The Dissertation Committee for Hyo Hun Choi
certifies that this is the approved version of the following dissertation:

Automatic Segmentation and Classification of Multiplex-Fluorescence In-Situ Hybridization Chromosome Images

Committee:

Alan C. Bovik, Supervisor

Kenneth R. Castleman

Joydeep Ghosh

Mia K. Markey

Thomas E. Milner

Automatic Segmentation and Classification of Multiplex-Fluorescence In-Situ Hybridization Chromosome Images

by

Hyo Hun Choi, B.S., M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2006

Dedicated to my parents and wife Myoungsun.

Acknowledgments

I would like to express my deepest gratitude to Dr. Alan Bovik and Dr. Kenneth Castleman for their support, patience, and encouragement throughout my graduate studies. I am grateful to meet and actually work with these two great persons, who have inspired me and affected my life deeply. I cannot express enough how much I appreciate their support. I also would like to thank Dr. Mia Markey, Dr. Thomas Milner, and Dr. Joydeep Ghosh for their support.

I am also grateful to spend time with all those bright people in our lab LIVE: Umesh, Mehul, Joonsoo, Yang, Sina, Kalpana, Sumohana, Farooq, Hamid, Abtine, Ragu, Shalini, and James. Especially, I would like to thank Umesh for giving me various help in the lab and sharing his time to discuss about anything.

Finally, I would like to thank my parents, my wife, my kids, and my brother and sister for their love and support. My life is meaningful because of them.

Automatic Segmentation and Classification of Multiplex-Fluorescence In-Situ Hybridization Chromosome Images

Publication No. _____

Hyo Hun Choi, Ph.D.
The University of Texas at Austin, 2006

Supervisor: Alan C. Bovik

Multicolor fluorescence *in-situ* hybridization (M-FISH) techniques provide color karyotyping that allows simultaneous analysis of numerical and structural abnormalities of whole human chromosomes. Chromosomes are stained combinatorially in M-FISH. By analyzing the intensity combinations of each pixel, all chromosome pixels in an image are classified. Often, the intensity distributions between different images are found to be considerably different and the difference becomes the source of misclassifications of the pixels. Improved pixel classification accuracy is the most important task to ensure the success of the M-FISH technique. Along with a reliable pixel classification method, automation of the karyotyping process is another important goal. The automation requires segmentation of chromosomes, which not only involves object/background separation but also involves separating touching and

overlapping chromosomes. While automating the segmentation of partially occluded chromosomes is an extremely challenging problem, a pixel classification method that satisfies both high accuracy and minimum human intervention has not been realized.

The main contributions of this dissertation include development of a new feature normalization method for M-FISH images that reduces the difference in the feature distributions among different images, and development of a new decomposition method for clusters of overlapping and touching chromosomes. A significant improvement was achieved in pixel classification accuracy after the new feature normalization. The overall pixel classification accuracy improved by 40% after normalization. Given a cluster, a number of hypotheses was formed utilizing the geometry of a cluster, pixel classification results, and chromosome sizes, and a hypotheis that maximized the likelihood function was chosen as the correct decomposition. Superior decomposition results were obtained using the new method compared to the previous methods.

Contributions also include development of a color compensation method for combinatorially stained FISH images (including M-FISH images) based on a new signal model for multicolor/multichannel FISH images. The true signal was recovered based on the signal model after color compensation. The resulting true signal does not have color spreading (channel crosstalk) among different color channels. Two new unsupervised nonparametric classification methods for M-FISH images are also introduced in this dissertation: a fuzzy logic classifier and a template matching method (a minimum distance clas-

sifier). While both methods produce an equivalent accuracy compared to a supervised classification method, their computation time is significantly less than a Bayes classifier.

Highly sophisticated and practical algorithms have been developed through this research. Using the developed methods, the amount of human intervention required will be significantly reduced: chromosomes are reliably and accurately segmented from the background, pixels are accurately classified, and clusters of overlapping and touching chromosomes are automatically decomposed.

