Ανάλυση Πολυκαναλικών Εικόνων Χρωμοσωμάτων

Η
ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από τον

Πέτρο Καρβέλη

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Νοέμβριος 2012

Τριμελής Συμβουλευτική Επιτροπή

Αριστείδης Λύκας, Αναπληρωτής Καθηγητής Τμήματος Πληροφορικής του Πανεπιστήμιου Ιωαννίνων (Επιβλέπων).

Χριστόφορος Νίκου, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Πανεπιστήμιου Ιωαννίνων.

Δημήτριος Φωτιάδης, Καθηγητής Τμήματος Μηχανικών Επιστήμης Υλικών του Πανεπιστημίου Ιωαννίνων.

Επταμελής Συμβουλευτική Επιτροπή

Αριστείδης Λύκας, Αναπληρωτής Καθηγητής Τμήματος Πληροφορικής του Πανεπιστήμιου Ιωαννίνων (Επιβλέπων).

Χριστόφορος Νίκου, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Πανεπιστήμιου Ιωαννίνων.

Δημήτριος Φωτιάδης, Καθηγητής Τμήματος Μηχανικών Επιστήμης Υλικών του Πανεπιστημίου Ιωαννίνων.

Κωνσταντίνος Μπλέκας, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Πανεπιστήμιου Ιωαννίνων.

Γεώργιος Μανής, Επίκουρος Καθηγητής Τμήματος Πληροφορικής του Πανεπιστήμιου Ιωαννίνων.

Κωνσταντίνα Νικήτα, Καθηγήτρια του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου.

Ανδρέας Σταφυλοπάτης, Καθηγητής του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου.

DEDICATION

To my Wife Argiro.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Petros Karvelis, S. M.

PhD, Department of Computer Science, University of Ioannina, Greece, 2012.

Title: Multichannel Chromosome Image Analysis

Supervisor: Prof. Aristidis Likas.

The study of chromosomes is one of the major areas of study for modern genetics because the chromosomes are the carriers of all the genetic material (DNA) of an organism that are transferred from generation to generation by means of reproduction. The assignment of each chromosome to each class from a chromosome image takes time and demands great experience to avoid mistakes that can lead to misdiagnosis. For this reason there have been developed algorithms for image processing and automated analysis of chromosomes.

There are several methods and techniques for the cultivation of chromosomes each of which leads to a different type of image. For example, if the chromosomes are cultured according to the protocol G-Banding the resulting image is a gray level image. In this thesis we deal with M-FISH protocol which leads to a multichannel image (6 channels). In this technique the biological experiment has been constructed so that each of the 24 chromosome types (1-22, X, Y) would be reflected in a different color.

The purpose of this thesis is the identification and classification of human chromosomes from multichannel M-FISH images. Initially, we developed a method based on the Watershed transform for the region segmentation (grouping pixels with similar characteristics) of chromosomes. The Watershed transform requires a measure of separability between similar areas and for this reason we chose to calculate the multichannel gradient. In this way we achieve a clear separation between areas with different color corresponding to a different chromosome class. The segmentation results are quite satisfactory compared to other methods reported in the literature on the same M-FISH basis images.

After segmentation we perform region classification using a statistical classifier that employs the Bayes rule. This classifier is simple to develop and implement and provides satisfactory classification performance. Compared with existing approaches that use Pixel by Pixel classification the proposed region-based method showed better results. We also study the effectiveness image smoothing using Vector Median Filtering and its variants and provide comparative experimental results.

One of the problems in the chromosome classification methodologies using multi-channel M-FISH images is the fact that they demand a labeled training set to build the classifier. For example a Bayes classifier requires estimating the parameters such as mean and covariance for each of the 24 chromosomes classes. The existence of a methodology that does not require a labeled training set is therefore essential. Such an unsupervised methodology is presented in this thesis. First, we segment the M-FISH image using the Watershed transform to remove the background. Then we estimate which of the remaining pixels have been hybridized or not using the EM algorithm in each of the 5 channels of the image. Then we use a Gaussian Mixture Model to classify each pixel into one of the 24 classes of chromosomes. To build this model for the first time we exploit prior information about which chromosome class emits to each of the five channels. The adaptation of the parameters of Gaussian Mixture Model by using the Maximum a Posterior Expectation Maximization method (MAP EM) results in an increase in the rate of correct classification. It is noteworthy that the proposed unsupervised methodology achieves higher classification rates when compared to supervised classification methodologies.

One of the problems for automatic chromosome segmentation methods is the problem of the occlusion of chromosomes. In particular, two important factors influencing the segmentation are the following:

• Chromosomes that overlap with one another,

• Chromosomes which adjoin one another.

We have developed a method that addresses both these problems successfully. Initially we apply a recursive Watershed transform to get an initial assessment of areas of chromosomes. Then for each area of the Watershed transform we determine high curvature points around the perimeter of the chromosomal region. From these points will begin a gradient path which crosses the region and separates the chromosome region where two chromosomes are tangent to each other. If two or more chromosomes overlap each other, then the path splits the chromosomes into pieces. Then we form the Region Adjacency Graph and

categorize each area using a region Bayes classifier. If a pair of neighboring regions shares the same class then they are joined together. The method was tested on chromosome images and the success rate of the method was satisfactory. In addition we compared the method with other segmentation methodologies such as the pale paths and the results were much better, especially for the case of overlapping chromosomes.

ΕΚΤΕΝΗΣ ΠΕΡΙΛΗΨΗ ΣΤΑ ΕΛΛΗΝΙΚΑ

Πέτρος Καρβέλης του Σταύρου και της Μαρίας.

Διδακτορικό Δίπλωμα, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, 2012.

Τίτλος: Ανάλυση Πολυκαναλικών Εικόνων Χρωμοσωμάτων

Επιβλέποντας: Αριστείδης Λύκας.

Η μελέτη των χρωμοσωμάτων αποτελεί έναν από τους σημαντικότερους τομείς μελέτης για τη σύγχρονη Γενετική διότι τα χρωμοσώματα αποτελούν τους φορείς όλου του γενετικού υλικού (DNA) ενός οργανισμού που μεταβιβάζονται από γενιά σε γενιά με την βοήθεια της αναπαραγωγής. Τα χρωμοσώματα ανήκουν σε κατηγορίες και η ανάθεση κάθε χρωμοσώματος στην κατηγορία του από μια εικόνα χρωμοσωμάτων απαιτεί χρόνο αλλά και μεγάλη εμπειρία για την αποφυγή λαθών που μπορούν να οδηγήσουν σε εσφαλμένη διάγνωση. Για το λόγο αυτό αναπτύχθηκαν τεχνικές για την επεξεργασία και ανάλυση εικόνων χρωμοσωμάτων και τον αυτόματο χαρακτηρισμό τους.

Υπάρχουν αρκετοί τρόποι και τεχνικές για την καλλιέργεια των χρωμοσωμάτων η κάθε μια από τις οποίες οδηγεί και σε διαφορετικό τύπο εικόνας. Για παράδειγμα αν τα χρωμοσώματα καλλιεργηθούν σύμφωνα με το πρωτόκολλο G-Banding η εικόνα που προκύπτει είναι μια γκρι (grey-scale) εικόνα. Στην παρούσα διατριβή ασχολούμαστε με το πρωτόκολλο M-FISH το οποίο οδηγεί σε μια πολυκαναλική εικόνα (6 καναλιών). Στην τεχνική αυτή το βιολογικό πείραμα έχει κατασκευαστεί έτσι ώστε η κάθε μια από τις 24 κατηγορίες χρωμοσωμάτων (1-22,X,Y) να αποτυπώνεται με διαφορετικό χρώμα.

Αντικείμενο της παρούσας διατριβής είναι η αναγνώριση και κατηγοριοποίηση των ανθρωπίνων χρωμοσωμάτων από πολυκαναλικές εικόνες M-FISH. Αρχικά αναπτύξαμε μια μέθοδο βασισμένη στον μετασχηματισμό Watershed για την κατάτμηση (ομαδοποίηση εικονοστοιχείων με παρόμοια χαρακτηριστικά) των χρωμοσωμάτων σε περιοχές. Ο μετασχηματισμός Watershed απαιτεί ένα μέτρο διαχωρισημότητας μεταξύ όμοιων περιοχών και γι' αυτό το λόγο επιλέξαμε τον υπολογισμό της πολυκαναλικής παραγώγου. Με αυτό τον

τρόπο εξασφαλίζεται ο σαφής διαχωρισμός μεταξύ περιοχών με διαφορετικό χρώμα και επιπλέον αφαιρούμε το υπόβαθρο (background). Τα αποτελέσματα της κατάτμησης είναι αρκετά ικανοποιητικά συγκρινόμενα με αυτά της βιβλιογραφίας για την ίδια βάση εικόνων M-FISH.

Για την ταξινόμηση κάθε περιοχής χρησιμοποιήσαμε έναν στατιστικό ταξινομητή βασισμένο στον κανόνα του Bayes. Ο ταξινομητής αυτός είναι απλός στην υλοποίηση του και έχει χρησιμοποιηθεί και σε άλλες μελέτες. Σε σύγκριση με ήδη υπάρχουσες μεθοδολογίες οι οποίες χρησιμοποιούν Pixel by Pixel κατηγοριοποίηση η μέθοδος μας (που βασίζεται σε ταξινόμηση περιοχών) εμφάνισε καλύτερα αποτελέσματα. Τέλος, μελετούμε την αποτελεσματικότητα των φίλτρων διανυσματικού διαμέσου (Vector Median Filtering) και παραλλαγών του εάν εφαρμοστούν στην εικόνα πριν την ταξινόμηση. Η αποτελεσματικότητα των φίλτρων Διαμέσου εξετάζεται συγκρίνοντας το ποσοστό σωστής ταξινόμησης πριν και μετά την χρήση των φίλτρων αυτών.

Ένα από τα προβλήματα που εμφανίζουν όλες οι μεθοδολογίες κατηγοριοποίησης χρωμοσωμάτων από πολυκαναλικές εικόνες M-FISH είναι η προϋπόθεση ύπαρξης ενός συνόλου εκπαίδευσης για την εκπαίδευση του ταξινομητή. Στην περίπτωση για παράδειγμα ενός ταξινομητή Bayes απαιτείται η εκτίμηση των παραμέτρων όπως της μέσης τιμής και του πίνακα συμμεταβλητότητας για κάθε μια από τις 24 κατηγορίες χρωμοσωμάτων. Η ύπαρξη μιας μεθοδολογίας που θα είναι ανεξάρτητη από το σύνολο εκπαίδευσης που επιλέγουμε έχει σημαντική αξία. Μια τέτοια μεθοδολογία παρουσιάζεται στην διατριβή αυτή. Αρχικά λαμβάνουμε μια κατάτμηση της εικόνας M-FISH με την χρήση της μεθοδολογίας Watershed (απομακρύνοντας το υπόβαθρο) και κατόπιν εκτιμούμε ποια από τα εικονοστοιχεία έχουν υβριδοποιηθεί ή όχι με την χρήση του αλγορίθμου EM σε κάθε ένα από τα 5 κανάλια της εικόνας μας. Στην συνέχεια χρησιμοποιούμε ένα πολυκαναλικό Gaussian Mixture Model για την κατηγοριοποίηση κάθε εικονοστοιχείου σε μια από τις 24 κατηγορίες χρωμοσωμάτων. Στο μοντέλο αυτό χρησιμοποιείται για πρώτη φορά εκ των προτέρων πληροφορία σχετικά με το σε ποιο κανάλι εκπέμπει κάθε κατηγορία χρωμοσώματος. Η περαιτέρω εκπαίδευση των παραμέτρων του Gaussian Mixture Model από τον αλγόριθμο Maximum A Posterior Expectation Maximization (MAP EM) επιτρέπει την αύξηση του ποσοστού σωστής κατηγοριοποίησης. Η μεθοδολογία αυτή επιτυγχάνει ακόμη καλύτερα ποσοστά συγκρινόμενη ακόμη με μεθοδολογίες ταξινόμησης με επίβλεψη.

Δύο σημαντικοί παράγοντες επηρεάζουν την κατάτμηση των εικόνων χρωμοσωμάτων και είναι οι εξής :

- Χρωμοσώματα που επικαλύπτουν το ένα το άλλο,

- Χρωμοσώματα που εφάπτονται το ένα στο άλλο.

Προτείνουμε μια μέθοδο που αντιμετωπίζει και τα δύο αυτά προβλήματα με επιτυχία. Αρχικά εφαρμόζουμε έναν επαναληπτικό μετασχηματισμό Watershed ώστε να πάρουμε μια αρχική εκτίμηση των περιοχών των χρωμοσωμάτων. Στην συνέχεια εντοπίζουμε σημεία υψηλής κύρτωσης πάνω στην περίμετρο της χρωμοσωματικής περιοχής. Ξεκινώντας από τα σημεία αυτά, θα δημιουργούμε ένα μονοπάτι παραγώγου (Gradient Path) το οποίο διασχίζει την περιοχή χρωμοσώματος και διαχωρίζει την περιοχή όταν δύο χρωμοσώματα εφάπτονται το ένα στο άλλο. Αν δυο ή και περισσότερα χρωμοσώματα επικαλύπτονται τότε το μονοπάτι διαχωρίζει σε δύο τμήματα την περιοχή των χρωμοσωμάτων. Στην συνέχεια σχηματίζουμε τον γράφο γειτνίασης περιοχών (Region Adjacency Graph) και κατηγοριοποιούμε κάθε περιοχή κάνοντας χρήση ενός ταξινομητή περιοχών Bayes. Για κάθε ζεύγος γειτονικών περιοχών που έχουν την ίδια κατηγορία ενώνουμε τις δύο αυτές περιοχές. Συγκριτικά αποτελέσματα με άλλες μεθόδους δείχνουν την ανωτερότητα της μεθόδου ιδιαίτερα στην περίπτωση των επικαλυπτόμενων χρωμοσωμάτων.

CHAPTER 1:

INTRODUCTION

1.1   Multichannel Chromosome Images

Chromosomes are the structures in cells that contain genetic information [1]. The study of chromosomes is made possible by staining techniques since chromosomes are colorless. Usually, a dye is applied during cell division by a solution of colchicine. Thus the chromosomes are colored and can be captured by a microscope producing an image. The images of these chromosomes contain significant information about the health of an individual. Chromosome images are useful for diagnosing genetic disorders and for studying various diseases, such as cancer.

Normally, the procedure of assigning each chromosome to a class (karyotyping) is based on the visual scanning of chromosome images by experts (biologists, cytogeneticists). This manual process of locating, classifying, and evaluating the chromosomes in these images could be lengthy and tedious. Thus a need for automated methods that could classify each chromosome to each class naturally arises.

The whole process can be divided in two main stages:

a) Segmentation: grouping the pixels of the multichannel chromosome image into two different classes; the chromosomes and the background. This is not a trivial task since chromosomes are not always clearly visible in the chromosome image due a number of parameters of the biological experiment [2].

b) Classification: classifying each chromosome to 1-24 classes (1,…,22,X,Y). Usually such methods extract features for each chromosome of the image such as the length, color, geometry etc. and a classifier assigns them into one of the 24 chromosome classes.

   However, if the classification method requires a training set (supervised classification) there is no guarantee that the set would be a representative one containing sufficient training pixels of each class [2].

   Furthermore, for the unsupervised type of classification, two main problems have to be overcome. First the number of classes ($K = 24$) is large and second how to define a proper initialization of the clusters which represent each chromosome class.

Another major problem for chromosome images is the fact that chromosomes could overlap or at least partially occlude each other. Thus segmentation and classification in the overlap areas could fail unless a dedicated procedure for this type of problem is used.

Next, we introduce the multichannel chromosome images and provide specific details for the procedure concerning the production these types of images. Different staining techniques allow analysis of different kinds of abnormalities. Table 1.1 presents the most popular chromosome imaging staining techniques. In the mid-1990's, a new technique for staining chromosomes was introduced. It produced an image in which each chromosome type appears to have a distinct color [3], [4]. This multi-spectral staining technique made the analysis of chromosome images easier, not only for visual inspection, but also for computer analysis of the images. This multichannel staining technique is called M-FISH (Multiplex Fluorescence In-Situ Hybridization). M-FISH uses five color dyes that attach to various chromosomes differently to produce a multichannel image. Also a DNA stain called DAPI (4', 6-Diamidino-2-phenylindole) attaches to DNA and thus labels all chromosomes. An example of an M-FISH image is shown in Figure 1.1.

(a) (b)

(c) (d)

(e) (f)

(g)
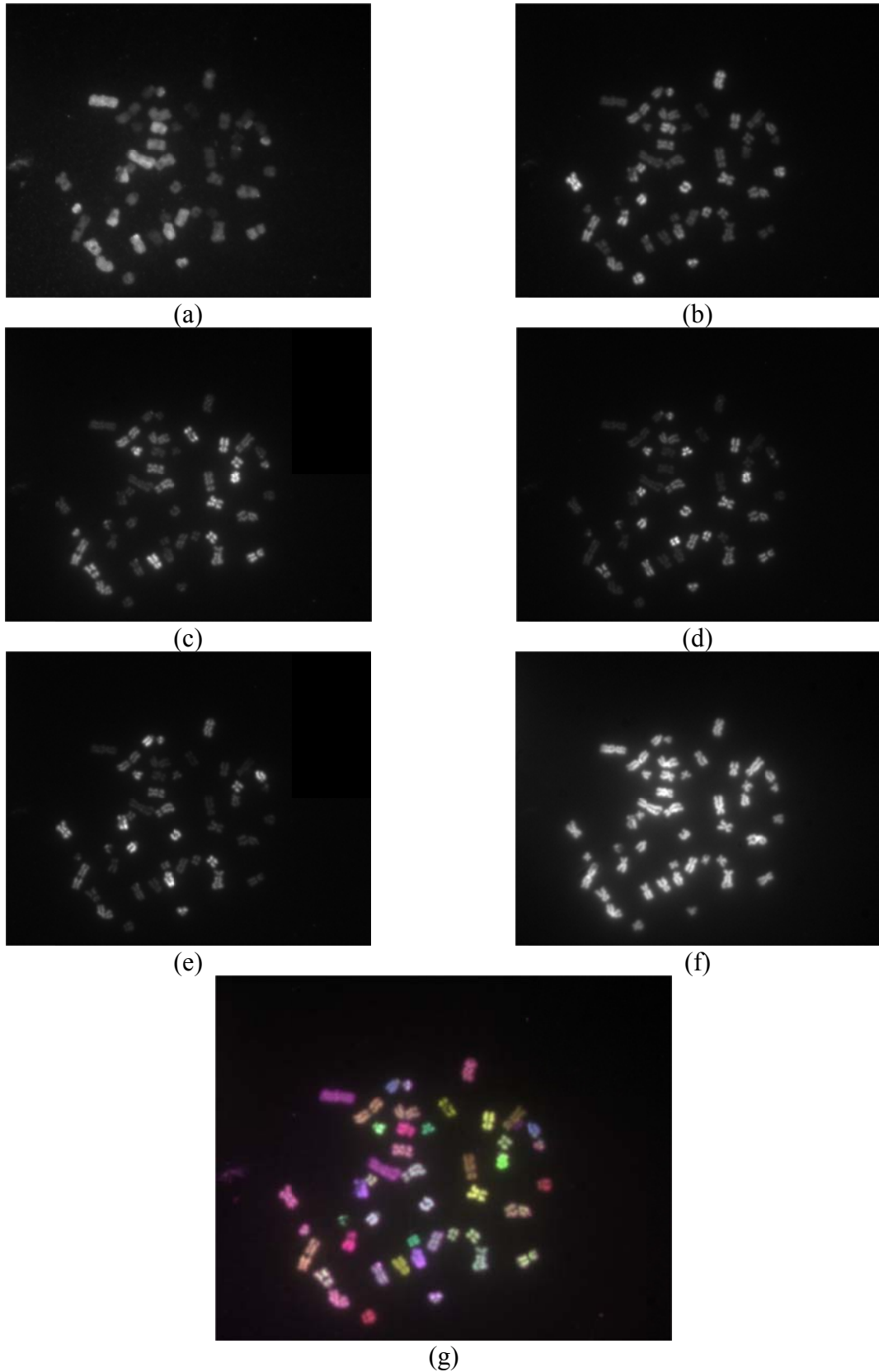
Figure 1.1: An M-FISH image with its channels. (a) Channel 1, (b) Channel 2, (c) Channel 3, (d) Channel 4, (e) Channel 5, (f) Channel 6.

3

Table 1.1: Different chromosome staining techniques.

Researcher		Method	Year	Types of Images Produced
Arrighi *et al.*	[5]	C-banding	1971	Greyscale
Sumner *et al.*	[6]	G-banding	1971	Greyscale
Bauman *et al.*	[7]	FISH	1980	Multicolor
Speicher *et al.*	[3]	M-FISH	1996	Multicolor
Schröck *et al.*	[4]	SKY	1996	Multicolor

The combinations of the five fluorophores that are used to label each class of chromosomes are shown in the Appendix A for three different M-FISH fluorophore sets. However, these tables are somewhat of an oversimplification because, in practice, fluorophore absorption is hardly binary. Table 1.2 shows the actual mean values of pixels of each class from a real set of M-FISH images.

This new imaging technique introduces several advantages:

- Chromosome classification is simplified [3], [8]. Only the spectral information from the multispectral image is used and no features such as length, centromere position and band pattern are used.

- Subtle chromosomal aberrations are detected [9]. Traditional monochrome imaging techniques failed to detect rearrangements of genetic material such as the translocation of telemetric chromatin, because it is difficult to detect them with banding alone. M-FISH (color karyotyping) is able to sufficiently depict these anomalies.

- It can be used for the identification of small genetic markers that remain elusive after banding [10].

- Chromosome aberrations are more easily detected in M-FISH. For a normal cell, all the pixels in each chromosome should be represented with one identical color. However, for a cancerous cell, different colors might show up in a chromosome as a result of the chromosomal rearrangements or the exchange of DNA material between chromosomes. An example of two chromosome abnormalities using grayscale and M-FISH technique is shown in Figure 1.2. One might be confused searching for the abnormality of the extra DNA material of chromosome 6 attached to chromosome 9 Figure 1.2(a). In Figure 1.2(b) the extra DNA material of chromosome 9 attached to

4

chromosome 14 is shown with different colors on the M-FISH image. By looking at the M-FISH image, even a non-expert can easily determine the chromosome abnormalities by searching for chromosomes with two colors.

Table 1.2: Average fluorophore magnitude for each class from a real set of M-FISH images. Bold denotes the classes to which each fluorophore is predicted to bind.

Chromosome Class	M-FISH Channels				
	Spectrum Green	Spectrum Orange	Texas Red	Cy5	Cy5.5
1	**0.5483**	0.2946	**0.4928**	**0.5171**	0.2554
2	0.4681	0.3596	0.4117	0.4311	**0.4978**
3	**0.5059**	**0.4549**	0.3350	**0.5200**	0.3434
4	**0.5852**	0.3447	0.3882	**0.5119**	0.3041
5	**0.5372**	**0.4523**	**0.5299**	0.3077	**0.3114**
6	**0.5390**	0.2577	**0.5019**	**0.4823**	**0.3469**
7	0.3244	0.2560	**0.5794**	**0.6313**	0.2453
8	**0.8027**	0.2842	0.3140	0.3034	0.2304
9	**0.6379**	**0.4764**	0.3160	0.2863	**0.3796**
10	0.3563	0.2809	0.3219	**0.6858**	**0.4257**
11	**0.5913**	**0.4987**	0.2877	**0.4994**	0.2066
12	0.3127	0.2479	**0.7338**	0.3297	**0.3945**
13	**0.6590**	**0.5828**	0.3083	0.2367	0.1849
14	0.3266	0.2946	**0.7695**	0.3396	0.2279
15	0.2590	**0.5066**	**0.6101**	**0.4857**	0.1936
16	**0.6752**	0.2025	**0.6194**	0.2544	0.1698
17	0.3739	0.2928	0.3339	**0.7386**	0.2823
18	**0.6085**	**0.5011**	**0.5353**	0.2151	0.1576
19	0.2917	**0.6466**	0.3369	**0.5539**	0.2019
20	0.2746	**0.8125**	0.3551	0.2596	0.1988
21	**0.5994**	0.3411	0.3636	0.3547	**0.4403**
22	0.2603	**0.4837**	**0.5860**	**0.4697**	**0.3041**
X	0.4014	**0.5829**	0.3966	0.3793	**0.3913**
Y	**0.6486**	0.2267	0.2274	**0.5492**	**0.3632**

(a)	(b)
Case A: translocation between chromosomes 6 and 9.	**Case B**: translocation between chromosomes 9 and 14.
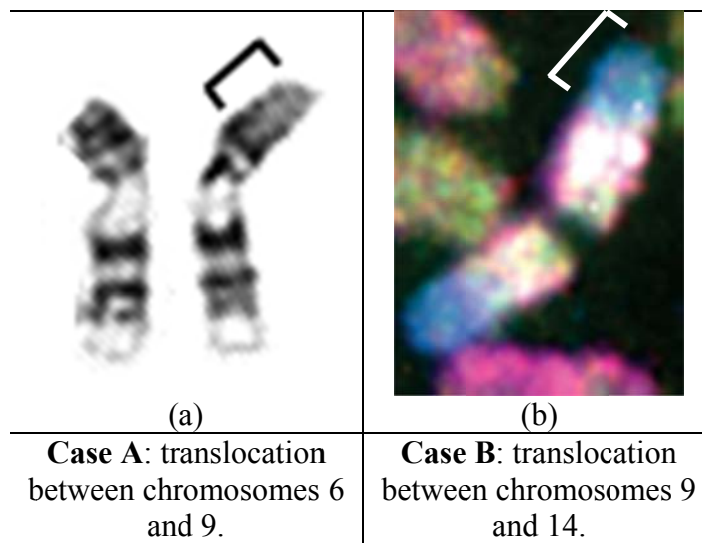
Figure 1.2: Chromosome abnormalities depicted using two different staining techniques: (a) Giemsa grayscale banding, and (b) M-FISH image.

Finally, chromosome analysis is the procedure for detecting genetic abnormalities in cells. Traditionally, cells are classified according to their karyotype, which is a tabular array where the chromosomes are aligned in pairs. Karyotyping is a useful tool to detect deviations from normal cell structure since abnormal cells may have an excess or a deficit of chromosomes [8]. Normal cells contain 46 chromosomes which consist of 22 pairs of similar, homologous chromosomes and two sex-determinative chromosomes (XY: male and XX: female). An example of a karyotype is shown in Figure 1.3.

(a)   (b)

Figure 1.3: (a) M-FISH chromosome image of a woman missing one chromosome from class 13, and (b) karyotype of the M-FISH image: 45 XX -13 (A female karyotype (XX) where one chromosome of class 13 is missing thus the number of the chromosomes are 45.).

1.2 M-FISH Image Database

The ADIR M-FISH chromosome image database consists of 200 multispectral images having dimension 517×645 pixels. In this database, there is no annotation for 17 images which are reported as "difficult to karyotype". These images were also extremely difficult to karyotype even by experienced cytogeneticists due to tightly packed chromosomes. The database contains five-channel image sets recorded at different wavelengths. In addition, a DAPI image file is included for each M-FISH image. The specimens were prepared with probe sets from four different laboratories:

- Applied Spectral Imaging (Migdal HaEmek, Israel),
- Advanced Digital Imaging Research (ADIR; League City, Texas, USA),
- Cytocell Technologies (Cambridge, UK), and
- Vysis (Downers Grove, IL, USA).

Three different probe sets (ASI, PSI and Vysis, Appendix A) were used in order to produce M-FISH images. Different wavelengths for each M-FISH image channel were used. The labeling charts for each chromosome are described presented Appendix A. Figure 1.4, presents the different wavelengths that were used for each probe set.

(a)

(b)

(c)
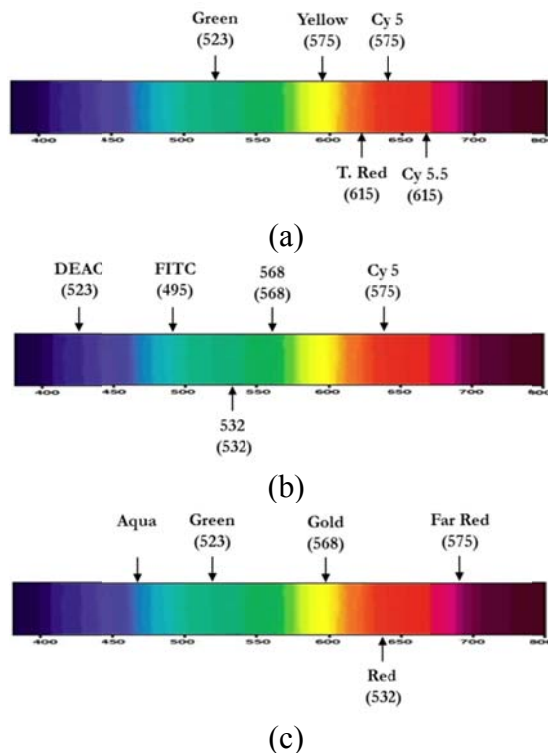
Figure 1.4: The different wavelengths used for each probe set. (a) ASI kit, (b) PSI kit, and (c) Vysis kit.

7

The ADIR dataset includes also a classification map, stored as an image file established by experienced cytogeneticists. This image is labeled so that the grey level of each pixel represents its class number (chromosome class). In addition, background pixels are labeled 0, and pixels in a region of overlap are labeled -1. This data file serves as "ground truth" to test the accuracy of the M-FISH image segmentation and classification methods. An example of the ground truth for an M-FISH image is shown in Figure 1.5(c).

(a)

(b)

(c)
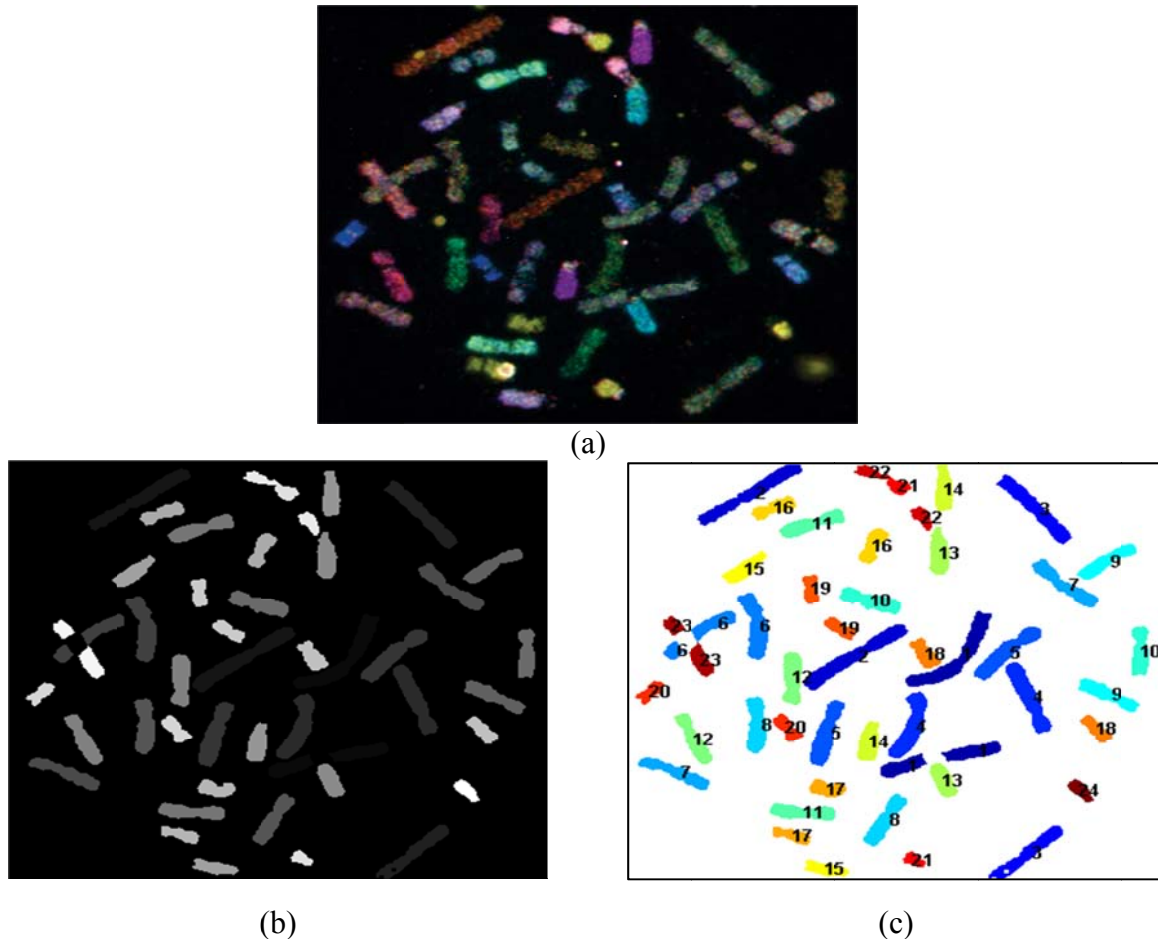
Figure 1.5: The annotation of each M-FISH image of the ADIR database. (a) The M-FISH image, (b) the greyscale annotated image, each pixel is represented with grey level intensity equal to the number of the class which it belongs to and (c) the color coded annotated image were each color represents a chromosome class.

1.3   M-FISH Image Segmentation and Classification

1.3.1   M-FISH Image Segmentation

Image segmentation is used to define a partition of image pixels into clusters (groups) that share similar characteristics. The common characteristic used as basis for the segmentation

could be a simple pixel property such as grey level or color. However, an image can also be segmented according to more complex characteristics such as texture features (for reviews of the subject see [11]).

One way of defining image segmentation is as follows [12]. Formally, a set of regions $\{R_1, R_2, \ldots, R_n\}$ is a segmentation of the image R into n regions if:

1. $\bigcup_{i=1}^{n} R_i = R$,

2. $R_i \cap R_k = \varnothing$, $i \neq k$,

3. R_i is connected, $i = 1, \ldots, n$,

Segmenting an M-FISH chromosome image into background and chromosome pixels is not a straightforward task since the image consists of six image channels where each chromosome class appears differently in each of the channels. The goal of a segmentation method is to develop an algorithm such that its input will be the M-FISH image and its output will be a set of regions where each region will contain the pixels of a chromosome. An example of a manual segmentation of an M-FISH image is shown below:
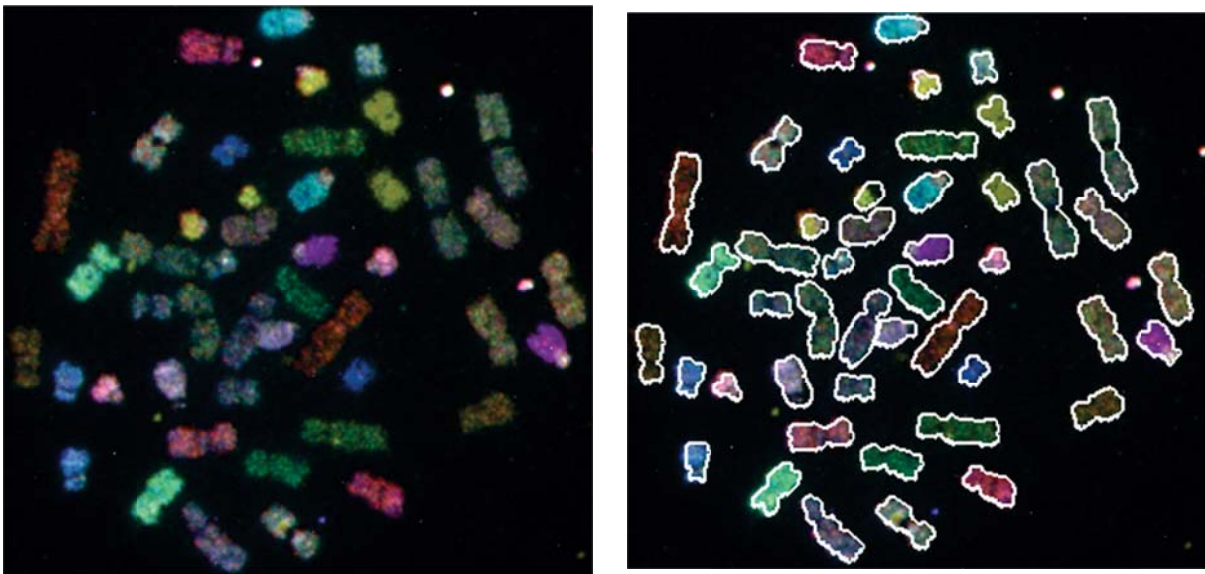
(a)   (b)

Figure 1.6: An M-FISH image and its segmentation. (a) The M-FISH image, and (b) Hand segmentation of the M-FISH image.

