| \leq a$. Zernike polynomials are a complete set of complex-valued function orthogonal over the unit disk, i.e., $x^2 + y^2 = 1$. The complex Zernike moments of order a with repetition b [Teague, 1980] are defined as

$$\begin{aligned}
A_{ab} &= \frac{a+1}{\pi} \sum_x \sum_y f(x, y) V_{ab}^*(x, y) \\
&= \frac{a+1}{\pi} \sum_r \sum_\theta f(r\cos\theta, r\sin\theta) R_{ab}(r) \times \mathrm{e}^{(jb\theta)}, r \leq 1
\end{aligned}$$

where $f(x, y)$ is a binary shape function and $V^*(x, y)$ is the complex conjugate of $V(x, y)$. The magnitudes of the acquired Zernike moments normalized by the mass of the ROI are used as shape-based descriptors [Zhang and Lu, 2002].

For our particular case, we decided to work with the first 36 Zernike moments up to order 10. Therefore, this descriptor contributes with 36 values to the whole feature vector.

Convexity Measure Descriptor

There are many proposals of convexity measures that have been applied in object classification and recognition [Rahtu et al., 2006, Zunić and Rosin, 2004, Rosin and Mumford, 2006]. Here we review a region-based convexity measure developed in [Rahtu et al., 2006]. For a given ROI S, the convexity measure is based on the simple idea of generating pairs of points from S and then checking if certain points on the corresponding line segments belong to S. In this sense, the convexity measure is defined as

$$C_\alpha(S) = P(\alpha X + (1-\alpha)Y \in S)$$

where X and Y are independent random variables drawn uniformly from S. The parameter α determines the location of the point on the line segment between two random points from S. By varying α, it is possible to obtain many different measures [Rahtu et al., 2006].

In this particular case, we vary $\alpha = 1/2^b$ where $b = 1 \dots 5$. In this sense, this descriptor contributes with 5 values to the whole feature vector.

Ellipticity Measure Descriptor

When one needs to assess how much a given ROI S differs from a perfect ellipse, the ROI ellipticity is a key feature that can be quantified. There are many studies in the literature that address this issue, using different techniques [Aktas and Žunić, 2011, Schleicher and Zagar, 2008, Aktas and Žunić, 2013, Aktas and Žunić, 2012]. In this section, we review a family of ellipticity measures developed in [Aktas and Žunić, 2012] that is invariant to translation, scaling and rotation, with values in the range $(0, 1]$ where 1 indicates that the ROI S is an ellipse.

Let S be a ROI defined by the points (x, y). The auxiliary ellipse $E(S)$ is defined for a given ROI S as follows

$$E(S) = \left\{ (x,y) \,|\, \frac{\pi}{\rho(S)}\, x^2 \;+\; (\pi\, \rho(S))\, y^2 \;\leq\; 1 \right\}$$

where $\rho(S)$ denotes the ratio between the major axis and the minor axis of S. The area of $E(S)$ is 1.

Let $\lambda > 0$ and let the function $\Phi_\lambda(x, y)$ be defined as

$$\Phi_\lambda(x,y) = \left(\frac{\pi}{\rho(S)}\, x^2 \;+\; (\pi\, \rho(S))\, y^2 \right)^\lambda$$

the ellipse $E(S)$ can then be expressed as

$$E(S) = \{(x,y) \,|\, \Phi_\lambda(x,y) \;\leq\; 1\}, \quad for \;\; all \;\; \lambda \;>\; 0$$

Let a given ROI S and let $\lambda > 0$. The ellipticity $\varepsilon_\lambda(S)$ of S is defined as

$$\varepsilon_\lambda(S) = \frac{1}{1+\lambda} \frac{Area(S)^{1+\lambda}}{\min_{\alpha\in[0,2\pi]} \iint_{S(\alpha)} \Phi_\lambda(x,y)\,\mathrm{d}x\,\mathrm{d}y}$$

where $\Phi_\lambda(x,y)$ is defined as before and $S(\alpha)$ denotes the ROI S rotated around the origin for an angle α.

For our case, we vary λ from 0.5 to 5, with step size of 0.5. In this sense, this descriptor contributes with 10 values to the whole feature vector.

6.2.3 Proposed Morphological Descriptor

It is a good idea to start designing a descriptor by considering how experts in the field would describe objects of study (ROIs in our case). Thus, rather than build a general theory of shape, a popular approach is to design shape-based descriptors sensitive to various aspects of shape. The importance of finding shape-based measures that are simple to compute, with intuitive meanings, has already been noted by Peura and Iivarinen [Peura and Iivarinen, 1997]. Common simple global shape-based measures are area, diameter, perimeter, and eccentricity, among others (see Section 6.2.1 for a detailed list). The majority of these measures not only have linear computational complexity in the number of (boundary or region) points, but also tend to be designed to be invariant to rotation, translation, and uniform scaling, and often have an intuitive meaning since they describe a single aspect of the object of study. If a ROI is described by a combination of shape-based measures, this should be sufficient to provide discrimination between different classes of shapes. Deciding on the most appropiate measures depends on their suitability for a particular application.

In this section, we introduce a morphological descriptor designed as a combination of simple global shape-based measures. Although no single shape-based measure in the combination is descriptive enough to distinguish sperm heads from different classes, they contain enough information when combining with other shape-based descriptors to discriminate sperm head classes. All these measures that are calculated from the boundary obtained after applying active contours, ignoring intensity information, are sumarized in Table 6.1.

It is important to point out specific comments for two measures. First, *quadrant fitness* measures how close each quadrant of the ROI defined by contour B is to the corresponding quadrant of an ellipse with the same centroid, major and minor axis length, and orientation (see Figure 6.2 for a graphical explanation). We are interested not only in keeping the absolute fitness in each quadrant, but also in the relationships between them. In this sense, let cf_1 be the fitness value in the first quadrant, cf_2 the fitness value in the second quadrant and so on. We include the following feature values in the descriptor: cf_1, cf_2, cf_3, cf_4, $\frac{cf_1}{cf_2}$, $\frac{cf_2}{cf_4}$, $\frac{cf_1}{cf_3}$ and $\frac{cf_3}{cf_4}$. For instance, for normal sperm heads, $\frac{cf_1}{cf_2}$, $\frac{cf_2}{cf_4}$, $\frac{cf_1}{cf_3}$ and $\frac{cf_3}{cf_4}$ should have values close to 1, while for pyriform sperm heads only $\frac{cf_1}{cf_3}$ and $\frac{cf_2}{cf_4}$ should have values close to 1.

Second, *bilateral symmetry* measures the normalized area of overlap between the ROI defined by contour B and a reflected version of itself, in both directions, horizontal and

Shape Measure	Discriminative class	Number of features
area	small	1
perimeter	small	1
eccentricity	tapered and pyriform	1
regularity	amorphous	1
circularity	pyriform	1
rectangularity	tapered	1
maximum & minimum curvature	pyriform	2
ellipticity	normal	1
quadrant fitness	normal, tapered, pyriform, amorphous	8
bilateral symmetry	pyriform	2

Table 6.1: Summary of shape-based measures included in the proposed morphological descriptor

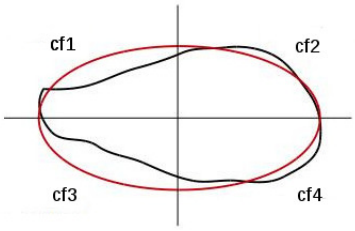


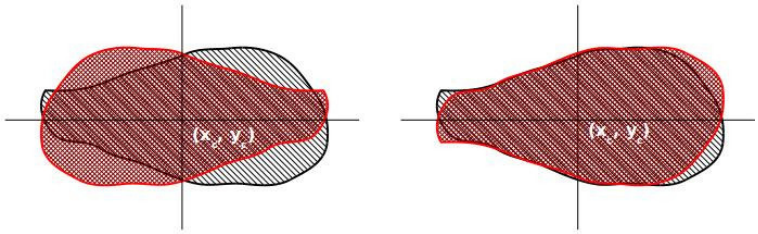
Figure 6.2: **Quadrant fitness**. We show an overlay of the contour shape (black) with an ellipse with the same centroid, axis length and orientation (red). In each quadrant i = 1 . . . 4, the fitness cf_{i} is calculated.

vertical with respect to its centroid (see Figure 6.3 for a graphical explanation). Let R_V be a vertical reflected version of ROI S and R_H be a horizontal reflected version of ROI S, then, the feature values that we propose to include in the descriptor are: $\frac{area(S \bigcap R_H)}{area(S)}$ and $\frac{area(S \bigcap R_V)}{area(S)}$. For instance, for pyriform sperm heads, the value of bilateral symmetry in horizontal and vertical directions should be quite different, while in the case of tapered or normal sperm heads, both directions should yield similar values.

The morphological descriptor contributes with 19 values to the whole feature vector.

6.3 Feature Selection

It is important to know how to constructively utilize all the information provided by all the features from the descriptors reviewed in Section 6.2. Concatenating all the features into a single feature vector does not guarantee an optimum performance and it exacerbates the *curse of dimensionality* problem [Damoulas and Girolami, 2009]. In Table 6.2, a summary



a) b)

Figure 6.3: **Bilateral symmetry**. We show an overlay between a ROI defined by contour B (black) and a reflected version of itself (red), in both directions, horizontal (a) and vertical (b) with respect to its centroid.

of all used descriptors is presented.

	Shape Descriptor	Type	Dimensionality
1	Morphological	Contour and region based	19
2	Fourier	Contour based	15
3	Geometric moments	Region based	7
4	Zernike moments	Region based	36
5	Convexity measures	Region based	5
6	Ellipticity measures	Region based	10

Table 6.2: Summary of used shape-based descriptors

Feature selection is an important stage in many applications where it is critical to find a small subset of features that maximizes the performance of the classification model. Feature selection could be done following one of two approaches: feature ranking and feature subset selection. Feature ranking calculates the score of each feature and then sorts them according to their scores. Feature subset selection selects a subset of features which collectively increases the performance of the model. Our work emphasis is on feature subset selection.

Feature selection needs to be combined with a classification model in order to get an estimate of the performance of the feature selection - classification combination. We propose to use an ensemble feature selection strategy defined as follows. As we have six feature families, we will combine them in every possible way and choose the combination with the best performance according to a particular evaluation criteria.

To perform this task, we can choose any base classifier (see Section 2.3.3). The k-NN classifier is a simple one, that not require parameter optimization: we only need to set the value of k [Okun and Valentini, 2008]. Furthermore, with this classifier is possible to use the leave-one-out (LOO) technique if we count with a limited dataset [Arlot and Celisse, 2010]. In this sense, for each feature family we will have a k-NN classifier using the LOO technique because of the limited data, creating a set of different k-NN classifiers, each providing their

output. Thus, for example, for the combination of feature families $< 2, 4, 5, 6 >$, we will have four k-NN classifiers; one for Fourier descriptor, one for Zernike moment descriptor, one for convexity measure descriptor and, finally, one for ellipticity measure descriptor.

Being k-NN LOO, each of the classifiers will return the k nearest neighbors to each element of the dataset. Aggregating the different classifiers results can be done by majority vote, unanimity voting, among others [Kuncheva, 2004]. In Chapter 7, we present the feature families selected for our classification scheme.

6.4 Summary

In this chapter, we have presented a proposal for representing a sperm head in terms of a curve, and describing it in terms of a feature vector. Special emphasis was given to a continuous representation of the curve defining the outline of a head rather than a discrete representation based on pixels. We have also reviewed common shape-based measures and descriptors and discussed the characterization of sperm heads using them. In this sense, we have introduced a new sperm head descriptor, named *morphological descriptor*. In addition, we have discussed about approaches for selecting the best combination of shape-based descriptors, according to their impact in discriminating each of the specific classes in study.

Chapter 7

Sperm Head Classification

This chapter is devoted to introduce and discuss the design and validation of a framework for morphological human sperm classification. Classification is the process of assigning objects into classes. There are many approaches used for classification purposes and are categorized as supervised and unsupervised methods.

Supervised classification techniques involve the participation of an expert who is responsible for teaching the classifier with examples. After training, the classifier is expected to classify similar objects, that are previously unseen, to the correct classes.

A classification paradigm uses a set of training examples of the form $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ for the projection of the function $f(x)$. The values x are usually vectors of real or discrete values of the form $< x_{\text{i1}}, x_{\text{i2}}, \ldots, x_{\text{i}n} >$. The values y are the expected outputs for given x values, and usually obtained from a discrete set of classes. Consequently, the task of a learning paradigm involves the approximation of a function $f(x)$ to produce a classifier.

It has been experimentally observed that creating a perfect monolithic classifier for a particular application is somewhat unfeasible for various reasons [Ranawana and Palade, 2006, Britto Jr. et al., 2014], for instance; the presence of noise within the dataset and the variability inherent to most pattern recognition problems. In particular, in this thesis we have trained several monolithic classifiers that demonstrate poor performace results (see Apendix B for details). Therefore, it is a good idea to use more classifiers and paradigms that provide close approximations to the function of the global classifier. Intuitively, it makes sense that a combination of classifiers provide better results than a single classifier. In order to design the proposed classification scheme, we tried with different cascade schemes, using different base learners and combination rules (see Appendix B). During this process, the results show that to separate amorphous sperm heads from the rest contributes to minimize the whole classification error. A detailed justification of design decisions of the classification scheme is presented in Appendix B.

In this chapter we propose a two-stage classification scheme. The basic idea beyond our scheme is a cascade ensemble of SVMs. We designed a two-stage approach following two main goals: a) discard many amorphous sperm heads as possible (it was shown in Chapter 4

that it is the most noisy class), and b) maximize the correct classification in the other four classes. We explain in detail both stages of our classification scheme in Section 7.1, while in Section 7.2 we present the results of applying each stage and the whole scheme to the classification gold-standard presented in Chapter 4.

7.1 Classification Scheme

As mentioned above, the key for automatic classification of sperm heads consists of separating amorphous sperm heads because this contributes to maximize the whole classification rate. This is due to the high level of noise (or heterogeneity) present in the amorphous class. With this in mind, we designed a classification scheme with two stages. The first stage acts as a filter for amorphous sperm heads, as well as a prior four-class classifier. In this sense, the second stage acts only as a confirmation step. In Figure 7.1 we show the high-level architecture of the proposed classification approach.

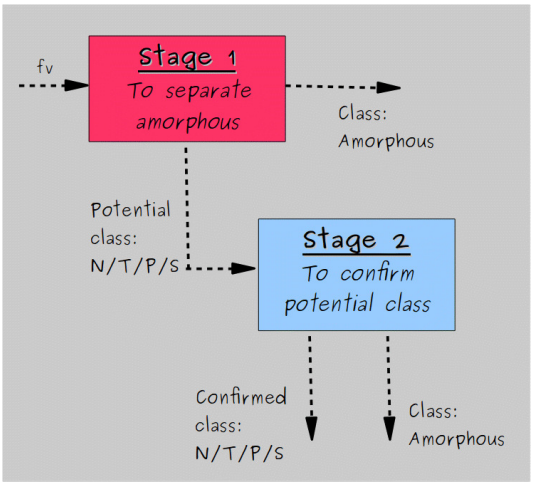


Figure 7.1: **High-level architecture of the proposed classification approach**. *fv* stands for input feature vector, *N* stands for normal, *T* stands for tapered, *P* stands for pyriform and *S* stands for small. Stage 1 is devoted to separate amorphous sperm heads and to assign a potential class label to *fv* (different from amorphous). Stage 2 is aimed to accept (confirm) or reject the assigned class label. In case of rejection, *fv* is regarded as an amorphous spem head.