Table of Contents

Acknowledgments **v**

Abstract **vi**

List of Tables **xii**

List of Figures **xiv**

Chapter 1. Introduction **1**
1.1 Motivation 1
1.2 Contributions 4
1.3 Organization 7

Chapter 2. Background **8**
2.1 Chromosomes 8
2.1.1 Chromosomal Aberrations 9
2.1.2 Brief History of Karyotyping 12
2.1.3 Multicolor FISH Karyotyping systems: M-FISH and SKY 14
2.2 Conventional Karyotyping 15
2.3 Multichannel/Multispectral Image Classification 16
2.4 Previous M-FISH Pixel Classification Methods 19
2.5 ADIR M-FISH Database 21

Chapter 3. Normalization Methods for M-FISH Images **22**
3.1 Introduction 22
3.2 Motivation 23
3.3 Image Registration 25
3.3.1 Source of Misalignment 27
3.3.2 Image Registration of M-FISH Images 28

3.4 Background Correction and Color compensation 30
3.4.1 Signal Model 33
3.4.2 Background Correction 36
3.4.3 Color Compensation 38
3.5 Results of Color Compensation of M-FISH Images 43
3.5.1 Example of Computing the Color Spread Matrix 44
3.5.2 Quantification of Image Quality Improvement After Color Compensation 46
3.5.3 Background Correction and Color Compensation Results 47
3.6 Expectation Maximization Normalization 53
3.7 Results of EM Normalization 62
3.8 Conclusion 63

Chapter 4. Pixel Classification Methods For M-FISH Images 67
4.1 Foreground-background segmentation 69
4.1.1 Detailed Procedure for Cell Removal 71
4.2 Supervised Classification Methods 74
4.2.1 Maximum-Likelihood Classifier 74
4.2.2 k Nearest Neighbor Classification 77
4.3 Unsupervised Classification Methods 78
4.3.1 Minimum Distance Classifier 79
4.3.2 Fuzzy Logic Classifier 81
4.4 Postprocessing Methods 82
4.4.1 Majority and Plurality Filtering 82
4.4.2 Prior Adjusted Reclassification 83
4.5 Accuracies of Classification Methods Before and After Normalization 85

Chapter 5. Decomposition of overlapping and touching M-FISH chromosomes 96
5.1 Introduction 96
5.2 Background 97
5.2.1 G-banded Chromosome Decomposition 97
5.2.2 M-FISH Chromosome Decomposition 99

5.3 Methods 102
5.3.1 Elements of clusters 102
5.3.2 Concave Points Detection 107
5.3.3 Evaluation of the hypothesis 111
5.3.4 Decomposition Steps 116
5.4 Results 117
5.5 Conclusion 120

Chapter 6. Conclusion and Future Work **126**
6.1 Conclusions 126
6.2 Future Work 127
6.2.1 Pixel classification accuracy 127
6.2.2 Automatic chromosome cluster decomposition 128

Bibliography **130**

Vita **139**

List of Tables

2.1 Nomenclature and chromosome classification [1]. 12

3.1 Pixel values of chromosome 1. Even though chromosome 1 is stained with DAPI and Gold, there is no obvious pattern in feature values because of channel crosstalk and independent integration time per channel. 25

3.2 Chromosome Labeling Chart of Vysis M-FISH Probe 26

3.3 Color map: object 1 is stained with red dye, object 2 is stained with green dye, and object 3 is stained with blue dye 35

3.4 Example. Color labeling table (L), color spread matrix (M), and exposure times (R) 44

3.5 Image quality improvement. Subindex B represents before color compensation, and subindex A represents after color compensation 47

3.6 Pixel values numbered on Fig. 3.7 (b). 52

4.1 Training images 88

4.2 The overall classification accuracy. NP = no preprocessing, BC = background correction, and EM = expectation maximization normalization. 89

4.3 Classification accuracies [%] of the commonly cited images. Images with empty values in the ML method are used as training. 90

4.4 List of bad quality images. PQ = poor quality due to either ill-hybridization or wrong exposure times, CT = channel crosstalk, MA = misalignment, WP = wrong probe. 94

5.1 The number of occurrences of the basic shapes. 106

5.2 Training images used for the size parameter estimation. 114

5.3 Normalized mean chromosome sizes. NS_NCBI represents normalized mean chromosome sizes calculated using the known chromosome lengths (obtained from NCBI's website). NS_database represents the normalized mean chromosome sizes calculated from the ADIR M-FISH database. 115

5.4 Decomposition results. Ncc = number of chromosomes in a cluster, NC = number of clusters, and N_{WD} = number of wrong decomposition 120

List of Figures

1.1 An M-FISH image. Chromosomes are combinatorially labeled using 5 fluorophores and counterstained using DAPI. Each gray scale image corresponds to the intensity of the emission wavelength of each fluorophore. 2