1.3.2 M-FISH Image Classification

Generally, after segmentation, the next step in chromosome image analysis is the classification of the segmented chromosomes. Classifying each segmented chromosome is the process of assigning each chromosome to one of 24 classes (1,2,…,X,Y). Figure 1.7, presents an M-FISH image and its manual classification map. Each chromosome is assigned to a class. Note that each image contains two chromosomes of the same class.

<table>
<tr><td>(a)</td><td>(b)</td></tr>
</table>
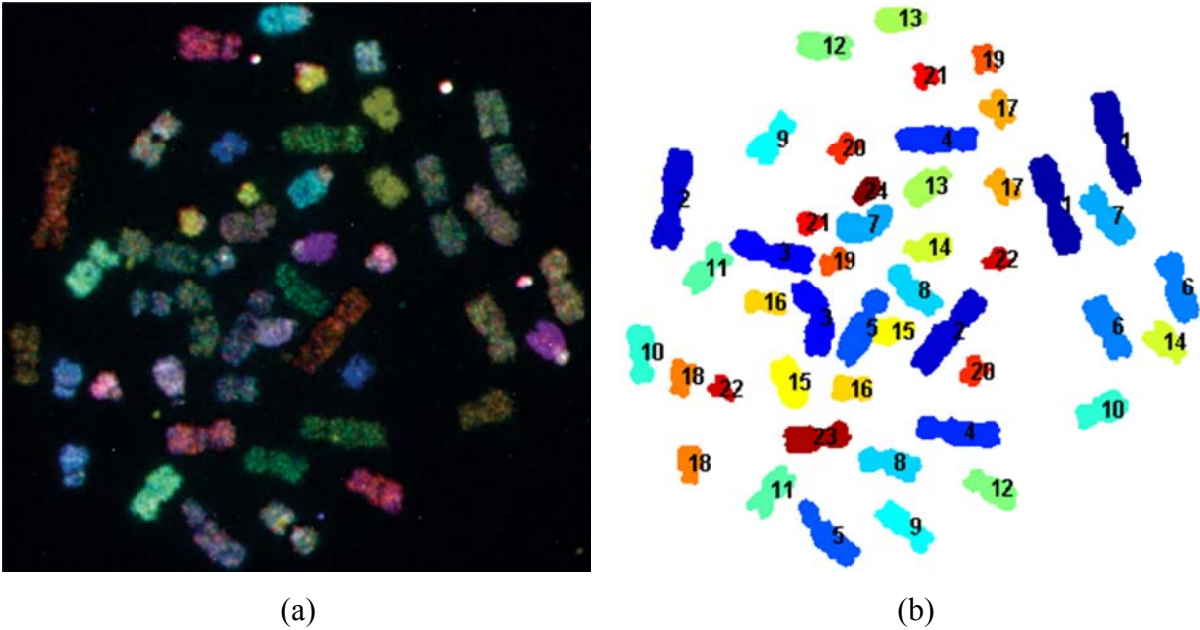
Figure 1.7: An M-FISH image and its classification. (a) The M-FISH image, (b) Manual classification of the chromosomes (chromosomes classes are printed for each chromosome).

1.4 Image Processing Algorithms

In the following we briefly describe the main image processing and analysis methods employed in this work.

1.4.1 Geodesic Dilation

A geodesic dilation [13] involves two images: a marker image f and a mask image g, with $f \leq g$, meaning that the image f is less than or equal to the image g if the value of f is less than or equal the value of g at all pixels p :

$$f \leq g \Leftrightarrow \forall p, f(p) \leq g(p).$$

(1.1)

Suppose a binary image with a pixel p having coordinates (i,j), its four neighbors are $N_4^p = \{(i+1,i),(i-1,i),(i,i+1),(i,i-1)\}$. The elementary dilation (Figure 1.8) of size 1 for the image f is defined as:

$$\delta^{(1)}(f) = \{q \mid q \in f\} \cup \{p \mid p \in N_4^q\}. \tag{1.2}$$
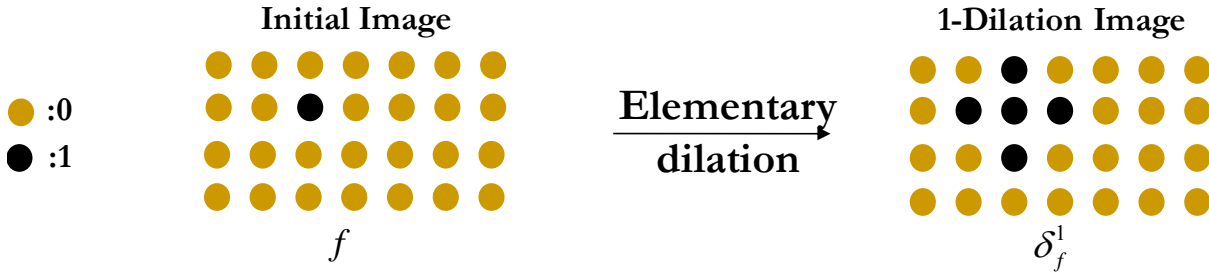
Figure 1.8: An example of the elementary dilation on a set of pixels of an image f.

The geodesic dilation of size 1 of the marker image f with respect to mask image g is denoted by $\delta_g^1(f)$ and is defined as the point-wise minimum between the mask image and the elementary dilation $\delta^1(f)$ of the marker image:

$$\delta_g^{(1)}(f) = \delta^{(1)}(f) \wedge g, \tag{1.3}$$

where the point-wise minimum $\wedge$ of two images f, g is defined as:

$$(f \wedge g)(p) = \min\left[f(p), g(p)\right] \tag{1.4}$$

An example of the application of the geodesic dilation to an image f is shown below:
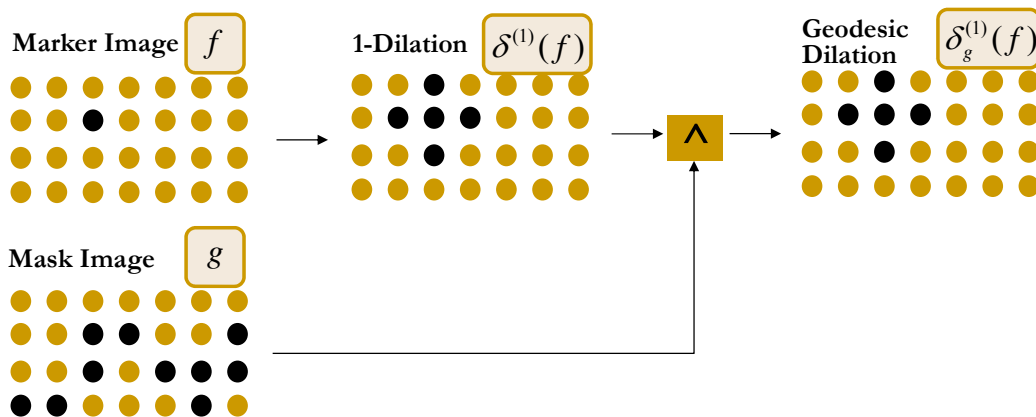
Figure 1.9: Geodesic dilation of size 1 of the marker image f with respect to mask image g

1.4.2 Greyscale Reconstruction

The reconstruction by dilation of a mask image g from a marker image f [13], [14] is defined as the geodesic dilation of f with respect to g iterated until stability and is denoted by $R_g^{\delta}(f)$:

$$R_g^{\delta}(f) = \delta_g^{(i)}(f), \tag{1.5}$$

where i such that $\delta_g^{(i)}(f) = \delta_g^{(i+1)}(f)$.

An example of the greyscale reconstruction in 1-D is shown below:
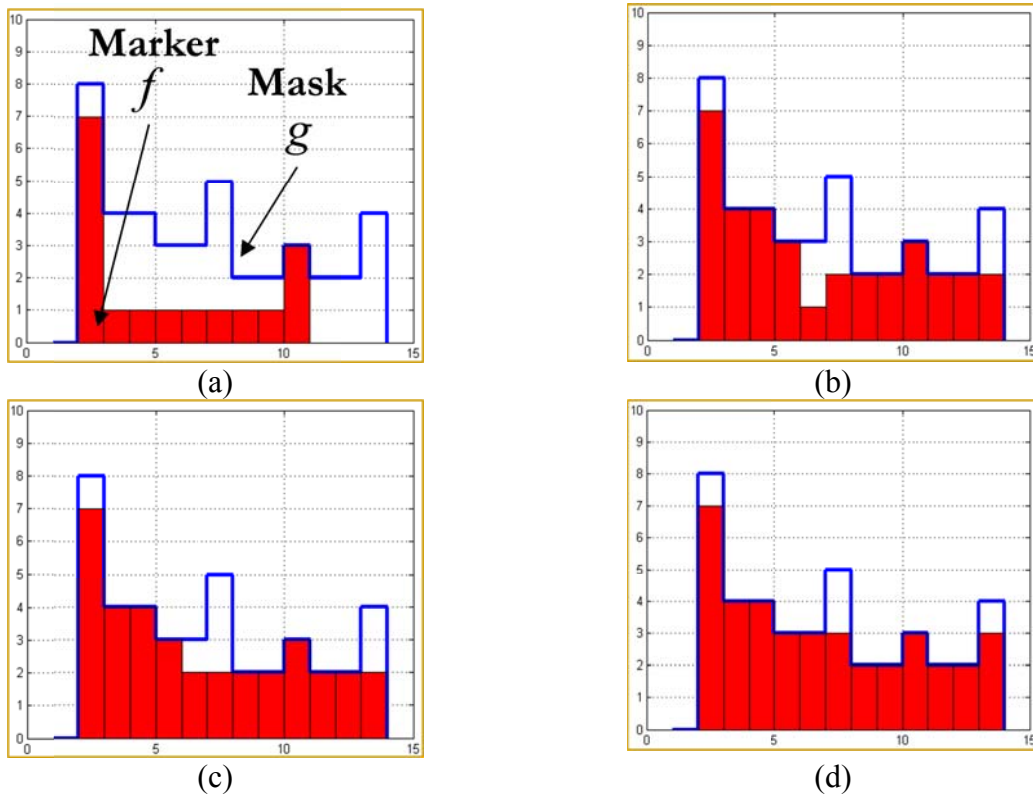
Figure 1.10: Greyscale Reconstruction of a mask g from a marker f. (a) The 1-D marker f and the mask g, (b) Geodesic Dilation of Size 3 $\delta_g^{(3)}(f)$, (c) Geodesic Dilation of Size 4 $\delta_g^{(4)}(f)$, (d) Greyscale reconstruction $R_g^{\delta}(f) = \delta_g^{(5)}(f)$.

1.4.3 Otsu method

Segmentation involves separating an image into regions corresponding to objects. We usually try to segment regions by identifying common properties. The simplest property that pixels in a region can share is intensity. So, a natural way to segment is through thresholding.

Thresholding creates binary images from grey-level ones by grouping all pixels below some threshold to class C_0 and all pixels above that threshold to class C_1.

Otsu's method is a thresholding method that separate objects from background. The threshold operation [15] at grey level l partitions the pixel values of an image I into two classes C_0 and C_1 (representing background and object respectively), i.e., $C_0 = \{1, 2, \ldots, l\}$ and $C_1 = \{l+1, l+2, \ldots, L\}$, where L is the total number of grey levels in the image. Let $\sigma_B^2(l)$ be the between-class variance for the threshold value l:

$$\sigma_B^2 = w_0 \left(\mu_0 - \mu_T \right)^2 + w_1 \left(\mu_1 - \mu_T \right)^2, \tag{1.6}$$

$$w_0 = \sum_{i=0}^{l} P_i, \ w_1 = 1 - w_0, \mu_1 = \frac{\mu_T - \mu_l}{1 - w_0}, \ \mu_0 = \frac{\mu_l}{w_0}, \ \mu_l = \sum_{i=0}^{l} i P_i, \mu_T = \sum_{i=0}^{L} i P_i, \ P_i = \frac{n_i}{n}, \tag{1.7}$$

where n_i is the number of pixels with grey-level i, n is the total number of pixels in a given image and P_i is the probability of the occurrence of grey level i.

An optimal threshold l^* can be determined by minimizing the between class variance:

$$l^* = \arg \min_l \sigma_B^2(l). \tag{1.8}$$

After the computation of the threshold l^* the binary image B can be computed:

$$B(x, y) = \begin{cases} 0, if \ I(x, y) \le l^* \\ 1, if \ I(x, y) > l^* \end{cases}. \tag{1.9}$$

1.4.4   The Watershed Transform

The concept of watersheds in image processing is based on considering an image in the three dimensional space, with two spatial coordinates versus intensity. The value of the intensity (e.g. of the gradient image) is assumed to be the elevation information. Pixels having the highest gradient magnitude intensities correspond to watershed lines, which represent the region boundaries. Water 'placed' on any pixel enclosed by a common watershed line flows downhill to a common local intensity minimum. Pixels draining to a common minimum form a catch basin, which represents a region. The result of the watershed transform is a tessellation of the image into regions.

An easy way to understand watershed is the definition by terms of the drainage patterns of rainfall. Generally, regions of terrain that drain to the same point are defined to be part of the same watershed. The same analysis can be applied to greyscale images based on the visualization of gray level into its topographic representation. This analysis includes three basic notions: minima, catchment basins and watershed lines. Let I be a greyscale image, a minima M_h is a connected set of pixels with intensity h, from which it is impossible to reach a pixel intensity h' without having to pass from a pixel of intensity h'', where $h' < h < h''$. The catchment basin $C(M)$ associated with the minima M is a set of pixels such that if a drop of water falls at any pixel in $C(M)$, then it will flow down to the minimum M Watersheds are defined as the lines separating the catchment basins which belong to different minima [16], [17].

Figure 1.11, illustrates the above features of the watershed transform.
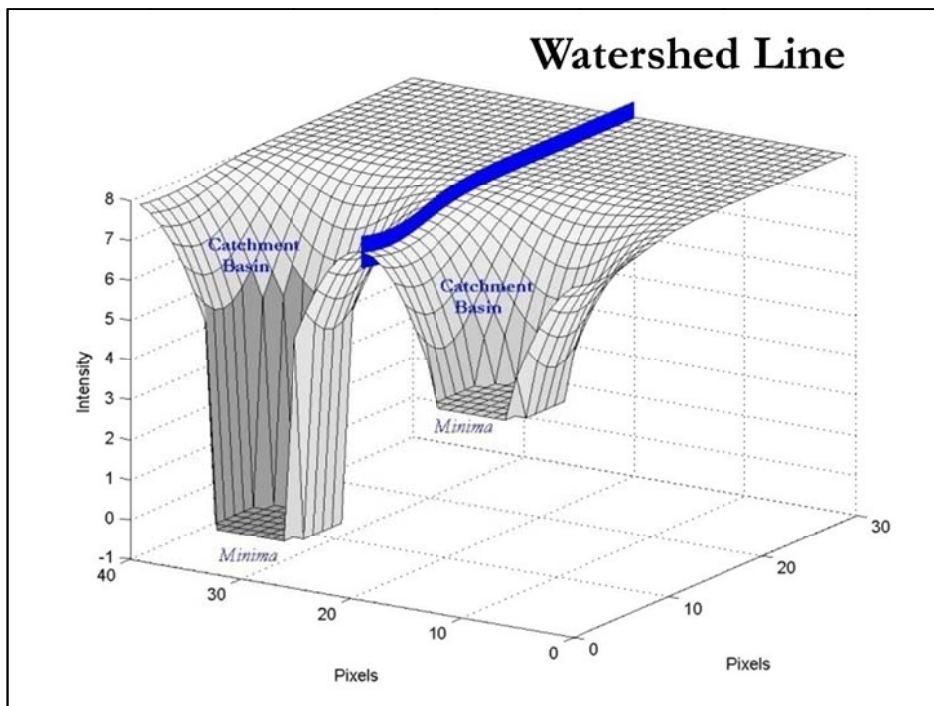
Figure 1.11: Example of regional minima, catchment basins, and watershed lines produced by the watershed transform.

Before introducing the algorithm for the computation of the Watershed Transform we must first define some terms. The Geodesic Distance between two pixels a and b in A is distance with the minimum length joining a and, and is included in A.
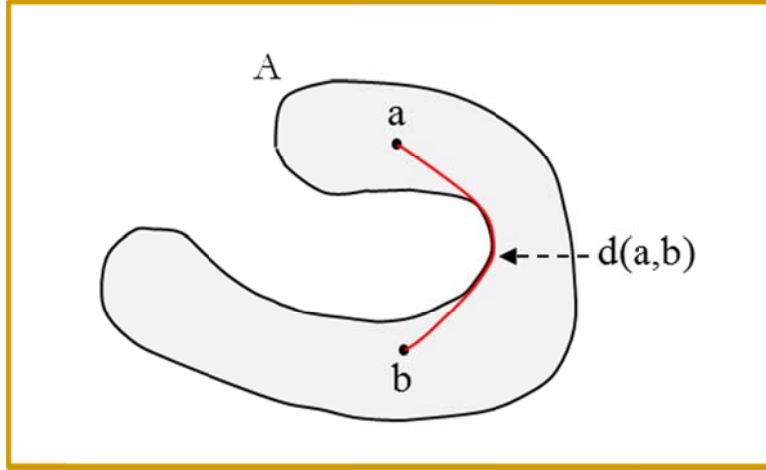
Figure 1.12: The geodesic distance between two points a and b.

Let B be a set composed of the union of components that are included in the set A.

$$B = \bigcup_{i=1}^{n} B_i \qquad (1.10)$$

The Geodesic Influence Zone $iz_A(B_i)$ of a connected component B_i in A is the locus of points of A whose geodesic distance to B_i is smaller than their geodesic distance to any other component of B [1]. We define the set of influence zones:

$$IZ_A(B) = \bigcup_{i=1}^{n} iz_A(B_i) \qquad (1.11)$$

Skeleton by Influence Zones are the pixels that belong to A and not to the set of influence zones [1].

$$SKIZ_A(B) = A / IZ_A(B) \qquad (1.12)$$
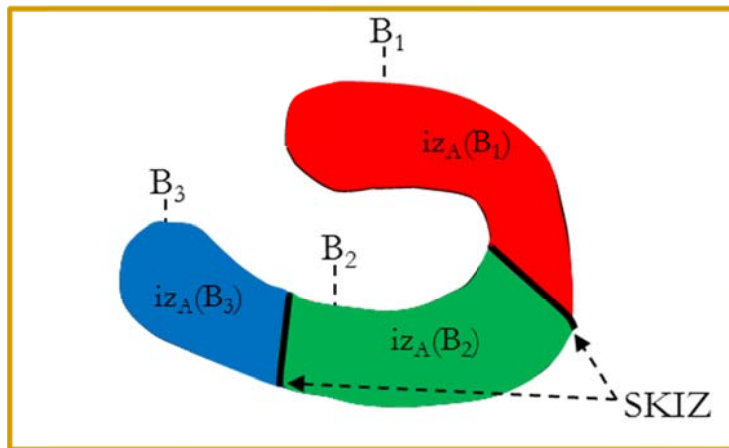
Figure 1.13: The geodesic influence zone of each connected component (B_1, B_2, B_3) and the skeleton by influence zones.

15

The purpose of the algorithm is to assign to each pixel p a label $lab[p]$ this label represents the number of the catchment basin-region which belongs to.

Initialization: All the image pixels get the label *INIT* (e.g. -1).

Sorting Step: Sort the image pixels based on their intensity.

Flooding Step: In this step we increase the height from h_{min} to h_{max}

For each step h we assume that the flooding step has been completed and all the pixels have been assigned to a catchment basin (e.g. assigned the label of each catchment basin).

Let's assume that we process the pixels of intensity $h+1$. We initialize these pixels with label *MASK* (e.g. -2).

Pixels that have a neighbor that has been assigned to a catchment basin are inserted to a queue and from these pixels the geodesic influence zones are expanded according to the following rules:

If a neighbor pixel q of the pixel p has been assigned to a catchment basin then this pixel p is assigned to the same catchment basin.

If a pixel p is has neighbors which belong to different catchment basins then this pixel is assigned to the watershed line and acquire a special label *WSHED _ LINE* (e.g. 0).

When the flooding step of pixels having intensity $h+1$ has been completed, the pixels that have the label *MASK* and have not been assigned to any catchment basin, are assigned to new catchment basins and they acquire new labels.

1.5 Machine Learning Algorithms

Machine learning is the area of artificial intelligence that attempts to provide machines with the ability to learn from examples [18]. More specifically, in machine learning problems we make use a set of observations (examples), which we call training set, in order to make predictions for unseen events. In the area of machine learning there are two major categories of problems; supervised learning and unsupervised learning. Unsupervised learning methods assume a training set that only consists of observed inputs in contrast to supervised learning where the input is also paired with a target (e.g. class label) that provides the desired output result.

16

1.5.1 K-means

The K-means can be used to partition a set of data points $X = \{x_1, x_2, \ldots, x_N\}$, into K clusters; where each point belongs to d-dimensional space, $x_i \in \Re^d$, $i = 1, \ldots, N$. Each cluster C_j is parameterized by a vector m_j, $(j = 1, \ldots, K)$ which is called its center.

To start the K-means algorithm the, K centers are initialized in some way, for example to randomly selected data points. K-means is then an iterative two-step algorithm. In the assignment step, each data point x_i is assigned to the nearest center according to the Euclidean distance. In the update step, the centers are adjusted to the sample means of the data assigned to the corresponding clusters.

Initialization: Set K clusters m_j, $j = 1, \ldots, K$ to randomly selected data points from the initial set of data points.

Assignment step: Assign x_i to cluster C_j when its distance to the center of this cluster is smaller than the distances to all other clusters centers:

$$\|x_i - m_j\| \le \|x_i - m_i\|, \, 1 \le i \le K \tag{1.13}$$

Update step: The centers are adjusted to match the sample means of the data points that they are responsible for:

$$m_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \tag{1.14}$$

where $|C_i|$ is the number of points belonging to cluster C_i.

1.5.2 Bayes Classifier

Suppose we wish to classify N objects $X = \{x_1, x_2, \ldots, x_N\}$ into K different classes $C_1, C_2, \ldots, C_K$ where each object belongs to d-dimensional space, $x_i \in \Re^d$. Let $P(C_i)$ denote the probability that an object belongs to class C_i, $1 \le i \le K$ with $\sum_{i=1}^{K} P(C_i) = 1$ This is called the a priori class probability. Let $p(x | C_i)$ denote the class-conditional probability distribution function. It represents the probability distribution of objects of class C_i. Let $P(C_i | x)$ be the class conditional probability which is the probability that the object belongs

17

to class C_i given its feature vector x. Given $P(C_i)$ and $p(x|C_i)$, the class conditional probability for an object represented by the feature vector x is given by the Bayes theorem [19]:

$$P(C_i \mid x) = \frac{p(x \mid C_i)P(C_i)}{p(x)},$$ (1.15)

where $p(x) = \sum_{i=1}^{N} p(x \mid C_i)P(C_i)$.

The Gaussian density function is often used to model the distribution of feature values of a particular class. The general multivariate Gaussian density function in d dimensions is given by:

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right],$$ (1.16)

where is x a d component feature vector, μ is the d component mean vector, Σ is the $d \times d$ covariance matrix, $|\Sigma|$ and Σ^{-1} are its determinant and inverse respectively. Also $(x-\mu)^T$ denotes the transpose of $(x-\mu)$. During the training phase, the mean vector μ, and the $d \times d$ covariance matrix Σ are calculated for each class from the training data:

$$\mu_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j, \, i = 1,\dots,K \, ,$$ (1.17)

$$\Sigma_i = \frac{1}{N_i - 1} \sum_{x_j \in C_i} \left(x_j - \mu_i\right) \cdot \left(x_j - \mu_i\right)^T, \, i = 1,\dots,K \, ,$$ (1.18)

where N_i the number of pixels of class C_i.

To classify an object described by the feature vector x, we calculate $P(C_i \mid x)$ for each class i and then use the the Bayes Decision Rule:

$$decide \, C_i \, if \, P(C_i \mid x) > P(C_j \mid x), \, \forall j \neq i.$$ (1.19)

1.5.3   Gaussian Mixture Models and the EM Algorithm

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features. Let x denote a feature vector. GMMs [20] represent density functions as a convex combination

18

of K Gaussian component densities $\varphi(x \mid \theta^j) = N(x \mid \mu_j, \Sigma_j)$, (where μ_j is the mean and Σ_j the covariance of the j th Gaussian), according to the equation:

$$p(x \mid \Theta) = \sum_{j=1}^{K} \pi_j \varphi(x \mid \theta^j) , \qquad (1.20)$$

where the parameters $0 \leq \pi_j \leq 1$ represent the mixing weights satisfying that $\sum_{j=1}^{K} \pi_j = 1$, while Θ is the vector of all unknown parameters of the model, i.e. $\Theta = [\pi_1, \pi_2, \ldots, \pi_k, \theta_1, \theta_2, \ldots, \theta_k]$, with $\theta_j = [\mu_j, \Sigma_j]$.

The EM (Expectation-Maximization) algorithm is used for training Gaussian mixtures given a training set of points $X = \{x_1, x_2, \ldots, x_N\}$. It is an iterative algorithm that starts from some initial estimate of Θ, and then proceeds to iteratively update Θ until convergence. Each iteration consists of an E-step and an M-step:

E-Step: Let $\theta^{(t)}$ denote the current parameter values. Compute the membership weights $w_{ik}^{(t)}$ for all data points x_i, $1 \leq i \leq N$ and all mixture components k, $(1 \leq k \leq K)$.

$$w_{ik}^{(t)} = \frac{\varphi(x_i \mid \theta_k^{(t)}) \pi_k}{\sum_{m=1}^{K} \varphi(x_i \mid \theta_m^{(t)}) \pi_m}, \ 1 \leq i \leq N, \ 1 \leq k \leq K \qquad (1.21)$$

M-Step: Use the membership weights to update the parameter values $\theta^{(t+1)}$. Let $N_k^{(t)} = \sum_{i=1}^{N} w_{ik}^{(t)}$, i.e., the sum of the membership weights for the k-th component. Then:

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{N}, \qquad (1.22)$$

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{N} w_{ik}^{(t)} x_i, \qquad (1.23)$$

$$\Sigma_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{N} \left(x_i - \mu_k^{(t+1)}\right)\left(x_i - \mu_k^{(t+1)}\right)^T . \qquad (1.24)$$

Both K-means and EM depend highly in the initialization of the parameters.

1.6   Thesis Objectives and Contribution

Most of the already published methods which deal with automated segmentation and classification of M-FISH images either first segment the image and then classify the

segmented region or they directly classify all the pixels of the M-FISH image including a class for the background. The main disadvantages of the methods are the following:

- Most of the methods that segment the M-FISH image include a class for the background and classify each pixel of the image into 1-24 chromosome classes including also one class for the background. This has the disadvantage that they do not take into account spatial information of neighborhood pixels. The segmentation method introduced in Chapter 2, is based on the multichannel watershed transform in order to define regions of similar spatial and spectral characteristics. *Thus the goal is to develop an image segmentation technique that will segment the M-FISH image using all the channels of the M-FISH image.*

- Automated classification methods for the M-FISH images reported in the literature; classify pixels of the M-FISH image instead of classifying regions. This usually produces noisy results. *The development of a method with high classification accuracy for all the chromosome classes will make the M-FISH technique widely used for the easily detection of chromosome anomalies.*

- Chromosome misclassification errors result from different factors such as uneven hybridization, spectral overlap among fluors, and biochemical noise. *However, no filtering method has been proposed that will take into consideration the information from all the channels in order to filter these multichannel images. A method such as the Vector Median Filtering could be the choice for this type of images.*

- Most of the methods that classify an M-FISH image use a number of labeled training images in order to train a supervised classification method such as Bayes, SVM, and k-NN. One of the key factors limiting the pixel classification accuracy of these methods is the variations between M-FISH images. This is due to the fact that the M-FISH imaging process is not always accomplished under the same conditions e.g. humidity, temperature, type of microscope, color spread [21,30] and these factors affects the quality of the produced M-FISH image. *An unsupervised method that does not require a set of labeled training images is a necessity for the wide application of the M-FISH image technique.*

- It is common for an M-FISH image that the chromosomes do very often partially occlude each other; hence, their segmentation is not trivial and requires special treatment. *Hence an automated method for the segmentation of touching and*

20

overlapping groups of chromosomes in M-FISH images would be very helpful for cytogeneticists.

The contribution of the thesis is fivefold:

- First, we focus on the efficient region segmentation of M-FISH images. We propose a new algorithm for the segmentation of multichannel M-FISH images into chromosome regions.

- Next, we propose a supervised method for the classification of the segmented regions of the M-FISH image based on the Bayes classifier.

- We study the effect of the Vector Median Filtering on the classification accuracy of the M-FISH images

- We propose an unsupervised method for labeling the chromosome regions.

- We propose a method for the disentangling of touching and overlapping chromosomes.

In Chapter 2, a method for the region segmentation of the multichannel M-FISH chromosome images [21], [22] is presented. The method uses the information from all the six channel M-FISH images in order to segment the M-FISH image. The main novelty of this method is that we segment effectively the M-FISH image into regions, without using training information, while until now all the methods tackle the segmentation problem as part of the classification problem considering the background as an extra class.

In Chapter 3, we employ a supervised statistical approach to classify the obtained segmented regions [22]. The proposed region classification approach is empirically shown to outperform already proposed pixel-by-pixel classification approaches. We further examine the effect on the classification accuracy using a multichannel filtering technique, the Vector Median Filtering [23].

In Chapter 4, we propose a fully automated unsupervised classification method for the classification of M-FISH images [24]. The method first uses the region segmentation method in order to segment the multichannel image. Then we employ the well-known Expectation Maximization algorithm in order to build single channel Gaussian Mixture Models (GMMs). Then we use those models to estimate the parameters of a multichannel Gaussian Mixture Model where each chromosome class is represented by a Gaussian component. More specifically, we build a single channel GMM for each channel of the M-FISH image

incorporating the emission information that is available. These singe channel GMMs are then used to estimate the multichannel Gaussian mixture that is used to train the parameters of the multichannel GMM using the MAP-EM algorithm. Finally, each region is classified using the multichannel GMM and the labeled regions are merged providing the final classification map. The main contribution of this method is that the whole procedure is unsupervised and does not require the specification of a labeled training set. Although unsupervised, the classification accuracy of this method seems to be superior compared to supervised ones.

Chapter 5, presents a novel method for the disentangling of touching and overlapping chromosomes of the M-FISH image using the watershed transform and gradient paths [25]. An efficient way to "cut" these groups of chromosomes is to find a path of pixels that have a relative low grey level intensity and run between touching groups of chromosomes. We expand the idea of pale paths for M-FISH images by defining gradient paths where the "cut" runs through pixels of high intensity (gradient pixels). The method uses the multichannel gradient as a measure of separability of touching and overlapping chromosomes. By defining gradient paths we manage to deal not only with touching groups of chromosomes but also for overlapping groups.

Finally, in Chapter 6 we provide a summary of the results of this thesis and suggest some interesting directions for further research.

CHAPTER 2:

A MULTICHANNEL WATERSHED-BASED SEGMENTATION METHOD FOR MULTISPECTRAL CHROMOSOME IMAGES

2.1   Introduction

Multiplex Fluorescent In Situ Hybridization (M-FISH) is a chromosome imaging technique where each chromosome class appears to have a distinct color. This technique not only facilitates the detection of subtle chromosomal aberrations but also makes the analysis of chromosome images easier; both for human inspection and computerized analysis. The proposed segmentation method is based on the multichannel watershed transform in order to define regions of similar spatial and spectral characteristics. The method consists of two steps: (a) computation of the gradient magnitude of the image, (b) application of the watershed transform to decompose the image into a set of homogenous regions.

2.2 Automated Chromosome Segmentation of M-FISH Images - Literature Review

Since M-FISH technology has been introduced, many attempts have been reported which tackle the problem of segmenting chromosomes [3], [26], [27], [28], [29], [30], [31], [32], [21]. Semi-automated analysis of M-FISH images was first introduced in the mid 90's [3]. The DAPI channel was used to create a binary mask. Then, for each pixel of the mask a threshold was applied in order to detect the presence or absence of a fluor in that pixel. Each pixel class was determined by comparing the response of the combined fluors to that of a labeling table.

Region based classification approaches were also introduced [26], [27]. The method of Eils *et al.* [26], consisted of two stages: (i) spectral calibration and (ii) adaptive region classification. During the calibration stage a five-dimensional optimal vector called adaptive feature vector, representing each class, was found by minimizing an energy function. The region classification stage was based on a Voronoi image tessellation algorithm [33]. Neighboring regions were merged if they belong to the same class or alternatively, when their color distance was below a preset threshold.

Saracoglu *et al.* [27], proposed a method consisting of three steps: image tessellation, clustering and classification. The image was tessellated into regions with similar properties using a region growing approach (tessellation step). Based on the "average" color information of the regions, clustering is performed. The region color vectors are grouped to form a known number of clusters (clustering step). Finally, each cluster is assigned to one of the color class vectors (classification step).

Methods using pixel-by-pixel classification algorithms have been used in M-FISH analysis. These methods either classify each pixel of the M-FISH image [28], [29], or create a binary mask of the DAPI image using edge detection algorithms, and classify each pixel of the mask [30], [31]. A method for joint segmentation-classification of chromosome M-FISH images was presented in [32]. A probabilistic model of M-FISH chromosomes was introduced which allows for simultaneous segmentation and classification. The additional information provided by multiple spectra in chromosome images made it feasible to distinguish chromosomes that overlap and are in touch within clusters.

Table 2.1, presents the advantages and limitations of related studies on chromosome segmentation appearing in the literature. Most of these methods [28], [29], [30], [31], [32] deal with the above problem using pixel-by-pixel classification techniques; without taking into account neighborhood information. On the other hand, only few region-based methods

[26], [27] have been proposed in the literature, that are based on a large number of parameters which are determined heuristically.

Table 2.1: Advantages and limitations of methods presented in the literature.