7.1.1 Stage 1: Separation of amorphous sperm heads

The first stage of the classification scheme aims to separate amorphous sperm heads (discarding and regarding them as rejected) from sperm heads labeled as other classes (regarded them as accepted), as well as to identify the potential class of accepted sperm heads.

To do this, it is necessary to select which feature families are more suitable to reach the two goals introduced in the paragraph above. As we are interested in identifying the

potencial class of sperm heads that are not amorphous, we set as the objective function of the process of feature selection in this stage, the minimization of confussion rate between the four remaining classes discarding amorphous (normal, tapered, pyriform and small). With that said, what we need to do is to assemble a feature selection strategy combining the six descriptors presented in Section 6.2 and using $1-$NN LOO technique for each descriptor, and any of different combination rules (see Chapter 2) to evaluate our objective function. The best combination of descriptors will be the feature families that will be used in the first stage of the classification scheme.

Let $DSC_1, DSC_2, \ldots, DSC_M$ be the best combination of descriptors selected in the previous step. For the classification process, we propose to use an independent SVM for each one of these descriptors. Therefore, we will have one SVM that receives features from descriptor DSC_1, another SVM that receives features from descriptor DSC_2, and so on. Each one of these SVMs will be trained using only four out of five classes (normal, tapered, pyriform and small), but they will be tested using five classes (including amorphous). To combine the outputs of all SVMs, we could use different combination rules designed for this purpose, such as unanimity, plurality, and majority voting, as well as one that considers the probability of the output from each classifier. It is important to realize that, for the latter kind of combination rule, it would be necessary to tune a threshold value (see Chapter 2 for a detailed explanation).

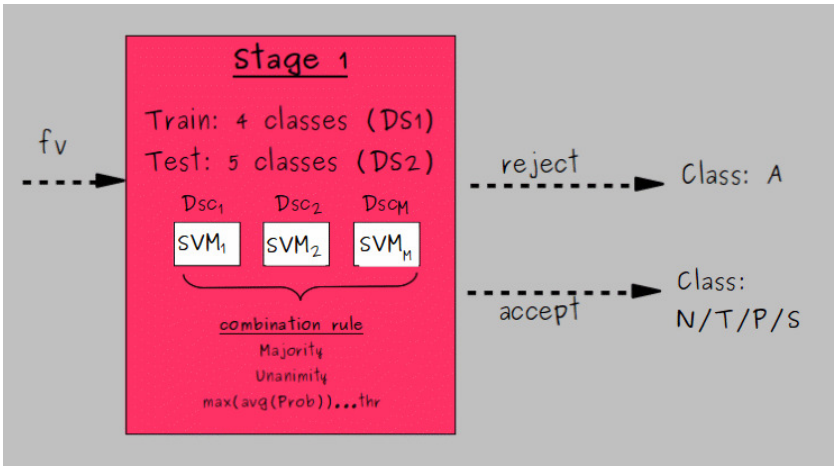


Figure 7.2: **First stage of the classification scheme**. *fv* stands for input feature vector, *DS1* stands for Dataset 1 *DS2* stands for Dataset 2, DSC_{i} $\mathrm{i} = 1 \ldots M$ stand for the best combination of descriptors, SVM_{i} $\mathrm{i} = 1 \ldots M$ stand for SVM trained with DSC_{i}, N stands for normal, T stands for tapered, P stands for pyriform, S stands for small and A stands for amorphous. Stage 1 is devoted to separate amorphous sperm heads (rejecting them) and to assign a potential class label to *fv* (different from amorphous).

See Figure 7.2 for a schematic explanation of the first stage of the classification approach.

7.1.2 Stage 2: Verification of potential classes

In the second stage of the classification scheme, the goal is to verify the potential class of the sperm heads, returned by the first stage. Thus, the model of classification is focused on combining the outputs of four verifiers. Each verifier is associated with to each of four classes (discarding amorphous). We will have four verifiers: V_1 verifies only normal sperm heads; V_2 verifies only tapered sperm heads; V_3 verifies only pyriform sperm heads; V_4 verifies only small sperm heads. If a certain sperm head with a given potential class in {normal, tapered, pyriform, small} does not verify its potential class, it will be rejected and could be considered as amorphous.

To do this, the descriptors that are most suitable for each class (normal, tapered, pyriform and small) need to be selected. We are interested in verifying the potencial class of sperm heads returned by the previous stage. Thus, the objective function of the process of feature selection in this stage is the maximization of mean precision rate between each of the four remaining classes discarting amorphous (normal, tapered, pyriform and small) versus amorphous (in all cases). What we need to do is to assemble a feature selection strategy combining the six descriptors presented in Section 6.2 and using $1-$NN LOO technique for each descriptor, and any of different combination rule (see Section 2) to evaluate our objective function in four scenaries (normal vs amorphous, tapered vs amorphous, pyriform vs amorphous, and small vs amorphous). The best combination of descriptors will be the feature families that will be used in each verifier of the second stage of the classification scheme.

Let $FS_{\mathrm{i}} = \{FS_{A\mathrm{i}}, FS_{B\mathrm{i}}, \ldots, FS_{M\mathrm{i}}\}$ be the best combination of feature families selected in the previous step for verifier V_{i}. For the verification process, we use an independent SVM for each one of these descriptors. Therefore, we will have one SVM that receives features from descriptor $FS_{A\mathrm{i}}$, another SVM that receives features from descriptor $FS_{B\mathrm{i}}$, and so on. Each one of these SVMs will be trained using only two out of five classes (normal, tapered, pyriform or small, and amorphous). They will be tested using the same two classes used for training. To combine the outputs of all SVMs, we could use different combination rules designed for this purpose, such as unanimity, majority voting, etc.

Having the four verifiers $V_{\mathrm{i}}, \mathrm{i} = 1 \ldots 4$, we do not need any combination rule in this stage, because only one verifier should be used, and its output should be considered as the output of the whole stage. See Figure 7.3 for a schematic explanation of the second stage of the classification scheme.

7.2 Experimental Results

7.2.1 Dataset

For the experimental results that we show in this section, we have used the classification gold-standard introduced in Chapter 4. We choose to work with the partial agreement dataset (further referenced only as *dataset*) with $1,132$ sperm heads with partial agreement among

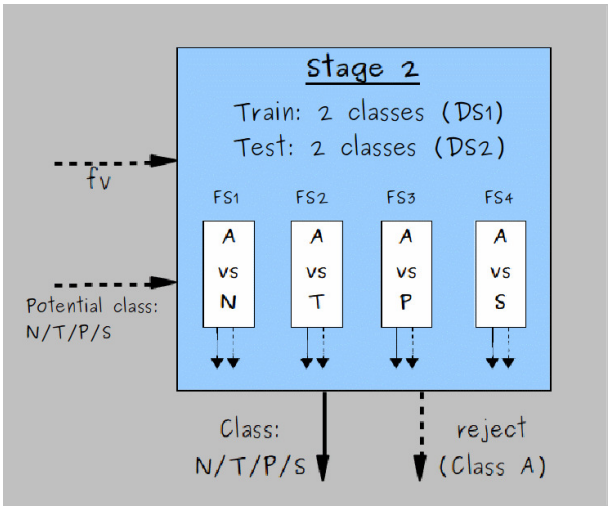


Figure 7.3: **Second stage of the classification scheme**. *fv* stands for input feature vector, *DS1* stands for Dataset 1, *DS2* stands for Dataset 2, FS_{i} i = 1 . . . 4 stand for the best combination of descriptors for verifier V_{i}, *N* stands for normal, *T* stands for tapered, *P* stands for pyriform, *S* stands for small and *A* stands for amorphous. Stage 2 is aimed to accept (confirm) or reject the assigned class label. In case of rejection, *fv* is regarded as an amorphous spem head.

experts and without overlapping, distributed in five classes.

The dataset has been partitioned in three subsets, named *Dataset 1 (DS1), Dataset 2 (DS2) and Dataset 3 (DS3)*, aiming to have a training (60% of the whole dataset), validating (20%) and testing (20%) dataset, respectively.

In addition, we have an extra testing dataset. Special Testing Dataset (DST) is a subset of the whole classification gold-standard, but with a particular feature: all the sperm heads that are contained in DST have been manually classified within the same class by all experts (total agreement between experts).

In Table 7.1, the size and distribution of classes in each partition and extra testing dataset are presented.

7.2.2 Stage 1: Separation of amorphous sperm heads

For each stage, the initial result consists of the selected features. As mentioned earlier, we had 6 different descriptors, and we needed to evaluate 63 different combinations.

To this end, we used 1-NN LOO for each descriptor looking to minimize the confusion rate among four classes (discarding amorphous). We evaluated two different combination rules: majority and unanimity. We used Dataset1 for selecting the best combination of descriptors. As we were using 1-NN LOO, we only need a single dataset. In Table 7.2, we show the best five combinations of descriptors with confusion rates using majority and unanimity as combination rule.

	Dataset	DS1	DS2	DS3	DST
Number of normal sperm heads	100	60	20	20	35
Number of tapered sperm heads	228	137	46	45	69
Number of pyriform sperm heads	76	44	15	16	7
Number of small sperm heads	72	45	14	14	11
Number of amorphous sperm heads	656	394	131	131	262
Total number of sperm heads	1132	680	226	226	384

Table 7.1: **Dataset partition**. $Dataset = DS1 \cup DS2 \cup DS3$. DST is a special dataset for testing purposes with total agreement between experts (subset of Dataset).

Combination of descriptors	Confusion rate using majority	Confusion rate using unanimity
MorphoD,FourierD,ZernikeD	**1.76**	3.13
MorphoD,GeomD,ZernikeD	1.84	3.07
MorphoD,FourierD,GeomD	1.85	3.07
MorphoD,FourierD,GeomD,ZernikeD,ConvD	1.88	3.69
MorphoD,ZernikeD,ConvD	1.88	3.24

Table 7.2: **Ranking of descriptor combinations for stage 1.** In column *Combination of descriptors*, *MorphoD* stands for morphological descriptor, *FourierD* stands for Fourier descriptor, *GeomD* stands for geometric moments descriptor, *ZernikeD* stands for Zernike moments descriptor, *ConvD* stands for convexity measures descriptor, and *EllipD* stands for ellipticity measures descriptor.

From Table 7.2, we made two decisions: 1) to use majority voting to combine the outputs of each 1-NN classifier as it achieved the lowest confusion rate, and 2) the best descriptor combination for this stage is the one that includes morphological descriptor, Fourier descriptor and Zernike moment descriptor.

Using these three selected descriptors, we apply the first stage of the classification process to our dataset. We used Dataset1 (DS1) for training and Dataset2 (DS2) for validating purposes. We did 10 runs and in each we balanced the training data at first. Let *minNoE* be the minimun number of elements per class. Let $c0$ be a random subset of *minNoE* normal sperm heads. Let $c1$ be a random subset of *minNoE* tapered sperm heads. Let $c2$ be a random subset of *minNoE* pyriform sperm heads. Let $c3$ be a random subset of *minNoE* small sperm heads. We create a balanced training set (*trBag*) in each run, containing sperm heads from four classes, as $\{c0 \cup c1 \cup c2 \cup c3\}$.

After creating *trBag*, for each selected descriptor (Morphological, Fourier and Zernike moments), we standardized the data and apply cross-validation to choose the best parameters for individual SVM. We trained SVM using *trBag* and tested using Dataset2. We used different combination rules for this stage (majority voting, unanimity voting, maximun probability using threshold 0.4, 0.5 and 0.6). As a final result, we calculated the mean value of 10 runs of True Positive Rate (tpr) of four classes: normal, tapered, pyriform and small. According to these datasets and applying only the first stage of our classification scheme, we get some interesting results (see Table 7.3). Due to the fact that the best compromise between the accuracy rate (mean of TPR of the four classes) and the rejection rate (percentage of amorphous sperm heads that are discarded) was reached using the maximum probability with a threshold of 0.4, we decided to use it as the combination rule of the outputs for our three different SVMs in this stage.

tpr(N)	tpr(T)	tpr(P)	tpr(S)	acc	rr	combination rule
0.59	0.69	0.75	0.83	0.71	0.11	majority voting
0.28	0.47	0.41	0.49	0.41	0.62	unanimity voting
0.50	0.63	0.69	0.75	**0.64**	**0.22**	max probability (thr: 0.4)
0.22	0.47	0.40	0.60	0.42	0.52	max probability (thr: 0.5)
0.04	0.33	0.21	0.40	0.25	0.73	max probability (thr: 0.6)

Table 7.3: **Results of the first stage of the classification scheme.** *tpr* stands for True Positive Rate, N stands for normal, T stands for tapered, P stands for pyriform, and S stands for small. *acc* stands for accuracy while *rr* stands for amorphous rejection rate. In bold face, we show the best compromise between accuracy and rejection rate.

7.2.3 Stage 2: Verification of potential classes

As in the previous case, the initial result consists of the selected features. However, in this stage, we need to perform four feature selection processes as we have four different verifiers (one for each class, different from amorphous). Thus, in each case, we had 6 different descriptors, and we needed to evaluate 63 different combinations of these descriptors.

To this end, we used 1-NN LOO for each descriptor looking to maximize the mean True Positive Rate of both classes (amorphous and one of {normal, tapered, pyriform, small}). We evaluated different combination rules, and decided to use majority voting to combine the outputs of each SVM. We used Dataset1 for selecting the best combination of descriptors in each case. As we were using 1-NN LOO, we only need a single dataset. We performed 10 runs for each experiment, and selected the most frequent descriptors that appear in the results of these 10 run results (disregarding descriptor combinations as a whole).

The best descriptor combination differs from one verifier to another, because the main goal of this section is to take advantage of the different features of classes versus amorphous

(See Table 7.4 for a summary).

Verifier	MorphoD	FourierD	GeomD	ZernikeD	ConvD	EllipD
Normal vs. Amorphous		✓		✓	✓	✓
Tapered vs. Amorphous	✓	✓	✓	✓		
Pyriform vs. Amorphous		✓	✓			✓
Small vs. Amorphous	✓			✓	✓	

Table 7.4: **Selected descriptors for four verifiers in stage 2**

Using these four selected descriptor combinations, we apply the second stage of the classification process to our dataset. We used Dataset1 (DS1) for training and Dataset2 (DS2) for validating purposes, considering only two classes in both situations. We need to train four verifiers, and the procedure to do this is the same in the four cases, changing only the composition of the training and testing dataset in order to consider only the two relevant classes in each case. For each verifier, we did 10 runs and in each of those runs, we balanced the training data at first. For a verifier V_n, let *minNoE* be the minimun number of elements per class (amorphous and class n). Let $c0$ be a random subset of *minNoE* sperm heads with label n. Let $c1$ be a random subset of *minNoE* amorphous sperm heads. We create a balanced training set (*trBag*) in each run, containing sperm heads from two classes, as $\{c0 \cup c1\}$.