2.1 Illustration of cell division and chromosome structure. Top row: chromosomes before cell division. Bottom row: chromosomes are duplicated for cell division. Different colors indicate that each copy of a chromosome (a homologous) is inherited from a parent. 9

2.2 Conventional G-banded chromosome image and its karyogram 17

3.1 Chromatic aberration among different channels is negligible in M-FISH images. 29

3.2 Image registration before and after. Channels 1, 4, and 5 are displayed as a color image. The misalignments of channel 4 and 6 are corrected, and as a result the classification accuracy increased from 43.45% to 66.81%. 31

3.3 SSIM values were computed between DAPI and Gold to find the amount of translation. Two images are registered where the SSIM value is the maximum. 32

3.4 Signal formation of M-FISH images. The measured signal $\mathbf{Y} = \mathbf{ECX}$. $\mathbf{x}_c$ is a true color channel and $\mathbf{y}_c$ is a captured color channel corresponding to a specific fluorophore. The matrices $\mathbf{C}$ and $\mathbf{X}$ are unknown. 39

3.5 Background correction. Elevated background intensity is removed after the background correction, but the channel crosstalk still remained. 48

3.6 Color compensation. The color compensation removed the channel crosstalk effectively. A significant increase in image quality is achieved on image (c). 49

3.7 Color compensation result on image V1301XY. 51

3.8 Segmentation result. Chromosomes are automatically segmented from background by utilizing 6 spectral information, global and local intensity, and edge information. Cells are also removed based on the size and circularity. 56

3.9 The mixture density distribution of $I_c(k)$ of V1401XX. 58

3.10 Distributions of training data and testing data. x and y axes are the feature values (thus, a feature vector forms a point in the figure). Each data set has its own fundamental error rate by its own distribution, but the classification accuracy for the testing data will be low because the distributions are different between the two data sets. The distributions should be normalized in order to obtain a high classification accuracy. 59

3.11 A marginal density function in (a) is normalized as in (c) by the piece-wise linear gray level mapping function in (b). The horizontal axes represent gray scale range. 62

3.12 Feature distribution (normalized histogram) of V1290562 before and after the EM normalization. x axis represents gray scale and y axis represents the normalized frequency of a gray level. The EM normalized images are shown in Fig. 3.13, and the classification result is shown in Fig 4.6. 64

3.13 EM normalization result of V1290562 before and after the EM normalization. 65

4.1 Segmentation result. Notice that the cell is effectively removed. 71

4.2 Segmentation steps. 72

4.3 Fuzzy logic classification and prior adjusted reclassification . . 86

4.4 Fuzzy logic classification and prior adjusted reclassification. The pixel classification accuracy improved from 86.15% to 93.26%. The chromosome 4 has a translocation of 9. The translocated segmented is not affected by the increased prior on chromosome 4. 87

4.5 Statistical significance of each classification method. The bootstrapping of each method. Left to right: NP_MD, NP_ML, BC_MD, BC_ML, EM_MD, and EM_ML. The error bars are drawn at the 95th percentile. 91

4.6 Classification result of V290562 (spectral channels are shown in Fig. 3.13). 92

4.7 Correct classification rate of individually self trained and tested images. Ten bins are used from 0 to 10, 10 to 20, . . . , 90 to 100. 93

4.8 Histogram of classification accuracies. x axis represents the classification accuracy [%], y axis represents the frequency. Ten bins are used from 0 to 10, 10 to 20, . . . , 90 to 100. Top to bottom: NP_MD, NP_ML, BC_MD, BC_ML, EM_MD, and EM_ML respectively. 95