METHOD	YEAR	ADVANTAGES	LIMITATIONS
Speicher *et al.* [3]	1996	• First time use of M-FISH images. • Simple classification method.	• Depends on optimal optical and experimental conditions. • Lack of classification results.
Eils *et al.* [26]	1998	• Region based segmentation-classification.	• Lack of classification accuracy for various M-FISH sets.
Saracoglu *et al.* [27]	1998	• Region based segmentation-classification.	• Several threshold and parameters are heuristically set.
Sampat *et al.* [28]	2002	• Pixel-by-pixel classification methodology. • High classification rate.	• Does not handle touching/overlapping chromosomes. • Small number of testing images.
Choi *et al.* [29]	2004	• Use of background correction, color compensation and filtering techniques as preprocessing step. • High classification rate.	• Does not handle touching/overlapping chromosomes. • Small number of testing images.
Sampat *et al.* [30]	2005	• Employment of different classification methods (MLE, k-NN).	• Does not handle touching/overlapping chromosomes. • Small number of testing images. • Segmentation based on edge detection only on DAPI image.
Wang *et al.* [31]	2005	• Use of background correction, feature selection and image registration techniques as preprocessing step. • High classification rate.	• Does not handle touching/overlapping chromosomes. • Small number of testing images. • Segmentation based on edge detection only on DAPI image.
Schwartzkopf *et al.* [32]	2005	• Handle overlapping/touching chromosomes. • Use of large number and various cases of M-FISH images.	• Complicated method. • Low pixel-by-pixel classification accuracy.

2.3 Motivation – Goals

As already noted the methods introduced for M-FISH image segmentation use a plethora of parameters and have been tested on a small number of M-FISH images. Furthermore the segmentation is based only on the greyscale DAPI image without using the information from the other five channels. Finally, the segmentation methodologies employ image classification techniques (e.g. pixel-by-pixel technique) for the image segmentation without taking into

25

consideration spatial information. Thus a segmentation method using all the M-FISH channels and taking into consideration spatial information is expected to be effective [22]. The approach uses the Watershed transform which is popular image segmentation used both for greyscale and color images.

2.4   Watershed Based M-FISH Image Segmentation

Segmentation of multispectral images using the watershed transform is performed in four steps (Figure 2.1). More specifically, the gradient magnitude of the multispectral image is computed combining the contrast information from the different spectral channels. Due to the high sensitivity of the watershed algorithm in the variations of the gradient an automatic selection of significant minima is realized in the next step, where the watershed transform is applied and a large number of homogenous regions is produced. A binary mask of the DAPI channel is computed and superimposed to the tessellation in order to further reduce regions that do not belong to chromosomes.
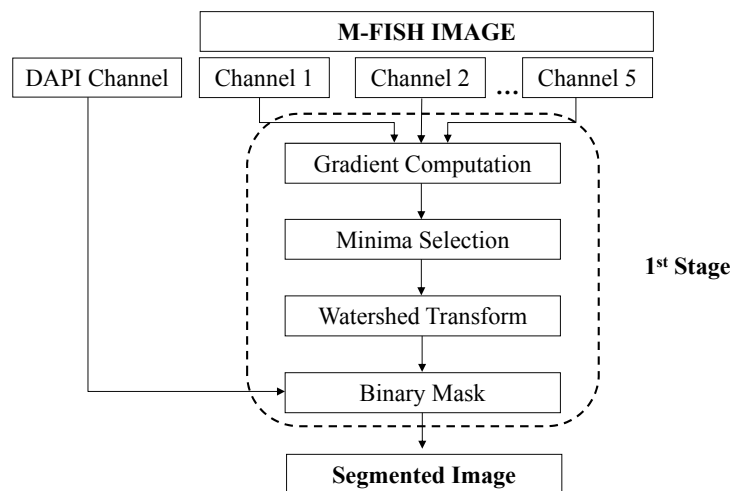
Figure 2.1: Flowchart of the proposed method.

2.4.1   Multichannel Gradient Computation

To apply the watershed based segmentation algorithm to the multichannel data, the gradient of the multichannel image must be defined. The computation of a tensor gradient was introduced by DiZenzo et al. [34], instead of separately computing the scalar gradient for each channel [35]. Drewniok [36] extended this work to multispectral images. Assuming a

26

multichannel image $I(i, j) : \mathbb{Z}^2 \to \mathbb{Z}^m$ ($m = 5$ for M-FISH images), the direction n is defined by the angle φ:

$$I(x, y) = \begin{bmatrix} I_1(x, y) \\ I_2(x, y) \\ \vdots \\ I_m(x, y) \end{bmatrix},$$ (2.1)

$$n = \begin{bmatrix} \cos \varphi \\ \sin \varphi \end{bmatrix},$$ (2.2)

where $I_i(x, y)$, $1 \le i \le m$ are the components (channels) of the M-FISH image.

The directional derivative of the function $I(x, y)$ consists of the directional derivatives of each component of $I(x, y)$: $\dfrac{\partial I}{\partial n} = \left[\dfrac{\partial I_1}{\partial n}, \dfrac{\partial I_2}{\partial n}, \cdots, \dfrac{\partial I_m}{\partial n} \right]^T$.

Projecting each directional derivative in the direction n, we have:

$$\frac{\partial I}{\partial n} = \begin{bmatrix} \nabla I_1 \cdot n \\ \nabla I_2 \cdot n \\ \vdots \\ \nabla I_m \cdot n \end{bmatrix} = \begin{bmatrix} I_1^x & I_1^y \\ I_2^x & I_2^y \\ \vdots & \vdots \\ I_m^x & I_m^y \end{bmatrix} \cdot n = J \cdot n,$$ (2.3)

where $\nabla I_i = \begin{bmatrix} I_i^x & I_i^y \end{bmatrix} : 1 \le i \le m$, J is the Jacobian matrix and I_i^x and I_i^y are the derivatives of the i-th component in the x and y direction, respectively.

Next, the direction n which corresponds to the maximum of the directional derivative $I(x, y)$ is found, by maximizing the Euclidean norm:

$$\frac{\partial I}{\partial n} = \begin{bmatrix} \nabla I_1 \cdot n \\ \nabla I_2 \cdot n \\ \vdots \\ \nabla I_m \cdot n \end{bmatrix} = \begin{bmatrix} I_1^x & I_1^y \\ I_2^x & I_2^y \\ \vdots & \vdots \\ I_m^x & I_m^y \end{bmatrix} \cdot n = J \cdot n,$$ (2.4)

$$\| J \cdot n \|^2 = (J \cdot n)^T (J \cdot n) = n^T (J^T J) n.$$ (2.5)

The symmetric matrix $J^T J$ can be written as:

$$J^T J = \begin{bmatrix} \displaystyle\sum_{i=1}^{m} \left(I_i^x \right)^2 & \displaystyle\sum_{i=1}^{m} I_i^x \cdot I_i^y \\ \displaystyle\sum_{i=1}^{m} I_i^x \cdot I_i^y & \displaystyle\sum_{i=1}^{m} \left(I_i^y \right)^2 \end{bmatrix},$$ (2.6)

27

The extrema of the quantity $n^T(J^T J)n$, are given by the eigenvalues of the matrix $J^T J$ [36]. Sobel operators [37] are used to compute the directional derivatives $I_i^x, I_i^y : 1 \le i \le m$ in the x and y directions, respectively.

2.4.2 *Minima Selection with the H-minima transform*

Direct application of the watershed algorithm to a gradient image usually leads to over segmentation due to noise and other local irregularities of the gradient. A practical solution to this problem is to limit the number of allowable regions by reducing the number of irrelevant minima. Several algorithms have been proposed for minima selection. The simplest is interactive selection by the user [38] or by using a priori knowledge for the image [39]. The dynamics approach [40], [41] orders all minima and selects only those below a threshold.

We have used the dynamics approach [41] to reduce the number of unwanted minima, as it provides an intuitive selection scheme controlled by a single parameter (h) using the greyscale reconstruction.

Image minima and maxima are important morphological features because they often mark relevant image objects: minima for dark objects and maxima for bright objects. In morphology, the term minimum is used in the sense of regional minimum, i.e., a minimum whose extent is not necessarily restricted to a unique pixel. A regional minimum M of an image f at elevation h is a connected component of pixels with the value h whose external boundary pixels have a value strictly greater than h . The h -minima transform of an image f , $HMIN_h(f)$ suppresses all minima whose depth is lower or equal to a given threshold level h . This is achieved by performing the reconstruction by dilation of f from a new image $f^c - h$, where f^c is the complement image of f , [37]:

$$HMIN_h(f) = \left[R_{f^c}^{\delta}(f^c - h) \right]^c . \tag{2.7}$$

An example of the application of the h -minima transform for 1-D signal is shown below.
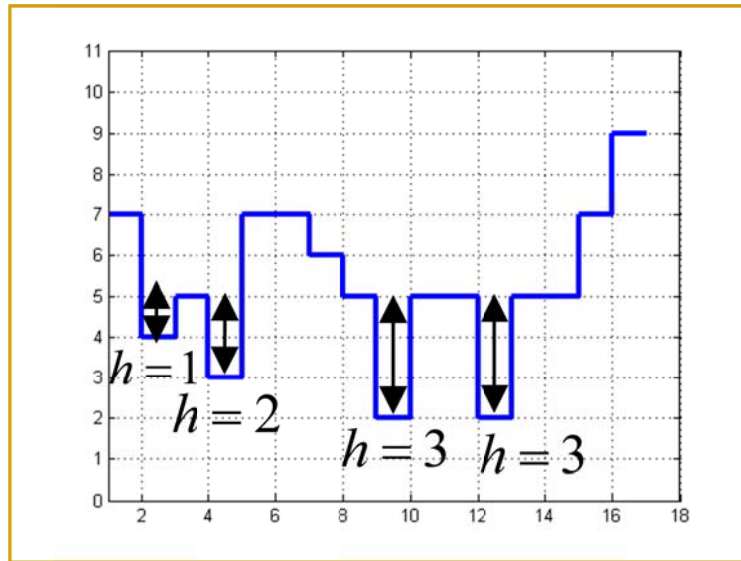
(a)
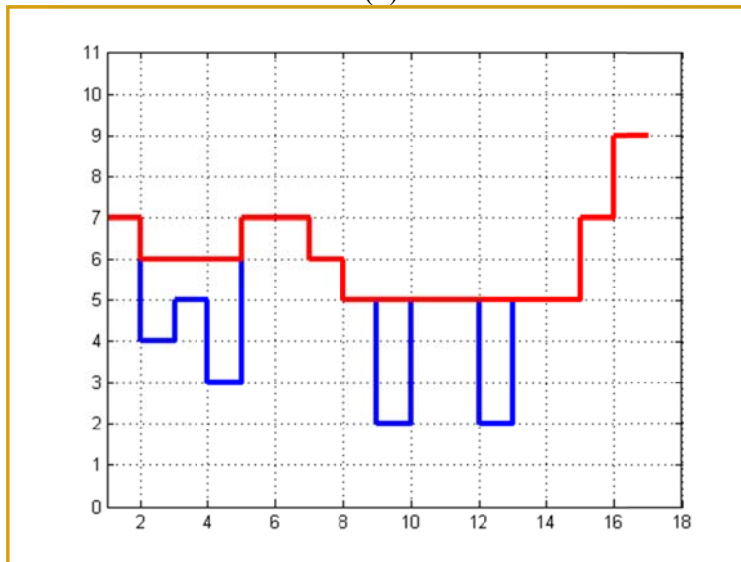
(b)

Figure 2.2: Removal of all the minima with depth $h = 3$, (a) Initial signal, (b) The signal after the h-minima transform, $HMIN_3(f)$ (red line).

2.4.3 Segmentation using the Watershed Transform

The watershed transform presents some advantages over other segmentation methods.

- The watershed lines form closed and connected regions, where edge based techniques usually define disconnected boundaries that need post-processing to produce closed regions.

- The watershed lines always correspond to obvious contours of objects which appear in the image.

The main problem of over-segmentation can be usually overcome by the use of preprocessing or post-processing, producing a segmentation that better reflects the arrangement of objects within the image. Such preprocessing or post-processing methods include region merging [21], multiscale watershed [22], [23], marker-based watershed segmentation [24], [25], and watershed-based deformable models [26], [27].

We adopted an efficient implementation of the watershed transform [16]. The watershed computation algorithm used here is based on the Immersion Approach: imagine that a hole is drilled in each minimum of the surface, and water is flooded into different catchment basins from the holes. As a result, the water starts filling all catchment basins, which have minima under the water level. If two catchment basins would merge as a result of further immersion, a dam is built all the way to the highest surface altitude and the dam represents the watershed lines. This flooding process will eventually reach a stage when only the top of the dam is visible above the water line.

The output of the watershed transform for an image I is a tessellation R_I of the image into its different regions $R_i, 1 \le i \le C_R$, each one characterized by a unique label $l_{R_i}, 1 \le i \le C_R$:

$$R_I = \left\{ (R_1, l_{R_1}), (R_1, l_{R_2}), \ldots, (R_1, l_{R_k}) \right\}, \tag{2.8}$$

where C_R the number of regions produced by the watershed transform.

Thus a new label image L_W is defined where each pixel is assigned the label of the region where it belongs to. Pixels belonging to the watershed lines are assigned the special label 0.

$$L_W(i,j) = \begin{cases} l_{R_i}, & \text{if the pixel } (i,j) \in R_i \\ 0, & \text{if the pixel } (i,j) \in \text{Watershed Line} \end{cases}. \tag{2.9}$$

Also a new image W_{Lines} is defined as:

$$W_{Lines}(i,j) = \begin{cases} 0, & \text{if } L_W(i,j) \in \text{Wateshed Line} \\ 1, & \text{otherwise} \end{cases}. \tag{2.10}$$

Figure 2.3(a) illustrates the initial segmentation produced by the watershed transform. In Figure 2.3(b) all the watershed regions are represented by the average color of each region, which is valuable information for the cytogeneticist.

(a)

(b)
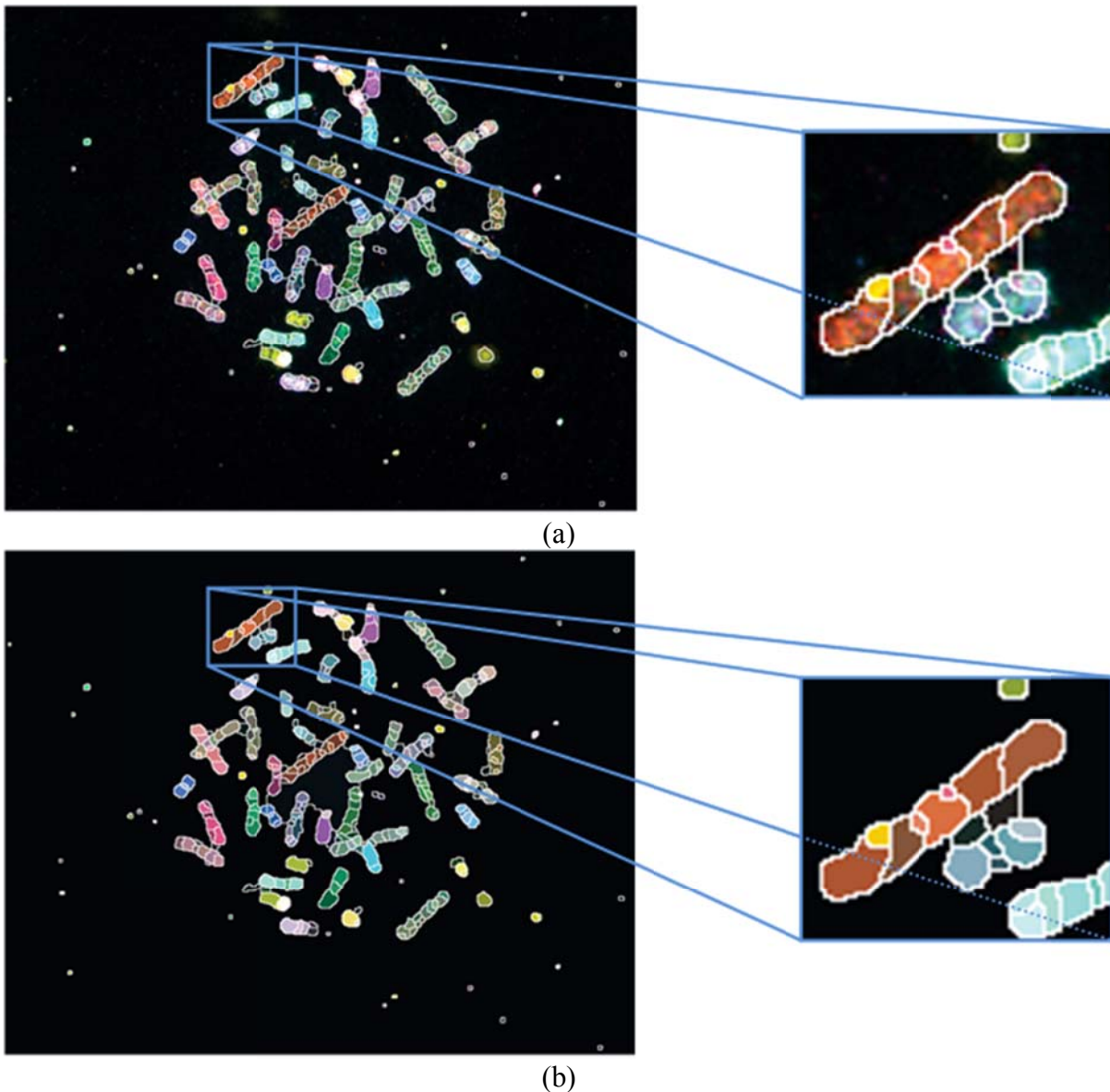
Figure 2.3: Watershed segmentation of an M-FISH image: (a) Watershed regions superimposed (white line) on the M-FISH image, and (b) watershed regions are labeled with the average color of each region.

2.4.4 Creation of the Binary Mask

After applying the WT on the greyscale reconstructed multichannel gradient image a number of regions is produced. However, two major problems are emerging which must be handled efficiently.

1. The M-FISH image often contains artifacts that appear in some channels but not in the greyscale DAPI channel. These artifacts often contain dust, unattached dye, salt deposits from evaporated solvents, fibbers, various airborne debris and fluorescent

31

artifacts [42]. An example of such an artifact is shown in Figure 2.4(a) with purple stain.
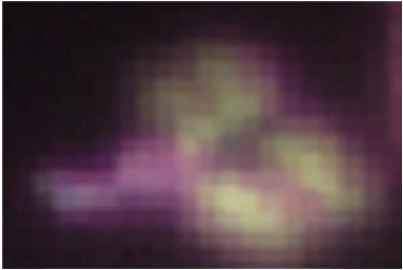
	M-FISH IMAGE	WATERSHED SEGMENTATION
(a)		
(b)		

Figure 2.4: (a) An example of an artifact shown in M-FISH image (purple region). (b) An example of failure of hybridization close to the centromere.

2. Regions on central areas of chromosomes (centromeres) usually fail to appear in the M-FISH image (failure of hybridization) [43]. An example of this effect is shown in Figure 2.4(b), where the centromere of the chromosome has not been hybridized and the watershed segmentation fails to segment it. Furthermore, in Figure 2.5 the average fluor signals for the five different M-FISH channels are shown along a chromosome. As it can be observed close to the center of the chromosome (centromere) the fluor signal of all the channels drops significantly making the detection of this region extremely difficult.
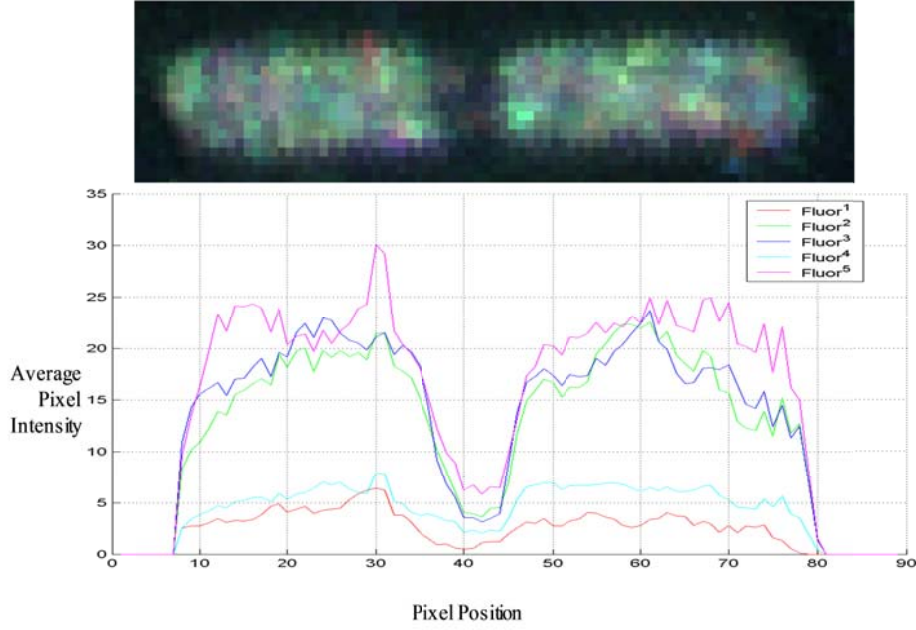
Figure 2.5: Averaged fluor signals along the chromosomal axis.

Thus, the goal of this step is to refine the watershed segmentation to exclude regions corresponding to artifacts (Figure 2.4(a)) and to include regions corresponding to centromeres (Figure 2.4(b)). To achieve this, the DAPI image is used. First, the DAPI chromosome image is converted to binary using Otus's threshold selection process [15]:

$$B_{Otsu}(i,j) = \begin{cases} 0, & if\ DAPI(i,j) \le l^* \\ 1, & if\ DAPI(i,j) > l^* \end{cases}, \tag{2.11}$$

where l^* the threshold found by the Otsu method [15].

Then the watershed lines are used to segment the regions produced from the binary image B_{Otsu}:

$$B_{Regions} = B_{Otsu} \bullet W_{Lines} \tag{2.12}$$

where $\bullet$ is the logical AND operator. At the end of this step all connected components which are not "0" are relabelled with a unique label.

Two indicative examples of the application of this step on the watershed segmentation method are presented in Figure 2.6. As it is shown in Figure 2.6, the contribution of this step is crucial for the segmentation stage of this method as two aspects are handled effectively. First, image artefacts contained in the M-FISH image are removed successfully as it is shown in Figure 2.6 (Case A). Second, in the central chromosome region (centromere) where the M-

33

FISH hybridization fails, this step uses the DAPI channel to segment correctly these regions, Figure 2.6 (Case B).

M-FISH IMAGE	DAPI IMAGE	WATERSHED SEGMENTATION	MASK $W_{watershed}$	BINARY MASK B	BINARY MASK WB	
						CASE A
						CASE B
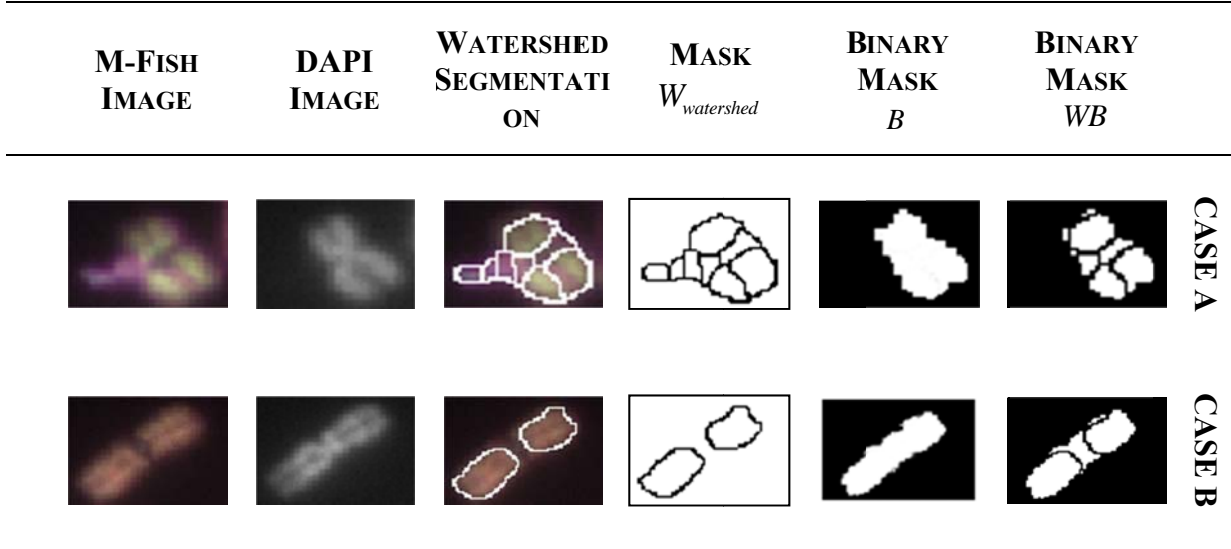
Figure 2.6: (Case A) Elimination of artifact (purple region) in an M-FISH image. The artifact is detected as chromosome region by the multichannel watershed transform and eliminated using the binary mask. (Case B) Detection of un-hybridized centromere using the binary mask, when the watershed segmentation fails to detect it.

2.5   Segmentation Results

The proposed method is tested on the ADIR M-FISH chromosome image database [44]. The only tuneable parameter of the proposed method is h in the greyscale reconstruction function. Several experiments were performed varying h from 0-250 (with step 10). $h = 0$ corresponds to the finest tessellation and $h = 250$ corresponds to the coarsest tessellation. A good choice was found to be $h = 50$ since as the region size increases, adjacent chromosomes regions with different color information are connected, Figure 2.7.

The performance of the segmentation step is evaluated using the overall and chromosome segmentation accuracy per image, defined as:

$$overall\_segmentation = \frac{\#\,pixels\,correctly\,segmented}{\#total\,image\,pixels} \qquad (2.13)$$

$$chromosome\_segmentation = \frac{\#chromosome\,pixels\,correctly\,segmented}{\#total\,chromosome\,pixels} \qquad (2.14)$$

34

(a)

(b)

(c)
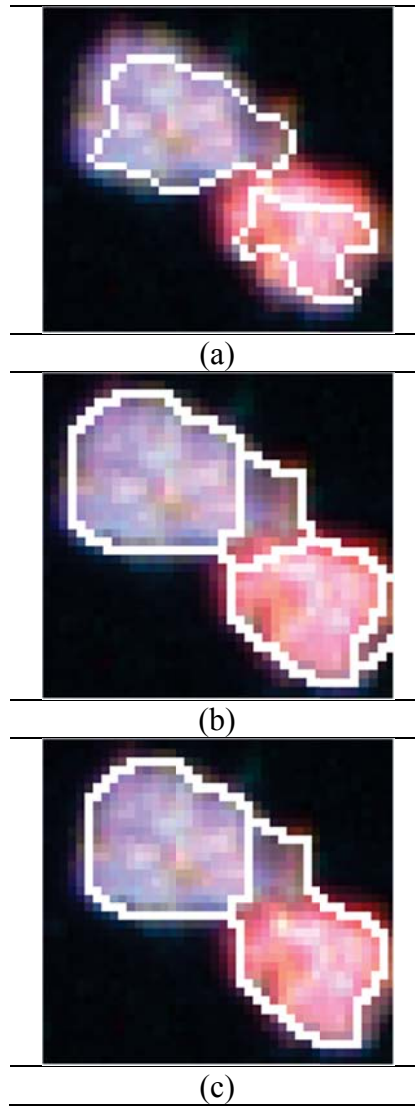
Figure 2.7: Example of region segmentation for two different values of the parameter h. (a) The ground truth classification map, (c) Image segmentation for $h = 190$ (e) Image segmentation for $h = 200$.

The overall segmentation (averaged over all images) accuracy is 98% with standard deviation 1% and the chromosome segmentation accuracy is 82.5% with standard deviation 12%. Since a majority of the pixels are background pixels, the chromosome segmentation accuracy mainly reflects chromosome segmentation. For this reason, the overall segmentation accuracy was substantially higher than chromosome segmentation accuracy.

Translocation is the most significant rearrangement. It involves two non homologous chromosomes which result from a break in each of the chromosomes, and subsequent reunion [45]. The detection of translocations is difficult even for an expert. A change in the colour of

a chromosome tip may be due to noise, staining, or an actual translocation. Three examples of translocations are shown in Figure 2.8. The translocations are accentuated due to the variation of the colour between two regions of the same chromosome. It is clear from Figure 2.8 that the segmentation method succeeds in defining regions of the same colour information which correspond to the translocation. Thus, a more acceptable segmentation map is provided to the expert in order to recognize more easily and accurately chromosome rearrangements.
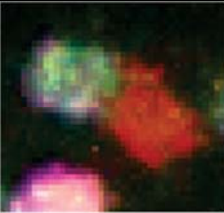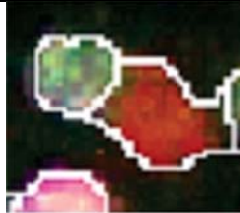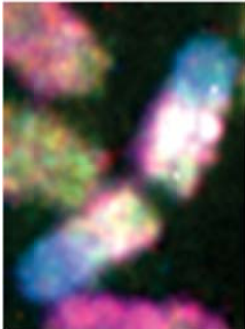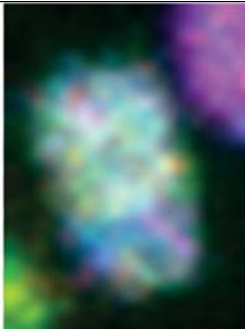
	M-FISH	Region Segmentation
CASE A		
CASE B		
CASE C		

Figure 2.8: Three examples of exchange of genetic material (translocation) between two different chromosomes and regions produced by the multichannel watershed algorithm. Case A: Translocation t(5;17) , Case B: Translocation t(9;14), and Case C: Small translocation t(14;15).

2.6 Conclusions

In this chapter we have presented a new fully automated chromosome segmentation method for M-FISH images. The method utilizes a multichannel watershed segmentation algorithm. Initially, the chromosome image is decomposed into a set of homogeneous regions using the multichannel watershed algorithm. The method uses all the M-FISH channels and creates a chromosome mask from the DAPI channel to further eliminate artefacts as also to segment correctly the centromere areas of the chromosome. To evaluate the method we used the ADIR M-FISH database and evaluation results are very promising, resulting in overall accuracy 82.5%.

The proposed method uses a multichannel segmentation method to segment the M-FISH image into homogenous spectral regions which combines spectral information from different channels. This is advantageous since it is an effective way to incorporate spatial characteristics into the analysis, which leads to superior performance in terms of classification accuracy.

The segmentation of each chromosome into regions emulates the procedure followed by an expert to identify chromosome rearrangements (anomalies). As we have already shown in Figure 2.8, regions with different colour information (translocations) are accurately defined by the proposed method. It is important to be mentioned that the segmentation by itself already gives the cytogeneticist an advantage in his/her medical assessment. On the other hand, the employment of the Otsu binarization method greatly simplifies the detection of chromosome regions that have not been hybridized (Figure 2.6), providing a more accurate segmentation of the M-FISH image.

In [21] an analogous approach has been proposed having however significant differences with the proposed method. The most important is that the gradient computation was based only on the DAPI channel, not taking into consideration information from the other five channels. Thus, chromosome anomalies could not be effectively detected.

Figure 2.9, presents indicative segmentation examples concerning the use of multichannel gradient magnitude versus the use of monochannel gradient computation used in [10]. Note that chromosome anomalies, shown in cases B and C, are not satisfactorily segmented using 1-D gradient by the watershed algorithm.
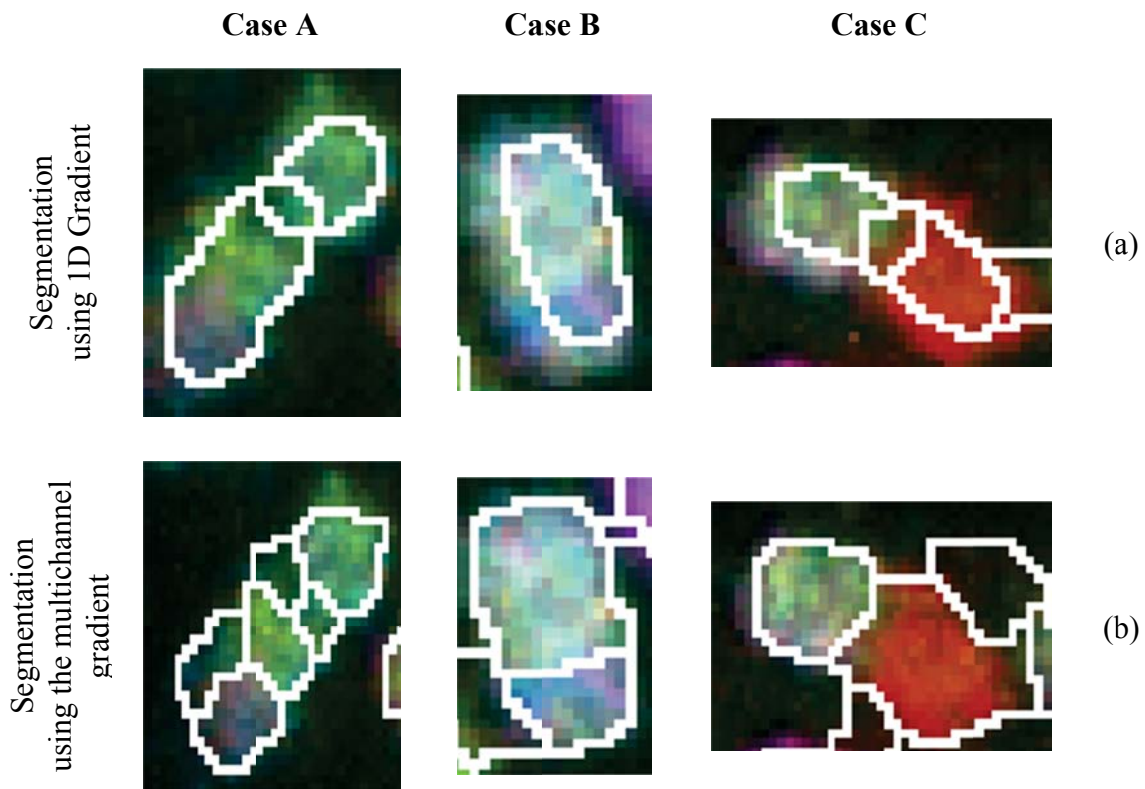
Figure 2.9: Three indicative cases related to the watershed segmentation map; produced using (a) 1D-gradient [27] of the DAPI channel and (b) the proposed multichannel gradient magnitude.

CHAPTER 3:

SUPERVISED REGION CLASSIFICATION

3.1   Introduction

Along with the emergence of M-FISH imaging, automated methods for the classification of this type of image were developed. Methods using pixel-by-pixel classification algorithms have been introduced in M-FISH analysis. These methods either classify each pixel of the M-FISH image [28], [29] or create a binary mask of the DAPI image using edge detection algorithms, and classify each pixel of the mask [30], [31]. A method for joint segmentation-classification of chromosome M-FISH images was presented [32] that build a probabilistic model of M-FISH chromosomes which allows for simultaneous segmentation and classification. The additional information provided by multiple spectra in chromosome images made it feasible to distinguish chromosomes that overlap and are in touch within clusters.

Pixel by Pixel classification techniques were the first methods which introduced for classifying the pixels of the M-FISH image. These techniques do not take into account neighborhood information, and also consider the background as an additional class. The goal is to develop a region classification method which will take into account spatial information.

39

This spatial information has been already available from the region segmentation step which was described in the previous Chapter. The region classification could also aid the cytogeneticist, since he usually is interested in chromosome regions instead of pixels.