After creating *trBag*, for each selected descriptor for verifier V_n, we standardized the data and apply cross-validation to choose the best parameters for individual SVM. We train SVM using *trBag* and test using Dataset2 considering only sperm heads with label n and amorphous. We use different combination rules for this stage (majority voting, unanimity voting, maximun probability using threshold 0.6, 0.7 and 0.8). As final result, we calculate the mean value of 10 runs of True Positive Rate (tpr) of two classes: amorphous and one of {normal, tapered, pyriform, small}.

Normal	Amorphous	Accuracy	Combination rule
0.81	0.60	0.70	majority voting
0.54	0.85	0.70	unanimity voting
0.25	0.56	0.40	max probability (thr: 0.6)
0.04	0.86	0.45	max probability (thr: 0.7)
0.00	0.97	0.48	max probability (thr: 0.8)

Table 7.5: **Results of verifier 1 : normal vs. amorphous.** True positive rate for normal and amorphous sperm heads classification is shown, as well as the accuracy reached using different combination rules.

According to these datasets and applying only the second stage of our classification scheme, we get performance results for each verifier (see Tables 7.5, 7.6, 7.7 and 7.8). Due to the fact that the best accuracy rate (mean of TPR of the two classes) is reached using majority

voting in almost all cases, we decided to use it as the combination rule of the outputs for the different SVMs of each verifier in this stage.

Tapered	Amorphous	Accuracy	Combination rule
0.63	0.81	0.72	majority voting
0.45	0.86	0.65	unanimity voting
0.59	0.17	0.38	max probability (thr: 0.6)
0.37	0.36	0.37	max probability (thr: 0.7)
0.23	0.59	0.41	max probability (thr: 0.8)

Table 7.6: **Results of verifier 2 : tapered vs. amorphous.** True positive rate for tapered and amorphous sperm heads classification is shown, as well as the accuracy reached using different combination rules.

Pyriform	Amorphous	Accuracy	Combination rule
0.83	0.64	0.73	majority voting
0.67	0.83	0.75	unanimity voting
0.41	0.51	0.46	max probability (thr: 0.6)
0.12	0.91	0.51	max probability (thr: 0.7)
0.01	0.99	0.50	max probability (thr: 0.8)

Table 7.7: **Results of verifier 3 : pyriform vs. amorphous.** True positive rate for pyriform and amorphous sperm heads classification is shown, as well as the accuracy reached using different combination rules.

Small	Amorphous	Accuracy	Combination rule
0.81	0.71	0.76	majority voting
0.44	0.89	0.66	unanimity voting
0.46	0.31	0.39	max probability (thr: 0.6)
0.15	0.65	0.40	max probability (thr: 0.7)
0.01	0.90	0.46	max probability (thr: 0.8)

Table 7.8: **Results of verifier 4 : small vs. amorphous.** True positive rate for small and amorphous sperm heads classification is shown, as well as the accuracy reached using different combination rules.

7.2.4 Complete Classification Scheme

Regarding the selected features for each stage, we apply the whole classification scheme in a cascade approach.

To this end, we trained using DS1, and as a result of training we obtain five combined classifiers. One called *svm1* as the combination of three SVMs of Stage 1, each one for one descriptor (Morphological, Fourier and Zernike). The one called *v1* consists of the combination of four SVMs, each for one descriptor (see first row of Table 7.4), looking to distinguish between normal from amorphous sperm heads. The one called *v2* consists of the combination of four SVMs too, each for one descriptor (see second row of Table 7.4), looking to distinguish between tapered from amorphous sperm heads. The one called *v3* consists of the combination of three SVMs, each for one descriptor (see third row of Table 7.4), looking to distinguish between pyriform from amorphous sperm heads. Finally, the one called *v4* consists of the combination of three SVMs too, each for one descriptor (see last row of Table 7.4), looking to distinguish between small from amorphous sperm heads.

For testing purposes, we evaluated sperm by sperm in the cascade approach (See Algorithm 7). Let s_{i} be a given sperm head. We tested *svm1* with s_{i} as input (Algorithm 7, step 2). If s_{i} is rejected, then the testing process is finished and we considered sperm s_{i} classified as an amorphous sperm head (Algorithm 7, step 4). Otherwise, if s_{i} is accepted, we continued with Stage 2 considering the output *label1* as following. If *label1* is class label *normal*, then we tested *v1* with s_{i} as input (Algorithm 7, step 8). Analogously, we tested *v2* if *label1* is class label *tapered* (Algorithm 7, step 12), *v3* if *label1* is *pyriform* (Algorithm 7, step 16), and *v4* if *label1* is *small* (Algorithm 7, step 20). In any case, we considered the output of the corresponding verifier (*v1, v2, v3, v4*) as the final output of the whole scheme (Algorithm 7, step 22). That is, suppose that *label1* is class label *tapered*, thus, we tested only verifier *v2*. If s_{i} is accepted, then we considered sperm s_{i} classified as a tapered sperm head, otherwise s_{i} is classified as an amorphous one.

We performed two different experiments to evaluate the performance of our approach. In both scenarios, we used the same configuration of classification scheme, the same training dataset (DS1), the same selected features and the same combination rules. We only varied the testing dataset.

In the first case, we used Dataset3 (DS3) for testing purposes. In the second case, we used a special testing dataset (DST). It is important to realize that in the first scenario, the testing dataset has only partial agreement between experts, while in the last case, the testing dataset has total agreement among experts.

According to these datasets and applying the whole classification scheme, we calculated performance results (see Tables 7.9 and 7.10). We did 30 runs for each testing dataset. We presented the mean of the True Positive Rate (tpr) for each class in the 30 runs.

In order to evaluate the experimental results, we defined two classification problems: the five-class and he two-class classification problems. As the primary study goal of this thesis, we are going to analyze the *five-class classification*. In 7.9, we can see that the correct

Algorithm 7 Testing procedure of the whole classification scheme.

MorphoD: morphological descriptor calculated for a sperm head{Section 6.2.3}
FourierD: Fourier descriptor calculated for a sperm head{Section 6.2.2}
GeomD: Geometric moments descriptor calculated for a sperm head{Section 6.2.2}
ZernikeD: Zernike moments descriptor calculated for a sperm head {Section 6.2.2}
ConvD: Convexity measure descriptor calculated for a sperm head {Section 6.2.2}
EllipD: Ellipticity measure descriptor calculated for a sperm head {Section 6.2.2}
svm1: combination of three trained SVMs of Stage 1 {Section 7.1.1}
v1: trained verifier for normal and amorphous {Section 7.1.2}
v2: trained verifier for tapered and amorphous {Section 7.1.2}
v3: trained verifier for pyriform and amorphous {Section 7.1.2}
v4: trained verifier for small and amorphous {Section 7.1.2}

```
fv1 ← [MorphoD FourierD ZernikeD]
label1 ← testingSVMensembled(svm1,fv1)
if label1 = amorphous then
    label ← amorphous
else
    if label1 = normal then
        fv2 ← [FourierD ZernikeD ConvD EllipD]
        label2 ← testingSVMensembled(v1,fv2)
    end if
    if label1 = tapered then
        fv2 ← [MorphoD FourierD GeomD ZernikeD]
        label2 ← testingSVMensembled(v2,fv2)
    end if
    if label1 = pyriform then
        fv2 ← [FourierD GeomD EllipD]
        label2 ← testingSVMensembled(v3,fv2)
    end if
    if label1 = small then
        fv2 ← [MorphoD ZernikeD ConvD]
        label2 ← testingSVMensembled(v4,fv2)
    end if
    label=label2
end if
return label
```

Testing dataset	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
DS3	0.62	0.64	0.50	0.82	0.30	0.58
DST	0.76	0.72	0.95	0.93	0.32	0.74

Table 7.9: **Results of the whole classification scheme for five-class classification.** *tpr* stands for True Positive Rate, N means normal, T means tapered, P means pyriform, S means small, and A means amorphous. *acc* stands for accuracy understood as mean of tpr of five classes.

classification for normal sperm heads reaches 62% if we use DS3 as the testing dataset. This increases to 76% if we use DST as the testing dataset. A similar situation occurs with tapered and small sperm heads. In the case of tapered sperm heads, the correct classification ranges from 64% to 72%, using DS3 and DST, respectively. For small sperm heads, our proposed

method achieves the best compromise between DS3 and DST testing datatsets, reaching 82% and 93%, respectively. The situation is kind of different with pyriform sperm heads, because achieving around 50% of correct classification using DS3 becomes 95% when using DST. The last case is about amorphous sperm heads, where our proposal achieves around 30%, with no regard as to which testing dataset is used, showing the complexity of correctly identifying this type of sperm head. In Figure 7.4 we show a graphical representation of this comparison.

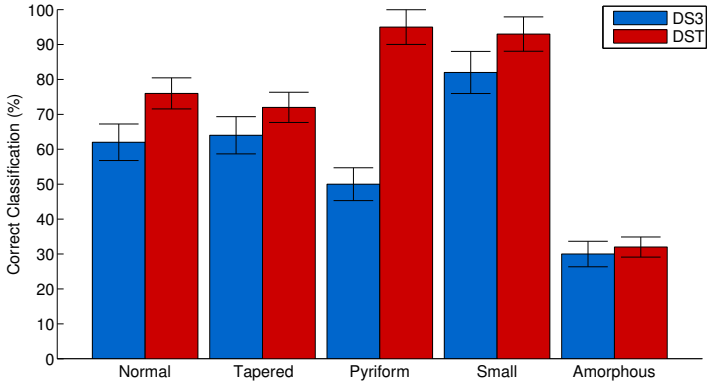


Figure 7.4: **Impact of testing datasets**. We compare the results obtained using two different datasets for testing: DS3 (Dataset 3 with partial agreement among experts) and DST (Special Testing Dataset with total agreement among experts). The percentage of correct classification ± SE for each class is shown.

If we consider only the *two-class classification*, regarding only normal and abnormal (including tapered, pyriform, small and amorphous sperm heads), the results are summarized in Table 7.10. In this case, we can see that the correct classification as normal sperm heads reaches 62% while for abnormal sperm heads reaches 57% when using DS3 as the testing dataset. The results are much better if we use DST as the testing dataset, getting 76% for normal sperm heads and 72% for abnormal ones.

Testing dataset	tpr(N)	tpr(Ab)	acc
DS3	0.62	0.57	0.60
DST	0.76	0.72	0.74

Table 7.10: **Results of the whole classification scheme for two-class classification.** *tpr* stands for True Positive Rate, *N* means normal, and *Ab* means abnormal. *acc* stands for accuracy understood as mean of tpr of two classes.

Our classification scheme outperforms a number of monolithic classifiers as $1-$NN, SVM, naive Bayes and decision trees (Tables B.1 and B.2). In Figures 7.5 and 7.6 we show a graphical comparison of achieved classification accuracy rates per class, using DS3 and DST as testing datasets, respectively. The feature vector representing each sperm head in training

and testing datasets consisted in the concatenation of all descriptors (Morphological, Fourier, Geometric moments, Zernike moments, Convexity and Ellipticity measures) for all monolithic classifiers.

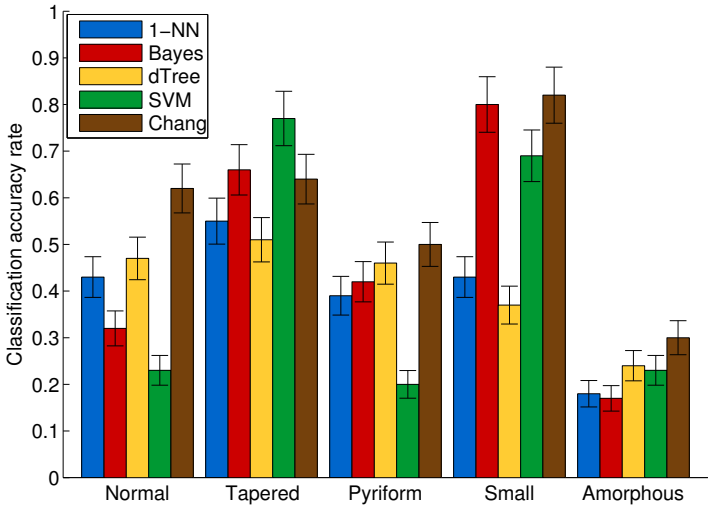


Figure 7.5: **Comparison of accuracy per class using DS3 achieved with the proposed classification scheme and different monolithic classifiers.** We compare four classic monolithic classifiers (1−NN, naive Bayes, decision trees and SVM) against the classification scheme proposed in this chapter (Chang). The classification accuracy rate ± SE for each class is shown.

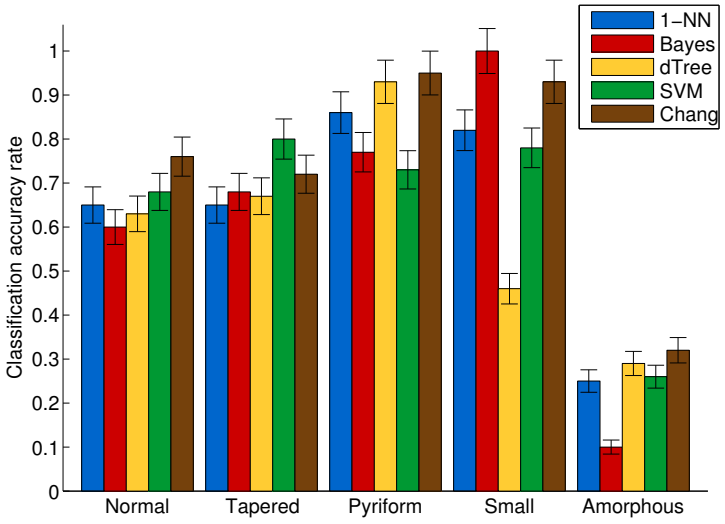


Figure 7.6: **Comparison of accuracy per class using DST achieved with the proposed classification scheme and different monolithic classifiers.** We compare four classic monolithic classifiers (1−NN, naive Bayes, decision trees and SVM) against the classification scheme proposed in this chapter (Chang). The classification accuracy rate ± SE for each class is shown.

In case of using DS3 as the testing dataset, our scheme achieves a mean accuracy rate of 0.58, while the best monolithic classifier (naive Bayes) achieves 0.47. For DST, the scenario remains similar: our scheme achieves a mean accuracy rate of 0.74, while the best monolithic classifiers for this case (SVM and $1-$NN) achieve 0.65. In both scenarios, the class amorphous is the most difficult class to identify. Using DS3, decision tree classifier achieves 0.24 versus 0.30 achieved by our scheme. While using DST, decision tree classifier achieves 0.29 versus 0.32 achieved by our scheme. Our experimental evaluation also shows that our proposed scheme outperforms different cascade classification schemes designed in the context of this research (See Appendix B for details). These comparisons are show in Figures B.1 and B.2.