5.1 Separation results of Ji's method [2]. 98
5.2 Possible separation lines of Agam and Dinstein's method [3] . 99
5.3 Several hypotheses for a cluster of three chromosomes [3]. . . . 99
5.4 Segmentation results of an M-FISH image by Schwartzkopf's method [4]. 100
5.5 (a) Schwartzkopf's method successfully decomposed touching chromosomes, whereas grayscale based method (using Cytovision software) could not since two chromosomes appear as a long chromosome. (b) Grayscale based method could decompose, whereas Schwartzkopf's method could not since two overlapping chromosomes belong to the same class [4]. 101
5.6 Elements of clusters. 103
5.7 Landmarks of a cross shape cluster 104
5.8 Landmarks of clusters 105
5.9 Boundary is smoothed using a wavelet denoising method. . . . 108
5.10 The derivative of the tangent of the boundary shown in Fig. 5.9. 109
5.11 Concave point detection. (a) Solid line is the original boundary, green circles are the smoothed boundary, ∗s represent the concave regions, and circles show the concave points of the boundary. (b) The concave points detected and are marked with red stars on the original boundary. 110
5.12 Probability density functions of the normalized chromosome sizes. 116
5.13 Landmarks of a cross shape cluster 117
5.14 Decomposition of a Cross shape cluster. Two chromosomes that belong to the same class crossing each other are successfully decomposed using the developed method. 118
5.15 Decomposition of a T-shape cluster. No previous methods could decompose this kind of partial overlaps correctly. 119
5.16 Landmarks of a cluster. 121
5.17 Decomposition results. (a) Cross case, (b) T case, (c) I case, (d) T and I case, (e) Cross and T case, and (f) Cross and T case 121
5.18 More results of chromosome cluster decomposition. The developed decomposition method is robust to misclassification errors. 122
5.19 Automatic karyotyping 123
5.20 Automatic karyotyping, continued from Fig. 5.19. In (b) translocation between 4 and 9 are shown, and overlapping chromosomes are segmented correctly and automatically. 124

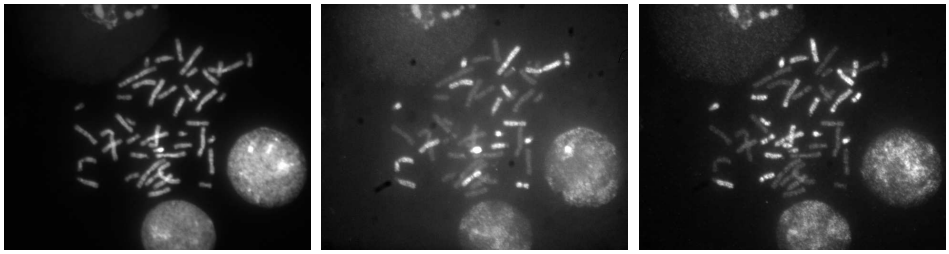
Chapter 1

Introduction

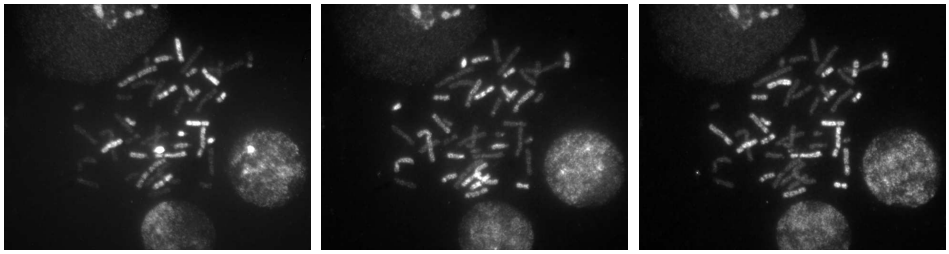
1.1 Motivation

The fluorescence in-situ hybridization (FISH) microscopic imaging modality has been widely used for the analysis of genes and chromosomes. Multiple fluorophores are often used combinatorially to visualize several biological specimens simultaneously. Using combinatorial labeling methods, $2^N - 1$ specimens can be discriminated using N fluorophores. When three fluorophores are used, seven specimens can be analyzed by binary combinations (presence or absence) of the fluorophores. N gray scale images of each specimen, stained with N fluorophores, can be obtained using a monochrome camera and a set of optical bandpass filters that are specifically designed for the excitation and emission wavelengths of the fluorophores.

In particular, multicolor (multiplex) fluorescence in-situ hybridization, called M-FISH, uses five fluorophores to uniquely identify all 24 chromosome types of the human genome. A sixth fluorophore, DAPI (4'-6-diamidino-2-phenylindole, a blue fluorescent dye), is used to counterstain the chromosomes [5, 6]. Thus, each pixel of an M-FISH image is composed of 6 values that correspond to the intensities of six fluorophores. Figure 1.1 shows an



(a) DAPI (b) Aqua (c) Green



(d) Gold (e) Red (f) Far Red

Figure 1.1: An M-FISH image. Chromosomes are combinatorially labeled using 5 fluorophores and counterstained using DAPI. Each gray scale image corresponds to the intensity of the emission wavelength of each fluorophore.

example of M-FISH images. By analyzing the combinations of the six spectral intensities, all of the chromosome pixels in an image are identified, and a pseudocolor is assigned based on the class the pixel belongs to [7, 8]. After the pixel classification, chromosomes are displayed in a standard format called the *karyogram*.