3.2   Region Bayes Classification

Assume that a segmented region, produced by the multispectral watershed segmentation, consists of N pixels. Let R is the set of N feature vectors of the region in the image $R = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in \Re^5$, $i = 1, \ldots, N$. Let also $p(x|C_i) = N(\mu_i, \Sigma_i)$ be the distribution for class C_i. Then the class conditional likelihood $p(R|C_i)$ of region R is computed as [46], [47]:

$$
\begin{aligned}
p(R|C_i) &= p(x_1, x_2, \ldots, x_\ell | \omega_i) = \prod_{j=1}^{N} p(x_j | C_i) = \\
&\left(\frac{1}{(2\pi)^{5/2} |\Sigma_i|^{1/2}} \right)^N \exp\left(-\frac{1}{2} \sum_{j=1}^{N} (x_j - \mu_i)^T \Sigma_i^{-1} (x_j - \mu_i) \right)
\end{aligned}
\tag{3.1}
$$

The mean vectors and the covariance matrixes for each class are computed by a training phase, from an annotated set of M-FISH images as follows:

$$
\mu_i = \frac{1}{N_i} \sum_{x_k \in C_i} x_k, \ i = 1, \ldots, K,
\tag{3.2}
$$

$$
\Sigma_i = \frac{1}{N_i - 1} \sum_{x_k \in C_i} (x_k - \mu_i) \cdot (x_k - \mu_i)^T, \ i = 1, \ldots, K.
\tag{3.3}
$$

Working with the natural logarithm and dropping all terms that are the same for all classes, the Bayes decision rule assigns the set of region pixels R to class C_i if:

$$
\forall i \neq j, \ DS_i(R) > DS_j(R),
\tag{3.4}
$$

where

$$
DS_i(R) = -\frac{N}{2} \ln |\Sigma_i| - \frac{1}{2} \sum_{j=1}^{N} (x_j - \mu_i)^T \Sigma_i^{-1} (x_j - \mu_i) + \ln P(\omega_i).
\tag{3.5}
$$

The a priori class probabilities for each class $P(\omega_i)$, are computed using the training set, as the percentage of all chromosome pixels in the training data that belong to class ω_i:

$$P(\omega_i) = \frac{(\# \, pixels \, belong \, to \, class \, \omega_i)}{\sum\limits_{k=1}^{24}(\# \, pixels \, belong \, to \, class \, \omega_k)} \; . \tag{3.6}$$

It is well known that the chromosome class reflects the size of each chromosome in descending order (i.e. chromosome 1 is the largest and chromosome 22 is the smallest). From Figure 3.1, it is obvious that as the chromosome size decreases its a priori probability also decreases.
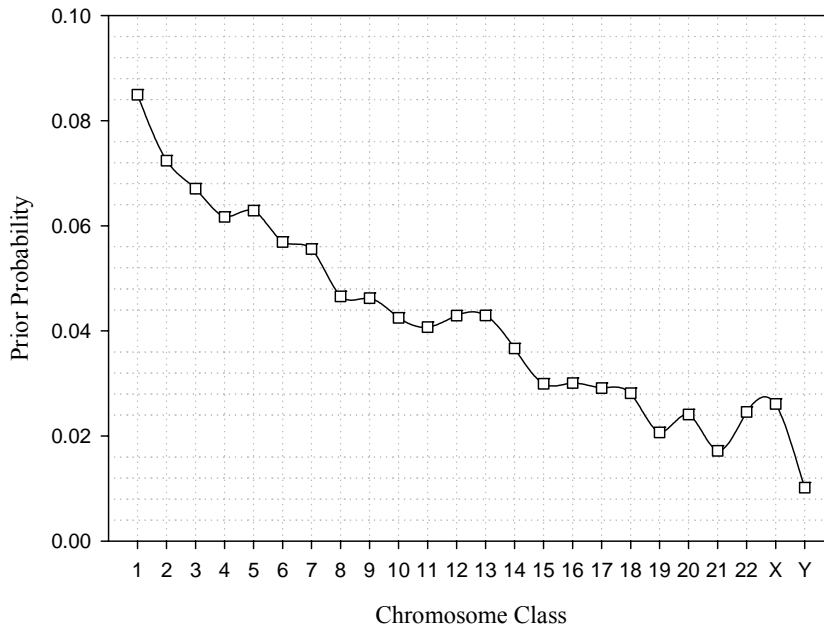
Figure 3.1: The a priori probabilities for the chromosome classes 1-22, X and Y.

3.3 Region Merging

After region classification there are still regions that could be merged resulting into a meaningful classification map based on the principle that adjacent regions of the same class could be merged to one region. Adjacency is a symmetric relationship which can be easily represented by the region adjacency graph (RAG), where two nodes (representing two distinct

41

regions) are connected if those two regions are adjacent in the image i.e. they have a common boundary. Consequently, all adjacent regions that share the same class are merged.

The application of the classification and merging step in an M-FISH image is shown in Figure 3.2. Initially the image is segmented with the multichannel watershed segmentation (Figure 3.2(b)) and then the segmented regions are classified based on the region Bayes classification method (Figure 3.2(c)). The final classification map after the region merging step is shown in Figure 3.2(d). A separate color was used to represent each chromosome class in the image.

(a)   (b)

(c)   (d)
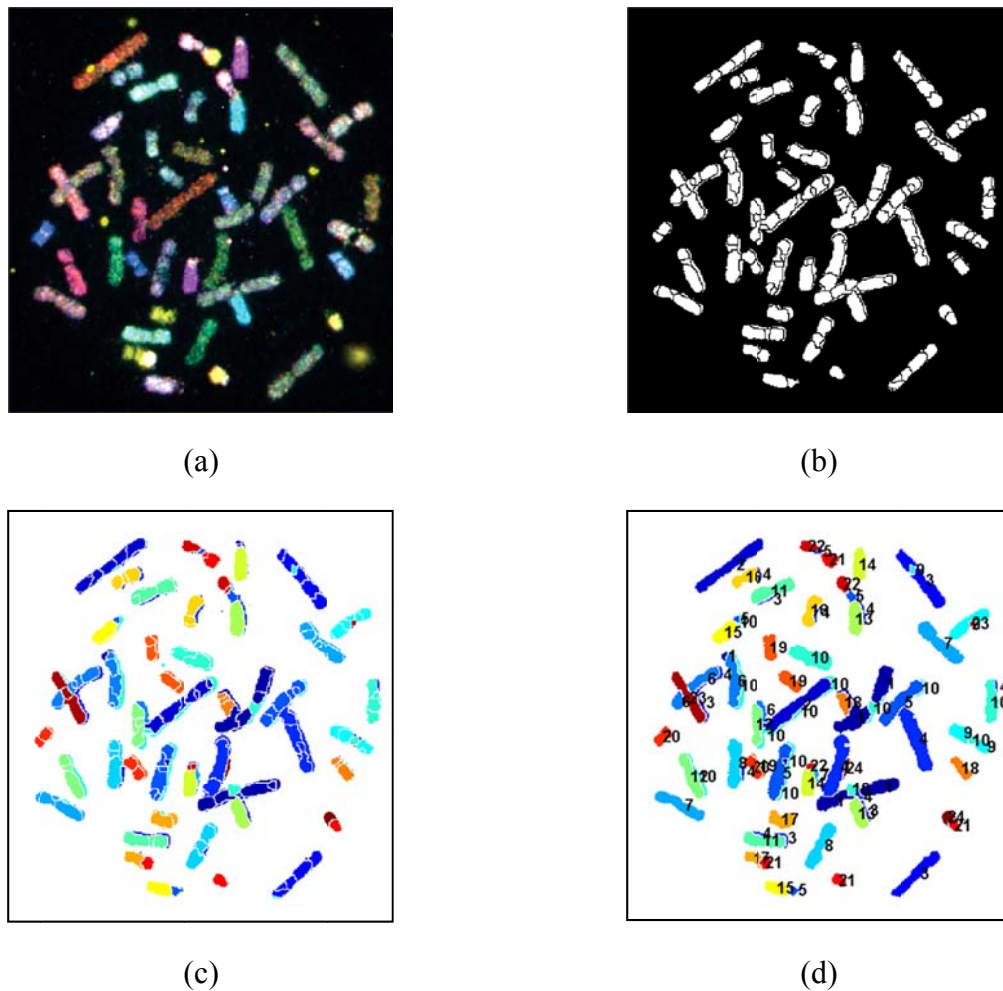
Figure 3.2: Example of an M-FISH image segmentation and classification: (a) original M-FISH image, (b) the segmented image using the multichannel watershed segmentation, (c) region classification, and (d) the final classification map using region merging.

42

3.4   Vector Median Filtering of M-FISH Images

In multichannel image filtering, each image pixel can be considered as a vector of features associated with the intensities of the channels. In order to filter multichannel images, the vector filtering algorithms are preferred [48], [49], [50]. We have chosen to filter M-FISH images using the vector median filtering (VMF) approach since the noise in M-FISH images is due to the microscope [49] (impulsive noise). Vector median filtering is the most popular and appropriate filtering technique for the removal of this type of noise since the impulse response of the VMF is zero [49].

In general, component wise (marginal) approaches produce new vector samples, i.e. color artefacts, caused by the composition of reordered channel samples. Vector filters represent a natural approach to the noise removal in multichannel images, since these filters utilize the correlation between color channels. The output of these filters is defined as the lowest ranked vector according to a specific ordering technique.

Suppose a square filter window with a set of input multichannel samples such that $X = \{x_i : i = 1, \ldots, N\}$, where $x \in \Re^5$ and N is an odd integer which represents the size of the window. Let us consider an input sample $x_i : 1 \leq i \leq N$, associated with the distance measure L_i and the angle distance A_i defined as [51], [50]:

$$L_i = \sum_{j=1}^{N} \left\| x_i - x_j \right\|_\gamma, \tag{3.7}$$

$$A_i = \sum_{j=1}^{N} \cos^{-1} \left(\frac{x_i x_j^T}{|x_i| |x_j|} \right), \tag{3.8}$$

where γ characterizes the employed norm, and $|\cdot|$ is the magnitude of the vector. Note that the well-known Euclidean distance corresponds to $\gamma = 2$ [49]. The ordering criterion is expressed using products of L_i and A_i [51]:

$$\Omega_i = L_i A_i = \sum_{j=1}^{N} \left\| x_i - x_j \right\|_\gamma \sum_{j=1}^{N} \cos^{-1} \left(\frac{x_i x_j^T}{|x_i| |x_j|} \right), \quad 1 \leq i \leq N \tag{3.9}$$

Then, the ordered set is given by, $\Omega_1 \leq \Omega_2 \leq \ldots \leq \Omega_N$. The same ordering scheme applied to the input set results in the ordered sequence, $x^{(\Omega_1)} \leq x^{(\Omega_2)} \leq \ldots \leq x^{(\Omega_N)}$. The sample $x^{(\Omega_1)}$ associated with Ω_1 represents the output of the directional distance filter (DDF). Let us assume the DDF with the power parameter p so that the power $1 - p$ is associated with the

sum of vector distances and the power $p \in [0,1]$ is associated with the sum of vector angles. Thus, Eq. (3.10) can be simply rewritten as:

$$\Omega_i = L_i^{1-p} A_i^p = \left(\sum_{j=1}^{N} \left\| x_i - x_j \right\|_\gamma \right)^{1-p} \left(\sum_{j=1}^{N} \cos^{-1} \left(\frac{x_i x_j^T}{|x_i| |x_j|} \right) \right)^p , \quad 1 \le i \le N \qquad (3.11)$$

If $p = 0$, the DDF operates as the vector median filter (VMF), whereas for $p = 1$, the DDF is equivalent to the basic vector directional filter (BVDF). The weighted vector median filter is defined through a set of weights. Assume a set of nonnegative integer weights $w_1, w_2, \ldots, w_N$ so that each weight $w_j, 1 \le j \le N$ is associated to each input sample z_j. Then, it is possible to express the weighted vector distance D_i as:

$$D_i = \sum_{j=1}^{N} w_j \left\| x_i - x_j \right\|_\gamma, \quad 1 \le i \le N . \qquad (3.12)$$

The sample $x^{(D_1)} \in \{x_1, x_2, \ldots, x_N\}$ associated with the minimal combined weighted distance D_1 is the sample which minimizes the sum of weighted vector distances and the output of the WVFM filter. The CWVMF [52], [53] framework is more adequate for adaptive filter design with validations of the smoothing levels in the filtering process. Consider that the weight vector is given by:

$$w_j = \begin{cases} N - 2k + 2, & j = (N+1)/2 \\ 1, & otherrwise \end{cases}, \qquad (3.13)$$

where $k = 1, \ldots, (N+1)/2$.

The above states that only the central weight $w_{(N+1)/2}$ associated with the central sample $x_{(N+1)/2}$ can be changed, whereas other weights associated with the neighboring samples remain equal to one [52]. If the smoothing parameter k is equal to one, then the CWVMF is equivalent to the identity operation and no smoothing is performed. In the case $k = (N+1)/2$, the maximum amount of smoothing is performed and the CWVMF filter is equivalent to the VMF.

3.5 Classification Results

To compute the classification accuracy four images were chosen randomly three times from the dataset and the test was performed with the remaining images. Thus three different training subsets (*Sub A*, *Sub B*, *Sub C*) were created. The training dataset consists of all chromosome classes and no overlap between the training and testing data exists. Also pixels belonging to two or more chromosomes (chromosome overlaps) were not considered for training and testing.

The proposed method was compared with a Bayes pixel-by-pixel classification technique [28], which is the main classification scheme for several related works in the literature. Pixel-by-pixel classification is performed for the pixels in the segmented regions of chromosomes. We have trained and evaluated both methods using the same training and testing set. The average chromosome classification accuracy obtained for each M-FISH training subset: *Sub A*, *Sub B*, and *Sub C* is shown in Table 3.1.

Table 3.1: Chromosome classification accuracy using the proposed method and a pixel-by-pixel classification method

SUBSETS	CLASSIFICATIONACCURACY	
	REGIONBASED	PIXEL-BY-PIXEL [7]
SUB A	82.2%($\pm$ 14.9%)	70.8%($\pm$ 16.2%)
SUB B	82.4%($\pm$ 14.8%)	70.6%($\pm$ 16.8%)
SUB C	82.6%($\pm$ 14.4%)	70.4%($\pm$ 16.5%)
OVERALL	**82.4%**	**70.6%**

The relationship between segmentation accuracy and region classification accuracy is shown in Figure 3.3. It should be mentioned, that the segmentation and classification stages are two independent methods. Therefore an almost perfect segmentation result cannot ensure the best classification accuracy.
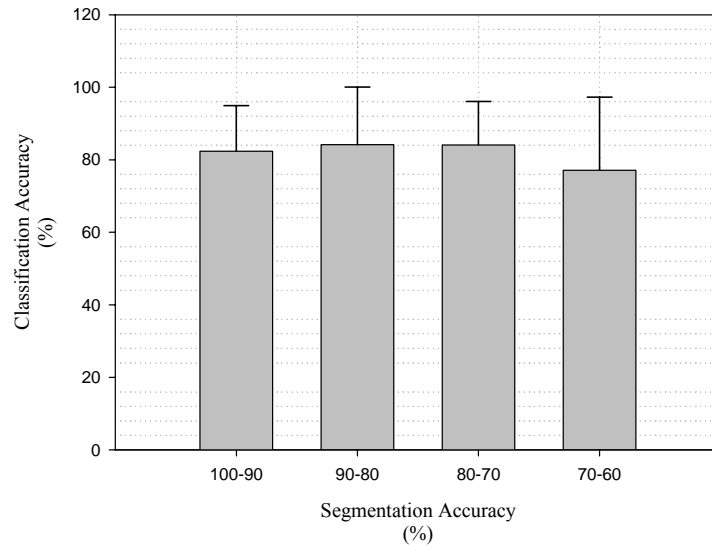
Figure 3.3: Region classification accuracy vs. segmentation accuracy.

Figure 3.4 depicts the classification accuracy difference between the proposed and the one presented in [28] for each chromosome class. From Figure 3.4(b) we can conclude that with the proposed method high accuracy is obtained for small chromosomes where the difference in classification performance increases.
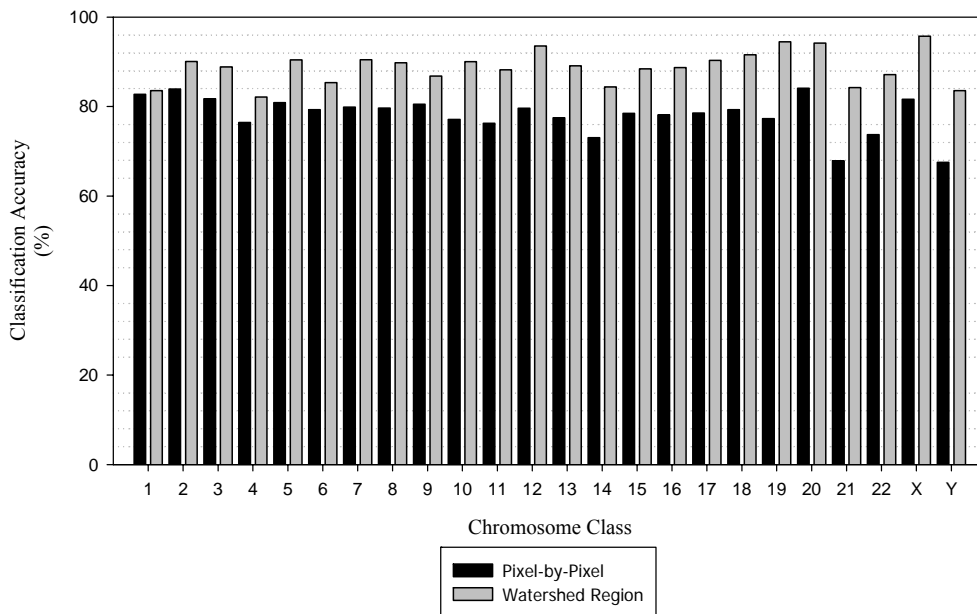
(a)

Figure 3.4: (a) Comparison of the classification accuracy of the two methods (Proposed vs. Pixel-by-Pixel) for each chromosome class.

46

3.5.1 *Vector Median Filtering Results*

We used the ground truth image in order to assess the performance of the region Bayes classifier for the different multichannel filtering schemes. Four images were chosen randomly five times from the database and the test was performed with the remaining images. Thus five different training subsets were created. The training dataset consists of all chromosome classes and no overlap between the training and testing data exists.

In order to measure the performance of the different multichannel filtering schemes we measured the chromosome classification accuracy of the region based classification using non filtered and filtered images. As it is shown in

Table 3.2, BVDF achieves the best results in terms of accuracy improving the classification accuracy by 3.34%. CWVMF attains the worst improvement by 2.47% while the DDF and VMF result in an improvement of 3.07% and 3.01%, respectively. The application of the different multichannel filtering schemes is illustrated in Figure 3.6.

(a)

(b)

| VMF | BVDF | DDF | WVMF |
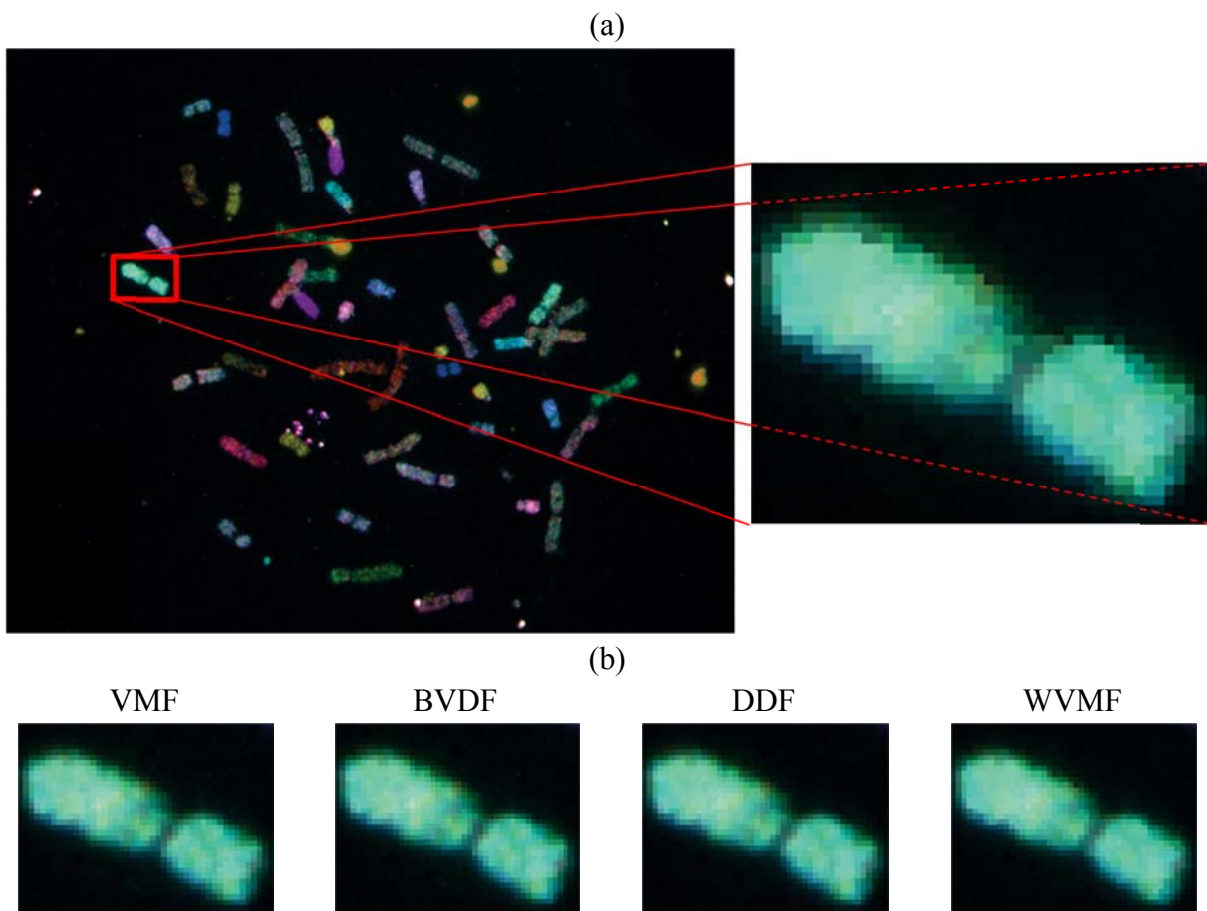
Figure 3.5: Application of different multichannel filtering enhancing schemes to an M-FISH image: (a) Original M-FISH image and (b) M-FISH image after the application of the different filtering schemes.

Table 3.2: Comparison of our research with other automated classification methods for M-FISH images.

#Dataset	Un filtered Images		Multichannel Filtering Schemes			
	Pixel Based Classification (%)	Region Based Classification (%)	VMF (%)	BVDF (%)	DDF (%)	CWVMF (%)
1	73.01	78.18	82.06	82.14	82.30	81.82
2	72.52	79.77	82.63	82.54	82.18	81.93
3	73.83	79.83	82.02	83.20	82.05	82.15
4	72.04	79.93	82.34	82.77	83.28	81.49
5	71.72	78.64	82.39	82.45	81.90	81.32
Overall	**72.62** ± **13.68**	**79.27** ± **13.04**	**82.28** ± **12.22**	**82.61** ± **11.12**	**82.34** ± **12.32**	**81.74** ± **12.44**

3.5.2 Pixel-by-Pixel Limitations

The method divides the M-FISH image into regions, i.e. groups of pixels which are assumed to be members of the same chromosome class. The method compares a set of pixels with the training class distributions instead of comparing a single feature vector (i.e. a pixel). Two indicative cases where the proposed method is superior compared to the pixel-by-pixel classification method are presented in Figure 3.6. In these two cases pixel-by-pixel classification produces noisy results making the decision of the expert difficult since these artefacts can be misinterpreted as chromosome abnormalities.
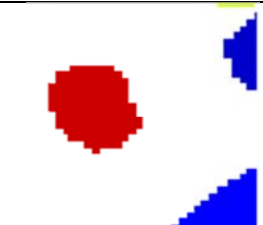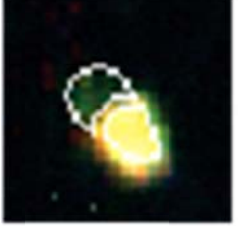
Figure 3.6: Examples where the pixel-by-pixel classification produces noisy results.

3.5.3 Influence of the classification accuracy due to region size

In order to describe the influence of the classification accuracy due to the region size several experiments were performed varying the tuneable parameter h as we have describe in the previous chapter. The mean classification accuracy was computed varying the values of h from 0-250 (with step 10) and is presented in Figure 3.7. Additionally this figure presents the region size versus h. The classification accuracy varies from 81% to 88%. Initially as h increases the classification accuracy increases but for values $h > 125$ it remains constant. This is due to the fact that when h is large, the mean size of the regions increases since regions with same spatial and color characteristics are merged.
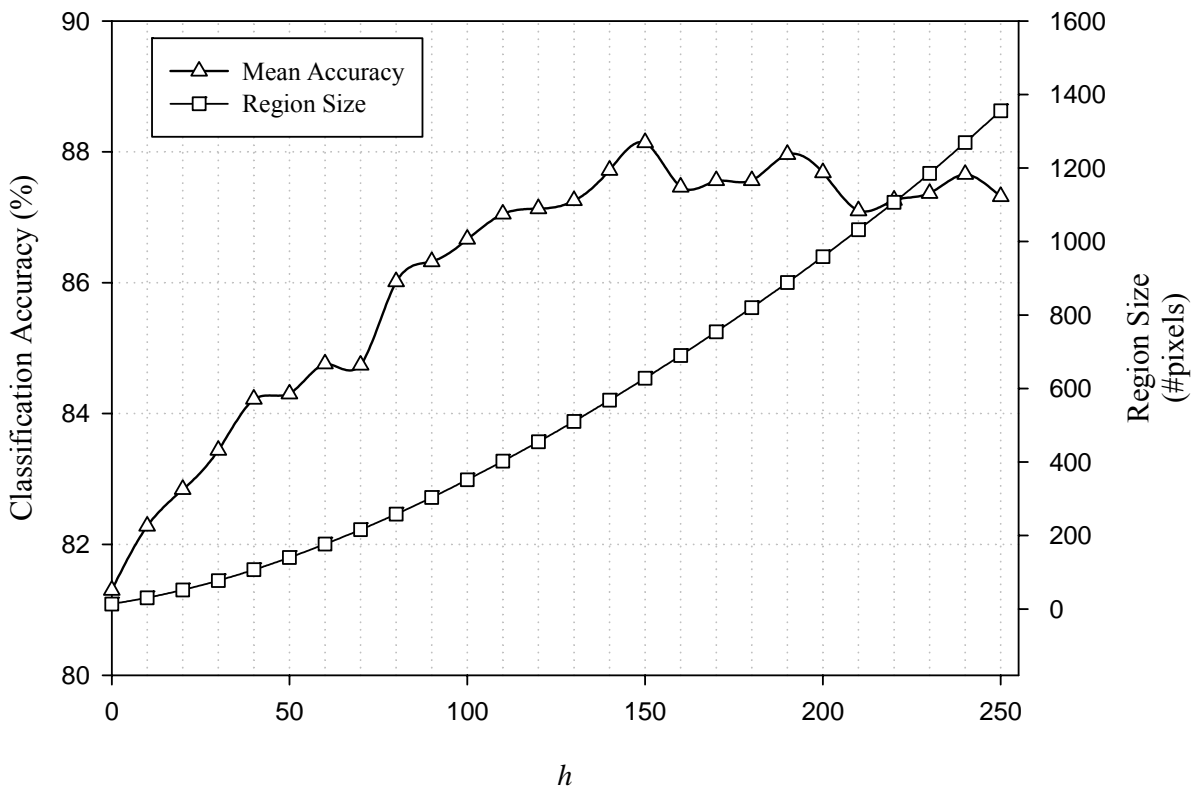
Figure 3.7: Classification accuracy vs. h, and region size vs. h.

3.6 Conclusions

If a classification is performed on a pixel-by-pixel basis, the classification will be dominated by noisy painting in homogeneities. This is obvious by the misclassifications errors produced by the pixel-by-pixel algorithm as it is shown in Figure 3.6. In contrary region-based

classification avoids these types of errors since pixels with similar spectral information contribute in the classification. Moreover, region-based classification provides better classification accuracy than the maximum a posterior pixel-by-pixel classifier. The increased classification accuracy of the proposed method in cases of small chromosomes is an important feature in clinical cytogenetics [54]. Small chromosomes are often involved in simple or complex rearrangements, either in genetic disorders or in cancer. In this way genetic analysis becomes more reliable and may explain unidentified aberrations in clinical cases.

Another important aspect is that the produced classification and segmentation map could be used as a decision support tool for cytogeneticists during their daily clinical practice. Figure 3.2(d), is an indicative example of the information that can be provided to the experts. It is noticeable that regions of the same class appear more than two times (e.g. regions of class "10") in the map. This can be interpreted either as a possible translocation, and thus its identification is very important, or as a classification error of the method; the final decision is made by the cytogeneticist.

CHAPTER 4:

FULLY UNSUPERVISED M-FISH CHROMOSOME

IMAGE CLASSIFICATION

4.1 Introduction

Unsupervised classification could be considered very important for the characterization of the M-FISH images. This is because of the relative high number of clusters (24 chromosome classes: 1-22, X,Y) and of the initialization of these cluster centres which is not a trivial task. Furthermore, there are significant chromosomes variations among M-FISH images. This could be explained due to the nature of the chromosome imaging since it is a biological experiment depending on a large number of parameters such as temperature and humidity of the place of the experiment. The goal was to try to model the biological problem by using a Gaussian Mixture Model. We accomplished that by representing the distribution of the greyscale values of the chromosome pixels by two Gaussians; one for the hybridized pixels and one for the non-hybridized pixels. Furthermore, we initialize the cluster centers by using

the emission information of each chromosome for each channel. This information is always available prior to the M-FISH experiment.

One of the key factors limiting the pixel classification accuracy is the variations between M-FISH images. This is due to the fact that the M-FISH imaging technique is not always accomplished under the same conditions e.g. humidity, temperature, type of microscope, color spread [2] and these factors affects the quality of the produced M-FISH image.

As one can see, the strength of absorption is not binary and varies widely across the chart. Both class 20 and class 3 are predicted to absorb Spectrum Orange, but Spectrum Orange is almost twice as strong in class 20. Also in this particular image set, the Cy5.5 fluorophore is weak; and its strength in classes that should absorb it is occasionally less than that of other dyes in classes that should not. Furthermore, the difference in magnitude of classes that should absorb Cy5.5 and classes that should not is not always great. The average magnitudes of Cy5.5 in classes 4 and 5 are nearly identical, although class 5 should bind Cy5.5, while class 4 should not. In addition, it is important to note that the characteristics in this table are valid only for this set of data, since fluorophore strength often varies by batch and by age of the fluorophore.

Furthermore, according to the color map of Kit-A, chromosome 3, should be ideally observed only in the Channel 1, 2, 4, and 5 and should not be visible in other channels. Figure 4.1 presents the color spread for two chromosomes of class 3 on two different M-FISH images (M-FISH$_1$, M-FISH$_2$). As one can observe the chromosome of Figure 4.1(a) presents a distribution of the fluor signal along the chromosome as the distribution described in the pattern of chromosome 3 (Figure 4.1(c)). However, Figure 4.1(b) presents the chromosome class 3 for a different M-FISH image where it is obvious that this chromosome class has failed to hybridize on channel 4 [2]. Specifically, the average fluor for chromosome 3 of the channel 4 has a similar distribution such as the one the channel 3 (Figure 4.1(b)). We can conclude that when the variation of the feature distribution across images is significant, which means the feature distribution of an unknown image is unpredictable, classification methods that rely on the estimation of class parameters (Supervised methods) will yield low accuracy [2].

M-FISH₁   M-FISH₂

(a)   (b)
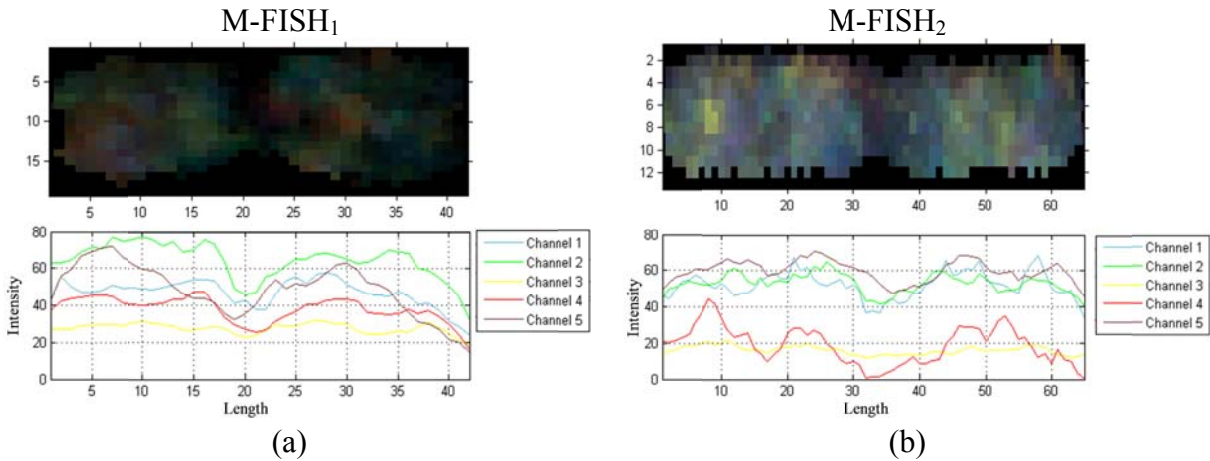
Chromosome 3	
Channel	Theoretic Pattern
1	1
2	1
3	0
4	1
5	1

(c)

Figure 4.1: Two chromosomes of class 3 of two different M-FISH images (M-FISH₁, M-FISH₂). (a) Chromosome class 3 and its graph of average fluor signal along the chromosome of M-FISH₁. (b) Chromosome class 4 and its graph of average fluor signal along the chromosome of M-FISH₂. (c)Theoretic emission for each of the channels of chromosome class 3.