Analyzing the case of five-class classification, Figure 7.7 shows the close results of our automatic approach compared against those classification by experts but not in all classes. The near results appear in pyriform and tapered classes, in which our proposal could be confused with a human expert. The confusing classes are normal, small and amorphous. These three classes have slight inter-class variations. For instance, the main difference between a small and a normal sperm head is often only the size, but the difference is poorly related to shape. Comparing normal against amorphous sperm heads, the slight shape variations are very notorious in a visual analysis. Furthermore, the intra-class variation in amorphous class is actually high, with many possibilities of confusion among experts (see Figure 4.5). In the case of two-class classification, the observable situation is expected, as the variability of experts is low. Our proposal results also show variability among experts (see Figue 7.8).

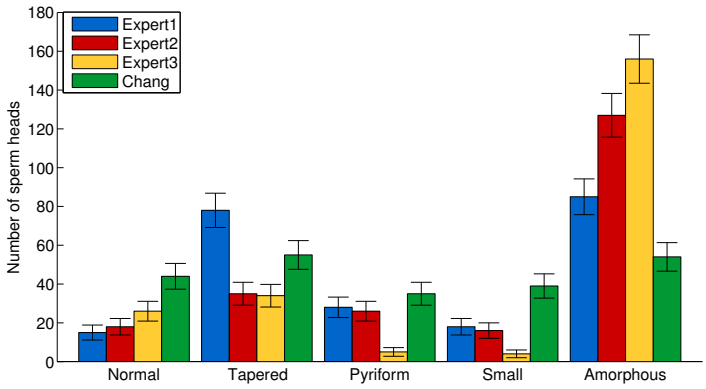


Figure 7.7: **Inter-expert and automatic classification variability in five-class classification.** For each class, the number of sperm heads $\pm$ SE for each class is shown.according to each expert and our approach is shown, using DS3 as testing dataset. One can observe that for certain classes (such as pyriform and tapered) our approach could be easily confused with a human expert. But it is not the case for classes such as amorphous.

Trying to show how difficult the discrimination of five classes is, we have performed some experiments to evaluate dimensionality reduction (DR) techniques. These experiments were conducted following two goals: a) evaluate the impact of using different feature spaces (according to DR technique), and b) identify the most discriminant features (no feature families) according to MLE instrinsic dimension. We worked with the selected features according to DR techniques for each stage of our classification scheme, using DS1 for training and DS3 for testing purposes. We have run 100 iterations in each case, using different dimension reduc-

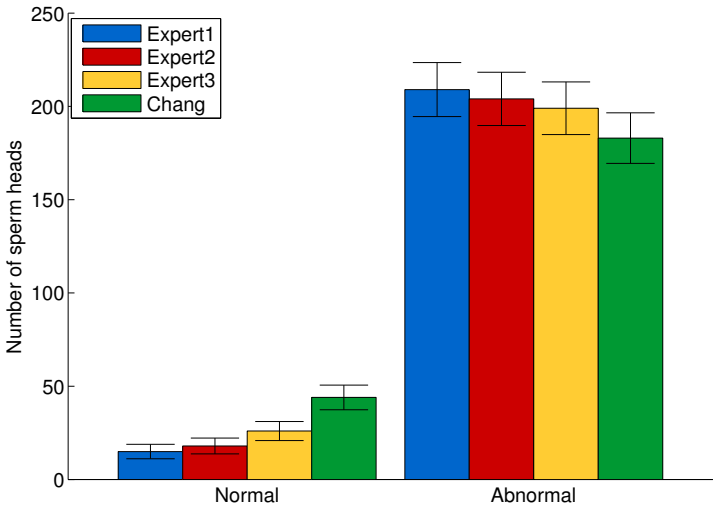


Figure 7.8: **Inter-expert and automatic classification variability in two-class classification.** For each class, the number of sperm heads ± SE according to each expert and our approach is shown, using DS3 as testing dataset. One can observe that as the variability of experts is low, our proposal results also demonstrate low variability against to experts.

tion techniques: PCA (Principal Components Analysis), MDS (Multidimensional Scaling), Kernel PCA and Diffusion Maps. The conclusion: The introduction of dimension reduction techniques has not yielded improved results. However these experiments allow us to confirm that even without DR technique, the problem is really difficult. Around $50-65\%$ can be reached for classes normal and pyriform, around $65-80\%$ for classes tapered and small, and around 30% for class amorphous. And with DR technique, around $25-30\%$ can be reached for classes normal and pyriform, around $65-90\%$ for classes tapered and small, and around 8% for class amorphous. In Tables 7.11 and 7.12, we present the True Positive Rate for each class and the mean accuracy using MLE intrinsic dimension and using the best number of dimensions (according to DR technique), respectively.

DR Technique	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
PCA	0.62	0.45	0.19	0.16	0.16	0.32
MDS	0.00	0.17	0.00	0.00	0.12	0.06
Kernel PCA	0.00	0.90	0.00	0.00	0.10	0.20
DiffusionMaps	0.00	0.60	0.20	0.00	0.20	0.20

Table 7.11: **Results of applying dimensionality reduction techniques using MLE intrinsic dimension.** *tpr* stands for True Positive Rate, N means normal, T means tapered, P means pyriform, S means small, and A means amorphous. *acc* stands for accuracy understood as mean of True Positive Rates of five classes.

DR Technique	nDims	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
PCA	46	0.21	0.67	0.30	0.93	0.08	0.44
MDS	35	0.35	0.54	0.25	0.86	0.06	0.43
Kernel PCA	46	0.36	0.05	0.00	0.04	0.67	0.23
DiffusionMaps	10	0.00	0.60	0.20	0.00	0.20	0.20

Table 7.12: **Results of applying dimensionality reduction techniques using the best number of dimensions.** *nDims* means number of dimensions, *tpr* stands for True Positive Rate, N means normal, T means tapered, P means pyriform, S means small, and A means amorphous. *acc* stands for accuracy understood as mean of True Positive Rates of five classes.

7.3 Discussion

In this chapter we have presented a novel clasification scheme for sperm heads. Our proposal is original and novel from those presented previously (see Chapter 3) because of the combination of ensembled methods for feature selection and classification as pointed below.

1. **Usage of ensemble method for feature selection.** Instead of selecting features for building an ensemble, it is possible to use an ensemble methodology to select features [Kuncheva, 2004]. It was shown that in large feature/small sample size domains several different feature subsets may yield equally optimal results [Saeys et al., 2008], and ensemble feature selection may reduce the risk of choosing an unstable subset from high-dimensional data [Boulesteix and Slawski, 2009]. Applied to mass spectrometry data, Dutkowski and Gambin [Dutkowski and Gambin, 2007] combined several filtering algorithms in a cross-validation framework where multiple classification algorithms are used to evaluate the biomarker selection. Zhang et al. [Zhang et al., 2009] used several classification and filtering algorithms to improve the accuracy prediction and stability of the gene ranking results in a genetic algorithm based wrapper procedure. Similar to the case of supervised ensemble methods for classification, where the idea is to construct a set of simple models called base learners or classifiers, and use their outcome to predict new data, the ensemble methods for feature extraction involves the combination of several feature selectors. That is, ensemble feature selection combines outputs from several base feature selectors to form a committee with improved performance. There are two primary approaches to ensemble construction: parallel and serial. We have chosen a parallel approach that combines independently constructed feature selectors (1NN) to exploit the small dataset for training purposes (DS1) as much as possible.
2. **Usage of a cascade ensemble of SVMs.** In ensemble classification, a set of single classification models is trained, and the output of the ensemble is obtained by aggregating the outputs of the single models, e.g., by majority voting. It was shown that ensemble methods often outperform any single base learner [Dietterich, 2000]. Generally, SVM is considered as the best off-the-shelf classifier and the potential improvement

gain could be significant if it can be successfully used as the base classifier of an ensemble. That is the reason why we choose to use SVM as base learner in our classification scheme after evaluating the impact of using SVM as a monolithic classifier. SVM has been used as the base classifier in many research studies, and it was always shown that the ensemble of SVMs consistently outperformed the single SVM classifier. The most simple and promising idea is to use a bagging procedure with SVM as the base classifier. For example, Caragea et al. [Caragea et al., 2007] applied a bagging ensemble using SVMs for glycosylation site prediction, training each SVM with a re-sampling of the balanced training set. Similarly, Guan et al. [Guan et al., 2008] applied the bagging approach to construct a SVM ensemble for gene function prediction.

Now, looking at the accuracy rates that we reached with our classification scheme and having five classes to consider, we see that the most difficult class is the amorphous one, using both testing datasets (see Table 7.9). There is some kind of consistency in four out of five classes (see Figure 7.4), but in class pyriform, the difference between using DS3 and DST as testing datasets is significant. This could be understood by looking at the large gap in agreement between experts while identifying pyriform sperm heads (see Chapter 4 for an explanation) if one takes into account total or partial agreement (DS3 has partial agreement between experts and DST has total agreement). An important fact to note is that in class small, there is the highest accuracy rate using DS3. While using DST, the best accuracy rate is reached in class pyriform, but followed very close by class small. Again, this could be explained according to what really happens with pyriform sperm heads in manual classification of experts. There are very few pyriform sperm heads in which all experts agree in manual classification, thus, it is supossed that these sperm heads are very well characterized, and easy to discriminate by our classification approach.

7.4 Summary

In this chapter, we have introduced a two-stage classification scheme for classifying human sperm head in five classes (normal, tapered, pyriform, small and amorphous), according to WHO criteria. The critical issues beyond the proposed classification scheme are: 1) emphasis on amorphous heads as an early separation of amorphous heads contributes to minimize the whole classification error, 2) usage of combining classifier approach, and 3) usage of an ensembled feature selection technique for each stage of our scheme. Experimental evaluation shows that our proposed scheme outperforms a number of monolithic classifiers, as well as different cascade classification schemes designed in the context of this research. Our results achieved more than 70% of classification accuracy on a dataset with total agreement among domain experts.

Chapter 8

Conclusions and Future Work

We have presented a two-stage framework for detection and segmentation of human sperm head characteristics (including acrosome and nucleus). The usage of color space combinations (RGB, L*a*b* and YCbCr), together with the usage of a clustering method, provides us a tool for illumination invariant segmentation approach. In addition, our method proposed an ellipse fitting based algorithm to identify the head front direction. This is a very relevant issue with regard to increasing the accuracy of the segmentation.

Our experimental evaluation shows that our proposed framework outperforms the state-of-the-art, with a higher Dice coefficient, lower Hausdorff distance and less dispersion with respect to the results achieved by Carrillos method [Carrillo et al., 2007b, Carrillo et al., 2007a]. Our results achieve notable improvement in the detection rate with fewer false positives and an accurate head, acrosome and nucleus segmentation achieving over 80% overlapping against hand-segmented mask.

To tackle the problem of lacking a public gold-standard for evaluating sperm head segmentation methods, we have introduced a gold-standard for head sperm parts segmentation, built with the cooperation of a referent expert in the field. This gold-standard has been used to evaluate and compare our results with the state-of-the-art method, and can be used to compare not only known techniques, but also future improvements to present approaches. This is a very significant contribution to the scientific community.

We have presented a new descriptor for human sperm heads, named *morphological descriptor*. This descriptor adopts and adapts a number of ROI shape-based measures focusing on ellipse fitness and symmetry.

We have introduced a two-stage classification scheme for classifying human sperm head in five classes (normal, tapered, pyriform, small and amorphous), according to WHO criteria. The approach of combining classifiers together with an ensembled feature selection technique, yields a method for characterizing and classifying sperm heads towards an accurate morphological sperm analysis. Our experimental evaluation shows that our proposed scheme outperforms a number of monolitic classifiers, as well as different cascade classification schemes designed in the context of this research. Our results achieved more than 70%

of classification accuracy on a dataset with total agreement among domain experts, showing that the results of our classification scheme could be confused with those of a human expert for particular classes.

Another significant contribution to the scientific community is the introduction of a gold-standard for evaluating morphological sperm analysis methods, built with the active collaboration of three referent experts in the field. The analysis of the data used to generate this gold-standard confirmed the high inter-expert variability in the sperm morphological classification.

Finally, this thesis demostrates the suitability of image processing, image analysis and machine learning methods for the computer-assisted morphological sperm analysis. We proposed a validated pipeline for detection, segmentation, charaterization and classification of sperm heads that demostrated very promising results.

8.1 Future Work

While this thesis has demonstrated the potential of image processing and machine learning techniques for detecting, segmenting, characterizing and classifying sperm heads towards an accurate morphological sperm analysis, many opportunities for extending the scope of this thesis remain. This section presents some of these directions.

About segmentation

A high correct segmentation rate is crucial for characterizing and classifying human sperm heads towards an accurate morphological sperm analysis. It is well known that segmentation is the hardest task in image processing research and is not difficult to find several up-to-date research works about novel approches promising to tackle this problem in different domains. One of the research directions in image segmentation is about the usage of machine learning methods with outstanding results. In this sense, deep learning techniques seem to be very suitable to extract high level information from images, to discriminate relevant from irrelevant regions. It will be interesting to study and explore the usage of deep learning, basically Convolutional Neural Networks, for segmentation of sperm heads.

About feature selection

One effective approach for generating an accurate combining classifier is the use of different feature subsets. In this thesis, we decided to select feature families (multivariate feature selection) instead of independent features (univariate feature selection). It will be interesting to explore and experiment with feature ranking as a way to select the features for each single classifier in our proposed classification scheme, instead of ranking whole families of features. This could be done by means of methods such as Principal Components Analysis, but also

using Random Subspace and Random Forest.

About base classifiers

A popular method for creating an accurate classifier from a training dataset is to construct several different classifiers and combine their predictions. The integration of multiple classifiers to improve classification results, is currently an active research area in the machine learning community. Therefore, the base classifier choice is a very important decision within the design of combining classifiers. In this thesis, we proposed SVM as the base classifier in both stages of our classification scheme. However, it would be interesting to explore the impact in classification rates when using different base classifiers, such as naive-Bayes classifier, decision tree and neural networks. Furthermore, to use different base classifiers in each stage would be a fascinating research direction.

About classification scheme

We designed a pipeline for detection, segmentation, characterization and clasification of human sperm heads. This pipeline was designed on the basis of a two-stage classification scheme by means of a combined classifier. We showed that the classification accuracy of other classification schemes designed in the context of this research was outperformed by the proposed scheme. However, combination of classifiers is not the only valid alternative to cope with a difficult classification task as morphological sperm analysis.

Future research focused on the applying of Convolutional Neural Networks (CNN) will be interesting. CNN has demonstrated its suitability in hard classification scenarios with slight intra-class variability, showing promising experimental results in different domain classification problems.