The method consists of four different stages. In the first stage (*Chromosome Mask*) the segmentation of the DAPI channel takes place using Otsu's threshold selection method. Using the binary image produced in the previous stage we extract all the pixels that belong to chromosomes. The *Region Segmentation* stage decomposes the M-FISH image into regions using the Watershed transform: the gradient magnitude of the multispectral image is computed. The goal of the next stage (*Region Characterization*) is the classification of each region of the M-FISH image. In order to achieve this we first estimate the parameters of Single Channel Gaussian Mixture Models. Each Single Channel GMM is used to describe the probability density of chromosome pixels from each channel of the M-FISH image. From this GMM we estimate the parameters (mean, covariance, mixture coefficient) of each chromosome class for a Multichannel Gaussian Mixture Model. Finally, we make use of the

MAP-EM algorithm in order to adapt the parameters of the Multichannel GMM for each M-FISH image. Having computed the class parameters, all chromosome regions are classified into 1-24 chromosome classes. The final stage (*Region Merging*) is used to merge the classified regions producing a final classification map for the cytogeneticist.
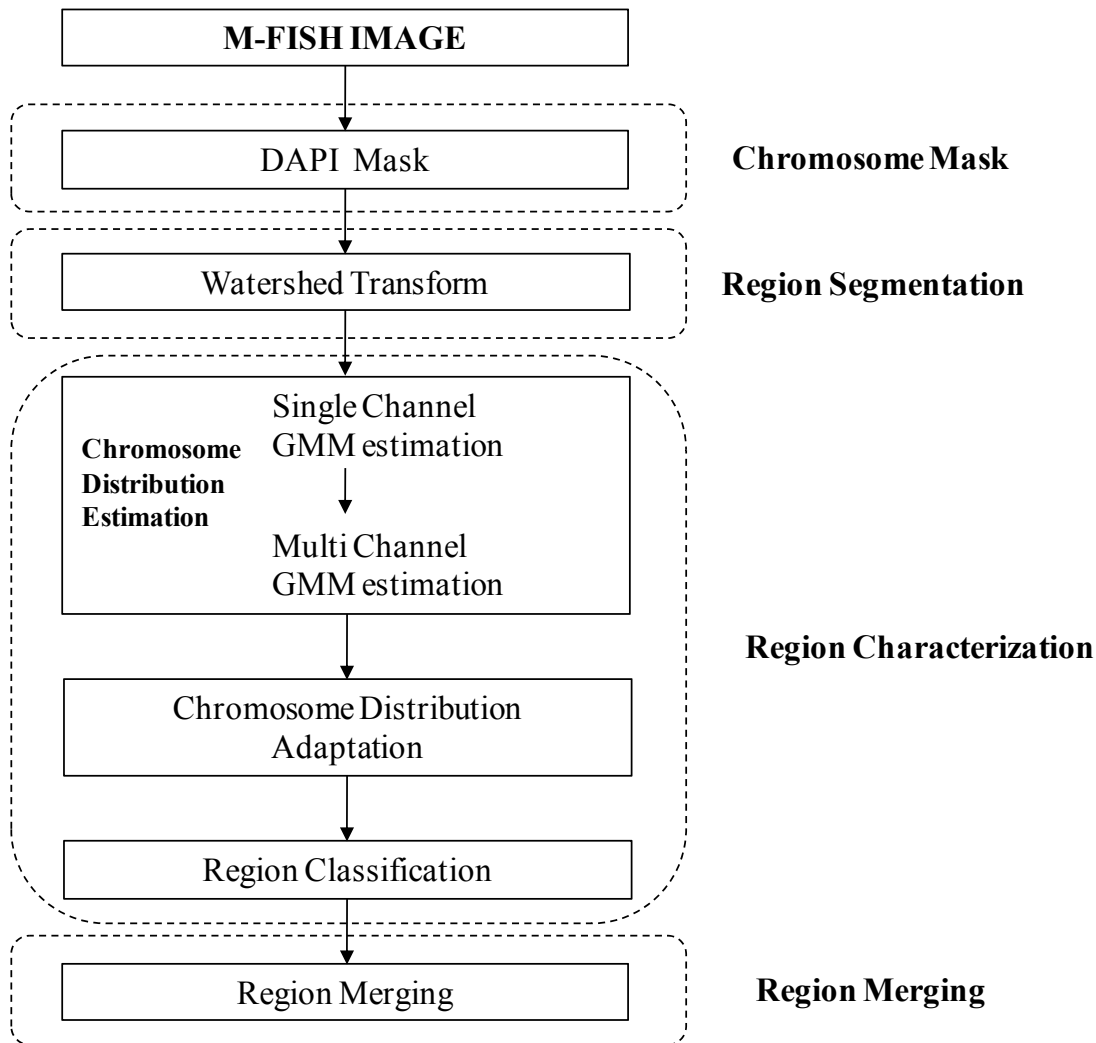
Figure 4.2: Flowchart of the proposed method

The main contribution is the proposal of a new fully unsupervised classification method for the M-FISH images. The unsupervised scheme is based on a multichannel Gaussian Mixture Model (GMM) with 24 components, one for each chromosome class. In order to overcome the problem of the initialization of the parameters of each GMM component, the first choice would be to randomly select parameter values or perform an initial clustering of the dataset. However, this would not incorporate the emission information of each chromosome class that we *a priori* had. In our case the initialization is based on estimating

the parameters of a single Gaussian Mixture Model via the Expectation Maximization (EM) algorithm. Combining the parameters of the single channel GMMs and the emission information of each chromosome class an effective initial estimation of the multichannel GMM can be derived.

Although this initial GMM could accurately describe the chromosome distribution for an M-FISH image with high Signal to Noise Ratio (SNR), this does not hold for images of low SNR. Thus, a further adaptation of the parameters of the multichannel GMM is needed. One straightforward approach is to use the classical EM algorithm starting from the initial estimation computed in the previous step. However, the large number of chromosome classes and the emission overlap could affect the mapping between each mixture component and each chromosome class. To overcome this problem, some kind of constraints should be applied on the GMM parameters. These constraints are naturally imported in the MAP-EM algorithm, which was used in the method. The estimation of the parameters using the MAP-EM method proves to be more effective in terms of classification accuracy over the classical EM algorithm. Furthermore, apart from effectively using the GMMs, the proposed method uses the Watershed transform in order to segment the M-FISH image into regions. It has been already shown [21] that by classifying regions instead of pixels a significant increase in the classification accuracy is obtained. Finally, the proposed method presents high classification accuracy, without any user interaction, even when compared to reported results using supervised classification methods.

4.2 Unsupervised M-FISH Image Classification – Literature Review

Although the M-FISH imaging ease the process of karyotyping [55], [3] visual inspection of these images is a laborious and time-consuming process. Also the characterization of chromosome anomalies is difficult since small rearrangements of chromosome material are difficult to identify for untrained personnel. For this reason many attempts have been to automate the whole or all part of the classification of M-FISH images process [26], [27], [28], [56], [29], [30], [31], [32], [21], [57], [58], [21], [2], [59].

The methods described in the literature either first segment the image and then they classify the pixels of the image or they directly classify all the pixels of the M-FISH image including a class for the background. In addition, these methods can be divided into two categories based on the use of images for training set:

- **Supervised** [28], [29], [30], [31], [32], [21]: in a typical supervised classification scheme, the goal is to train a classifier that can be used to predict previously unseen images. These methods use a small number of images from the dataset to train the classifier (e.g. Bayes [28]) and then they test its performance to rest dataset.

- **Unsupervised** [57], [58], [2]: these methods do not use a set of M-FISH images to train the classifier. They directly classify the M-FISH image using prior knowledge such as the emission information of each chromosome class (see Appendix A). Although, there have been proposed methods for unsupervised classification of M-FISH image they have been tested only to a small number of M-FISH images and they present low classification accuracy.

4.3 Chromosome Mask & Region Segmentation

The first step of the method is the segmentation of the DAPI channel of the M-FISH image. Otsu's method [23] computes automatically an optimal threshold value l^* by maximizing the between class variance. An example of the segmentation of the DAPI channel of an M-FISH image is shown in Figure 4.3.

(a)   (b)
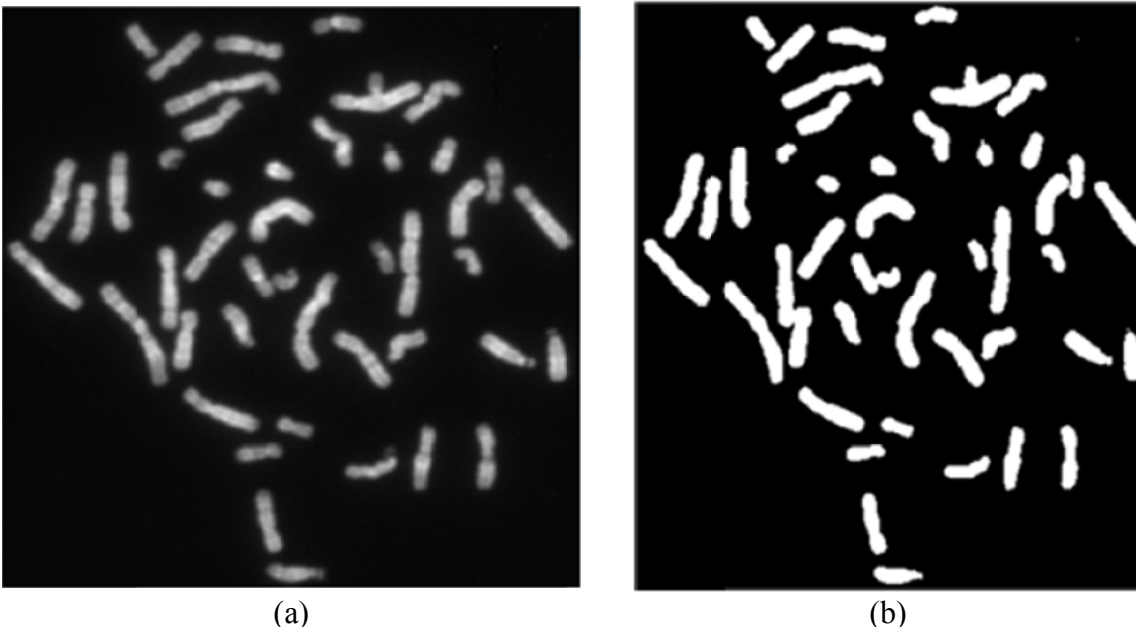
Figure 4.3: The segmentation mask of a DAPI image. (a) The DAPI image and (b) the segmentation mask B_{Otsu}.

As a result of this task the image pixels are classified as foreground pixels (called chromosome pixels) or background pixels. Suppose that the segmented image B_{Otsu} contains N chromosome pixels and let X be the set of these chromosome pixels $X = \{x_1, x_2, \ldots, x_N\}$. Each chromosome pixel is associated with a five dimensional vector x_k, $k = 1, \ldots, N$ containing the corresponding intensities of each of the five M-FISH image channels $x_k = \left(x_k^1, x_k^2, x_k^3, x_k^4, x_k^5\right)^T$. The intensities of the chromosome pixels are normalized using the standard method:

$$y_k^j = \frac{x_k^j - \mu^j}{\sigma^j},$$
$$k = 1, \ldots, N, j = 1, \ldots, 5$$

(4.1)

where μ^j, σ^j the mean and standard deviation of the chromosome pixels of the channel j. Thus the set $Y = \{y_1, y_2, \ldots, y_N\}$ is obtained (where $y_k \in \Re^5, k = 1, \ldots, N$ containing the corresponding normalized intensities of each of the five M-FISH image channels).

The Watershed Transform (WT) is a widely image segmentation algorithm that originated from the field of mathematical morphology. The image is considered as a topographical relief, where the height of each pixel is related to its grey level. Imaginary rain falls on the terrain. The watersheds are the lines separating the catchment basins [16]. In order to be able to apply the WT in a multichannel image one has first to define the gradient. Instead of separately computing the scalar gradient from each channel of the image we computed the tensor gradient [35], [36].

The output of the watershed transform for an image I is a tessellation T_I of the image into its different regions R_I, $1 \le i \le NR_i$ each one characterized by a unique label l_i:

$$T_I = \left\{\left(R_1, l_1\right), \left(R_2, l_2\right), \ldots, \left(R_{NR_I}, l_{NR_I}\right)\right\},$$

(4.2)

where NR_I is the number of regions.

(a)

(b)
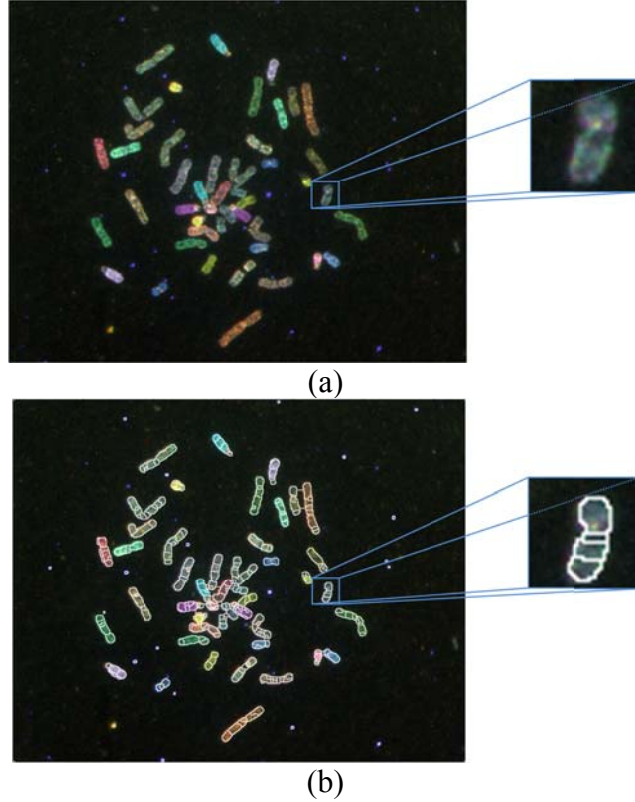
Figure 4.4: Watershed segmentation of an M-FISH image: (a) The M-FISH image, (b) watershed regions superimposed on the M-FISH image.

4.4 Chromosome Distribution Estimation

Each chromosome pixel $y \in \Re^5$ is associated with a five dimensional vector, containing the normalized intensities of each of the five M-FISH image channels. We wish to partition the chromosome pixels into $K = 24$ groups equal to the number of chromosome classes. This can be done using any clustering algorithm. The approach is based on Gaussian Mixture Model in which one assumes that the pixels were sampled from multiple Gaussian components such that each component corresponds to one chromosome class. The assignment of pixels to classes can then be easily performed by computing the posterior probability of a pixel to a class. Having described all that, the distribution of the chromosome pixels y could be modeled using a multidimensional Gaussian Mixture Model (called multichannel GMM) with K components, one for each chromosome class:

$$p(\mathrm{x}) = \sum_{i=1}^{K} \pi_i \cdot N(\mathrm{x}; \mu_i, \Sigma_i),$$ (4.3)

where π_i is the mixing probability or the prior probability of the i-th chromosome class, $N(y; \mu_i, \Sigma_i)$ is a Gaussian distribution with $\mu_i \in \mathfrak{R}^5$ the mean and Σ_i the 5×5 covariance of the i-th chromosome class, $i = 1, \ldots, K$. A standard approach to learn the parameters of each chromosome class ($\pi_i, \mu_i, \Sigma_i, i = 1, \ldots, K$) is the EM algorithm. However, a common problem of the aforementioned approach (EM algorithm) is the initialization of the parameters of each chromosome class. This is due to the fact that the EM algorithm converges to a local maximum of the likelihood that highly depends on the initial parameter values. Next, we describe a novel process for the initialization of the Multichannel GMM.

4.4.1   Initial Chromosome Distribution Estimation

At first we compute initial values ($\tilde{\pi}_i, \tilde{\mu}_i, \tilde{\Sigma}_i, i = 1, \ldots, K$) for the parameters of each chromosome class. The initialization step is composed of two procedures:

A. *Single channel GMM estimation*: For each channel a two Gaussian Mixture Model (GMM) is estimated: the first component corresponds to pixels belonging to hybridized chromosomes (brighter in the image) and the second component corresponds to pixels belonging to non-hybridized chromosomes (darker in the image).

B. *Multichannel GMM initialization*: Using the above single channel GMMs, we compute an initialization of the 24-component Multichannel GMM that models the distributions over the 5-D M-FISH image space.

4.4.1.1   Single Channel GMM estimation

Each channel j of the M-FISH image (except the DAPI channel) contains chromosome pixels that either belong to hybridized chromosome or to non-hybridized chromosomes. Thus the distribution of the grey scale values y_k^j is modeled using a mixture of two Gaussians [2]:

$$p(y_k^j) = p(y_k^j \mid C_h^j) P(C_h^j) + p(y_k^j \mid C_{nh}^j) P(C_{nh}^j), \qquad (4.4)$$

where y_k^j is the intensity of a chromosome pixel k in channel j, $p(y_k^j \mid C_h^j)$ is the probability of pixel k to be hybridized in channel j, $p(y_k^j \mid C_{nh}^j)$ is the probability of pixel k to be non-hybridized in channel j and C_h^j, C_{nh}^j are the hybridized and non-hybridized

classes, respectively. Also, $p(y_k^j | C_h^j) \sim N(\mu_h^j, (\sigma_h^j)^2)$ and $p(y_k^j | C_{nh}^j) \sim N(\mu_{nh}^j, (\sigma_{nh}^j)^2)$, μ_h^j, σ_h^j are the mean intensity and standard deviation of the hybridized pixels of channel j, respectively, whereas μ_{nh}^j, σ_{nh}^j are the mean intensity and standard deviation of the non-hybridized pixels of channel j, respectively. Finally, $P(C_h^j)$ and $P(C_{nh}^j)$ are the prior probabilities for the hybridized and non-hybridized classes, for channel j, respectively, that sum to 1.

For the estimation of the parameters $\left\{\left\{ \mu_h^j, \sigma_h^j, \mu_{nh}^j, \sigma_{nh}^j, P(C_h^j), P(C_{nh}^j) \right\}, j = 1,\ldots 5 \right\}$ of the GMMs we employed the well EM algorithm [60]. The EM algorithm is an iterative algorithm which at each iteration consists of two steps, the expectation (*E-step*) and the maximization step (*M-step*):

E-step: Given the estimation of parameters at iteration t, denoted as $\left\{\left\{ \mu_h^{j(t)}, \sigma_h^{j(t)}, \mu_{nh}^{j(t)}, \sigma_{nh}^{j(t)}, P(C_h^{j(t)}), P(C_{nh}^{j(t)}) \right\}, j = 1,\ldots 5 \right\}$, we define the sufficient statistics as:

$$P^{(t)}(C_h^j | y_k^j) = \frac{P^{(t)}(C_h^j) \cdot p^{(t)}(y_k^j | C_h^j)}{P^{(t)}(C_h^j) \cdot p^{(t)}(y_k^j | C_h^j) + P^{(t)}(C_{nh}^j) \cdot p^{(t)}(y_k^j | C_{nh}^j)}, \tag{4.5}$$

$$P^{(t)}(C_{nh}^j | y_k^j) = \frac{P^{(t)}(C_{nh}^j) \cdot p^{(t)}(y_k^j | C_{nh}^j)}{P^{(t)}(C_h^j) \cdot p^{(t)}(y_k^j | C_h^j) + P^{(t)}(C_{nh}^j) \cdot p^{(t)}(y_k^j | C_{nh}^j)}. \tag{4.6}$$

M-step: Update the parameters using:

$$\Lambda_h = \sum_{k=1}^{N} P^{(t)}(C_h^j | y_k^j), \tag{4.7}$$

$$\Lambda_{nh} = \sum_{k=1}^{N} P^{(t)}(C_{nh}^j | y_k^j), \tag{4.8}$$

$$P^{(t+1)}(C_h^j) = \frac{\Lambda_h}{N}, \tag{4.9}$$

$$P^{(t+1)}(C_{nh}^j) = \frac{\Lambda_{nh}}{N}, \tag{4.10}$$

$$\mu_h^{j(t)} = \frac{\sum_{k=1}^{N} P^{(t)}(C_h^j | y_k^j) \cdot y_k^j}{\Lambda_h}, \tag{4.11}$$

$$\mu_{nh}^{j^{(t)}} = \frac{\sum_{k=1}^{N} P^{(t)}(C_{nh}^{j} \mid y_{k}^{j}) \cdot y_{k}^{j}}{\Lambda_{nh}} , \qquad (4.12)$$

$$\sigma_{h}^{j^{(t)}} = \sqrt{\frac{1}{\Lambda_{h}} \sum_{k=1}^{N} P^{(t)}(C_{h}^{j} \mid y_{k}^{j}) \cdot (y_{k}^{j} - \mu_{h}^{j^{(t)}})^{2}} , \qquad (4.13)$$

$$\sigma_{nh}^{j^{(t)}} = \sqrt{\frac{1}{\Lambda_{nh}} \sum_{k=1}^{N} P^{(t)}(C_{nh}^{j} \mid y_{k}^{j}) \cdot (y_{k}^{j} - \mu_{nh}^{j^{(t)}})^{2}} . \qquad (4.14)$$

The EM algorithm is a local optimization algorithm, thus, it is sensitive to initial values of the parameters. In order to overcome this problem, we are going to exploit prior information about the problem. More specifically for the priors $P(C_{h}^{j})$ and $P(C_{nh}^{j})$ we exploit the emission information about the chromosome classes (Appendix A). It is well known that the chromosome class index reflects the size of each chromosome in descending order (i.e., chromosome 1 is the largest and chromosome 22 is the smallest [22]). Thus, it is easy to estimate the proportion A_i of pixels belonging to chromosome class i either from a small set of M-FISH images [22] or to define it using medical knowledge [8]. Consequently, we can define the initial prior $P^{(t=0)}(C_{h}^{j})$ as the proportion of the pixels that belong to the hybridized pixels for channel j as:

$$P^{(t=0)}(C_{h}^{j}) = \frac{\sum_{i=1}^{K} \Phi_{ij} A_{i}}{\sum_{i=1}^{K} A_{i}} , \qquad (4.15)$$

where the matrix Φ is defined as (Appendix A):

$$\Phi_{ij} = \begin{cases} 1, & if\ chromosome\ class\ i\ emmits\ on\ channel\ j \\ 0, & otherwise \end{cases} , \qquad (4.16)$$

and $P^{(t=0)}(C_{nh}^{j}) = 1 - P^{(t=0)}(C_{h}^{j})$.

Since the data have been normalized (Eq. (2)) the initial values for mean and standard deviation for the components of non-hybridized pixels are set to -1 and 1 respectively, and the mean and standard deviation for the components of the hybridized pixels are set to 1:

The initial values for mean and standard deviation of non-hybridized pixels are -1 and 1 respectively and the mean and standard deviation of the hybridized pixels are 1 and 1 respectively:

$$\mu_h^{j^{(t=0)}} = 1, \sigma_h^{j^{(t=0)}} = 1, \tag{4.17}$$

$$\mu_{nh}^{j^{(t=0)}} = -1, \sigma_{nh}^{j^{(t=0)}} = 1. \tag{4.18}$$

4.4.1.2 *Multichannel GMM estimation*

The goal of this step is to compute an initial estimation of the chromosome class parameters ($\tilde{\pi}_i, \tilde{\mu}_i, \tilde{\Sigma}_i, i = 1, \dots, K$) of the multichannel GMM using the emission information that we are going to exploit for each chromosome class.

First, we have to take into consideration some key points about the emission information of each chromosome class. If a chromosome of class C_i emits in channel j then the probability of a pixel k belonging to that chromosome would be equal to the probability $p(y_k^j | C_h^j)$ of that pixel to be hybridized in this channel. Vice versa, if a chromosome of class C_i does not emit in channel j then the probability of a pixel which has been computed from the single channel GMM belonging to that chromosome would be equal to the probability of that pixel to be non-hybridized $p(y_k^j | C_{nh}^j)$ in channel j. Based on the above, we can define the probability $\tilde{p}(y_k^j | C_i)$ of a chromosome pixel y_k^j in channel j belonging to chromosome class C_i using the emission matrix Φ as:

$$\tilde{p}(y_k^j | C_i) = \Phi_{ij} p(y_k^j | C_h^j) + (1 - \Phi_{ij}) p(y_k^j | C_{nh}^j), \tag{4.19}$$

where C_i, $1 \le i \le K$ are the 24 chromosome classes and $1 \le j \le 5$.

For example the probability that a chromosome pixel of channel $j = 4$ belonging to chromosome C_1 is:

$$\tilde{p}(y_k^4 | C_1) = \Phi_{14} p(y_k^4 | C_h^4) + (1 - \Phi_{14}) p(y_k^4 | C_{nh}^4). \tag{4.20}$$

Using the emission matrix Φ (from appendix A, we can see that chromosome 1 emits in channel 4 thus $\Phi_{14} = 1$) and substituting in Eq. (24) we get:

$$\tilde{p}(y_k^4 | C_1) = p(y_k^4 | C_1). \tag{4.21}$$

62

Using this assumption we are going to derive initial estimation for the prior, mean and covariance parameters of each class distribution ($\tilde{\pi}_i, \tilde{\mu}_i, \tilde{\Sigma}_i, i = 1, \ldots, K$). The initial mean $\tilde{\mu}_i = \left(\tilde{\mu}_i^1, \tilde{\mu}_i^2, \tilde{\mu}_i^3, \tilde{\mu}_i^4, \tilde{\mu}_i^5 \right)$ of each chromosome distribution, $1 \leq i \leq K$ is computed by estimating each $\tilde{\mu}_i^j, 1 \leq j \leq 5$ using Eq. (23):

$$\tilde{\mu}_i^j = \Phi_{ij} \mu_h^j + (1 - \Phi_{ij}) \mu_{nh}^j, j = 1, \ldots, 5.. \tag{4.22}$$

Furthermore the initial covariance $\tilde{\Sigma}_i$ of each chromosome's distribution, $1 \leq i \leq K$, is computed using Eq. (4.23) as:

$$\tilde{\Sigma}_i = diag\left(\left(\sigma_i^1\right)^2, \left(\sigma_i^2\right)^2, \left(\sigma_i^3\right)^2, \left(\sigma_i^4\right)^2, \left(\sigma_i^5\right)^2 \right) \tag{4.24}$$

where $\left(\sigma_i^j\right)^2 = \left(\Phi_{ij}\right)^2 \left(\sigma_h^j\right)^2 + \left(1 - \Phi_{ij}\right)^2 \left(\sigma_{nh}^j\right)^2, \ j = 1, \ldots, 5$.

Finally, the prior probability is the expected proportion of pixels belonging to a specific chromosome [22]:

$$\tilde{\pi}_t = A_i, i = 1, \ldots, K . \tag{4.25}$$

4.4.2 Chromosome Distribution Adaptation

Having estimated the initial parameters for the multichannel GMM, the goal is to allow the parameters to be further adapted. Instead of estimating the GMM parameters via the EM algorithm, we employ Maximum A Posteriori (MAP) for parameter estimation, since we already have incorporated prior knowledge (such as the emission information) to the initial model. More specifically we have used the MAP-EM algorithm [61] which exploits prior information. Like the EM algorithm, the MAP estimation is a two-step estimation process. The first step is identical to the "Expectation" step of the EM algorithm. Unlike the M-step of the EM algorithm, in the M-step of MAP-EM, parameters are obtained by taking into account the initial model.

The parameters $\left\{ \pi_1^{(t)}, \mu_1^{(t)}, \Sigma_1^{(t)}, \ldots, \pi_K^{(t)}, \mu_K^{(t)}, \Sigma_K^{(t)} \right\}$ are initialized using the values from the previous step e.g.:

$$\left.\begin{array}{l} \pi_i^{(t=0)} = \tilde{\pi}_i \\ \mu_i^{(t=0)} = \tilde{\mu}_i \\ \Sigma_i^{(t=0)} = \tilde{\Sigma}_i \end{array}\right\}, 1 \leq i \leq K . \tag{4.26}$$

The MAP-EM estimation of the GMM parameters is described below:

E-step: In the E-step similarly to the previous case, we calculate the sufficient statistics, which given the estimation of parameters at iteration t $\{\pi_1^{(t)}, \mu_1^{(t)}, \Sigma_1^{(t)}, \ldots, \pi_K^{(t)}, \mu_K^{(t)}, \Sigma_K^{(t)}\}$, they are calculated as:

$$v_i^{(t)}(y_k) = \frac{\sum_{k=1}^{N} p(y_k \mid C_i)}{\sum_{j=1}^{K} \sum_{k=1}^{N} p(y_k \mid C_j)}, \tag{4.27}$$

$$E_i^{(t)}(y_k) = \frac{\sum_{k=1}^{N} y_k p(y_k \mid C_i)}{\sum_{k=1}^{N} p(y_k \mid C_i)}, \tag{4.28}$$

$$E_i^{(t)}(y_k y_k^T) = \frac{\sum_{k=1}^{N} y_k y_k^T p(y_k \mid C_i)}{\sum_{k=1}^{N} p(y_k \mid C_i)}, \tag{4.29}$$

M-step: The M-step is described by the following update equations:

$$\pi_t^{(t+1)} = (1-a)v_i^{(t)}(y_k) + \alpha \tilde{\pi}_i, \tag{4.30}$$

$$\mu_i^{(t+1)} = (1-\beta)E_i^{(t)}(y_k) + \beta \tilde{\mu}_i, , \tag{4.31}$$

$$\Sigma_i^{(t+1)} = (1-\gamma)E_i^{(t)}(y_k y_k^T) + \gamma \left[\tilde{\Sigma}_i + \tilde{\mu}_i \tilde{\mu}_t^{\mathrm{T}} \right] - \mu_i^{(t+1)} \left[\mu_i^{(t+1)} \right]^T, \tag{4.32}$$

where the parameters a, β and γ are the learning rates which define how confident we are about the prior values $\tilde{\pi}_i$, $\tilde{\mu}_i$, $\tilde{\Sigma}_i$, $i = 1, \ldots, K$.

Note that in the M-step the updates of the parameters Eqs. (4.33)-(4.34) are made using a combination of the updates suggested by the typical EM (first term of the sum) and the initial model we have computed in the initialization phase (4.4 Chromosome Distribution Estimation).

4.5 Region Classification & Merging

Having estimated the parameters $(\pi_i, \mu_i, \Sigma_i, 1 \le i \le K)$ of each chromosome class we could classify a pixel $y \in \Re^5$ to a chromosome class C_i, $1 \le i \le K$ using the posterior probability $P(C_i \mid y)$ using the Region Bayes classifier described in previous chapter.

The next step of the method is region merging. Note that there are still regions that could be merged resulting into a meaningful classification map. First, adjacent regions that share the same regions class label are merged into one single region. In order to connect the adjacent regions, the Region Adjacency Graph (RAG) is computed for the image [17]. The RAG is a graph where two nodes (representing two regions) are connected if those two regions are adjacent in the image. Thus, each region of the image is connected to all regions that share the same class. An example is shown in Figure 4.5(a)-(b).

| (a) | (b) |
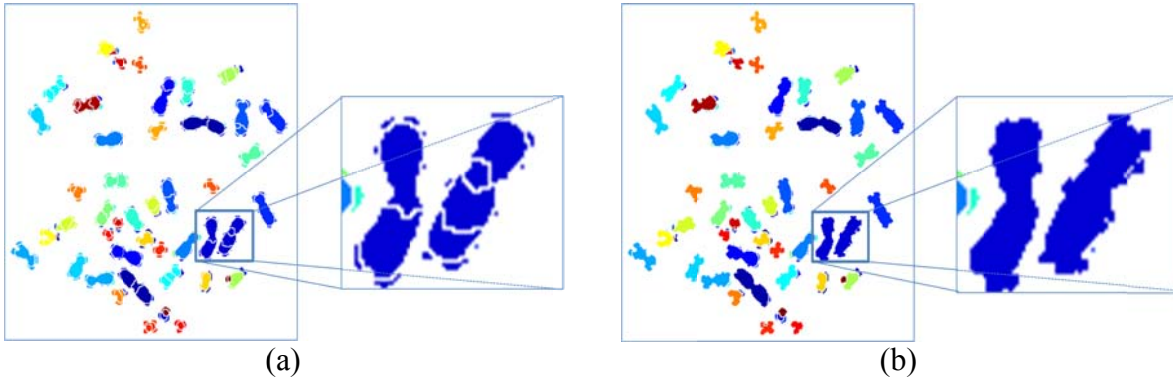
Figure 4.5: Region merging of chromosomes. (a) Initial classification map (we have used a separate color for each chromosome class), (b) Region merging of adjacent regions of the same class.

4.6   Small Region Merging

It has been observed that small regions are often misclassified by the region classification step. In order to tackle this problem we have adopted the following procedure. First, all small regions are identified. We define a small region as the region whose number of pixels is lower than a threshold. This threshold must be lower than the number of pixels of the smallest chromosome of the database and must be adapted for each M-FISH image $M_i, i = 1,\ldots,F$ of the database (where F is the total number of images of the M-FISH database). We define P_i the number of chromosome pixels of the i-th M-FISH image. Let's assume that the M-FISH image with the smallest chromosome is M_s and SCP the number of pixels of this chromosome. Thus, the proportion of pixels for the smallest chromosome for this image will be:

$$T_s = \frac{SCP}{P_s}.$$

(4.35)

where P_s, is the number of chromosome pixels of the image containing the smallest chromosome and T_s is the proportion of the pixels for the smallest chromosome of this image.

Finally, we compute the pixel threshold T_i for each M-FISH image $M_i, i = 1, \ldots, F$ of the database as:

$$T_i = T_s \cdot P_i. \tag{4.36}$$

Assume now a small region R of an M-FISH image where the $\# pixels\, of\, region\, R \leq T_i$. We can use the RAG to select the neighbors of R: $N_R = \{R_1, R_2, \ldots, R_G\}$. Let C_i the class of region $R_i \in N_R, i = 1, \ldots, G$ and $Q = \{C_1, C_2, \ldots, C_G\}$ the set of these classes. Then we compute the posterior probabilities $P_{Avg}(C_i \mid R)$ for each $C_i \in Q$ and select the class C_j with maximum posterior:

$$P_{Avg}(C_j \mid R) \geq P_{Avg}(C_i \mid R), \forall C_i \in Q. \tag{4.37}$$

An example of the application of this step is shown in Figure 4.6(a)-(b).

(a)   (b)
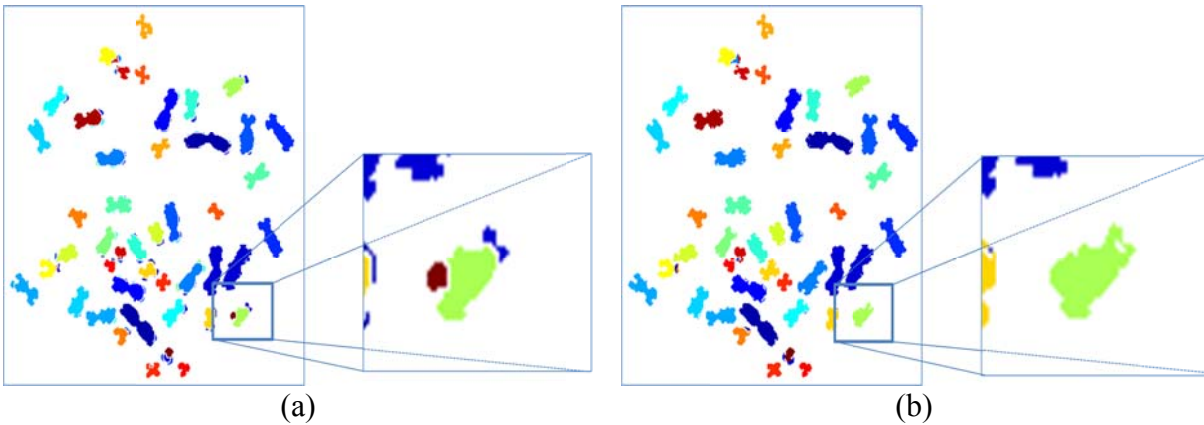
Figure 4.6: Small Region merging of chromosomes. (a) Initial classification map (we have used a separate color for each chromosome class) and (b) the final classification map.

4.7   Results

In order to test the method we have used the Advanced Digital Imaging Research (ADIR) M-FISH database. There are many images that give low classification accuracy [2], the common factor among those images is that the image quality is poor. These images were self-trained and tested using a Bayes pixel-by-pixel classifier. Those images that gave lower classification accuracy than 85% were identified (also visually confirmed) as Bad Images (a list of the images is provided by Choi *et al.* [2]).

4.7.1 Parameter Estimation

In order to estimate the learning rates α, β, and γ we conducted the following experiment. We varied the values of the learning rates as followed $\alpha = [0, 0.05, 0.1, ..., 1]$, $\beta = [0, 0.05, 0.1, ..., 1]$, $\gamma = [0, 0.05, 0.1, ..., 1]$ and computed the classification accuracy for these values. Figure 4.7, presents the accuracy of the method using these values for the ADIR M-FISH image database.
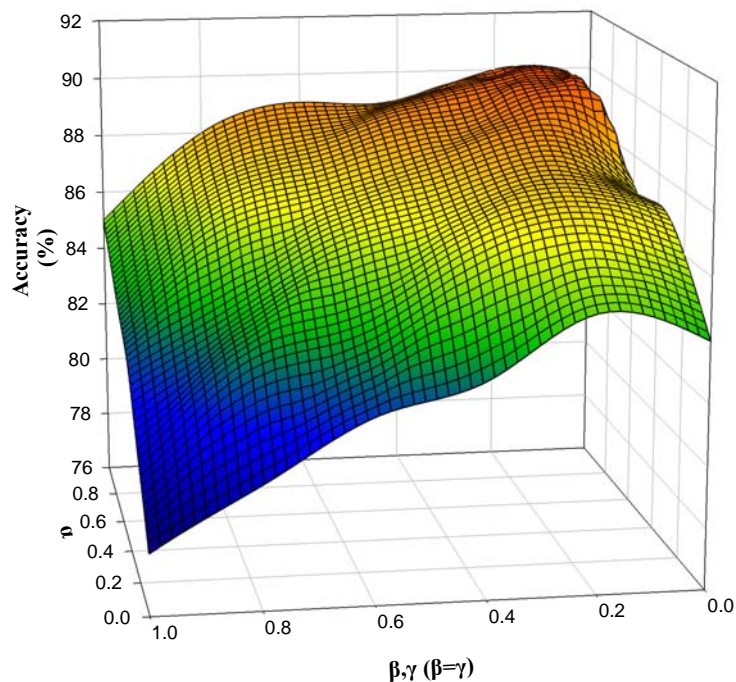
Figure 4.7: Classification accuracy using for different values of the learning rates α, β, and γ.

4.7.2 Classification Accuracy

The best overall classification accuracy for all the M-FISH images was found to be 89.95% for learning rates equal to $\alpha = 0.9$, $\beta = \gamma = 0.2$. The learning rates are computed for the ADIR M-FISH image dataset. However, this database contains 3 different M-FISH datasets produced by 3 different kits. Thus, the estimation is not required for a new kit. However, note that when the learning rates are set to $\alpha = \beta = \gamma = 0.0$, thus the MAP-EM algorithm degenerates to the classical EM estimation, the classification accuracy reduces to 83.62%. This proves the effectiveness of the MAP-EM algorithm over the classical EM algorithm.

67

Figure 4.8 presents the class classification accuracy for the best parameters values mentioned above.
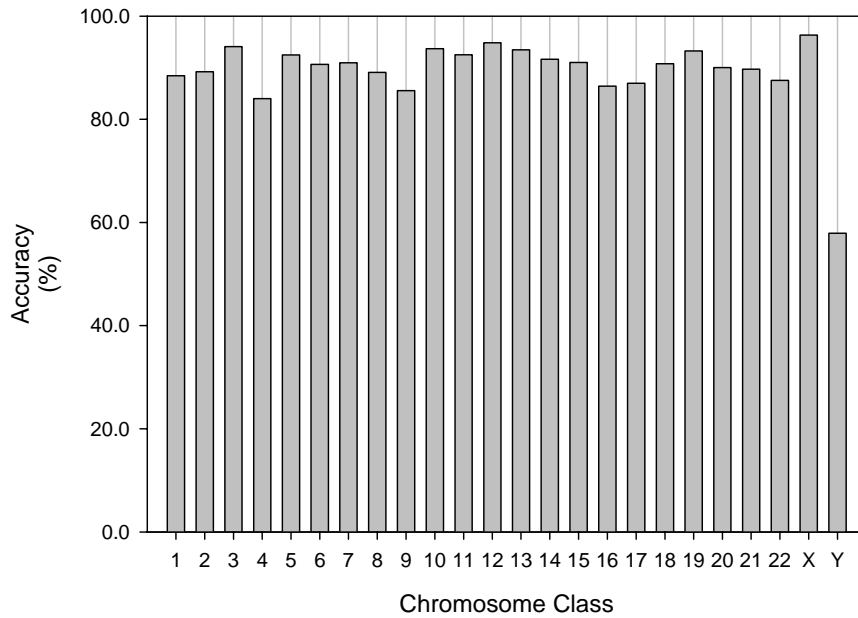
(a)

Figure 4.8: (a) Class classification accuracy for each of the chromosome classes (1,2,..,22,X,Y).

Finally, Figure 4.9 presents an example of the application of the method to an M-FISH image.
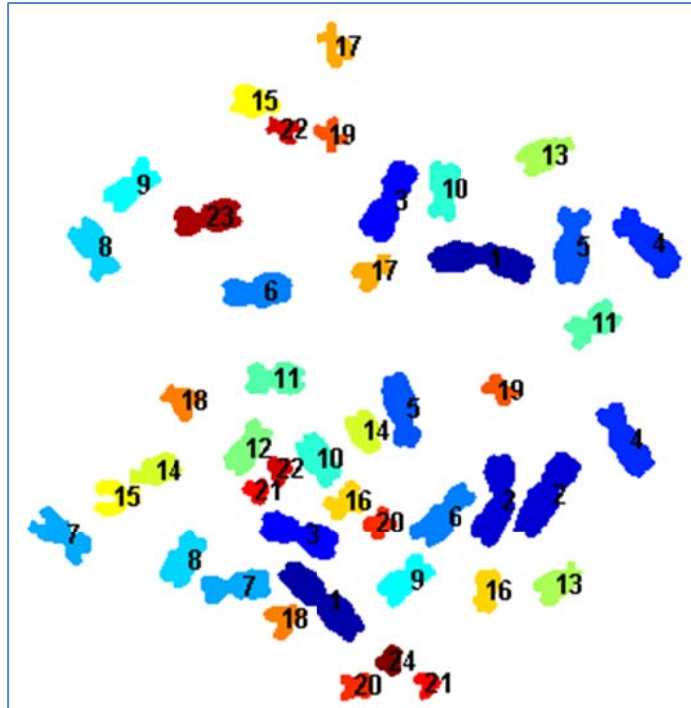
(a)
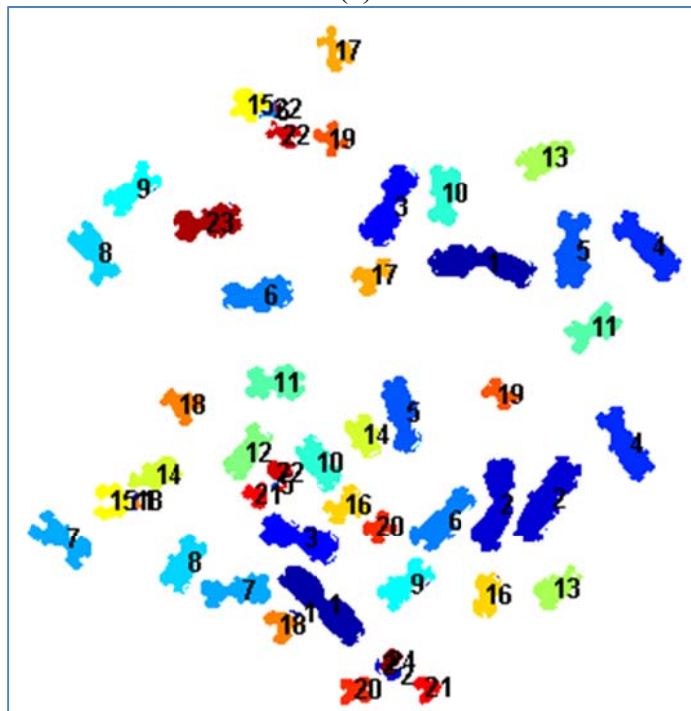
(b)

Figure 4.9: An example of the application of the method. (a) The annotated image, (b) Classification map of the method.

4.8   Conclusions

The method first uses a watershed based algorithm to segment the M-FISH image into regions. Next, the goal is to partition the chromosome pixels by a Gaussian Mixture Model with $K = 24$ components, one for each chromosome class. However two main problems had to be overcome:

A. The initialization of the multichannel Gaussian Mixture Model: We first estimate five, two component Single Channel GMMs where the first component corresponds to the hybridized class and the second to the non-hybridized class. We then combine those GMMs in order to estimate an initial multichannel Gaussian Mixture Model. Furthermore, we incorporate to this model the emission information that we had from the M-FISH experiment (presented in Appendix A).

B. The adaptation of the parameters of the multichannel GMM: Although we could adapt this multichannel Gaussian model using the EM algorithm we chose to employ the MAP-EM method which uses the initial model build in the previous step. This proves to be more efficient in terms of classification accuracy from the classical application of the EM algorithm. More specifically the MAP-EM method (89.95%) attains an increase 6.33%, over the application of the classical EM (83.62%).

Having estimated the multichannel Gaussian mixture model we then classify each region of the MFISH image. Finally a region merging step is utilized in order to produce a final classification map to the cytogeneticist.

Several methods have been proposed in the literature for the M-FISH chromosome image classification. Most of these methods are supervised requiring a small number of images to train the classifier. Supervised classification methods, such as the Bayes classifier and k-nearest neighbor require training data [14]. However, collecting and labeling a large set of samples can be extremely costly. Additionally significant variations have been observed between the M-FISH images. These variations are often due to a lot of factors such as long exposure times, humidity, temperature, type of microscope, color spread [2]. When a supervised classification method is used, the classification accuracy will be high when the sample distributions of both training and testing data are the same. However, this is often not the case making the need for a fully unsupervised M-FISH image classification method a necessity. The method requires only the knowledge of the emission matrixes (Appendix A)

which is available upon the purchase of the kit thus, making the M-FISH technique attractive for use when ground truth does not exist.

Table 4.1: Comparison of the proposed method and other methods reported in the literature in terms of chromosome classification accuracy

Method	Year	Type of Classification		Database Used[*]	Average Chromosome Pixel Classification Accuracy (%)
Sampat et al. [28]	2002	Pixel-by-Pixel	Supervised	A	91.4
Choi et al. [29]	2004	Pixel-by-Pixel	Supervised	A	97.1
Sampat et al. [30]	2005	Pixel-by-Pixel	Supervised	A	90.5
Wang et al. [31]	2005	Pixel-by-Pixel	Supervised	A	87.5
Schwartzkopf et al. [32]	2005	Pixel-by-Pixel	Supervised	C	68.0
Karvelis et al. [21]	2006	Region Based	Supervised	A	89.0
Karvelis et al. [22]	2008	Region Based	Supervised	C	82.5
Choi et al. [2]	2008		Unsupervised	B	77.8
This method	2012	Region Based	Unsupervised	B	89.95
		Region Based	Unsupervised	C	83.62
K-means	2012	Pixel-by-Pixel	Unsupervised	B	72.48

[*]A: Part of the database
 B: The whole ADIR database minus the Bad Images
 C: The whole ADIR database

71

Table 4.1, presents a comparison of several different classification algorithms presented in the literature. Most of the methods employ pixel-by-pixel classification schemes and use the whole or part of the ADIR M-FISH dataset. Supervised based methods were the first methods [28] that have been introduced for the classification of M-FISH images. The classification accuracy for the whole ADIR M-FISH database is 82.5%. Unsupervised based methods already have been tested for the whole database and their reported accuracy is 77.8% [2]. We have also tested the K-means algorithm for the ADIR M-FISH database, where we have used the emission information for each chromosome class in order to initialize the cluster centers [24]. The classification accuracy was 72.48%. The method is superior to both unsupervised and supervised methods as shown in Table 4.1.

Translocation is the most significant rearrangement. It involves two non-homologous chromosomes which result from a break in each of the chromosomes, and subsequent reunion [62]. A change in the color of a chromosome tip may be due to noise, staining, or an actual translocation. Figure 4.10, presents two translocations between chromosome classes 9 and 4. As it can be observed for this example, the method segments and classifies correctly the translocated areas. If a fragment of a chromosome belonging to a translocation is smaller than the smallest chromosome this will be probably merged to the neighbor chromosome. This is a drawback of the method. However, there is a tradeoff between reducing misclassifications and detecting translocations.

Fluorescent in Situ Hybridization (FISH) technology has been widely recognized as a promising molecular and biomedical optical imaging tool to screen and diagnose different types of chromosome anomalies that could be evolved in different types of cancer (e.g. trisomy of chromosomes 3, 7, X has a significant impact on cervical cancer development and prognosis [63]). One of the advantages of the method could be the application using different types of multichannel FISH images in order to correctly segment and classify not only chromosomes but also chromosome spots in general [63]. Furthermore, the method is independent of the number of channels used by the FISH technology (e.g. a 2 image channel image is used by Wang et al. [63] to detect cervical cancer).
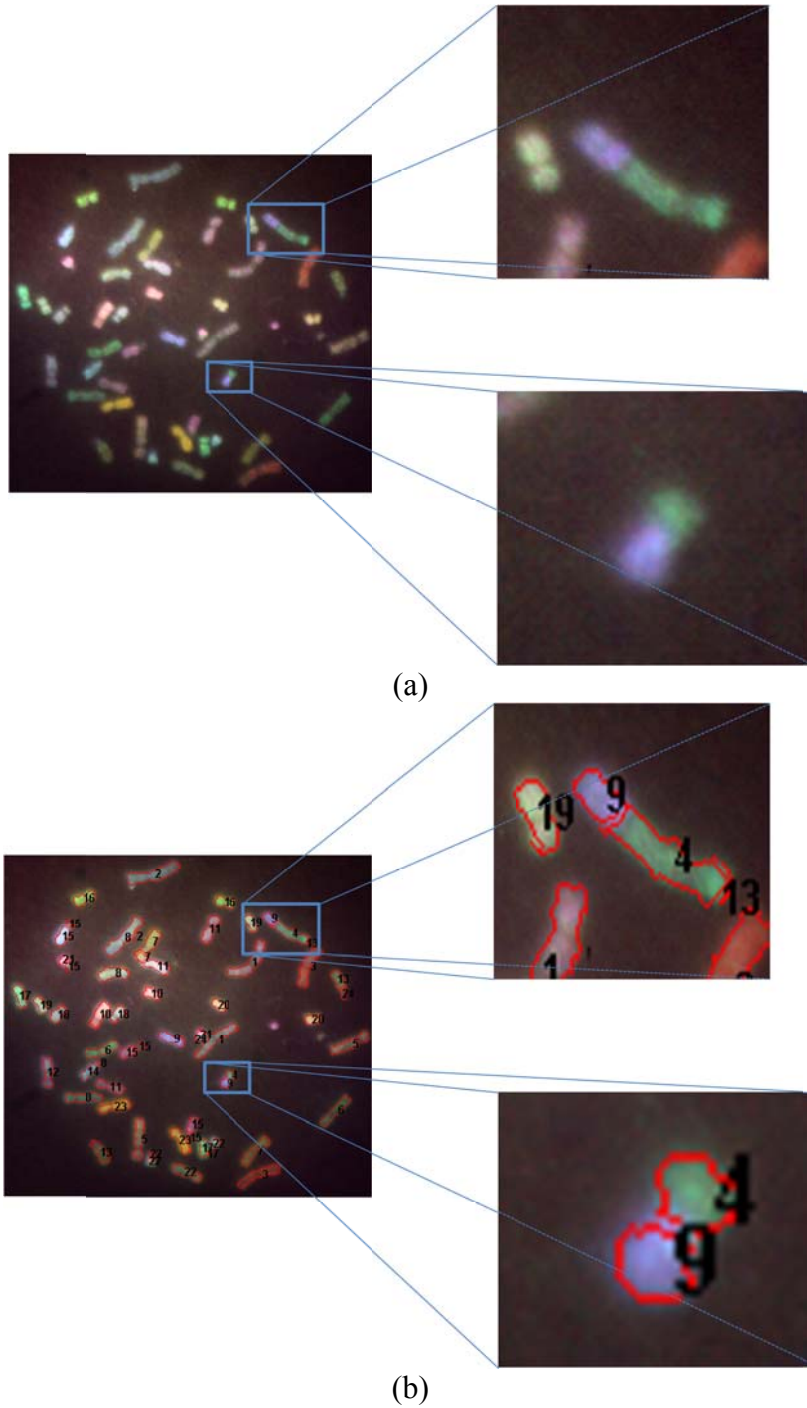
Figure 4.10: Two examples of accurately detection and classification of exchange of genetic material (translocation) between two different chromosomes (green chromosome 4 and blue chromosome 9). (a) The two translocated chromosomes and (b) The two translocated chromosomes segmented and classified correctly by the method.

CHAPTER 5:

IDENTIFYING TOUCHING AND OVERLAPPING CHROMOSOMES USING THE WATERSHED TRANSFORM AND GRADIENT PATHS

5.1   Introduction

Automation of chromosome analysis has long been considered as a difficult task. However, chromosomes in an M-FISH image do very often partially occlude each other; hence, their segmentation is not trivial and requires the application of a dedicated procedure. In this chapter a method is presented for the segmentation of touching and overlapping groups of chromosomes in M-FISH images.

Currently there is no method for disentangling touching and overlapping group of chromosomes for the M-FISH images. This created the necessity to develop such kind of method. However, the goal was to combine geometrical features and already developed methods for greyscale chromosome images. This was feasible by recursively applying the watershed transform to each watershed area and incorporating the idea of gradient paths for

disentangling overlapping groups of chromosome.

5.2   Automated Disentangling of Chromosomes – Literature Review

Many attempts have been made to automate parts of the chromosome M-FISH image analysis procedure. However, chromosome images are inherent with the partial occlusion and touching of chromosomes, as shown in Figure 5.1. This is one of the major factors hindering automatic analysis. Spectrum based methods use a pixel-by-pixel classifier to classify each pixel of the M-FISH image and this information may be sufficient to segment touching and overlapping chromosomes [32]. However the measured fluorescence at a pixel may be the combination of fluorescence in a neighbouring region leading many times to misclassification errors. These factors make the pixel spectral information of touching or overlapping chromosomes unreliable. Hence the spectral information alone cannot separate the touching and overlapping chromosomes efficiently.

On the other hand there is a variety of geometric separation based methods proposed in the literature for greyscale chromosome images [64], [65], [5], [66]. The main idea of these methods is that they split the chromosome groups into segments and then they try to combine these segments into chromosomes. Valley searching techniques [64], [65] attempt to find a "pale path" of grey values corresponding to a separation between touching-overlapping groups of chromosomes. Initially, all high concavity points (cut-points) are detected along the boundary of chromosomes. Next, a heuristic search is performed to detect the minimum density path between touching chromosomes. The chromosome group is split by the pale path and the segments are combined to form separate chromosomes. Agam et al. [5] used concave points to construct all the possible separation lines. In their work, they determined potential chromosomes using rectangle hypothesis testing. However this hypothesis does not always hold because of the existence of bended chromosomes that are touching or overlapping to each other and thus a straight line cannot split exactly the chromosomes.

We can conclude that when only the spectral information is used, the segmentation accuracy relies on the pixel-by-pixel classification accuracy. On the contrary, the geometry based methods assume that chromosome shape alone is sufficient for the purpose of separation. Thus both, geometry and spectral information, has to be merged in order to achieve better segmentation results for M-FISH chromosome images.
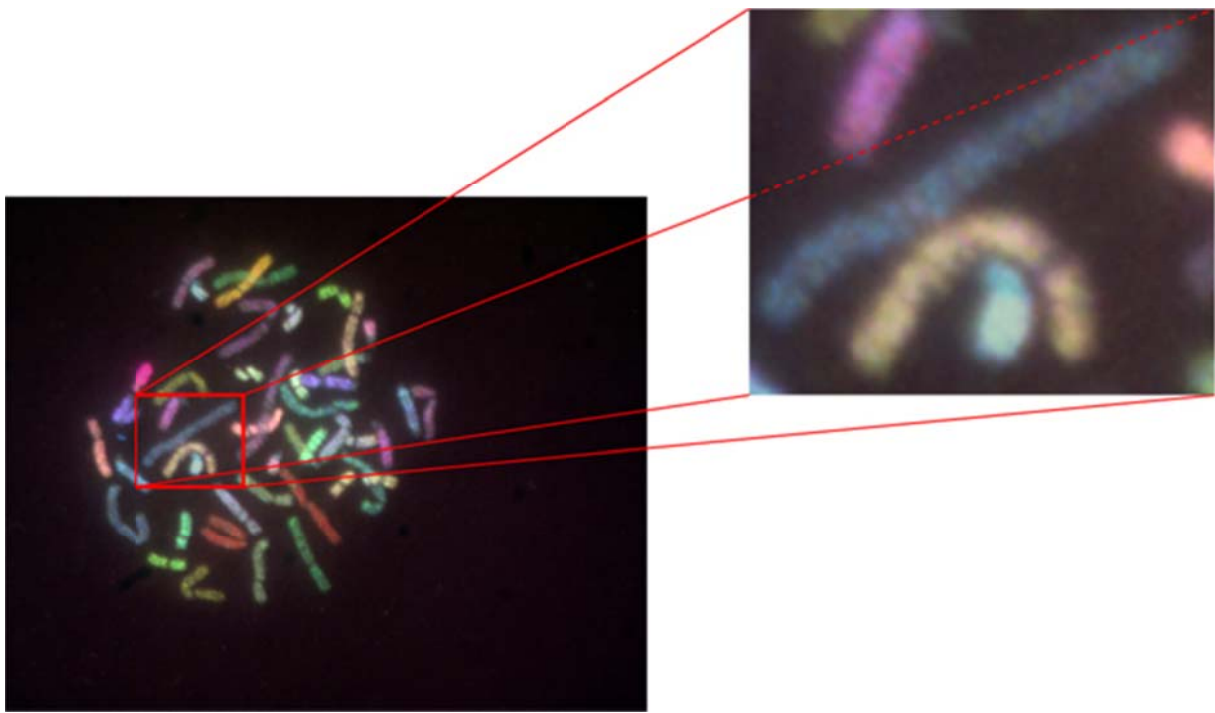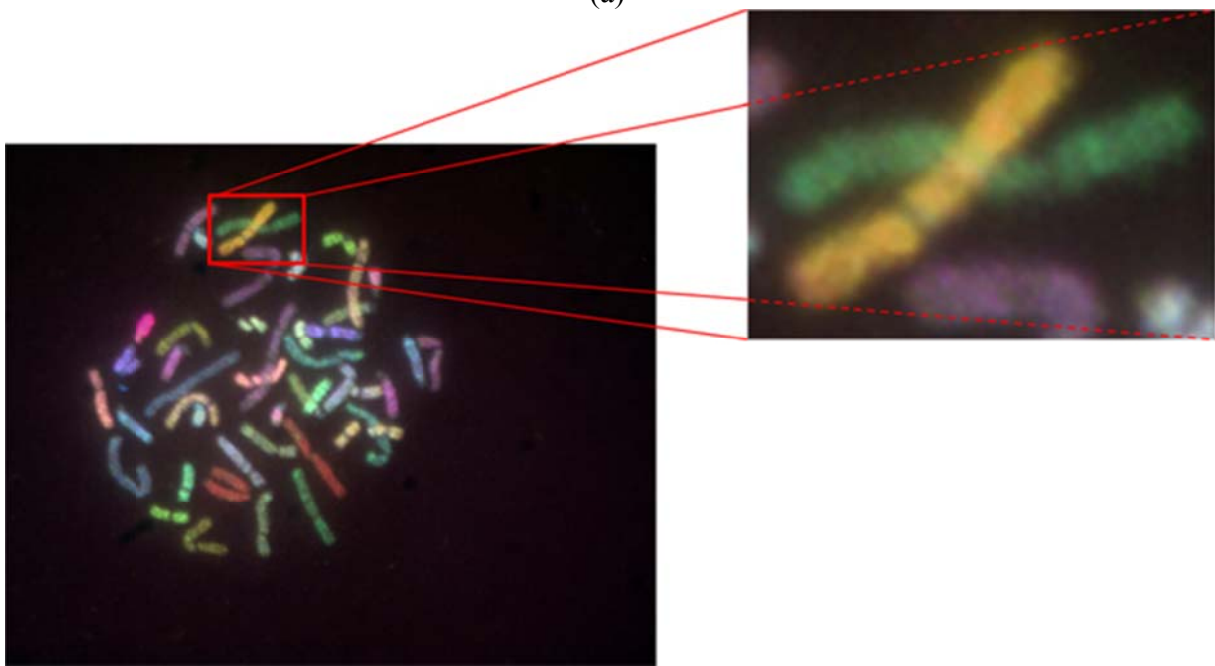
Figure 5.1: Touching and overlapping group of chromosomes. (a) Three chromosomes that are touching each other, (b) Two chromosomes that overlap.

We describe a novel method that tackles the problem of touching-overlapping group of chromosomes. Initially, the method uses the watershed transform to segment the DAPI image into watershed regions. The watershed transform has been widely used for the separation of

touching/overlapping groups of objects from images [39], [16], [67], [68]. In our case we propose the recursive application of the watershed transform to each watershed region. However there exist difficult cases of touching as also of overlapping groups of chromosomes that need separation. For this reason we use a geometry method such as the "gradient paths" to split each group of touching-overlapping chromosomes. However we do not compute the gradient paths using the intensity of pixels of the DAPI image, but we propose the computation of paths in the M-FISH image using pixels with high multichannel gradient magnitude values. This computation proves to be more efficient than the computation of the gradient path on the DAPI image since there are cases of touching or overlapping groups of chromosomes where the gradient path on the DAPI image is difficult to compute since the chromosomes are difficult to disentangle. Finally, after path computation, a region adjacency graph is computed and a region merging algorithm is used to merge all regions.

5.3   Recursive Watershed Segmentation

The proposed method consists of three stages as it is shown in Figure 5.2: (a) the recursive watershed transform computation, (b) the computation of each gradient path and (c) the region merging process. The first stage consists of a number of steps. The first step is the conversion of the initial DAPI chromosome image to binary. In the second step, the Euclidean distance transform of the binary image is computed. The watershed transform is applied in the next step and an initial estimation of the segmented chromosome areas is obtained. The watershed transform is further applied separately to every segmented area until no more new areas are created. The first step of the second stage is the computation of the high concavity points along the boundary of each chromosome area. Next, all gradient paths are computed and the binary chromosome area is split along the gradient path. All gradient paths are computed using the multichannel gradient magnitude. In the final stage a recursive region merging procedure is applied as follows. A region adjacency graph is computed and also each region is classified independently using a region Bayes classifier. Then we merge all neighbouring regions that share the same class. The identification of the overlapping chromosomes takes place in the final step.
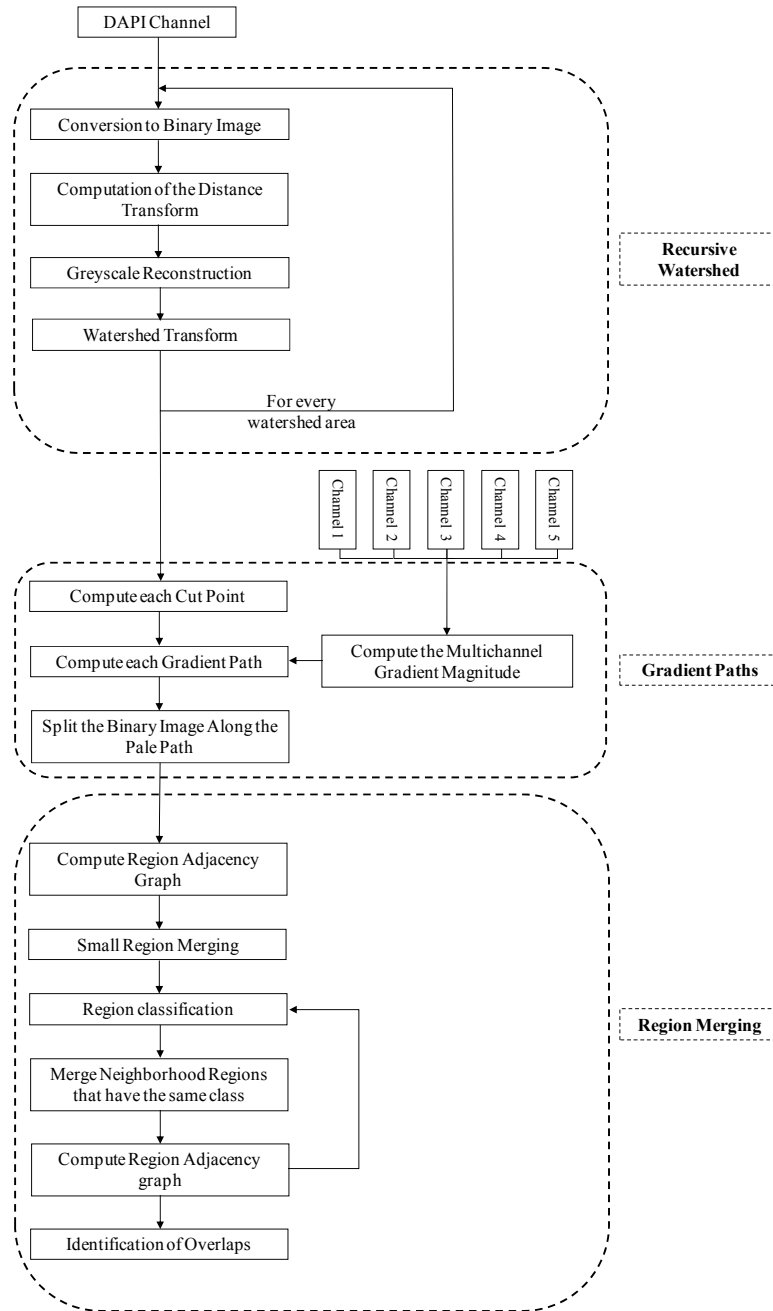
Figure 5.2: Flowchart of the method.

In the first step, the DAPI chromosome image is converted to binary using a well known automated threshold selection process [15]. Using the DAPI channel an initial estimation of the regions of the M-FISH image is produced. The threshold operation at grey level l partitions the pixel values of an image into two classes K_0 and K_1 (representing background and object respectively), i.e., $K_0 = \{1, 2, \ldots, l\}$ and $K_1 = \{l+1, l+2, \ldots, L\}$, where

L is the total number of grey levels in the image. An optimal threshold l^* can be determined by minimizing the following criterion function:

$$l^* = \arg\min_l \sigma_B^2(l), \qquad (5.1)$$

where $\sigma_B^2(l)$ is the between-class variance for the threshold value l.

After the computation of the threshold l^* the binary image B can be computed:

$$B(x, y) = \begin{cases} 0, & if\ DAPI(x, y) \le l^* \\ 1, & if\ DAPI(x, y) > l^* \end{cases}. \qquad (5.2)$$

An example of the application of the threshold operation to a DAPI image is shown in Figure 5.3.

(a)

(b)
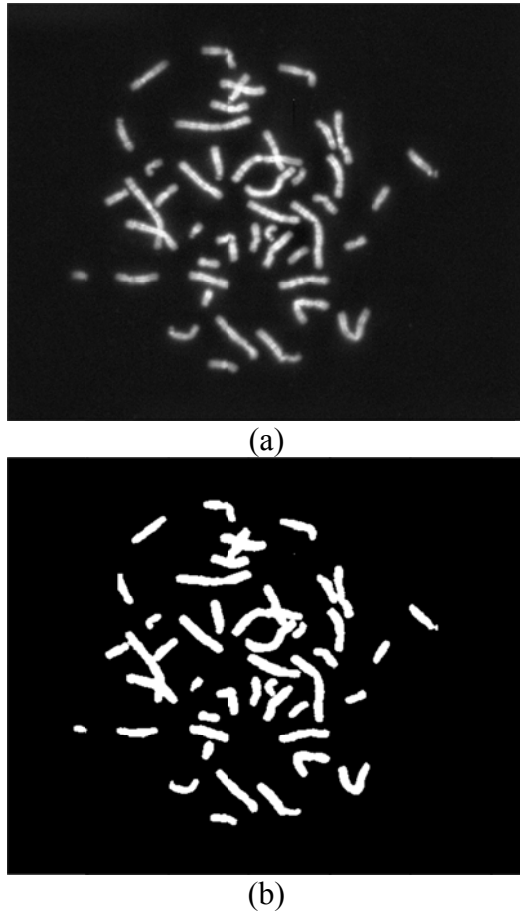
Figure 5.3: The thresholding procedure for a greyscale DAPI image. (a) The DAPI image, and (b) the binary image.

In order to apply the Watershed Transform (WT) [16] to the image B it is common to first compute the Distance Transform (DT) [67], [68]. Given an $m \times n$ binary image B, its distance transform is a map that assigns to each on-pixel (p_1) (with coordinates (x_1, y_1)) the

79

distance to the nearest off-pixel (p_2) (with coordinates (x_2, y_2)). The distance metric used is the Euclidean distance $D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

The distance transform provides important information for the application of the watershed algorithm [69]. The number of regional minima of the negative distance transform constitutes indication of the number of areas that will be segmented by the WT. However a common problem is that the distance transform contains a large number of such minima leading the WT to over segment the initial image. On the other hand the greyscale reconstruction [14] of the negative distance transform suppresses all minima whose depth are lower than or equal to a threshold $h \in \Re$. Thus we apply this procedure in order to alleviate the over segmentation problem. An alternative for the elimination of the over-segmentation effect could be the Gaussian blur of the gradient image [70], however the choice of the width of the Gaussian kernel is a key parameter for these approaches.

The next step is the application of the WT. The watershed transform is a popular segmentation method originated in the field of mathematical morphology. The image is considered as a topographical relief, where the height of each point is related to its grey level. Imaginary rain falls on the terrain and water begins to rise filling the different catchment basins. The watersheds are the lines separating the catchment basins that form.

In our case we apply the watershed method using the negative distance transform. The watershed algorithm produces a tessellation of the image into regions; these regions are called watershed regions and depicted in Figure 5.4(a)-(b). Whereas several methods start with an over-segmentation of the image and iteratively merge regions based on some measures of similarity [71], the method introduces a new region splitting technique based on the watershed transform. All the steps of the method –which do not require any a-priori knowledge–, are recursively applied to every watershed area until no more new areas are produced. The result of the recursive watershed transform is shown Figure 5.4(c)-(f).
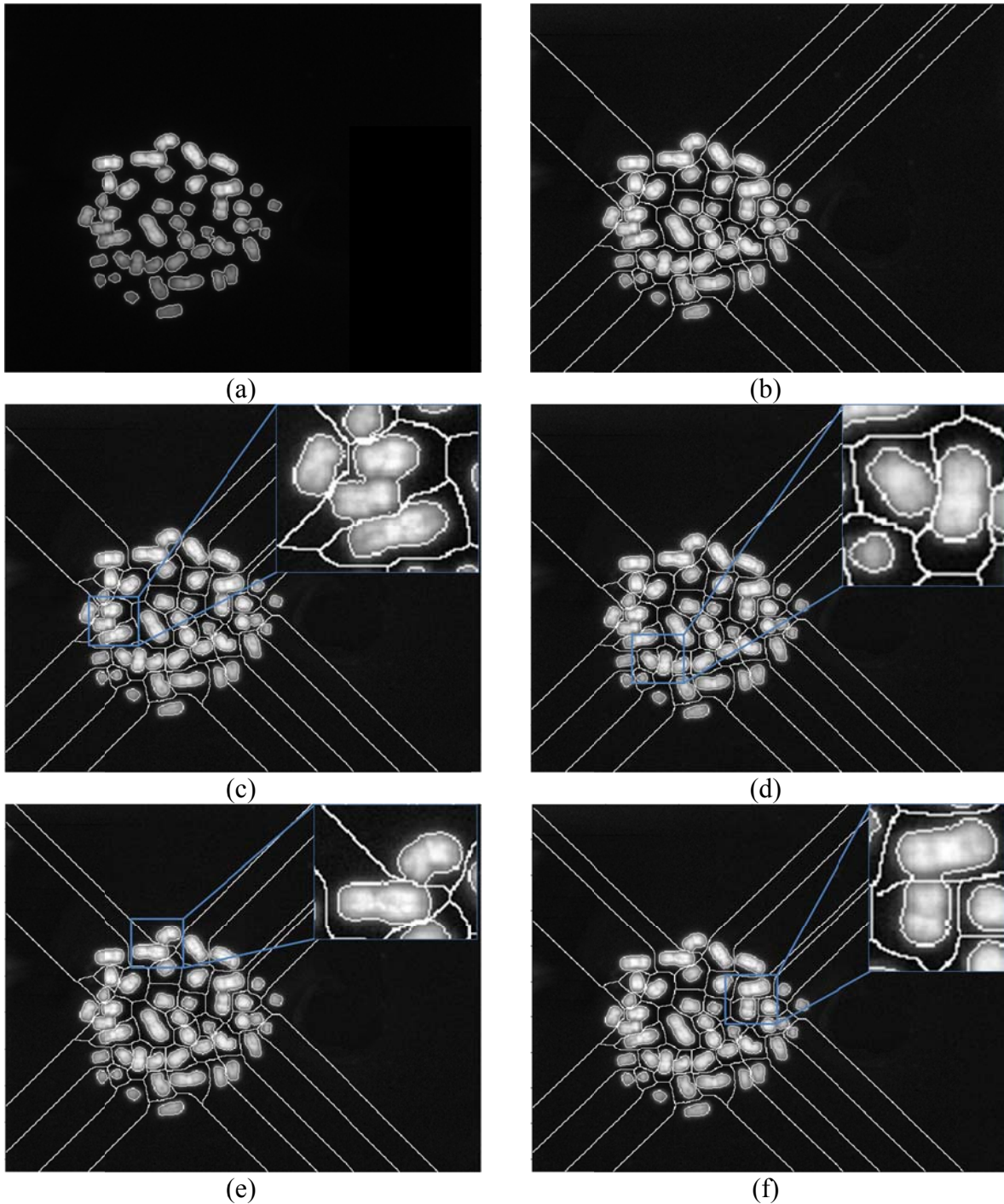
Figure 5.4: An example of the application of the recursive watershed transform for a DAPI chromosome image. (a) DAPI image, (b) 1 iteration, (c) 2 iteration, (d) 3 iteration (e) 4 iteration and (f) 5 iteration.