Another approach that will be challenging to apply regarding this kind of data is the multi-label classification. When there is not an absolute ground-truth (like in the case of morphological sperm analysis) and a machine learning approach depends on the heterogeneous experience of domain experts, combining the *knowledge* of different experts remains as an open problem. The multi-label classification approach is aimed to train a classifier under such variability situations taking into account different training labels from a number of experts with distinct levels of experience.

Bibliography

[Abbiramy and Tamilarasi, 2011] Abbiramy, V. and Tamilarasi, A. (2011). A comparative study on human spermatozoa images classification with artificial neural network based on FOS, GLCM and morphological features. In *Advances in Digital Image Processing and Information Technology*, volume 205 of *Communications in Computer and Information Science*, pages 220–228. Springer Berlin Heidelberg.

[Aktas and Žunić, 2011] Aktas, M. and Žunić, J. (2011). Measuring shape ellipticity. In *Computer Analysis of Images and Patterns*, volume 6854 of *Lecture Notes in Computer Science*, pages 170–177.

[Aktas and Žunić, 2012] Aktas, M. and Žunić, J. (2012). Sensitivity/robustness flexible ellipticity measures. In *Pattern Recognition*, volume 7476 of *Lecture Notes in Computer Science*, pages 307–316.

[Aktas and Žunić, 2013] Aktas, M. and Žunić, J. (2013). A family of shape ellipticity measures for galaxy classification. *SIAM Journal on Imaging Sciences*, 6(2):765–781.

[Alegre et al., 2013] Alegre, E., Biehl, M., Petkov, N., and Sánchez, L. (2013). Assessment of acrosome state in boar spermatozoa heads using n-contours descriptor and RLVQ. *Computer Methods and Programs in Biomedicine*, 111(3):525–536.

[Alegre et al., 2012] Alegre, E., González-Castro, V., Alaiz-Rodríguez, R., and García-Ordás, M. (2012). Texture and moments-based classification of the acrosome integrity of boar spermatozoa images. *Computer Methods and Programs in Biomedicine*, 108(2):873–881.

[Ali et al., 2012] Ali, R., Gooding, M., Szilágyi, T., Vojnovic, B., Christlieb, M., and Brady, M. (2012). Automatic segmentation of adherent biological cell boundaries and nuclei from brightfield microscopy images. *Machine Vision and Applications*, 23(4):607–621.

[Allen et al., 2008] Allen, G., Hodgson, R., Marsland, S., and Flenley, J. (2008). Machine vision for automated optical recognition and classification of pollen grains or other singulated microscopic objects. In *Proceedings of the 15th International Conference on Mechatronics and Machine Vision in Practice (M2VIP 2008)*, pages 221–226.

[Alt and Guibas, 2000] Alt, H. and Guibas, L. J. (2000). *Handbook of Computational Geometry*, chapter Discrete Geometric Shapes: Matching, Interpolation, and Approximation, pages 121 – 153. North-Holland.

[Amanatiadis et al., 2011] Amanatiadis, A., Kaburlasos, V., Gasteratos, A., and Papadakis, S. (2011). Evaluation of shape descriptors for shape-based image retrieval. *IET Image Processing*, 5(5):493–499.

[Anami et al., 2005] Anami, B., Savakar, D., Makandar, A., and Unki, P. (2005). A neural network model for classification of bulk grain samples based on color and texture. In *Proceedings of the International Conference on Cognition and Recognition*, pages 359–368.

[Arlot and Celisse, 2010] Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

[Atallah, 1985] Atallah, M. J. (1985). On symmetry detection. *IEEE Transactions on Computers*, C-34(7):663–666.

[Atkins et al., 2003] Atkins, E., Hupp, H., Grimes, L., Thompson, C., and Higdon, H. (2003). Evaluating bovine spermatozoa using the Hamilton-Thorne Integrated Visual Optical System (IVOS). *Journal of Animal and Veterinary Advances*, 2(7):385–391.

[Auger, 2010] Auger, J. (2010). Assessing human sperm morphology: top models, underdogs or biometrics? *Assian Journal of Andrology*, 12(1):36–46.

[Auger et al., 2000] Auger, J., Eustache, F., Ducot, B., Blandin, T., Daudin, M., Diaz, I., Matribi, S., Gony, B., Keskes, L., Kolbezen, M., Lamarte, A., Lornage, J., Nomal, N., Pitaval, G., Simon, O., Virant-Klun, I., Spira, A., and Jouannet, P. (2000). Intra- and inter-individual variability in human sperm concentration, motility and vitality assessment during a workshop involving ten laboratories. *Human Reproduction*, 15(11):2360–2368.

[Barroso et al., 1999] Barroso, G., Mercan, R., Ozgur, K., Morshedi, M., Kolm, P., Coetzee, K., Kruger, T., and Oehninger, S. (1999). Intra- and inter-laboratory variability in the assessment of sperm morphology by strict criteria: impact of semen preparation, staining techniques and manual versus computerized analysis. *Human Reproduction*, 14(8):2036–40.

[Beletti et al., 2005a] Beletti, M., Costa, L., and Viana, M. (2005a). A comparison of morphometric characteristics of sperm from fertile Bos taurus and Bos indicus bulls in Brazil. *Animal Reproduction Science*, 85:105 – 116.

[Beletti et al., 2005b] Beletti, M., Costa, L., and Viana, M. (2005b). A spectral framework for sperm shape characterization. *Computers in Biology and Medicine*, 35(6):463 – 473.

[Belongie et al., 2002] Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

[Bhanu Prakash et al., 2002] Bhanu Prakash, K., Ramakrishnan, A., Suresh, S., and Chow, T. (2002). Fetal lung maturity analysis using ultrasound image features. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):38–45.

[Bijar and Mikaeili, 2011] Bijar, A. and Mikaeili, M. (2011). Sperm's tail identification and discrimination in microscopic images of stained human semen smear. In *Proceedings of the*

7th International Symposium on Image and Signal Processing and Analysis (ISPA 2011), pages 709–714.

[Bijar et al., 2012] Bijar, A., Peñalver Benavent, A., Mikaeili, M., and Khayati, R. (2012). Fully automatic identification and discrimination of sperms parts in microscopic images of stained human semen smear. *Journal of Biomedical Science and Engineering*, 5(7):384–395.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 1st edition.

[Blake and Isard, 1998] Blake, A. and Isard, M. (1998). *Active Contours: The Application of Techniques from Graphics, Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer-Verlag, 1st edition.

[Boulesteix and Slawski, 2009] Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–568.

[Bradbury, 2009] Bradbury, L. (2009). *Segmentation of bright-field cell images*. PhD thesis, University of Waterloo, Ontario, Canada.

[Braga-Neto and Goutsias, 2003] Braga-Neto, U. and Goutsias, J. (2003). A theoretical tour of connectivity in image processing and analysis. *Journal of Mathematical Imaging and Vision*, 19(1):5–31.

[Britto Jr. et al., 2014] Britto Jr., A., Sabourin, R., and Oliveira, L. (2014). Dynamic selection of classifiersa comprehensive review. *Pattern Recognition*, 47(11):3665 – 3680.

[Caragea et al., 2007] Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., and Honavar, V. (2007). Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinformatics*, 8(1).

[Carrillo, 2013] Carrillo, H. (2013). Private communication.

[Carrillo and Villarreal, 2007] Carrillo, H. and Villarreal, J. (2007). Software prototipo para la clasificación morfológica de espermatozoides humanos mediante técnicas de procesamiento digital de imágenes. Technical report, Fundación Universidad del Norte, Barranquilla, Colombia.

[Carrillo et al., 2007a] Carrillo, H., Villarreal, J., Sotaquira, M., Goelkel, M., and Gutiérrez, R. (2007a). A computer aided tool for the assessment of human sperm morphology. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2007)*, pages 1152–1157.

[Carrillo et al., 2007b] Carrillo, H., Villarreal, J., Sotaquira, M., Goelkel, M., and Gutiérrez, R. (2007b). Spermatozoon segmentation towards an objective analysis of human sperm morphology. In *Proceedings of the 5th International Symposium on Image and Signal Processing and Analysis*, pages 522–527.

[Casas et al., 1994] Casas, J., Esteban, P., Moreno, A., and Carrera, M. (1994). *Mathematical*

Morphology and its Applications to Image Processing, chapter Morphological scheme for morphometric analysis of epidermal biopsy images, pages 325–331. Kluwer Academic Publishers.

[Caselles V. et al., 1993] Caselles V., Catté F., Coll T., and Dibos F. (1993). A geometric model for active contours. *Numerische Mathematik*, 66(1):1–31.

[Celenk, 1990] Celenk, M. (1990). A color clustering technique for image segmentation. *Computer Vision, Graphics, and Image Processing*, 52(2):145 – 170.

[Chang et al., 2015] Chang, V., García, A., Hitschfeld, N., and Härtel, S. (2015). Gold-standard for computer-assisted morphological sperm analysis. *Computer Methods and Programs in Biomedicine*. Submitted.

[Chang et al., 2014] Chang, V., Saavedra, J., Castañeda, V., Sarabia, L., Hitschfeld, N., and Härtel, S. (2014). Gold-standard and improved framework for sperm head segmentation. *Computer Methods and Programs in Biomedicine*, 117(2):225 – 237.

[Chen et al., 2010] Chen, X., Xun, Y., Li, W., and Zhang, J. (2010). Combining discriminant analysis and neural networks for corn variety identification. *Computers and Electronics in Agriculture*, 71(S1):S48–S53.

[Cheng et al., 2001] Cheng, H., Jiang, X., Sun, Y., and Wang, J. (2001). Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259 – 2281.

[Chica, 2012] Chica, M. (2012). Authentication of bee pollen grains in bright-field microscopy by combining one-class classification techniques and image processing. *Microscopy Research and Technique*, 75(11):1475 – 1485.

[Cipak et al., 2009] Cipak, A., Stanić, P., Durić, K., Serdar, T., and Suchanek, E. (2009). Sperm morphology assessment according to WHO and strict criteria: method comparison and intra-laboratory variability. *Biochemia Medica*, 19(1):87–94.

[Coetzee et al., 1999] Coetzee, K., Kruger, T., and Lombard, C. (1999). Repeatability and variance analysis on multiple computer-assisted (IVOS) sperm morphology readings. *Andrologia*, 31(3):163–168.

[Cohen and Cohen, 1993] Cohen, L. and Cohen, I. (1993). Finite-element methods for active contour models and balloons for 2D and 3D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1131–1147.

[Coifman and Lafon, 2006] Coifman, R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30.

[Conan et al., 1992] Conan, V., Gesbert, S., Howard, C., Jeulin, D., Meyer, F., and Renard, D. (1992). Geostatistical and morphological methods applied to three-dimensional microscopy. *Journal of Microscopy*, 166(2):169–184.

[Costa and Cesar, 2009] Costa, L. and Cesar, R. (2009). *Shape Classification and Analysis:*

Theory and Practice. CRC Press, 2nd edition.

[Cox and Cox, 2008] Cox, M. and Cox, T. (2008). Multidimensional scaling. In *Handbook of Data Visualization*, pages 315–347.

[Damoulas and Girolami, 2009] Damoulas, T. and Girolami, M. (2009). Combining feature spaces for classification. *Pattern Recognition*, 42(11):2671–2683.

[Dasarathy, 1991] Dasarathy, B. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, California, USA, 1st edition.

[Davis et al., 1992] Davis, R., Bain, D., Siemers, R., Thal, D., Andrew, J., and Gravance, C. (1992). Accuracy and precision of the CellForm-Human automated sperm morphometry instrument. *Fertility and Sterility*, 58(4):763–769.

[Davis and Gravance, 1993] Davis, R. and Gravance, C. (1993). Standardization of specimen preparation, staining, and sampling methods improves automated sperm-head morphometry analysis. *Fertility and Sterility*, 59(2):412–417.

[Dearing et al., 2014] Dearing, C., Kilburn, S., and Lindsay, K. (2014). Validation of the Sperm Class Analyzer CASA system for sperm counting in a busy diagnostic semen analysis laboratory. *Human Fertility*, 17(1):37–44.

[Dice, 1945] Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

[Diestel, 2010] Diestel, R. (2010). *Graph Theory*. Springer-Verlag, 4th edition.

[Dietterich, 2000] Dietterich, T. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

[Domar et al., 1992] Domar, A., Broome, A., Zuttermeister, P., Seibel, M., and Friedman, R. (1992). The prevalence and predictability of depression in infertile women. *Fertility and Sterility*, 58(6):1158–1163.

[Douglas-Hamilton, 1995] Douglas-Hamilton, D. (1995). Validation procedures for the Hamilton-Thorne Integrated Visual Optical System sperm and cell analyzer. *Quality Assurance*, 4(4):340–347.

[Duda et al., 2000] Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley-Interscience, 1st edition.

[Dugundji, 1966] Dugundji, J. (1966). *Topology*. Allyn and Bacon, 1st edition.

[Duin, 1996] Duin, R. (1996). A note on comparing classifiers. *Pattern Recognition Letters*, 17(5):529–536.

[Dutkowski and Gambin, 2007] Dutkowski, J. and Gambin, A. (2007). On consensus

biomarker selection. *BMC Bioinformatics*, 8(Suppl 5).

[Eckhardt and Latecki, 2008] Eckhardt, U. and Latecki, L. (2008). Digital topology. In *Current Topics in Pattern Recognition Research, Research Trends*. Council of Scientific Information, Vilayil Gardens, Trivandrum, India.

[El-Naqa et al., 2002] El-Naqa, I., Yongyi, Y., Wernick, M., Galatsanos, N., and Nishikawa, R. (2002). A support vector machine approach for detection of microcalcifications. *IEEE Transactions on Medical Imaging*, 21(12):1552–1563.

[Enginsu et al., 1991] Enginsu, M., Dumoulin, J., Pieters, M., Bras, M., Evers, J., and Geraedts, J. (1991). Evaluation of human sperm morphology using strict criteria after diff-quik staining: correlation of morphology with fertilization in vitro. *Human Reproduction*, 6(6):854–858.

[Fleiss, 1971] Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

[Ford and Roberts, 1998] Ford, A. and Roberts, A. (1998). Colour space conversions. Technical report, Westminster University.

[Freund, 1966] Freund, C. (1966). Standards for the rating of human sperm morphology. A cooperative study. *International Journal of Fertility*, 11(1):97–180.

[Fuchs and Buhmann, 2011] Fuchs, T. and Buhmann, J. (2011). Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, 35(78):515 – 530.

[Garcia and Tziritas, 1999] Garcia, C. and Tziritas, G. (1999). Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277.

[García-Olalla et al., 2015] García-Olalla, O., Alegre, E., Fernández-Robles, L., Malm, P., and Bengtsson, E. (2015). Acrosome integrity assessment of boar spermatozoa images using an early fusion of texture and contour descriptors. *Computer Methods and Programs in Biomedicine*, 120(1):49–64.

[Giraldi and Oliveira, 2000] Giraldi, G. and Oliveira, A. (2000). Snake models. Technical report, Universidade de Rio de Janeiro.