5.4 Gradient Path Computation

The idea of paths has been introduced in early 90's [64], [65] in order to separate touching groups of chromosomes for greyscale images such as the G-banded chromosome images [72].

It is based on two assumptions: (a) where chromosomes touch the cluster boundary tends to form an acute angle and (b) at points where chromosomes touch, the optical density is relatively low. The detection of the paths is computed via a search algorithm. The search begins at a cut-point and proceeds in the direction of the normal vector. A cut-point is a boundary point at which the boundary is highly concave. It then proceeds until another boundary point is found as follows: At the current point a list of candidates is found as it is shown in Figure 5.5(a). A new trace point is found by choosing the candidate with the smallest intensity value. Finally, the searching direction is updated every d points to allow the path to follow the shape of its trace points, as it is shown in Figure 5.5 (b). The path that starts from the cut-point and ends to a boundary point was called a pale path.
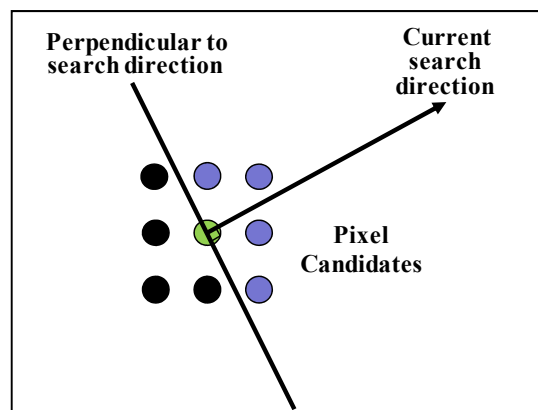
(a)
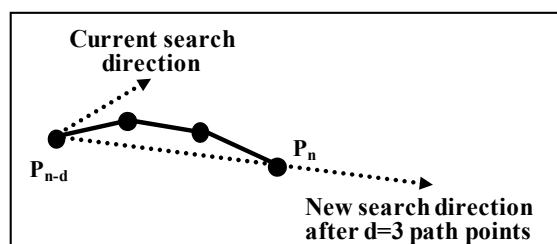
(b)

Figure 5.5: Pale path computation: (a) Candidates for the next path point, and (b) update of path's direction after $d = 3$ points.

The pale paths were used to cut only touching groups of chromosomes without addressing the case of overlapping chromosomes. Moreover, these studies computed the pale paths only for greyscale images. Using a low intensity path the separation of touching chromosomes is feasible, but fails particularly in overlapping cases. Indeed as it is shown in Figure 5.6(a) a pale path does not exist for the case of the overlapping group of chromosomes since the intensity of the overlapping region is homogenous and relatively high.

We propose a modification of the pale path approach in order to achieve separation of touching and overlapping chromosome groups in coloured M-FISH images. This modification uses the multichannel gradient of the M-FISH image [22], [36]. The basic idea is the following: instead of leading the path to follow low intensities pixels, the path now follows pixels of high multichannel gradient magnitude values. The computation of the multichannel gradient magnitude is based on the five channel coloured M-FISH image. This gives the advantage that the paths follow high gradient magnitude pixel values and these high values occur when chromosomes touch or overlap. The path that begins from a cut-point and follows pixels of high gradient magnitude values of the M-FISH image until it reaches a boundary point is now called a gradient path.

To compute the cut-points we first extract the boundary from the binary image B. Suppose that the pixels of the boundary of a segmented region define the set $(c_1, c_2, \ldots, c_{PB})$ where c_i, c_{i+1} are successive points of the boundary and PB the number of pixels of the region boundary. In order to compute the cut-points we compute the curvature of the boundary [65], [64], [66] since local maxima of the curvature indicate candidate positions of the cut-points. For each point of the boundary $(c_i : i = 1, \ldots, PB)$ we consider the triangle that is defined from the three points c_{i-k}, c_i, c_{i+k} ($k = 3$) and compute the angle $a(i)$ defined by the triangle:

$$a(i) = \arccos\left(\frac{(c_i - c_{i-k}) \cdot (c_{i+k} - c_i)}{\|c_i - c_{i-k}\| \cdot \|c_{i+k} - c_i\|}\right) \cdot \text{sgn}\left[\det\left(c_i - c_{i-k} \quad c_{i+k} - c_i\right)\right] \qquad (5.3)$$

In Figure 5.6, we demonstrate the steps for the computation of the cut-points in a group of touching and overlapping chromosomes. After the binarization of the chromosome group (Figure 5.6(a)-(b)) the curvature of the boundary points (Figure 5.6(c)) is computed and is illustrated in Figure 5.6(d). All the cut-points are automatically computed by choosing the boundary points that exceed an angle threshold: $a(i) \geq 210^o$, $i = 1, \ldots, PB$. The red points in Figure 5.6(d) and Figure 5.6(e) illustrate the cut-points that exceed this angle threshold. As we observe in Figure 5.6(e), several candidate cut-points are computed. To overcome this problem, the neighbouring candidate cut-points are automatically grouped and from each

group of candidate cut-points we extract the one that has the maximum angle as it is illustrated in Figure 5.6(f).

(a)

(b)

(c)

(d)

(e)

(f)

(g)
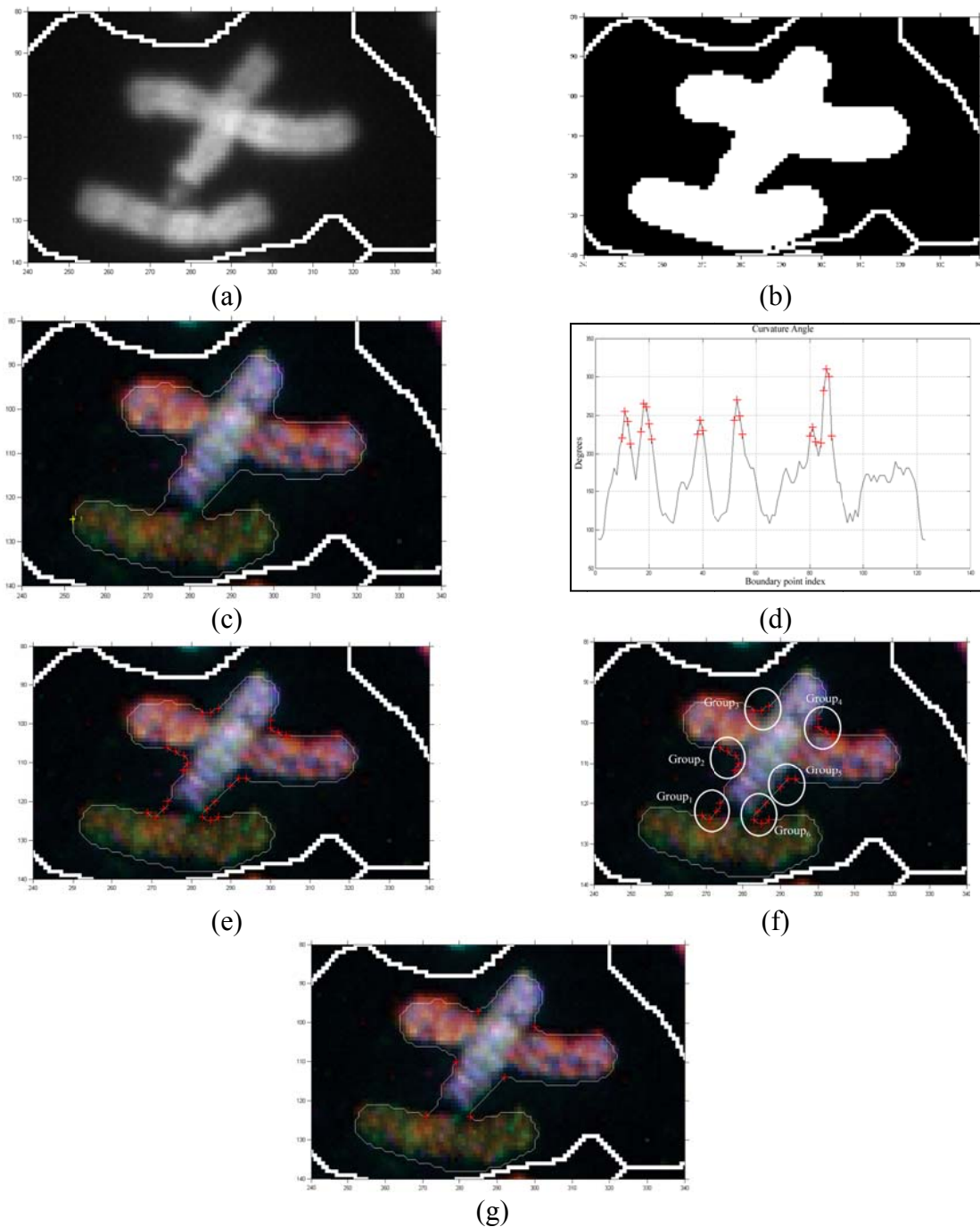
Figure 5.6: The computation of the cut-points for a touching-overlapping group of chromosomes. (a) The DAPI watershed area, (b) the binary image, (c) the boundary of the group of chromosomes over imposed on the M-FISH image (the yellow point depicts the first boundary point), (d) the curvature (in degrees) along the boundary points with red points are depicted the cut-points that exceed the angle threshold, (e) the cut-points (red points) over imposed on the M-FISH image, (f) the groups of the candidate cut-points and (g) the final computed cut-points.

The next step of the method is the computation of the multichannel gradient magnitude, as the gradient path will follow pixels having high multichannel gradient values. The multichannel gradient magnitude is computed as described in Section 2.4.1.

Next we proceed to compute the gradient path. The initial direction of the gradient path is set as the bisector of the angle $a(i) = \angle c_{i-k}, c_i, c_{i+k}$ at the starting points c_{i-k}, c_i, c_{i+k}, as it is shown in Figure 5.7(a) and Figure 5.7(b), where c_i is the initial cut-point. The computation of the gradient path proceeds as follows: we choose from the pixel-candidates the one that has the maximum gradient value. We then proceed to the next pixel updating the current search direction every $d = 3$ points until we reach a boundary point. Finally, we delete points, of the binary image, along the gradient path. We present the computation of the gradient path for a touching-overlapping chromosome group in Figure 5.7(c)-(d) and in Figure 5.7(e) the final regions produced from the binary image by cutting along the gradient paths.

| (a) | (b) |
| (c) | (d) |
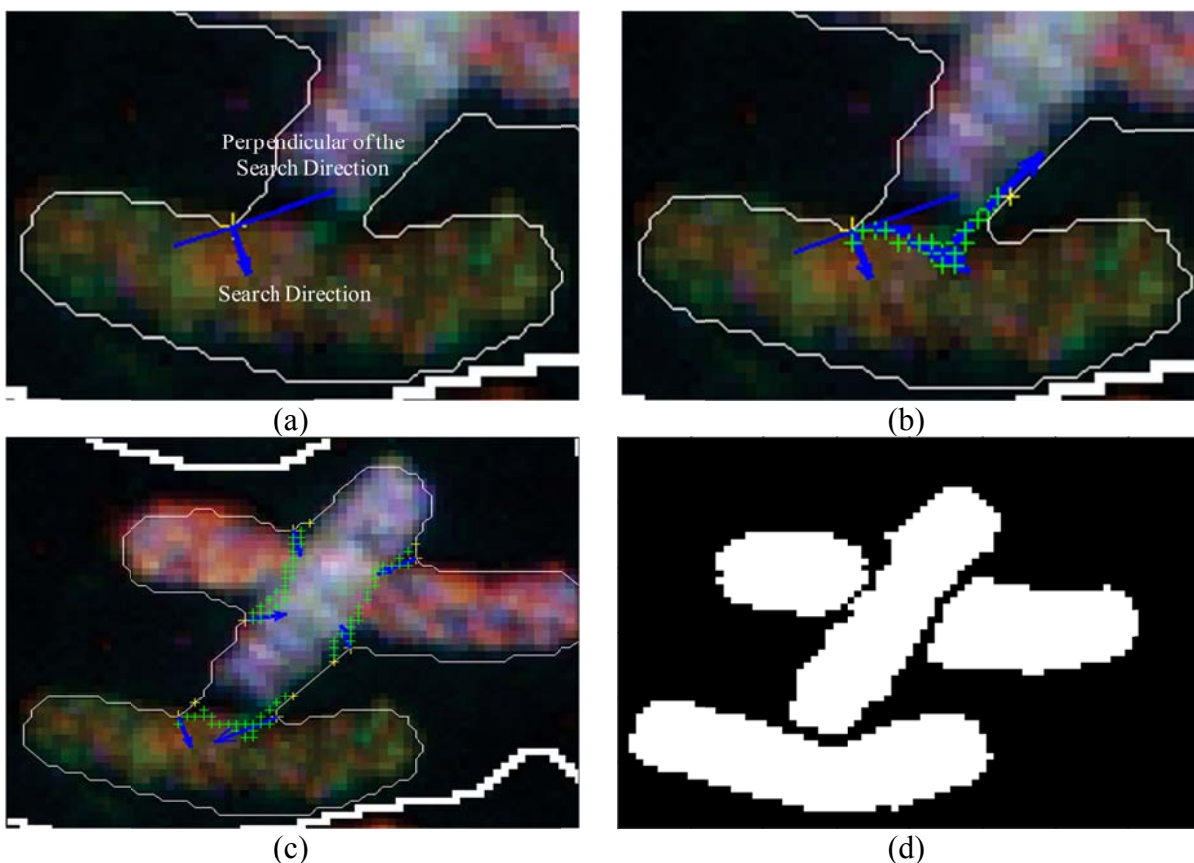
Figure 5.7: The computation of the gradient paths. (a) The initial search direction and the perpendicular of the search direction, (b) a gradient path reaching the other side of the boundary (the green points depict the points of the gradient path), (c) all the gradient paths

computed for all the cut-points of the chromosome group, and (d) the binary image after the binary image has been cut by the gradient paths.

5.5   Region Merging

The purpose of this stage is to connect regions that have been split by gradient paths. In our case we call a region small if it contains less than 25 pixels. This step was implemented by computing for each region of the binary image, the Region Adjacency Graph (RAG) [71], where two nodes (representing two distinct regions) are connected if the corresponding regions are adjacent. An example of a RAG is shown in Figure 5.8.

(a)   (b)

(c)
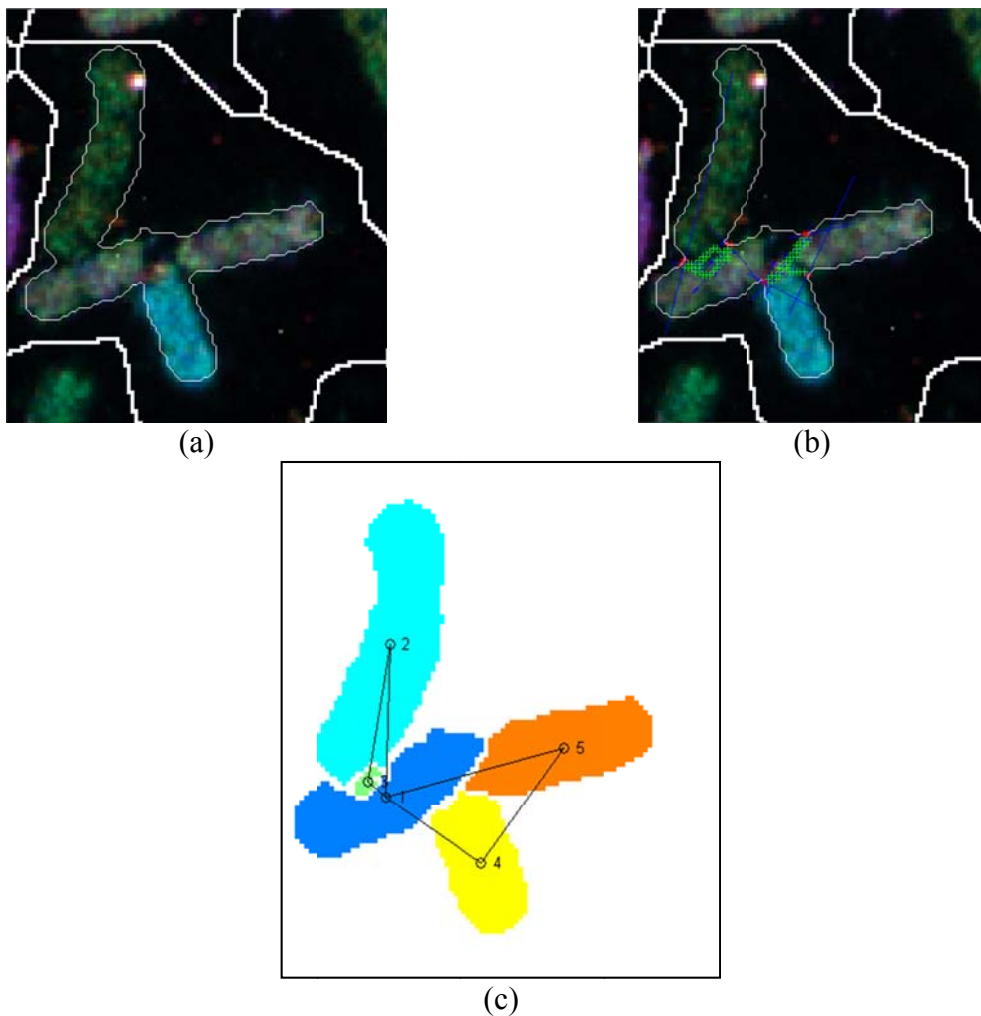
Figure 5.8: The Region Adjacency Graph after the gradient paths split the binary image. (a) The M-FISH image for a touching group of chromosomes (the thin white line depicts the boundary of the binary image), (b) the gradient paths overimposed on the M-FISH image, and (c) the Region Adjacency Graph after the gradient paths split the chromosomes.

Then the Region Bayes Classifier (RBC) [22], [47] was employed in order to classify all the regions of the watershed area as follows. An example of small region merging is shown in Figure 5.9(a)-(b).

The next step of the method is to classify all the regions (including the merged ones) of the binary image using the Region Bayes Classifier described previously. Then all regions that are adjacent and share the same class are connected. Finally, the RAG is computed and the procedure is repeated until no more regions are connected.

#Region	Class
R_1	ω_1
R_2	ω_4
R_3	ω_{16}
R_4	ω_{13}
R_5	ω_1

#Region	Class
R_1	ω_1
R_2	ω_4
R_3	ω_{13}
R_4	ω_1

#Region	Class
R_1	ω_1
R_2	ω_4
R_3	ω_1

(a)   (b)   (c)
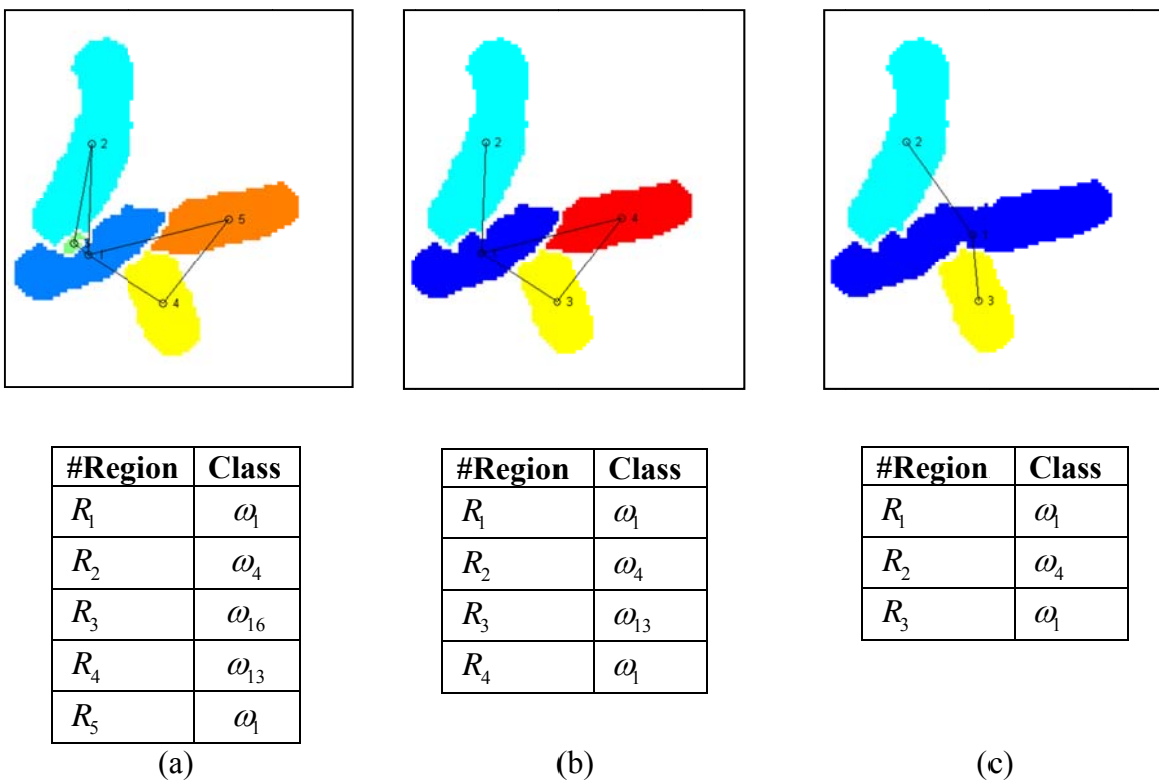
Figure 5.9: Region merging of the binary image. (a) Initial region adjacency graph and the classes of each of the regions of the binary image: the small region R_3 is merged with region R_1, since the posterior probability $P(\omega_1 | R_3) \geq P(\omega_i | R_3), i = \{1, 4\}$, (b) the region R_1 and R_4 are merged since they share the same class ω_1 and (c) the final region merging result.

The final step of the method is the identification of the overlapping chromosomes. The key idea in this step is that when two chromosomes overlap, a cross shaped object is formed. In this case the method splits the binary image in way that one region of the image separates two regions that share the same class. This is illustrated in Figure 5.10. Thus the final step of

87

the method is to identify these overlapping cases by checking for each region of the binary image whether it has two neighbouring regions of the same class.

	GRADIENT PATHS	RAG AFTER REGION MERGING
CASE A		
CASE B		
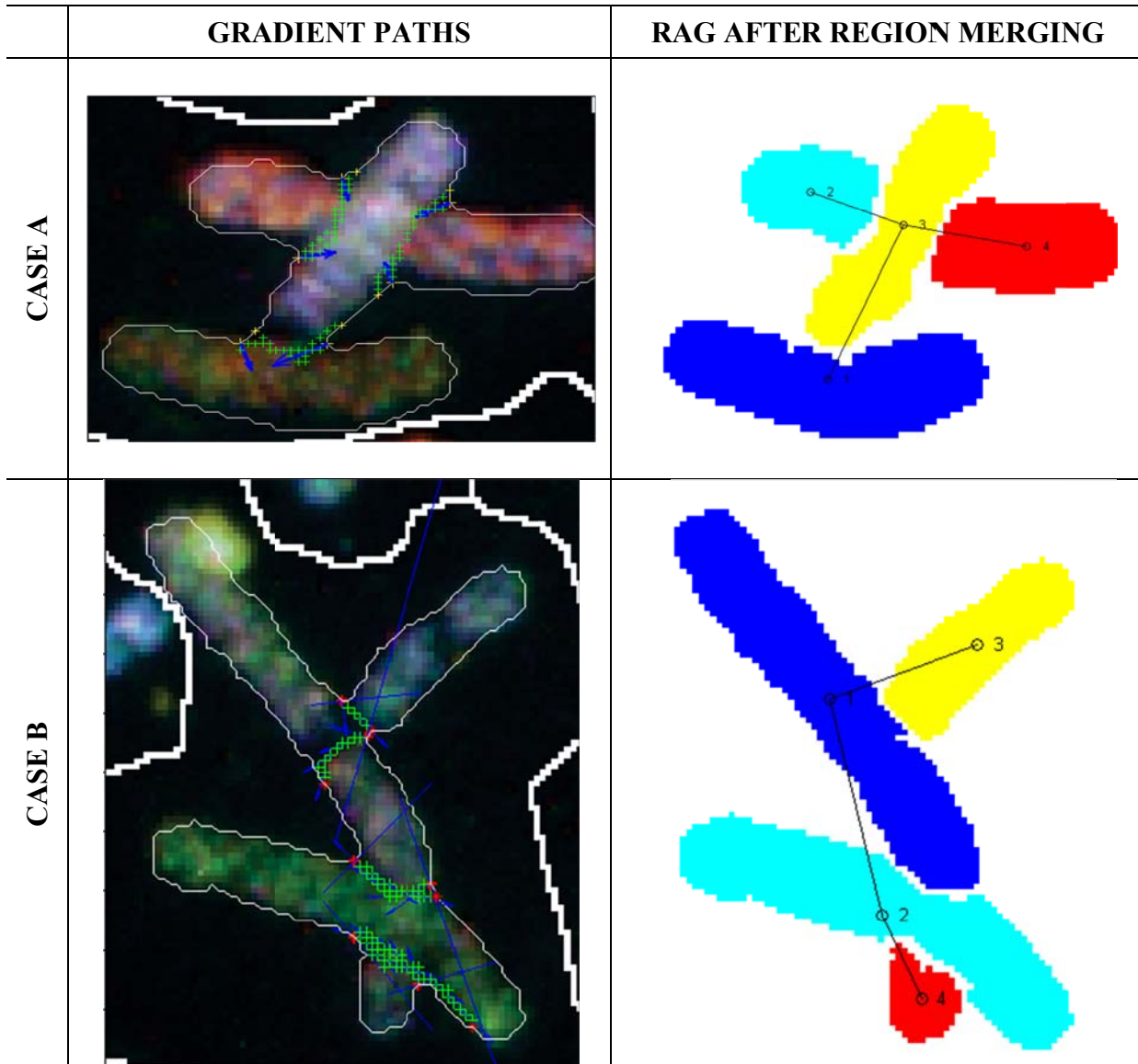
Figure 5.10: Two overlapping chromosome cases. The proposed method identifies them correctly. CASE A: Regions R_2 and R_4 are identified as an overlap since the region R_3 is connected with two regions of the class ω_2. CASE B: Regions R_1 and R_4 are identified as an overlap since the region R_2 is connected with two regions of the class ω_1.

5.6 Results

5.6.1 Dataset

To validate the method we used the ADIR M-FISH database. As a ground truth for the touching chromosomes, we used the binary image produced by the DAPI image to identify the cases of touching chromosomes in an M-FISH image. For each object produced by the binarization procedure we determined the cases of touching. Finally as a ground truth for the cases of overlapping we used the characterized karyotype image of the M-FISH database since an overlapping region is represented in that image by pixels having the value of -1. The number of touches and overlaps in the M-FISH database is shown in Table 5.1.

Table 5.1: Number of touching and overlapping chromosomes in the M-FISH database.

	Touching Chromosomes	Overlapping Chromosomes
Total Number	1178	189

5.6.2 Touching Chromosomes

The separation accuracy for the touching group of chromosomes was measured by the method. A correct separation occurs when two or more touching chromosomes are segmented correctly. The results of the method for the touching groups of chromosomes are shown in Table 5.2. We have also compared the method with the method of Pale Paths [64] for the touching groups of chromosomes as the method of pale paths cannot handle overlapping cases. In order to compute the pale path we have used the DAPI image since the pale path uses a greyscale image.

It is interesting to mention the robust behaviour of the method in the case of isolated bended chromosomes. It is common in the M-FISH chromosome database to find cases where isolated chromosomes bend, as shown in Figure 5.11. For these cases, cut-points are found and gradient paths begin to split the chromosome into two regions. However the region merging stage merges these regions to form one chromosome again.

89

Table 5.2: Comparison of our work with other works presented in the literature for the touching group of chromosomes.

		Schwartzkopf et al. [32]	Jiet al. [64]	The Proposed Method [25]
SEPARATION ACCURACY	Accuracy	77%	84.2%	90.6%
DATASET DECRIPTION	#Images	183	183	183
	#Touches	720	1178	1178

Gradient Path	RAG	Final Binary Image
CASE A		
CASE B		
CASE C		
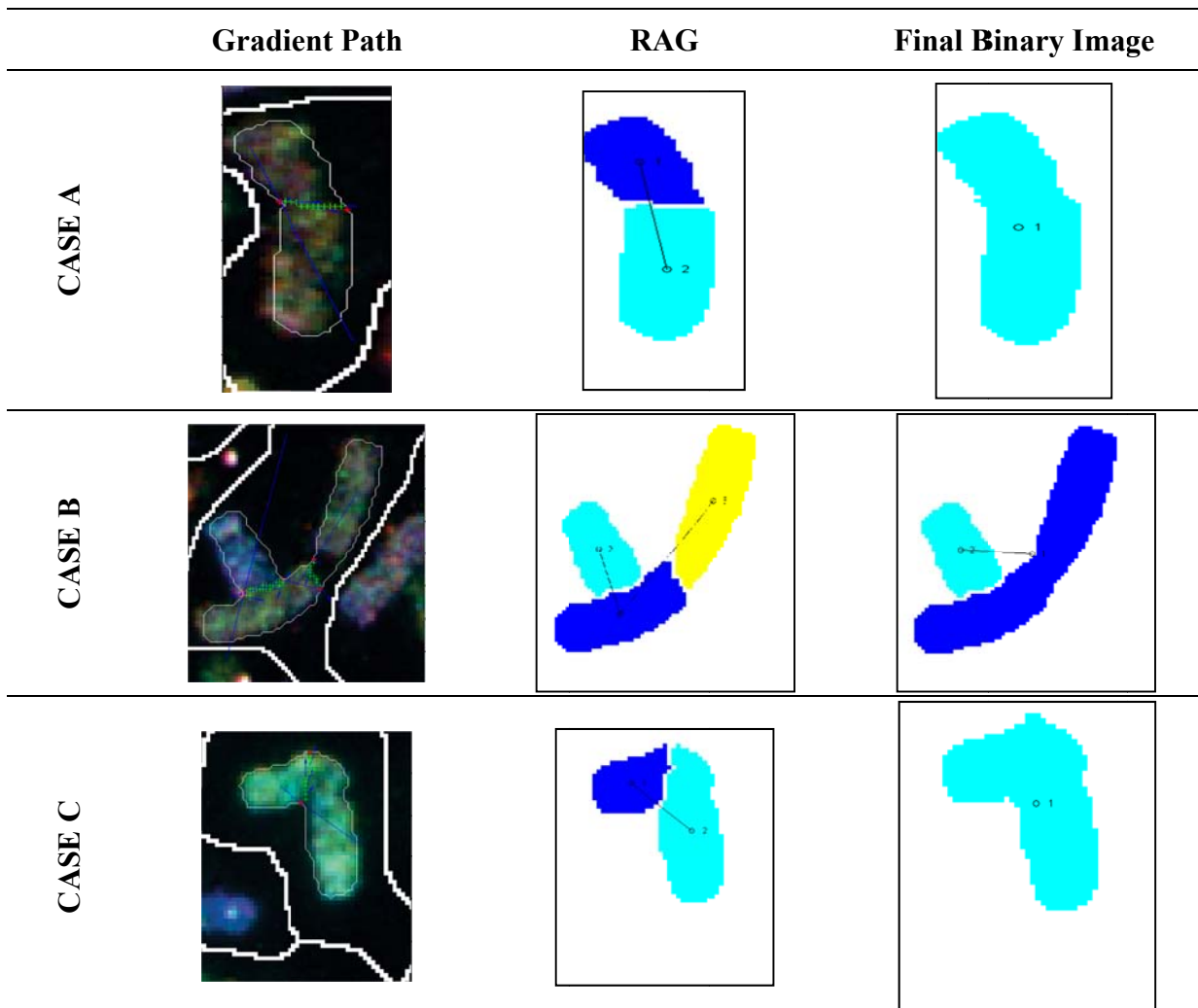
Figure 5.11: Examples of three different cases of bended chromosomes. The proposed method handles them successfully after the region merging stage.

5.6.3 Overlapping Chromosomes

The separation accuracy for the overlapping group of chromosomes is also measured. The results of the method for the overlapping groups of chromosomes are shown in Table 5.2.

Table 5.3: Comparison of our work with other works presented in the literature for the overlapping group of chromosomes.

		SCHWARTZKOPF et al. [32]	THE PROPOSED METHOD
SEPARATION ACCURACY	Accuracy	34%	80.4%
DATASET DECRIPTION	#Images	183	183
	#Overlaps	189	189

5.7 Conclusions

We have described a novel method for the separation of touching and overlapping groups of M-FISH chromosome images. The method is based on the recursive application of the watershed transform and the computation of gradient paths for each watershed area. A region merging stage is finally applied to merge regions that have been wrongly split by the gradient paths. The method is evaluated using an M-FISH chromosome image database and an overall separation accuracy of 90.6% and 80.4% for the touching and overlapping groups of chromosomes respectively has been found.

In fact, only one method has been presented in the literature for the separation of M-FISH images testing its ability to separate touching and overlapping groups of chromosomes for the whole M-FISH database [32]. The method uses the information from all the channels (the 5 channel M-FISH image including the DAPI image) whereas Schwartzkopf et al. [32] use only the information provided by the 5 channel M-FISH image.

To best of our knowledge the pale paths were able to separate only touching groups of chromosomes without handling overlapping chromosomes [65] [64]. We expand the idea of the paths in order to address also the case of overlapping groups. More specifically we introduce the gradient paths which more effectively segment not only touching but also overlapping groups of chromosomes for the M-FISH images. The gradient path is superior to other proposed splitting techniques for two reasons:

1. Unlike other methods [5], we do not assume that a path is a straight line between

two cut-points. Figure 5.12, depicts some examples of touching and overlapping chromosome groups, none of which can be split by a straight line without fragmenting a chromosome. Such cases usually happen where more than two chromosomes are involved in a group of touching-overlapping or one of the chromosomes is bent.

(a)　　　　　　　　　　　　　　　(b)

(c)　　　　　　　　　　　　　　　(d)

(e)　　　　　　　　　　　　　　　(f)
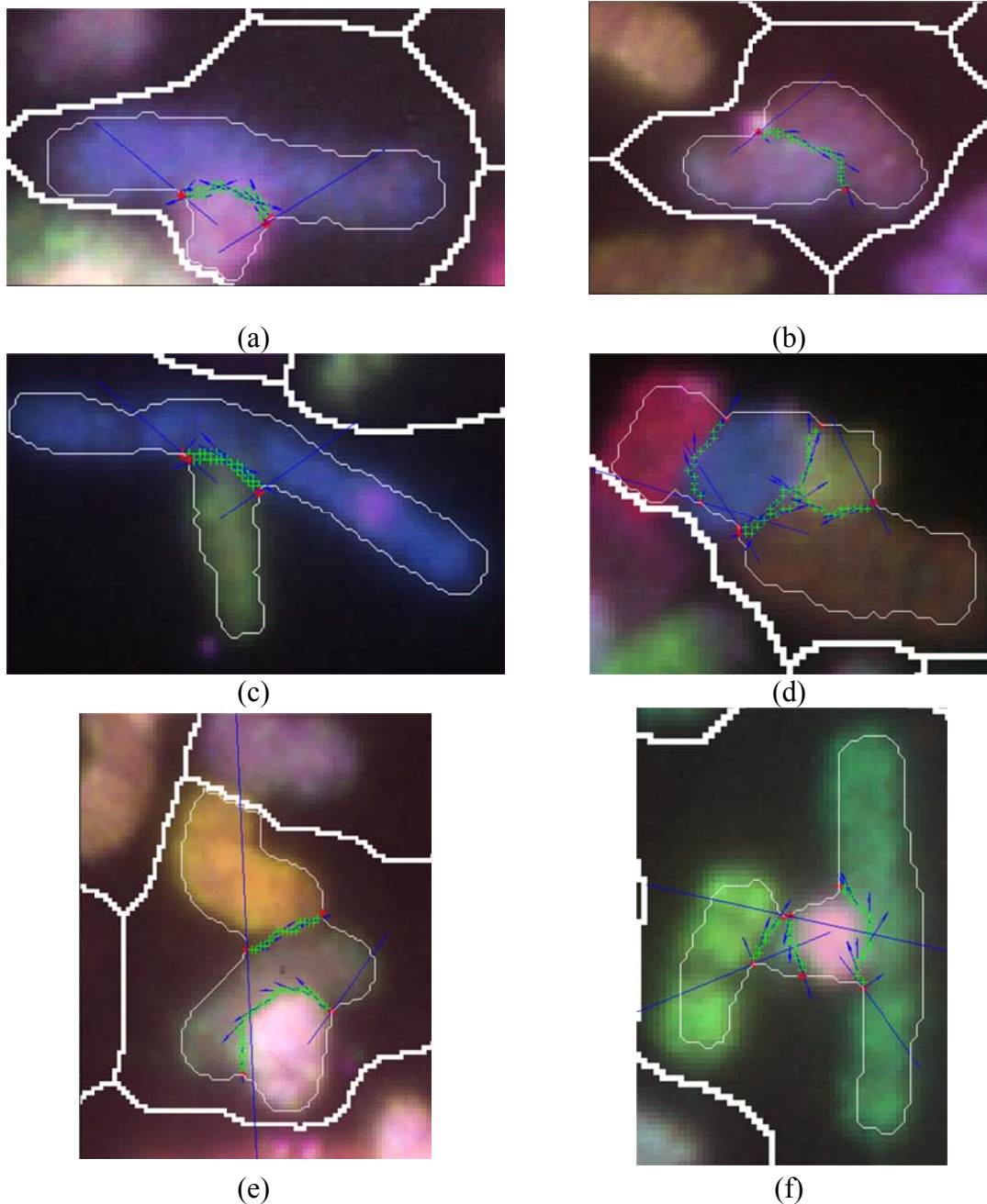
Figure 5.12: Six examples of touching and overlapping group of chromosomes where straight lines between cut points cannot separate correct the chromosomes.

2. The paths have been appropriately modified in order to separate overlapping and

touching groups of chromosomes in M-FISH images by computing the multichannel gradient and the running of the path through high gradient magnitude values. Figure 5.13, illustrates some examples of the computation of the pale path using the DAPI image versus the gradient path which uses the multichannel gradient magnitude of the M-FISH image.
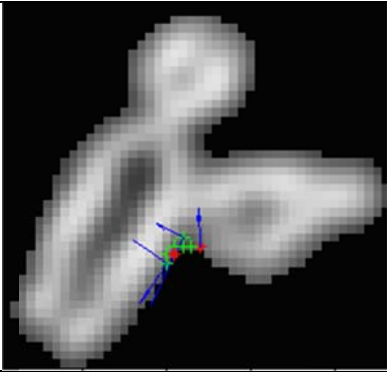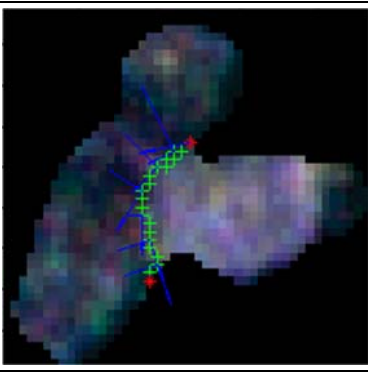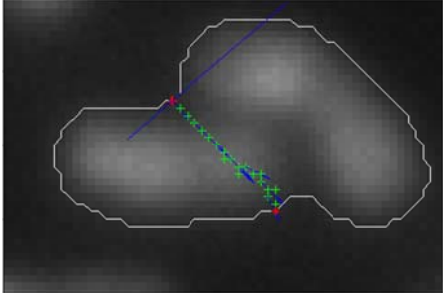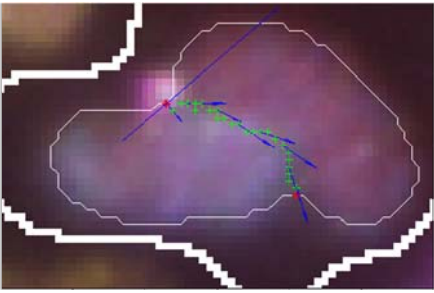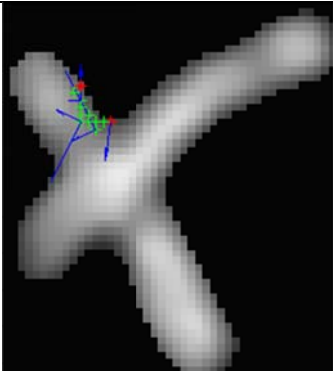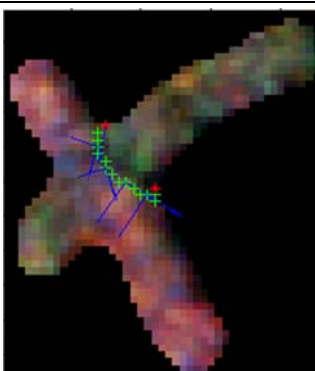
PALE PATH	GRADIENT PATH

Figure 5.13: Pale path versus Gradient Path. Three examples where the pale path fails to separate correctly the chromosomes while the gradient path correctly separates them.

Table 5.2 and Table 5.3 shows a comparison between the proposed study and the method which proposed by Schwartzkopf et al. [32]. While the number of overlapping cases is the same, the number of touches differs between the two methods. This is done due to the

different approaches employed by the two methods for the determination of touching groups of chromosomes. In our case we used the binary image produced by the DAPI image to identify the number of touching chromosomes in an M-FISH image whereas Schwartzkopf et al. [32] has manually chosen the number of touches. In general, it is difficult to compare the two methods directly since they are not handling the same number of touching chromosomes. However the method is employed in the same M-FISH database and the number of touches is higher than that reported in Schwartzkopf et al. [32].

CHAPTER 6:

CONCLUSIONS

6.1   Concluding Remarks

6.2   Directions for Future Research

6.1 Concluding Remarks

In this thesis we have proposed novel methods for multichannel chromosome image segmentation and classification. First, we introduced a region segmentation method for multichannel image segmentation. Then, a supervised region classification method employed for chromosome classification. A fully unsupervised classification method was also proposed

More specifically, in order to perform image segmentation we proposed in Chapter 2, a multichannel watershed-based segmentation method for multispectral chromosome images [21], [22]. This way we were able to segment the M-FISH image into regions. These regions contained pixels with same color characteristics. The method was tested on the ADIR M-FISH database and compared to another pixel-by-pixel classification method. Superior results were achieved.

In Chapter 3, we introduced a supervised region classification method for M-FISH images [73], [22]. After the region segmentation stage the classification of the regions takes place. The method is also tested and compared to a pixel-by-pixel methodology on the same dataset and higher classification accuracy is achieved when using the method. Finally, we use the concept of the Vector Median Filtering in order to enhance the classification accuracy of chromosome M-FISH image classification. The Direction Distance Filter (DDF), Basic Vector Directional Filter (BVDF) and Weighted Vector Median Filtering (WVMF) were also tested and prove the enhancement in the classification accuracy when using the ADIR M-FISH dataset.

In Chapter 4, an unsupervised classification method for M-FISH images was presented [74], [75]. The method incorporates prior information about the emission of each chromosome class to each M-FISH image channel. The classification accuracy of the method was higher when compared even to supervised methodologies.

In Chapter 5, a method for the disentangling of touching and overlapping chromosomes [76] was presented. The method introduces for the first time the idea of gradient paths a split path that is used to cut merged chromosomes. The Region Adjacency Graph and a region classifier are used to merge the parts of chromosomes that have been cut previously by the gradient paths. The method has been tested on a large number of touched and overlapping chromosomes achieving a good separation ratio when compared to other available methods such as pale paths.

6.2   Directions for Future Research

It will be important to move one step ahead by detecting the different types of chromosome anomalies. Currently our algorithm is able to detect only arithmetic and translocations anomalies which are the most significant type of anomaly in chromosomes. However it may be possible by incorporating medical knowledge to the problem to detect further anomalies such as deletions, duplications and inversion rings.

The combination of all the methods described in this thesis could result in the development of an integrated fully automated system for the analysis of M-FISH chromosome images, which would embody automated chromosome segmentation, separation of occluded chromosomes and finally classification.

It would be very interesting to detect the centromere (the center of the two arms of the chromosome) of each chromosome. This type of information is valuable for the cytogeneticist since the centromere plays an important role for the identification of chromosome abnormalities.

Fluorescent in Situ Hybridization (FISH) technology has been widely recognized as a promising molecular and biomedical optical imaging tool to screen and diagnose different types of chromosome anomalies that could be evolved in different types of cancer (e.g. trisomy of chromosomes 3, 7, X has a significant impact on cervical cancer development and prognosis [36]). One of the advantages of the method could be the application using different types of multichannel FISH images [77]. Furthermore, the unsupervised classification method

is independent of the number of channels used by the FISH technology (e.g. a 2 channel image is used by Wang et al. [77] to detect cervical cancer).

BIBLIOGRAPHY

[1] H. Tjio and Levan A., "The chromosome number of man," *Hereditas*, vol. 42, pp. 1-6, 1956.

[2] H. Choi, A. Bovik, and K. Castleman, "Feature normalization via expectation maximization and unsupervised nonparametric classification for M-FISH chromosome images," *IEEE Transactions on Medical Imaging*, vol. 27, no. 8, pp. 1107-1119, 2008.

[3] M. Speicher, S. Ballard, and D. Ward, "Karyotyping human chromosomes by combinatorial multi-fluor FISH," *Nature Genetics*, vol. 12, no. 4, pp. 368-375, 1996.

[4] E. Schrock et al., "Multicolor spectral karyotyping of human chromosomes," *Science*, vol. 273, pp. 494–497, 1996.

[5] G. Agam and I. Dinstein, "Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1212-1222, 1997.

[6] T. Sumner, J. Evans, and A. Buckland, "A new technique for distinguishing between human chromosomes," *Nat. New Biol.*, vol. 232, pp. 31-32, 1971.

[7] G. Bauman, J. Wiengant, P. Borst, and P. Duijin, "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome labelled RNA," *Exp. Cell Res.*, vol. 128, no. 2, pp. 485-490, 1980.

[8] M. Thompson, R. McInnes, and H. Willard, *Genetics in Medicine*. Saunders: Kingston, 2005.

[9] T. Veldman, C. Vignon, E. Schröck, J. Rowley, and T. Ried, "Hidden chromosome abnormalities in haematological malignancies detected by multicolor spectral karyotyping," *Nat. Genet.*, vol. 15, pp. 406-410, 1997.

[10] T. Ried et al., "Tumor cytogenetics revisited: Comparative genomic hybridization and spectral karyotyping," *J. Mol. Med.*, vol. 75, no. 11, pp. 801-814, 1997.

[11] R. Haralick and L. Shapiro, "Image segmentation techniques," *Comput. Graph. Image Process.*, vol. 29, pp. 100-132, 1985.

[12] K. Fu and J. Mui, "A survey on image segmentation," *Pattern Recognition*, vol. 13, no. 1, pp. 3-16, 1981.

[13] P. Soille, *Morphological Image Analysis*, 2nd ed. New York: Springer, 2002.

[14] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. Image Process.*, vol. 2, no. 2, pp. 176-201, 1993.

[15] N. Otsu, "A Threshold selection method from gray level histograms," *IEEE Trans. System Man Cybernet.*, vol. 9, pp. 62-66, 1979.

[16] L. Vincent and P. Soille, "Watershed in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp.

583-598, 1991.

[17] K. Haris et al., "Model-based morphological segmentation and labelling of coronary angiograms," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 1003-1015, 1999.

[18] C. Bishop, *Pattern Recognition and Machine Learning*.: Springer, 2006.

[19] R. Duda, P. Hart, and P. Stork, *Pattern Classification*. San Diego: Harcourt Brace Jovanovich, 2000.

[20] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley & Sons, 2001.

[21] P. Karvelis, D. Fotiadis, M. Syrrou, and I. Georgiou, "A watershed based segmentation method for multispectral chromosome images classification," in *Proc. 28th IEEE Ann. Intern. Conf. (EMBS)*, New York, 2006, pp. 3009-3012.

[22] P. Karvelis, D. Fotiadis, and I. Georgiou, "A multichannel watershed-based segmentation method for multispectral chromosome classification," *IEEE Trans. on Med. Imag.*, vol. 27, no. 5, pp. 1107-1119, 2008.

[23] P. Karvelis, D. Fotiadis, and G. Georgiou, "Enhancement of Multichannel Chromosome Classification Using a Region Based Classifier and Vector Median Filtering," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 561-570, 2009.

[24] P Karvelis, A. Likas, and D. Fotiadis, "Semi Unsupervised M-FISH chromosome image classification," in *Proc. Of 10th IEEE Intern. Conf. on Information Technology and Applications (ITAB)*, Corfu, Greece, 2010, pp. 1-4.

[25] P. Karvelis, A. Likas, and D. Fotiadis, "Identifying touching and overlapping chromosomes using the watershed transform and gradient paths," *Pattern Recognition Letters*, vol. 31, pp. 2474-2488, 2010.

[26] R. Eils et al., "An optimized, fully automated system for fast and accurate identification of chromosomal rearrangements by multiplex-FISH (M-FISH)," *Cytogenetics and Cell Genetics*, vol. 82, no. 3, pp. 160-171, 1998.

[27] K. Saracoglu et al., "New concepts to improve resolution and sensitivity of molecular cytogenetic diagnostics by multicolor fluorescence in situ hybridization," vol. 44, no. 1, pp. 7-15, 2001.

[28] M. Sampat, A. Bovik, and K. Castleman, "Pixel-by-pixel classification of MFISH images," in *in Proc. 24th IEEE Ann. Intern. Conf. (EMBS)*, Houston, 2002, pp. 999-1000.

[29] H. Choi, K. Castleman, and A. Bovik, "Joint segmentation and classification of M-FISH chromosome images," in *in Proc. 26th IEEE Ann. Intern. Conf. (EMBS)*, San Fransisco, 2004, pp. 1636-1639.

[30] M. Sampat, A. Bovik, J. Aggarwal, and K. Castleman, "Supervised parametric and non-parametric classification of chromosome images," *Pattern Recognition*, vol. 38, no. 8, pp. 1209-1223, 2005.

[31] Y. Wang and K. Castleman, "Normalization of multicolor fluorescence in situ hybridization (M-FISH) images for improving color karyotyping," *Cytometry*, vol. 64, no. 2, pp. 101-109, 2005.

[32] W. Schwartzkopf, A. Bovik, and B. Evans, "Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images," *IEEE Transactions on Medical Imaging*, vol. 24, no. 12, pp. 1593-1610, 2005.

[33] S. Fortune, "A sweepline algorithm for Voronoi diagrams," *Algorithmica*, vol. 2, no. 1-4, pp. 153-174, 1987.

[34] S. DiZenzo, "A note on the gradient of a multi-image," *Comput. Vis. Graph. Image Process.*, vol. 33, pp. 116-125, 1986.

[35] A. Cumani, P. Grattoni, and A. Giuducci, "An edge-based description of color images," *Comput. Vis. Graph. Image Process.*, vol. 53, pp. 313-323, 1991.

[36] C. Drewniok, "Multi-spectral edge detection—Some experiments on data from Landsat-TM," *Int. J. Remote. Sens.*, vol. 15, no. 18, pp. 3743-3765, 1994.

[37] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed.: Prentice Hall, 2001.

[38] Y. Yan, S. Zhao, L. Wang, A. Zelenetz, and H. Schwartz, "Marker-controlled watershed for lymphoma segmentation in sequential CT images," *Med. Phys.*, vol. 33, no. 7, pp. 2452-2460, 2006.

[39] S. Beucher and F. Meyer, "Morphological segmentation," *J. Vis. Comm. Image Representation*, vol. 1, no. 1, pp. 21-40, 1990.

[40] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1163-1173, 1996.

[41] M. Grimaud, "A new measure of contrast: Dynamics," *Image Algebra Morphological Processing III*, vol. 1769, pp. 292-305, 1995.

[42] B. Smolka and A. Chydzinski, "Fast detection and impulsive noise removal in color images," *Real-Time Imag.*, vol. 11, no. 5, pp. 389-402, 2005.

[43] O. Henegariu et al., "Small marker chromosome identification in metaphase and interphase using centromeric multiplex FISH (CM-FISH)," *Lab. Investing.*, vol. 81, pp. 475-481, 2001.

[44] The ADIR M-FISH Image Database. [Online]. http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml

[45] I. Belyaev, "Molecular targets and mechanisms in formation of chromosomal aberrations: Contributions of Soviet scientists," *Cytogenet. Genome Res.*, vol. 104, pp. 56-64, 2004.

[46] R. Kettig and D. Landgrebe, "Classification of multispectral image data by extraction and classification of homogenous objects," *IEEE Trans. Geosci. Electron.*, vol. 14, no. 1, pp. 19-26, 1976.

[47] D. Landgrebe, "The development of a spectral-spatial classifier for earth observational data," *Pattern Recognit.*, vol. 12, no. 3, pp. 165-175, 1980.

[48] B. Smolka and A. Chydzinski, "Fast detection and impulsive noise removal in color images," *Real-Time Imag.*, vol. 11, no. 5, pp. 389-402, 2005.

[49] R. Lukac, B. Smolka, L. Martin, K. Plataniotis, and A. Venetsanopoulos, "Vector filtering for color imaging," *IEEE Signal Process. Mag.*, vol. 22, no. 1, pp. 74–86, 2005.

[50] R. Lukac and K. Plataniotis, *Color Image Processing: Methods and Applications*. Canada: CRC Press/Taylor & Francis, 2006.

[51] R. Lukac, B. Smolka, K. Plataniotis, and A. Venetsanopoulos, "Selection weighted vector directional filter," *Comput. Vis. Image Understanding*, vol. 94, pp. 140–167, 2004.

[52] L. Alparone, M. Barni, F. Bartolini, and R. Caldelli, "Regularization of optic flow estimates by means of weighted vector median filtering," *IEEE Trans. Image Process.*, vol. 8, no. 10, pp. 1462–1467, 1999.

[53] R. Lukac, "Adaptive color image filtering based on center-weighted vector directional filters," *Multidimensional Syst. Signal Process.*, vol. 15, no. 2, pp. 169–196, 2004.

[54] T. Liehr, U. Claussen, and H. Starke, "Small supernumerary marker chromosomes (sSMC) in humans," *Cytogenet. Genome Res.*, vol. 107, pp. 55-67, 2004.

[55] S. Uhrig et al., "Multiplex-FISH for Pre and Postnatal Diagnostic Applications," *Am. J. Hum. Genet.*, vol. 65, pp. 448–462, 1999.

[56] Y. Wang and A. Dandpat, "Classification of M-FISH Images using Fuzzy C-means Clustering Algorithm and Normalization Approaches," in *Proc. 38th IEEE Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, California, 2004, pp. 41-44.

[57] Y. Wang and A. Dandpat, "A hybrid approach of using wavelets and fuzzy clustering for classifying multispectral florescence in situ hybridization images," *Intern. Journal of Biom. Imag.*, pp. 1-11, 2006.

[58] H. Choi, K. Castleman, and A. Bovik, "Segmentation and fuzzy-logic classification of M-FISH chromosome images," in *Proc. IEEE Intern. Conf. on Image Processing (ICIP)*, Atlanta, 2006, pp. 69-72.

[59] H. Choi, K. Castleman, and A. Bovik, "Color Compensation of Multicolor FISH Images," *IEEE Trans. on Med. Imag.*, vol. 28, no. 1, pp. 129-136, 2009.

[60] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1-38, 1977.

[61] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[62] I. Belyaev, "Molecular targets and mechanisms in formation of chromosomal aberrations: Contributions of Soviet scientists," *Cytogenet. Genome Res.*, vol. 104, pp. 56-64, 2004.

[63] X. Wang et al., "Automated detection and analysis of fluorescent in situ hybridization spots depicted in digital microscopic images of Pap-smear specimens," *J. of Biomedical Optics*, vol. 14, no. 2, pp. 1-8, 2009.

[64] L. Ji, "Intelligent splitting in the chromosome domain," *Pattern Recognition*, vol. 22, no. 5, pp. 519-532, 1989.

[65] L. Ji, "Fully automatic chromosome segmentation," *Cytometry*, vol. 17, pp. 196-208, 1994.

[66] G. Ritter and G. Schreib, "Using dominant points and variants for profile extraction from chromosomes," *Pattern Recognition*, vol. 34, no. 4, pp. 923-938, 2001.

[67] N. Malpica et al., "Applying watershed algorithms to the segmentation of clustered nuclei," *Cytometry*, vol. 28, pp. 289-297, 1997.

[68] Q. Chen, X. Yang, and E. Petriu, "Watershed segmentation for binary images with different distance transforms," in *3rd IEEE Internat. Workshop on Haptic, Audio and Visual Environments and their Applications*, Ottawa, Canada., 2003, pp. 20-21.

[69] S. Beucher, "The watershed transformation applied to image segmentation," *Scanning Microscope*, vol. 6, pp. 299-314, 1992.

[70] J. Gauch, "Image segmentation and analysis via multiscale gradient watershed hierarchies," *IEEE Trans. Image Process.*, vol. 8, no. 1, pp. 69-79, 1999.

[71] K. Haris, S. Efstratiadis, N. Maglaveras, and A. Katsaggelos, "Hybrid image segmentation using watersheds and fast region merging," *IEEE Trans. Image Process.*, vol. 7, no. 12, pp. 1684-1699, 1998.

[72] T. Sumner, J. Evans, and A. Buckland, "A new technique for distinguishing between human chromosomes," *Nature New Biol.*, vol. 232, pp. 31-32, 1971.

[73] P. Karvelis, D. Fotiadis, and I Georgiou, "Region Based Segmentation and Classification of Multispectral Chromosome Images," in *20th IEEE International Symposium on Computer-Based Medical Systems, (CBMS 2006)*, Maribor, 2007, pp. 251-256.

[74] Karvelis. P and A. Likas, "Fully Unsupervised M-FISH Chromosome Image Characterization," *submitted*.

[75] P Karvelis, A. Likas, and D. Fotiadis, "Semi unsupervised M-FISH chromosome image classification," in *10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB), ,* Corfu, Greece, 2010, pp. 1-4.

[76] P Karvelis, A. Likas, and D. Fotiadis, "Identifying touching and overlapping chromosomes using the watershed transform and gradient paths," *Pattern Recognition Letters*, vol. 31, pp. 2474-2488, 2010.

[77] X. Wang et al., "Automated detection and analysis of fluorescent in situ hybridization spots depicted in digital microscopic images of Pap-smear specimens," *J. of Biomedical Optics*, vol. 14, no. 2, pp. 1-8, 2009.

[78] M. Wertheimer, *A source book of Gestalt psychology*. London: Routledge & Kegan Paul, 1938.

[79] L. Shapiro and G. Stockman, *Computer Vision*. New Jersey: Prentice-Hall, 2001.

[80] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146-168, 2004.

[81] M. Sezan, "A peak detection algorithm and its application to histogram-based image data reduction," *Graphical Models Image Processing*, vol. 29, pp. 47–59, 1985.

[82] M. Seabright, "A reapid banding technique for human chromosomes," *Lancet*, vol. 2, pp. 971-972, 1971.

[83] T. Ridler and S. Calvard, "Picture Thresholding Using an Iterative Selection Method," *IEEE Transactions on In Systems, Man and Cybernetics*, vol. 8, no. 8, pp. 630-632, 1978.

[84] M. Popescu et al., "Automatic karyotyping of metaphase cells with overlapping chromosomes," *Computers in Biology and Medicine*, vol. 29, no. 1, pp. 61-82, 1999.

[85] J. Piper, "On fully automatic feature measurement for banded chromosome classification," *Cytometry*, vol. 10, pp. 242-255, 1989.

[86] M. Petrou and C. Petrou, *Image Processing The Fundamentals*. Guildford: John Willey, 2003.

[87] N. Pala and S. Pala, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277-1294, 1993.

[88] C. Lundsteen, T. Gerdes, J. Maahr, and Philip J., "Clinical Performance of a system for semiautomated chromosome analysis," *Am. Journal of Human Genetics*, vol. 41, pp. 493-502, 1989.

[89] B. Lerner, H. Guterman, I. Dinstein, and Romem Y., "Medial axis transform-based

features and a neural network for human chromosome classification," *Pattern Recognition*, vol. 28, no. 11, pp. 1673-1683, 1995.

[90] B. Lerner, H. Guterman, and Dinstein I., "A classification-driven partially occluded object segmentation (cpoos) method with application to chromosome analysis," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2841-2847, 1998.

[91] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria," *IEEE Trans. System Man Cybernet.*, vol. 15, pp. 652-655, 1985.

[92] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, pp. 41-47, 1986.

[93] L. Hertz and R. Schafer, "'Multilevel thresholding using edge matching," *Computer Vision, Graphics, and Image Processing*, vol. 44, no. 3, pp. 279 - 295, 1988.

[94] K. Chen, D. Wang, and Liu X., "Weight Adaptation and Oscillatory Correlation for Image Segmentation," *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1106-1123, 2000.

[95] T. Caspersson, L. Zech, and C. Johansson, "Differential banding of alkylating fluorochromes in human chromosomes," *Experimental Cell Research*, vol. 60, pp. 315-319, 1970.

[96] G. Carters and J. Graham, "Trainable grey-level models for disentangling overlapping chromosomes," *Pattern Recognition*, vol. 32, no. 8, pp. 1335-1349, 1999.

[97] G. Carters and J. Graham, "Disentangling chromosome overlaps by combining trainable shape models with classification evidence," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 2080-2085, 2002.

[98] E. Arrighi and Hsu C., "Localization of heterochromatin in human chromosomes," *Cytogenetics*, vol. 10, pp. 81-86, 1971.

[99] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, 1990.

[100] J. Astola, P. Haavisto, and Y. Neuov, "Vector median filter," *Proc. IEEE.*, vol. 78, no. 4, pp. 678-689.

[101] P. Trahanias and A. Venetsanopoulos, "Vector directional filters - A new class of multichannel image processing filters," *IEEE Trans. Image Process.*, vol. 2, no. 4, pp. 528-534, 1993.

[102] P. Trahanias, D. Karakos, and A. Venetsanopoulos, "Directional processing of color images: Theory and experimental results," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 868-880, 1996.

[103] T. Viero, L. Oistamo, and Y. Neuvo, "Three-dimensional median related filters for color image sequence filtering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, no. 2, pp. 129-142, 1994.

[104] R. Lukac and B. Smolka, "Application of the adaptive center-weighted vector median framework for the enhancement of cDNA microarray images," *Int. J. Appl. Math. Comput. Sci.*, vol. 13, no. 3, pp. 369–383, 2003.

[105] Z. Ma, H. Wu, and D. Feng, "Partition-based vector filtering technique for suppression of noise in digital color images," *IEEE Trans. Image Process*, vol. 15, no. 8, pp. 2324–2342, 2006.

APPENDIX A

M-FISH emission charts for each chromosome class.

PSI-Kit

Chromosome Class	Channel 1	Channel 2	Channel 3	Channel 4	Channel 4
1	1	0	1	1	0
2	0	0	0	0	1
3	1	1	0	1	1
4	1	0	0	1	0
5	1	1	1	0	1
6	1	0	1	1	1
7	0	0	1	1	0
8	1	0	0	0	0
9	1	1	0	0	1
10	0	0	0	1	1
11	1	1	0	1	0
12	0	0	1	0	1
13	1	1	0	0	0
14	0	0	1	0	0
15	0	1	1	1	0
16	1	0	1	0	0
17	0	0	0	1	0
18	1	1	1	0	0
19	0	1	0	1	0
20	0	1	0	0	0
21	1	0	0	0	1
22	0	1	1	1	1
X	0	1	0	0	1
Y	1	0	0	1	1

PSI-Kit

Chromosome Class	Channel 1	Channel 2	Channel 3	Channel 4	Channel 4
1	0	0	1	0	0
2	0	0	0	1	0
3	1	0	0	0	0
4	0	1	0	1	0
5	0	0	1	0	1
6	0	1	0	0	0
7	0	0	1	1	0
8	0	0	0	1	1
9	0	0	0	0	1
10	1	0	1	0	1
11	1	0	0	1	0
12	0	1	1	0	0
13	1	1	0	0	0
14	0	1	1	1	0
15	1	0	1	1	0
16	0	1	0	0	1
17	0	1	0	1	1
18	0	0	1	1	1
19	0	1	1	0	1
20	1	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
X	1	0	0	0	1
Y	1	0	1	0	0

Vysis-Kit

Chromosome Class	Channel 1	Channel 2	Channel 3	Channel 4	Channel 4
1	0	0	1	0	0
2	0	0	0	1	0
3	1	0	0	0	0
4	0	1	0	1	0
5	0	0	1	0	1
6	0	1	0	0	0
7	0	0	0	0	1
8	0	0	0	1	1
9	0	0	1	1	0
10	1	0	1	0	0
11	1	0	0	1	0
12	0	1	1	0	0
13	1	1	0	0	0
14	0	1	1	1	0
15	1	0	1	1	0
16	0	1	0	0	1
17	0	1	0	1	1
18	0	0	1	1	1
19	0	1	1	0	1
20	1	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
X	1	0	0	0	1
Y	1	0	1	0	1

AUTHOR'S PUBLICATIONS

Journal Publications related to this thesis

1. P. Karvelis, D. Fotiadis, and I. Georgiou, "A Multichannel Watershed-Based Segmentation Method for Multispectral Chromosome Classification", *IEEE Transactions on Medical Imaging*, vol. 27, no. 5, pp. 697-708, 2008.

2. P. Karvelis, D. Fotiadis, D. Tsalikakis and I. Georgiou, "Enhancement of Multichannel Chromosome Classification Using a Region Based Classifier and Vector Median Filtering," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp.: 561-570, 2009.

3. P. Karvelis, A. Likas, D. Fotiadis "Identifying touching and overlapping chromosomes using the watershed transform and gradient paths," *Pattern Recognition Letters*, vol. 3, no. 16, pp. 2474-2488, 2010.

4. P. Karvelis, A. Likas "Fully Unsupervised M-FISH Chromosome Image Characterization," (*submitted IEEE Transactions on Information Technology in Biomedicine*).

Other Journal Publications

5. A. Tzallas, P. Karvelis, C. Katsis, D. Fotiadis, S. Konitsiotis, S. Giannopoulos, "An automated method for the classification of transient events in EEG recordings", *Methods of Information in Medicine*, vol. 45, no. 6, pp. 610-621, Dec. 2006.

6. M. Bourantas, M. Papafaklis, P. Karvelis, C. Katsouras, L. Michalis, "Comparison of Quantitative Coronary Angiography (QCA) with Intracoronary Ultrasound (ICUS). Can QCA accurately estimate the severity of a luminal stenosis?", *Angiology*, vol. 60, pp. 169 – 179, Apr 2009.

7. L. Athanasiou, P. Karvelis, V. Tsakanikas, K. Naka, L. Michalis, C. Bourantas, and D. Fotiadis, "A novel semi-automated atherosclerotic plaque characterization method using grayscale intravascular ultrasound images. Comparison with Virtual Histology," *IEEE Transactions on Information Technology in Biomedicine*, (accepted).

8. L. Athanasiou, P. Karvelis, T. Exarchos, V. Tsakanikas, K. Naka, L. Michalis, and D. Fotiadis, A hybrid plaque characterization method using intravascular ultrasound images, *Computers in Biology and Medicine*, (submitted).

Book chapters

1. Petros Karvelis and Dimitrios Fotiadis, "Recent Advances in Automated Chromosome Image Analysis," *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*.

Conference Publications

1. P. Karvelis, D. Fotiadis, M. Syrrou, I. Georgiou, Segmentation of chromosomes images based on a recursive watershed transform, *Proceedings of the 3rd European Medical & Biological Engineering Conference, (EMBEC 2005)*, pp. 497-502, Prague, 20-25 November 2005.

2. P. Karvelis, D. Fotiadis, I. Georgiou, and M. Syrrou, A Watershed Based Segmentation Method for Multispectral Chromosome Images Classification, *Proceedings of the 28th IEEE EMBS Annual International Conference, (EMBS 2006)*, pp. 3009-3012, New York City, USA, Aug 30-Sept 3, 2006.

3. P. Karvelis, D. Fotiadis, and I. Georgiou, "Region Based Segmentation and Classification of Multispectral Chromosome Images," *20th IEEE International Symposium on Computer-Based Medical Systems*, (CBMS 2006), June 20-22, Maribor, pp. 251-256, 2007.

4. P. Karvelis, A. Likas, and D. Fotiadis, "Semi Unsupervised M-FISH chromosome image classification," ITAB 2010: Corfu, *10th IEEE International Conference on Information Technology and Applications*.

5. L. Athanasiou, P. Karvelis, V. Tsakanikas, K. Stefanou, K. Naka, L. Michalis, G. Rigas and D. Fotiadis, "Atherosclerotic Plaque Characterization Using Geometrical Features from Virtual Histology Intravascular Ultrasound Images," *ITAB 2010 : 10th IEEE International Conference on Information Technology and Application*.

SHORTVITA

Petros Karvelis was born in Mesologi 1978. He received his BSc and MSc in Computer Science in 2001 and 2004 from the Computer Science Department, University of Ioannina, Greece. Since 2005 has been a Ph.D. Candidate in the Department of Computer Science, University of Ioannina, Greece.

He has been dealt with several issues concerning the processing of biomedical images, such as:

- chromosome image processing (MSc, Thesis & PhD Thesis).
- Intravascular Ultrasound (IVUS) image segmentation and classification,
- 3-D reconstruction of coronary arteries through IVUS images and angiography,

His research interests include medical image processing and color image processing.