[Gokturk et al., 2001] Gokturk, S., Tomasi, C., Acar, B., Beaulieu, C., Paik, D., Jeffrey, R., Yee, J., and Napel, S. (2001). A statistical 3D pattern processing method for computer-aided detection of polyps in CT colonography. *IEEE Transactions on Medical Imaging*, 20(12):1251–1260.

[González-Castro et al., 2012] González-Castro, V., Alegre, E., García-Olalla, O., García-Ordás, D., García-Ordás, M., and Fernández-Robles, L. (2012). Curvelet-based texture description to classify intact and damaged boar spermatozoa. In *Image Analysis and Recognition*, volume 7325 of *Lecture Notes in Computer Science*, pages 448–455.

[González-Castro et al., 2009] González-Castro, V., Alegre, E., Morala-Argüello, P., and Suárez, S. (2009). A combined and intelligent new segmentation method for boar semen based on thresholding and watershed transform. *International Journal of Imaging*, 2(S09):70–80.

[Guan et al., 2008] Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., and Troyanskaya, O. (2008). Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome biology*, 9(Suppl 1):S3.

[Hall et al., 1992] Hall, L., Bensaid, A., Clarke, L., Velthuizen, R., Silbiger, M., and Bezdek, J. (1992). A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. *IEEE Transactions on Neural Networks*, 3(5):672–682.

[Hartigan, 1975] Hartigan, J. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

[Hidalgo and Dorado, 2009] Hidalgo, M. and Dorado, J. (2009). Objective assessment of goat sperm head size by computer-assisted sperm morphometry analysis (ASMA). *Small Ruminant Research*, 87(13):108 – 110.

[Hidalgo et al., 2008] Hidalgo, M., Rodríguez, I., Dorado, J., and Soler, C. (2008). Morphometric classification of Spanish thoroughbred stallion sperm heads. *Animal Reproduction Science*, 103(34):374 – 378.

[Hsu et al., 2002] Hsu, R.-L., Abdel-Mottaleb, M., and Jain, A. (2002). Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706.

[Hu, 1962] Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187.

[Isa et al., 2009] Isa, N., Salamah, S., and Ngah, U. (2009). Adaptive fuzzy moving k-means clustering algorithm for image segmentation. *IEEE Transactions on Consumer Electronics*, 55(4):2145–2153.

[Jain et al., 2000] Jain, A., Duin, R., and Jianchang, M. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal Component Analysis*. Springer-Verlag New York, New York, USA, 2nd edition.

[Jurio et al., 2009] Jurio, A., Pagola, M., Paternain, D., Lopez-Molina, C., and Melo-Pinto, P. (2009). Interval-valued restricted equivalence functions applied on clustering techniques. In *Proceedings of the 13rd International Fuzzy Systems Association World Congress and 6th European Society for Fuzzy Logic and Technology Conference (IFSA-EUSFLAT)*, pages 831–836.

[Karp, 2013] Karp, G. (2013). *Cell and Molecular Biology: Concepts and Experiments*. John Wiley and Sons, 7th edition.

[Kass M. et al., 1988] Kass M., Witkin A., and Terzopoulos D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331.

[Katz et al., 1986] Katz, D., Overstreet, J., Samuels, S., Niswander, P., Bloom, T., and Lewis, E. (1986). Morphometric analysis of spermatozoa in the assessment of human male fertility. *Journal of Andrology*, 7(4):203–210.

[Kaur and Kranthi, 2012] Kaur, A. and Kranthi, B. (2012). Comparison between YCbCr color space and CIELab color space for skin color segmentation. *International Journal of Applied Information Systems*, 3(4):30–33.

[Kaur and Singh, 2013] Kaur, H. and Singh, B. (2013). Classification and grading rice using multi-class SVM. *International Journal of Scientific and Research Publications*, 3(4):1–5.

[Kaya et al., 2013] Kaya, Y., Pnar, S., Erez, M., and Fidan, M. (2013). An expert classification system of pollen of onopordum using a rough set approach. *Review of Palaeobotany and Palynology*, 189(0):50 – 56.

[Kim and Kim, 2000] Kim, H. and Kim, J. (2000). Region-based shape descriptor invariant to rotation, scale and translation. *Signal Processing: Image Communication*, 16(12):87 – 93.

[Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

[Kobayashi et al., 1991] Kobayashi, T., Jinno, M., Sugimura, K., Nozawa, S., Sugiyama, T., and Iida, E. (1991). Sperm morphological assessment based on strict criteria and in-vitro fertilization outcome. *Human Reproduction*, 6(7):983–986.

[Kohavi and Provost, 1998] Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(2-3):271–274.

[Korzynska et al., 2007] Korzynska, A., Strojny, W., Hoppe, A., Wertheim, D., and Hoser, P. (2007). Segmentation of microscope images of living cells. *Pattern Analysis and Applications*, 10(4):301–319.

[Kruger et al., 1993] Kruger, T., du Toit, T., Franken, D., Acosta, A., Oehninger, S., Menkveld, R., and Lombard, C. (1993). A new computerized method of reading sperm morphology (strict criteria) is as efficient as technician reading. *Fertility and Sterility*, 59(1):202–209.

[Kruger et al., 1986] Kruger, T., Menkveld, R., Stander, F., Lombard, C., Van der Merwe, J., van Zyl, J., and Smith, K. (1986). Sperm morphologic features as a prognostic factor in in-vitro fertilization. *Fertility and Sterility*, 46(6):1118–1123.

[Kuncheva, 2004] Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 1st edition.

[Lacquet et al., 1996] Lacquet, F., Kruger, T., du Toit, T., Lombard, C., Sánchez Sarmiento, C., de Villiers, A., and Coetzee, K. (1996). Slide preparation and staining procedures for reliable results using computerized morphology. *Archives of Andrology*, 36(2):133–138.

[Leou and Tsai, 1987] Leou, J.-J. and Tsai, W.-H. (1987). Automatic rotational symmetry determination for shape analysis. *Pattern Recognition*, 20(6):571 – 582.

[Littmann and Ritter, 1997] Littmann, E. and Ritter, H. (1997). Adaptive color segmentation – A comparison of neural and statistical methods. *IEEE Transactions on Neural Networks*, 8(1):175–185.

[Liu et al., 2005] Liu, Z., Cheng, F., Ying, Y., and Rao, X. (2005). Identification of rice seed varieties using neural network. *Journal of Zhejiang University - Science B*, 6(11):1095–1100.

[MacLeod and Gold, 1951] MacLeod, J. and Gold, R. (1951). The male factor in fertility and infertility – Sperm morphology in fertile and infertile marriage. *Fertility and Sterility*, 2(1):394–414.

[Marcos et al., 2015] Marcos, J., Nava, R., Cristóbal, G., Redondo, R., Escalante-Ramírez, B., Bueno, G., Déniz, O., González-Porto, A., Pardo, C., Chung, F., and Rodríguez, T. (2015). Automated pollen identification using microscopic imaging and texture analysis. *Micron*, 68(0):36 – 46.

[Maree et al., 2010] Maree, L., du Plessis, S., Menkveld, R., and van der Horst, G. (2010). Morphometric dimensions of the human sperm head depend on the staining method used. *Human Reproduction*, 25(6):1369–1382.

[Maree and van der Horst, 2013] Maree, L. and van der Horst, G. (2013). Quantification and identification of sperm subpopulations using computer-aided sperm analysis and species-specific cut-off values for swimming speed. *Biotechnic & Histochemistry*, 88(3-4):181–193.

[Masuda et al., 1993] Masuda, T., Yamamoto, K., and Yamada, H. (1993). Detection of partial symmetry using correlation with rotated-reflected images. *Pattern Recognition*, 26(8):1245 – 1253.

[Mebatsion et al., 2013] Mebatsion, H., Paliwal, J., and Jayas, D. (2013). Automatic classification of non-touching cereal grains in digital images using limited morphological and color features. *Computers and Electronics in Agriculture*, 90:99–105.

[Menkveld et al., 2011] Menkveld, R., Holleboom, C., and Rhemrev, J. (2011). Measurement and significance of sperm morphology. *Asian Journal of Andrology*, 13(1):59–68.

[Menkveld et al., 2001] Menkveld, R., Wong, W., Lombard, C., Wetzels, A., Thomas, C., Merkus, H., and Steegers-Theunissen, R. (2001). Semen parameters, including WHO and strict criteria morphology, in a fertile and subfertile population: an effort towards

standardization of in-vivo thresholds. *Human Reproduction*, 16(6):1165–1171.

[Metzler et al., 2000] Metzler, V., Lehmann, T., Bienert, H., Mottaghy, K., and Spitzer, K. (2000). Scale-independent shape analysis for quantitative cytology using mathematical morphology. *Computers in Biology and Medicine*, 30(3):135–151.

[Mitsumoto et al., 2009] Mitsumoto, K., Yabusaki, K., and Aoyagi, H. (2009). Classification of pollen species using autofluorescence image analysis. *Journal of Bioscience and Bioengineering*, 107(1):90 – 94.

[Moench and Holt, 1931] Moench, G. and Holt, H. (1931). Sperm morphology in relation to fertility. *American Journal of Obstetrics and Gynecology*, 22(1):199–210.

[Mortimer and Menkveld, 2001] Mortimer, D. and Menkveld, R. (2001). Sperm morphology assessment - historical perspectives and current opinions. *Journal of Andrology*, 22(2):192–205.

[Mortimer et al., 2015] Mortimer, S. T., Horst, G. v. d., and Mortimer, D. (2015). The future of computer-aided sperm analysis. *Asian journal of andrology*, 17(4):545–553.

[Nafisi et al., 2005] Nafisi, V., Moradi, M., and Nasr-Esfahani, M. (2005). Sperm identification using elliptic model and tail detection. *Transactions on Engineering, Computing and Technology*, 7(1):419–421.

[Nisbet et al., 2009] Nisbet, R., Elder, J., and Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, 1st edition.

[Nixon and Aguado, 2008] Nixon, M. and Aguado, A. (2008). *Feature Extraction & Image Processing*. Academic Press, 2nd edition.

[Ohta et al., 1980] Ohta, Y.-I., Kanade, T., and Sakai, T. (1980). Color information for region segmentation. *Computer Graphics and Image Processing*, 13(3):222 – 241.

[Okun and Valentini, 2008] Okun, O. and Valentini, G. (2008). Dataset complexity can help to generate accurate ensembles of k-nearest neighbors. In *IEEE International Joint Conference on Neural Networks (IJCNN 2008): IEEE World Congress on Computational Intelligence*, pages 450–457.

[Otsu, 1979] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66.

[Park et al., 1997] Park, K., Yi, W., and Paick, J. (1997). Segmentation of sperms using the strategic Hough transform. *Annals of Biomedical Engineering*, 25:294–302.

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639.

[Persoon and Fu, 1977] Persoon, E. and Fu, K.-S. (1977). Shape discrimination using Fourier

descriptors. *IEEE Transactions on Systems, Man and Cybernetics*, 7(3):170–179.

[Peura and Iivarinen, 1997] Peura, M. and Iivarinen, J. (1997). Efficiency of simple shape descriptors. In *Proceedings of the 3rd International Workshop on Visual Form*, pages 443–451.

[Phetudomsinsuk et al., 2008] Phetudomsinsuk, K., Sirinarumitr, K., Laikul, A., Pinyopummin, A., et al. (2008). Morphology and head morphometric characters of sperm in Thai native crossbred stallions. *Acta Veterinaria Scandinavica*, 50(1):41.

[Pietikainen et al., 1996] Pietikainen, M., Nieminen, S., Marszalec, E., and Ojala, T. (1996). Accurate color discrimination with classification based on feature distributions. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 3, pages 833–838.

[Punyasena et al., 2012] Punyasena, S., Tcheng, D., Wesseln, C., and Mueller, P. (2012). Classifying black and white spruce pollen using layered machine learning. *The New phytologist*, 196(3):937–944.

[Rahtu et al., 2006] Rahtu, E., Salo, M., and Heikkila, J. (2006). A new convexity measure based on a probabilistic interpretation of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1501–1512.

[Ramakrishnan and Emary, 2010] Ramakrishnan, S. and Emary, I. (2010). *Computational Techniques and Algorithms for Image Processing*. Lambert Academic Publishing, 1st edition.

[Ranawana and Palade, 2006] Ranawana, R. and Palade, V. (2006). Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, 3(1):35–61.

[Raykar et al., 2009] Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., and Moy, L. (2009). Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 889–896.

[Redondo et al., 2015] Redondo, R., Bueno, G., Chung, F., Nava, R., Marcos, J., Cristóbal, G., Rodríguez, T., González-Porto, A., Pardo, C., Déniz, O., and Escalante-Ramírez, B. (2015). Pollen segmentation and feature evaluation for automatic classification in bright-field microscopy. *Computers and Electronics in Agriculture*, 110:56 – 69.

[Rivera-Montes et al., 2013] Rivera-Montes, A., Rivera-Gallegos, A., Rodríguez-Villasana, E., Juárez-Bengoa, A., Díaz-Pérez, M., and Hernández-Valencia, M. (2013). Estimate of the variability in the evaluation of semen analysis. *Ginecologia y Obstetricia de Mexico*, 81(11):639–644.

[Rodríguez-Damián et al., 2006] Rodríguez-Damián, M., Cernadas, E., Formella, A., Fernández-Delgado, M., and De Sá-Otero, P. (2006). Automatic detection and classification of grains of pollen based on shape and texture. *IEEE Transactions on Systems,*

Man, and Cybernetics, Part C: Applications and Reviews, 36(4):531–542.

[Rosin, 2003] Rosin, P. (2003). Measuring shape: ellipticity, rectangularity, and triangularity. *Machine Vision and Applications*, 14(3):172–184.

[Rosin, 2005] Rosin, P. (2005). *Handbook of Pattern Recognition and Computer Vision*, chapter Computing global shape measures. World Scientific.

[Rosin, 2009] Rosin, P. (2009). Classification of pathological shapes using convexity measures. *Pattern Recognition Letters*, 30(5):570 – 578.

[Rosin and Mumford, 2006] Rosin, P. and Mumford, C. (2006). A symmetric convexity measure. *Computer Vision and Image Understanding*, 103(2):101 – 111.

[Saavedra and Bustos, 2014] Saavedra, J. and Bustos, B. (2014). Sketch-based image retrieval using keyshapes. *Multimedia Tools and Applications*, 73(3):2033–2062.

[Saeys et al., 2008] Saeys, Y., Abeel, T., and Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pages 313–325.

[Sánchez and Petkov, 2009] Sánchez, L. and Petkov, N. (2009). *Similarity-Based Clustering*, chapter Estimation of Boar Sperm Status Using Intracellular Density Distribution in Grey Level Images, pages 169–184. Springer-Verlag.

[Schleicher and Zagar, 2008] Schleicher, D. and Zagar, B. (2008). Image processing to estimate the ellipticity of steel coils using a concentric ellipse fitting algorithm. In *Proceedings of the 9th International Conference on Signal Processing (ICSP 2008)*, pages 884–890.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computing*, 10(5):1299–1319.

[Serra, 1986] Serra, J. (1986). Introduction to mathematical morphology. *Computer Vision, Graphics, and Image Processing*, 35(3):283–305.

[Severa et al., 2010] Severa, L., Máchal, L., Švábová, L., and Mamica, O. (2010). Evaluation of shape variability of stallion sperm heads by means of image analysis and Fourier descriptors. *Animal Reproduction Science*, 119(12):50 – 55.

[Shah, 2008] Shah, S. (2008). Automatic cell segmentation using a shape-classification model in immunohistochemically stained cytological images. *IEICE - Transactions on Information and Systems*, E91-D(7):1955–1962.

[Shih, 2009] Shih, F. (2009). *Image Processing and Mathematical Morphology: Fundamentals and Applications*. CRC Press, 1st edition.

[Shojaedini and Heydari, 2014] Shojaedini, S. and Heydari, M. (2014). Automatic sperm analysis in microscopic images of human semen: Segmentation using minimization of information distance. *Irian Journal of Medical Physics*, 11(2-3):284 – 293.

[Soille, 2010] Soille, P. (2010). *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, 2nd edition.

[Soler et al., 2003] Soler, C., de Monserrat, J., Gutiérrez, R., Nuñez, J., Nuñez, M., Sancho, M., Pérez-Sánchez, F., and Cooper, T. (2003). Use of the Sperm-Class Analyser for objective assessment of human sperm morphology. *International Journal of Andrology*, 26(5):262–270.

[Sonka et al., 2008] Sonka, M., Hlavac, V., and Boyle, R. (2008). *Image Processing, Analysis and Machine Vision*. Thomson-Engineering, 3th edition.

[Szeliski, 2011] Szeliski, R. (2011). *Computer Vision: Algorithms and Applications*. Springer-Verlag, 1st edition.

[Teague, 1980] Teague, M. (1980). Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70(8):920–930.

[Theodoridis and Koutroumbas, 2009] Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*. Academic Press, 4th edition.

[Treloar et al., 2004] Treloar, W., Taylor, G., and Flenley, J. (2004). Towards automation of palynology 1: analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning electron microscope images. *Journal of Quaternary Science*, 19(8):745–754.

[Tulyakov et al., 2008] Tulyakov, S., Jaeger, S., Govindaraju, V., and Doermann, D. (2008). Review of classifier combination methods. In *Machine Learning in Document Analysis and Recognition*, volume 90 of *Studies in Computational Intelligence*, pages 361–386.

[Ukil, 2007] Ukil, A. (2007). *Intelligent Systems and Signal Processing in Power Engineering*. Springer-Verlag, 1st edition.

[Vege, 2012] Vege, S. (2012). Ensemble of feature selection techniques for high dimensional data. Master's thesis, Department of Mathematics and Computer Science, Western Kentucky University, Kentucky, USA.

[Vezhnevets et al., 2003] Vezhnevets, V., Sazonov, V., and Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. In *Proceedings of the 13th International Conference on Computer Graphics and Vision (Graphicon 2003)*, pages 85–92.

[Viola and Jones, 2002] Viola, P. and Jones, M. (2002). Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154.

[Vorobyov, 2011] Vorobyov, M. (2011). Shape classification using zernike moments. Technical report, iCamp-University of California Irvine.

[Walczak-Jedrzejowska et al., 2013] Walczak-Jedrzejowska, R., Marchlewska, K., Oszukowska, E., Filipiak, E., Bergier, L., and Slowikowska-Hilczer, J. (2013). Semen analysis standardization: is there any problem in polish laboratories? *Asian Journal of Andrology*,

15:616–621.

[Wang et al., 1991] Wang, C., Leung, A., Tsoi, W., Leung, J., Ng, V., Lee, K., and Chan, S. (1991). Computer-assisted assessment of human sperm morphology: comparison with visual assessment. *Fertility and Sterility*, 55(5):983–988.

[Weickert, 1998] Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*. ECMI Series, Teubner-Verlag, 1st edition.

[WHO, 2010a] WHO (2010a). Mother or nothing: the agony of infertility. *World Health Organization Bulletin*, 88(12):881–882.

[WHO, 2010b] WHO (2010b). *World Health Organization - Laboratory Manual for the Examination and Processing of Human Semen*. Cambridge University Press, 5th edition.

[Yan et al., 2010] Yan, Y., Rosales, R., Fung, G., Schmidt, M. W., Valadez, G. H., Bogoni, L., Moy, L., and Dy, J. G. (2010). Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, volume 9, pages 932–939.

[Yao et al., 2005] Yao, J., Kharma, N., and Grogono, P. (2005). A multi-population genetic algorithm for robust and fast ellipse detection. *Pattern Analysis and Applications*, 8(1-2):149–162.

[Yi et al., 1998] Yi, W., Park, K., and Paick, J. (1998). Parameterized characterization of elliptic sperm heads using Fourier representation and wavelet transform. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 974–977.

[Zahn and Roskies, 1972] Zahn, C. T. and Roskies, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, C-21(3):269–281.

[Zhang and Lu, 2001] Zhang, D. and Lu, G. (2001). Content-based shape retrieval using different shape descriptors: a comparative study. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2001)*, pages 1139–1142.

[Zhang and Lu, 2002] Zhang, D. and Lu, G. (2002). Shape-based image retrieval using generic Fourier descriptor. *Signal Processing: Image Communication*, 17(10):825 – 848.

[Zhang and Lu, 2004] Zhang, D. and Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37:1–19.

[Zhang et al., 2004] Zhang, Y., Fountain, D. W., Hodgson, R. M., Flenley, J. R., and Gunetileke, S. (2004). Towards automation of palynology 3: pollen pattern recognition using Gabor transforms and digital moments. *Journal of Quaternary Science*, 19(8):763–768.

[Zhang et al., 2009] Zhang, Z., Yang, P., Wu, X., and Zhang, C. (2009). An agent-based hybrid system for microarray data analysis. *IEEE Intelligent Systems*, 24(5):53–63.

[Zheng and Xue, 2009] Zheng, N. and Xue, J. (2009). *Statistical Learning and Pattern Analysis for Image and Video Processing*. Springer-Verlag, 1st edition.

[Zunić and Rosin, 2004] Zunić, J. and Rosin, P. (2004). A new convexity measure for polygons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):923–934.

Appendix A

Sperm Segmentation Method by Carrillo et al.

Carrillo et al. [Carrillo et al., 2007b, Carrillo et al., 2007a] presented a two-stage method for detection and segmentation of sperm heads and mid-pieces, looking for an objective analysis of human sperm morphology at future. At the first stage (detection), the main goal is to detect and extract a sperm cell from a microscopic image with multiple sperm cells. At the second stage (segmentation), the authors proposed a segmentation method based on an information fusion technique. We will discuss the whole method in the next section, as well as implementation details.

It is important to mention that the images used for testing the method poposed by Carrillo et al. were obtained from semen smears using Hematoxylin/Eosin staining and an adapter of $600x$. The input images had JPG format and size of 3072×2304 pixels.

A.1 Method Description

A.1.1 Detection and extraction of sperm cells

As an initial step, the RGB image is transformed to a gray scale image followed by image thresholding using Otsu method [Otsu, 1979]. The objects discriminated after the thresholding process are discarded or not according to their area. The remaining objects are again discriminated by histogram analysis and only objects that have a specific number of dark pixels are finally segmented (this leads to complete differentiation between spermatozoa and artifacts of similar size).

A.1.2 Segmentation of sperm cells

The main purpose of the segmentation stage is to subdivide a sperm cell into its constituent regions of interest: acrosome, nucleus and mid-piece. The proposed method is based on nth-level thresholding of an image (in this case a spermatozoon image) followed by intersection with n special masks, this leads to some partial results (isolated regions of interest) that are joined or fusioned to obtain the desired segmentation results. This method is aimed to segment objects in microscopic images with a prior knowledge of the morphology of the object.

Image pre-processing

The image of each sperm cell is processed with a power law transformation over each component of the RGB cube; the most important components for the transformation were the red and green. The red component contains most of the information associated with the darkest color, which domains the head, and with its enhancement is possible to get a better differentiation of the nucleus from the others parts. The transformation of the green component gives uniformity to the background and allows the spermatozoon's contour discrimination from the seminal plasma.

Nth-fusion method

The procedure of the nth-fusion method is as follows:

1. **Nth-thresholding segmentation**: the input image is modeled as an object formed by m regions, where n of these regions define the object of interest in the image. In this sense, n different thresholded images are obtained by applying the same number of thresholds to the initial image. Each one of these thresholded images represents correctly at least one region of the object to be segmented. The n different threshold values are given by the prior knowledge of the object's morphological structure. The result of this block is a set of n images each one correctly representing at least one region of the object to be segmented.
2. **Building growing masks**: The isolation of the n regions of interest needed for the fusion process is accomplished by performing morphological operations between the set of n images and their corresponding growing masks. The growing mask enhances the regions to be detected and discards those that are not of interest. In other words, is necessary that the mask only grows over the region of interest and never over other regions. This block shows the importance of bringing information to the algorithm with the information fusion technique, because a priori morphological model makes easier and accurate the construction of the growing masks.
3. **Morphological operation**: Several morphological operations between each of the nth-thresholded images and its corresponding growing mask are performed. This process leads to n isolated objects representing the details of interest of the region.

4. **Region Fusion**: The last step is to fusion the n isolated regions (from previous step) into one. This is accomplished by means of morphological operations (opening and closing procedures, followed by erosions and unions). The final result is the desired segmented object.

Segmentation procedure

The design of the segmentation procedure was based on the nth-fusion method, discussed before. This procedure starts with a RBG image containing a single sperm cell (as returned by the detection stage - Section A.1.1), and applies a transformation from RGB color space to gray intensities. With this ideal input, the segmentation procedure continue as follows:

1. **Getting reference points**: This step works with the green component from the original RGB image. This green component is used because it gives the best contrast between the background and the spermatozoon's contour. The Otsu's thresholding method is applied and the ROI with the biggest area is selected. The object obtained represents the complete spermatozoon , while the centroid and points that define the major axis of the basic rectangle are obtained from this object. The main goals are getting a point situated in the top part of the head and belonging to its main axis, and getting a point situated in the bottom of the head and belonging to its main axis. The idea is to compare the Euclidean distance between this two points and the centroid is possible to determine which one is situated in the top or bottom part of the head, this is possible because a prior knowledge of the morphology of the spermatozoon stands that the mass center of the head is near of its top.
2. **Nucleus segmentation**: In this step, the authors obtained the first image of the set of thresholded images for the nth fusion method. In this first image, it is important to mark the bottom part of the sperm head (nucleus region). The thresholding process in this region is done by means of the usage of the Euclidian distance with the following parameters: center of the sphere $[10, 0, 100]$ and radius 100.
3. **Building of growing masks**: Specifically in this algorithm two growing masks are used, one used to segment the head and the other for acrosome segmentation. The first mask, applied to the nucleus, enhances the bottom part of the head. The second mask is a polygon that encloses the top region of the head and excludes the bottom part; this is done first by dividing the head in two planes with a line passing through the mass center of the head and perpendicular to the major axis too. Besides, the polygon has to include the reference point located in the top of the head and exclude the one located at the bottom.
4. **Head segmentation**: The head is segmented following the nth-fusion method for segmentation and uses two level thresholds: one obtained with the Otsu's method and one as a fixed threshold. The first threshold enhances the details of the bottom part of the head and the second the top part. The masks are obtained in the previous step.
5. **Acrosome segmentation**: The acrosome is segmented by intersecting previous segmented objects. The acrosome is obtained by intersecting the segmented head with the isolated object that defines the bottom part of the head (obtained in the previous step).

A.2 Algorithms

Algorithm 8 Sperm detection and segmentation by Carrillo et al.

```
    imRgb: original image
    aMin: minimum size in pixels for a sperm head
    aMax: maximum size in pixels for a sperm head
    gIntensity: maximum intensity (threshold value)
    tMin: threshold value for minimum intensity sum
    tMax: threshold value for minimum intensity sum
 1: imGray ← transformRGBtoGray(imRgb)
 2: threshold ← OtsuThreshold(imGray)
 3: imBin ← pixels in imGray whose value is > threshold
 4: imOpen ← opening(imBin)
 5: imArea ← eraseBySize(imOpen,aMin,aMax)
 6: regions ← labelRegions(imArea)
 7: for each region i do
 8:     subimg ← regions(i) cropped according to its MBR
 9:     thr ← sum(pixels in subimg with intensity in [0,gIntensity])
10:     if (thr > tMin) AND (thr < tMax) then
11:         subimRgb ← imRgb(subimg)
12:         subimPro ← transformRGBtoGrayChannel(subimRgb)
13:         [head,nucleus,acrosome]=nthFusion(subimRgb,subimPro)
14:     end if
15: end for
```

Algorithm 9 Nth-Fusion Segmentation by Carrillo et al.

```
    imgPro: processed image containing only one sperm cell
 1: imGray ← transformRGBtoGray(imgPro)
 2: threshold ← OtsuThreshold(imGray)
 3: imBin ← pixels in imGray whose value is > threshold
 4: region ← chooseBiggerRegion(imBin)
 5: [xc,yc] ← getCentroid(region)
 6: [pd1,pd2] ← getMajorAxisPoints(region,xc,yc)
 7: imRad ← pixels with Euclidean distance to fixed-central-point < fixed-radius
 8: mask1 ← getGrowingMask1(imRad,xc,yc)
 9: mask2 ← getGrowingMask2(imBin,xc,yc,pd1,pd2)
10: imRed ← transformRGBtoGrayChannel(imPro,1)
11: imBinR ← pixels in imRed whose value is > fixed-threshold
12: headUp ← opening(imBin AND mask2)
13: nucleus ← opening(imBinR AND mask1)
14: head ← nucleus OR headUp
15: acrosome ← head - nucleus
```

A.3 Implementation Details

We implemented Carrillo's method [Carrillo et al., 2007b, Carrillo et al., 2007a, Carrillo and Villarreal, 2007, Carrillo, 2013] following the algorithms presented before using Matlab[1], since

[1] Matlab *R*2013*a* 8.1.0.604

it is not available as a source code by the authors to compare our detection and segmentation precision.

In addition, we evaluated parameter values for Carrillo's method. There are five free parameters in our implementation. The first parameter, *aMin*, is the minimum size in pixels for a sperm head, while *aMax* is the maximum size. The parameter *gIntensity* is used as a maximum threshold value. The two last parameters, *tMin* and *tMax*, represent a range for sperm cell validation. In this sense, if the sum of pixel values of a candidate sperm cell falls within the range $[tMin, tMax]$, the candidate is considered a sperm cell, in any other case it is discarded. In Table A.1, we show the variation of parameters for Carrillo's method in our experiments.

Parameter name	Range	Best value
aMin	50 : 25 : 400	200
aMax	500 : 100 : 1500	500
gIntensity	90 : 10 : 150	120
tMin	2000 : 500 : 6000	5500
tMax	2300 : 500 : 27000	24000

Table A.1: Variation of parameters for Carrillo's method.

There some parameters showed in the algorithms presented in the Section A.2 whose value was kept as set by the authors. In this sense, *fixed-central-point* = [100100], *fixed-radius* = 100, and *fixed-threshold* = 130.

Appendix B

Discarded Classification Schemes

In this appendix, we aim to describe the classification schemes that were regarded along the development of this PhD thesis, and to explain the reasons for choosing the classification scheme introduced in Section 7.1.

It is important to note that we set out to classify sperm heads in five classes: normal, tapered, pyriform, small and amorphous. We designed the following classification schemes with this goal in mind, evaluating four base classifiers in each classification scheme: $1-$NN, Bayes classifier, decision trees and SVM.

For training purposes, we used dataset DS1, and for testing purposes we used datasets DS3 and DST (see Section 7.2 for details). The feature vector representing each sperm head in training and testing datasets consisted in the concatenation of all descriptors (Morphological, Fourier, Geometric moments, Zernike moments, Convexity and Ellipticity measures).

B.1 Monolithic multi-class classifiers

- Training procedure:
  1. To standardize data (mean=0 and std=1 for each feature)
  2. To balance training data by randomly taking the same number of samples in each class
  3. To train the base classifier with balanced training data
- Testing procedure: For each feature vector in testing dataset, the predicted class from the base classifier will be the prediction of the whole scheme
- Classification results: are shown in Table B.1 using DS3 for testing purposes and Table B.2 using DST for testing purposes.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.43	0.55	0.39	0.43	0.18	0.40
Bayes	0.32	0.66	0.42	0.80	0.17	0.47
dTree	0.47	0.51	0.46	0.37	0.24	0.41
SVM	0.23	0.77	0.20	0.69	0.23	0.42

Table B.1: **Results of monolithic multi-class classifiers using testing dataset DS3.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.65	0.65	0.86	0.82	0.25	0.65
Bayes	0.60	0.68	0.77	1.00	0.10	0.63
dTree	0.63	0.67	0.93	0.46	0.29	0.59
SVM	0.68	0.80	0.73	0.78	0.26	0.65

Table B.2: **Results of monolithic multi-class classifiers using testing dataset DST.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

B.2 First cascade of two-class classifiers

- Training procedure:
  1. To standardize data (mean=0 and std=1 for each feature)
  2. To balance training data by randomly taking the same number of samples in each class
  3. To train five base classifiers with balanced training data as follows:
     - C1: One classifier to separate normal from abnormal sperm heads
     - C2: One classifier to separate tapered from other abnormal sperm heads
     - C3: One classifier to separate pyriform from other abnormal sperm heads
     - C4: One classifier to separate amorphous from other abnormal sperm heads
     - C5: One classifier to separate small from other abnormal sperm heads
- Testing procedure: For each feature vector fv in testing dataset
  1. $y1 \leftarrow$ predicted class after testing classifier C1 with input fv
  2. If $y1 = abnormal$, to test with the other four classifiers
     (a) $[y2, score2] \leftarrow$ predicted class and score after testing classifier C2 with input fv
     (b) $[y3, score3] \leftarrow$ predicted class and score after testing classifier C3 with input fv
     (c) $[y4, score4] \leftarrow$ predicted class and score after testing classifier C4 with input fv
     (d) $[y5, score5] \leftarrow$ predicted class and score after testing classifier C5 with input fv
     (e) To choose the predicted class of the whole scheme from $y2, y3, y4$, and $y5$ with the maximum score
  3. Otherwise, the predicted class of the whole scheme will be *normal*.
- Classification results: are showen in Table B.3 using DS3 for testing purposes and Table B.4 using DST for testing purposes.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.81	0.58	0.14	0.23	0.08	0.37
Bayes	0.57	0.62	0.39	0.56	0.22	0.47
dTree	0.64	0.33	0.29	0.44	0.21	0.38
SVM	0.68	0.33	0.14	0.29	0.55	0.39

Table B.3: **Results of first cascade of two-class classifiers using testing dataset DS3.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.90	0.73	0.21	0.34	0.12	0.46
Bayes	0.97	0.73	0.57	0.51	0.06	0.57
dTree	0.57	0.63	0.44	0.31	0.35	0.46
SVM	0.90	0.73	0.54	0.48	0.71	0.67

Table B.4: **Results of first cascade of two-class classifiers using testing dataset DST.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

B.3 Second cascade of two-class classifiers

- Training procedure:
  1. To standardize data (mean=0 and std=1 for each feature)
  2. To balance training data by randomly taking the same number of samples in each class
  3. To train five base classifiers with balanced training data as follows:
     - C1: One classifier to separate normal from abnormal sperm heads
     - C2: One classifier to separate small from other abnormal sperm heads
     - C3: One classifier to separate tapered from pyriform sperm heads
     - C4: One classifier to separate tapered from amorphous sperm heads
     - C5: One classifier to separate pyriform from amorphous sperm heads
- Testing procedure: For each feature vector fv in testing dataset
  1. $y1 \leftarrow$ predicted class after testing classifier C1 with input fv
  2. If $y1 =$*abnormal*, to test with classifier C2
     (a) $y2 \leftarrow$ predicted class after testing classifier C2 with input fv
     (b) If $y2 =$*small*, to test with the other three classifiers
        i. $[y3, score3] \leftarrow$ predicted class and score after testing classifier C3 with input fv
        ii. $[y4, score4] \leftarrow$ predicted class and score after testing classifier C4 with input fv
        iii. $[y5, score5] \leftarrow$ predicted class and score after testing classifier C5 with input fv
        iv. To choose the predicted class of the whole scheme from $y3, y4$, and $y5$ with the maximum score
     (c) Otherwise, the predicted class of the whole scheme will be *small*.
  3. Otherwise, the predicted class of the whole scheme will be *normal*.

- Classification results: showed in Table B.5 using DS3 for testing purposes and Table B.6 using DST for testing purposes.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.81	0.56	0.29	0.38	0.00	0.41
Bayes	0.62	0.60	0.60	0.51	0.07	0.48
dTree	0.79	0.24	0.24	0.41	0.16	0.37
SVM	0.70	0.54	0.17	0.49	0.42	0.47

Table B.5: **Results of second cascade of two-class classifiers using testing dataset DS3.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.89	0.68	0.54	0.46	0.00	0.51
Bayes	0.97	0.71	0.51	0.45	0.03	0.53
dTree	0.75	0.45	0.34	0.48	0.17	0.44
SVM	0.90	0.82	0.59	0.64	0.48	0.68

Table B.6: **Results of second cascade of two-class classifiers using testing dataset DST.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

B.4 Third cascade of two-class classifiers

- Training procedure:
  1. To standardize data (mean=0 and std=1 for each feature)
  2. To balance training data by randomly taking the same number of samples in each class
  3. To train four base classifiers with balanced training data as follows:
     - C1: One classifier to separate small from other sperm heads
     - C2: One classifier to separate normal from abnormal sperm heads (different from small sperm heads)
     - C3: One classifier to separate pyriform from tapered and amorphous sperm heads

  - C4: One classifier to separate tapered from amorphous sperm heads
- Testing procedure: For each feature vector fv in testing dataset
  1. $y1 \leftarrow$ predicted class after testing classifier C1 with input fv
  2. If $y1$ =*small*, to test with classifier C2
     (a) $y2 \leftarrow$ predicted class after testing classifier C2 with input fv
     (b) If $y2$ =*normal*, to test with classifier C3
        i. $y3 \leftarrow$ predicted class after testing classifier C3 with input fv
        ii. If $y3$ =*pyriform*, to test with classifier C4
           A. $y4 \leftarrow$ predicted class after testing classifier C4 with input fv
           B. If $y4$ =*tapered*, the predicted class of the whole scheme will be *tapered*.
           C. Otherwise, the predicted class of the whole scheme will be *amorphous*.
        iii. Otherwise, the predicted class of the whole scheme will be *pyriform*.
     (c) Otherwise, the predicted class of the whole scheme will be *normal*.
  3. Otherwise, the predicted class of the whole scheme will be *small*.
- Classification results: showed in Table B.7 using DS3 for testing purposes and Table B.8 using DST for testing purposes.

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.27	0.35	0.28	0.73	0.08	0.34
Bayes	0.12	0.42	0.55	0.89	0.12	0.42
dTree	0.30	0.34	0.28	0.83	0.11	0.37
SVM	0.30	0.47	0.26	0.68	0.42	0.42

Table B.7: **Results of third cascade of two-class classifiers using testing dataset DS3.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

B.5 Justification on choosing *the* classification scheme

There are three basic reasons that justify the choice of the architecture of the classification scheme proposed in Section 7.1. We will disscuss those reasons in the remainder of this appendix.

1. **With respect to the emphasis on amorphous sperm heads**. From the variability analysis presented in Chapter 4, a valid conclusion is that the class of amorphous sperm heads shows a high degree of discrepancy among all experts. A natural consequence is the notorious difficulty in the classification of sperm heads of this class. Formally,

Classifier	tpr(N)	tpr(T)	tpr(P)	tpr(S)	tpr(A)	acc
1−NN	0.46	0.46	0.53	0.87	0.13	0.49
Bayes	0.25	0.50	0.76	1.00	0.05	0.51
dTree	0.42	0.52	0.61	0.78	0.23	0.51
SVM	0.48	0.70	0.64	0.76	0.48	0.61

Table B.8: **Results of third cascade of two-class classifiers using testing dataset DST.** *tpr* means True Positive Rate, *N* means Normal, *T* means Tapered, *P* means Pyriform, *S* means Small, *A* means Amorphous, and *acc* means mean accuracy.

a way to measure the complexity of a dataset is by means of using a monolitic 1NN classifier. Looking at Tables B.1 and B.2, we can confirm that class *amorphous* is, in fact, the most complex class, achieving only 25% of classification accuracy tested with a total expert-agreement dataset, against at least 65% of classification accuracy of the remaining classes. The nature of *noisy class* observed in class Amorphous, motivated us to focus on the whole classification scheme in the identification of sperm heads of this particular class as a main goal. In this sense, the first stage of our proposed scheme acts as a filter for amorphous sperm heads, as well as a prior four-class classifier. The second stage acts only as a confirmation step of the potential class versus amorphous class.

2. **With respect to the usage of SVM as base classifier**. Among different valid alternatives of choosing the base classifier for our final proposed scheme, we chose to use SVM according to the experimental evaluation presented before in this appendix. For the decision making, we considered three parameters for each base classifier: mean accuracy, variance of accuracy per class and standard deviation of accuracy per class. These evaluation metrics were also regarded using both testing datasets (DS3 and DST). By looking at Tables from B.1 to B.8, one can observe that the base classifier with the best compromise of the highest mean accuracy and the lowest variance and standard deviation is SVM.
3. **With respect to the composition of feature vector**. All the classification schemes presented in this appendix were tested using a concatenation of all descriptors as feature vector. None of the discarded classification schemes presented in this appendix reached classification accuracy rates correlated with the complexity of the data. As mentioned before, this complexity can be reliably evaluated by means of using a 1NN classifier (see Tables B.1 and B.2). According to this, the classes with the highest accuracy rate are expected to be *pyriform* and *small*, followed by *normal* and *tapered*, while the expected class with lowest accuracy rate will be *amorphous*. In this sense, we suspected that this discrepancy between dataset complexity and classification accuracy rates is due to the curse of dimensionality. Following this concern, we decided to select a subset of specialized features for each class in each stage of our proposed classification scheme, as explained in Section 7.1.

As a tool to confirm that our decision was sound in proposing the classification scheme introduced in Section 7.1, we present a comparison of the rates of classification accuracy per class achieved by the four discarded classification schemes and the proposed classification scheme in Figures B.1 and B.2. The achieved rates of accuracy per class using the proposed classification scheme (see Section 7.2) are fully consistent with the complexity of the data in each class. In addition, the proposed classification scheme outperformed the four discarded classification schemes presented in this appendix.

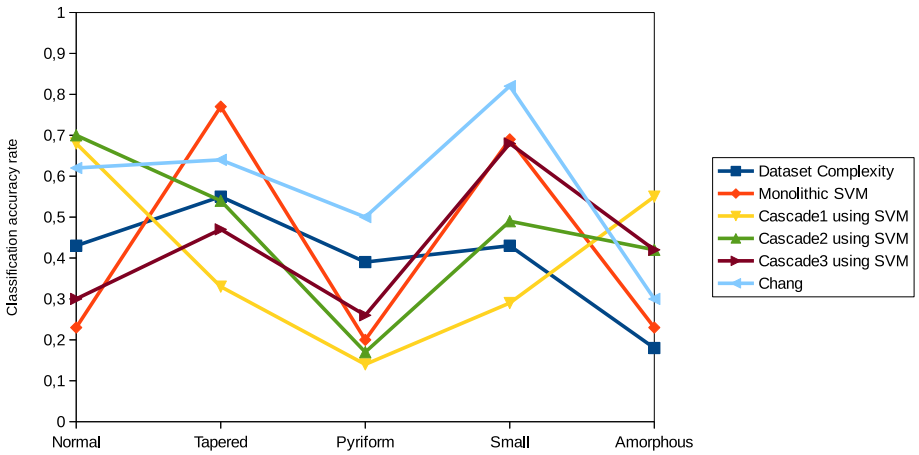


Figure B.1: **Comparison of accuracy per class using DS3 with different classification schemes.** We compare the four classification schemes presented in this appendix (Monolitic SVM, Cascade1, Cascade2 and Cascade3) against the classification scheme proposed in Section 7.1 (Chang) and the complexity of the dataset DS3.

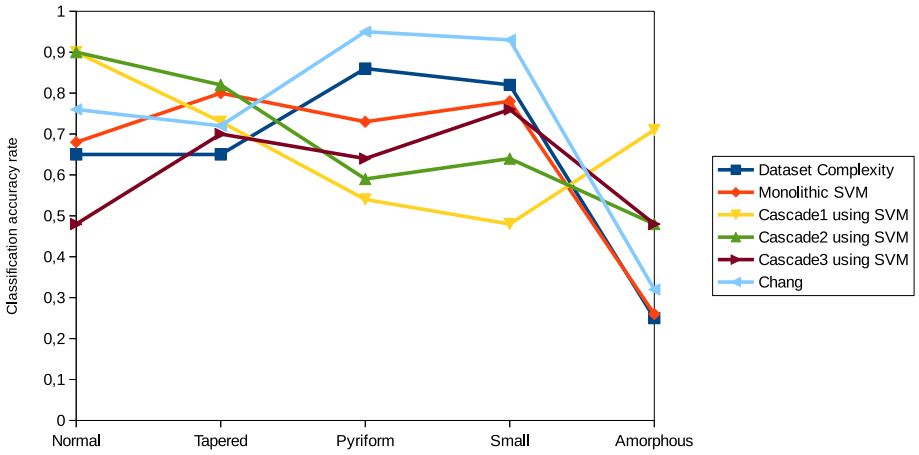


Figure B.2: **Comparison of accuracy per class using DST with different classification schemes.** We compare the four classification schemes presented in this appendix (Monolitic SVM, Cascade1, Cascade2 and Cascade3) against the classification scheme proposed in Section 7.1 (Chang) and the complexity of the dataset DST.