The M-FISH technique has been used for the characterization of structural rearrangements, such as translocations, to search for cryptic rearrangements, to study mutagenesis, tumors, and radiobiology [9]. In cancerous cells, translocations, or exchanges of chromosomal material between chromosomes, are extremely common.

Currently available M-FISH systems still exhibit misclassifications of multiple pixel regions due to a number of factors, including non-homogeneity of staining, variations of intensity levels within and between image sets, and emission spectra overlaps between fluorophores. The size of the misclassified regions are often larger than the actual chromosomal rearrangement. To reliably detect subtle and cryptic chromosomal aberrations, a highly accurate pixel classification method has to be developed. Along with a reliable pixel classification method, automation of the karyotyping process is another important goal. The automation requires segmentation of chromosomes, which not only involves object/background separation but also involves separating touching and overlapping chromosomes. While automating the segmentation of partially occluded chromosomes is an extremely challenging problem, a pixel classification method that satisfies both high accuracy and minimum human intervention has not been realized.

The grand goal of this research is to develop a completely automated and accurate chromosome analysis system. The scope of automation will expand as the technology progresses. However, based on currently available technologies, the envisioned automated system will be composed of three major parts: 1) automatic image capture, 2) automatic image analysis, and 3) statistical data analysis from the collection of previously analyzed data. Once a large batch of specimens is placed under a microscope, high quality infocused images will be captured automatically for all specimens. During image capture, each specimen will be automatically identified (using e.g. bar

code) and registered with the patients' medical information. In parallel with the image capture, images will be analyzed automatically, producing accurate karyotypes. A user can verify the results at this point and interpret the data, which will be stored along with the patients' information. After a substantial amount of information is accumulated, various statistical analyses can be performed, which will lead to a better understanding of the genes and the diseases.

My research objectives were to develop algorithms for improved pixel classification and algorithms for automated segmentation of touching and overlapping chromosomes.

1.2 Contributions

The contributions of this dissertation are as follows:

1. A new **feature normalization** method for M-FISH images [10], that reduces the difference in the feature distributions among different images using the expectation-maximization algorithm, is developed. In order to obtain a high classification accuracy in pattern recognition, feature normalization is a crucial part of classification after feature selection. In particular, when features are obtained independently, the normalization must be performed in order to reduce the intra-variance of the feature distribution among different images. In M-FISH, each channel is captured independently, and each channel has a different integration time

due to different signal strengths of fluorophores. As the relative intensity values across the six channels are used as features, intensity variations should be normalized prior to the pixel classification. The developed normalization method significantly increased the classification accuracy.

2. **Color compensation** method [11] for combinatorially stained FISH images is developed. FISH images including M-FISH images contain a certain amount of crosstalk between the color channels due to the overlap of excitation and emission spectra and the broad sensitivity of image sensors. This phenomenon is called color spread. Thus in M-FISH images, all chromosomes are visible on all channels with different intensity levels (Fig. 1.1). Furthermore, each fluorophore has a different sensitivity to the excitation wavelength. Thus some fluorophores require a short integration time while others require a long exposure time. A new **signal model** for M-FISH images is proposed and the true signal without the crosstalk is recovered based on the model.

3. Two new **unsupervised nonparametric classification methods** [10, 12] are developed. In an early stage of investigation regarding the structure of the data based on some features, an unsupervised method is desired since the samples are unlabeled. A fuzzy logic classifier and a template matching algorithm are the two methods. Both methods provide a significant advantage in terms of computation time compared to supervised methods, and their accuracies are comparable to that of a maximum-likelihood classifier.

4. A new **decomposition** method [13, 14] for overlapping and touching M-FISH chromosomes is developed. Automatic segmentation of partially occluded and/or touching objects is an extremely challenging task. Chromosome images are inherent with the partial occlusion and touching of chromosomes. This is one of the major factors that hinders automating the analysis. Previous chromosome decomposition methods utilized partial information of chromosome clusters resulting in limited success. A cluster was better decomposed by incorporating more knowledge. Multiple hypotheses were formed utilizing the geometry of a cluster, pixel classification results, and chromosome sizes. **Basic elements** of overlap and touching cases are proposed. These basic elements yield hypotheses of possible overlapping and/or touching cases. Given a cluster, multiple hypotheses are evaluated and the most likely hypothesis is chosen as the correct decomposition.

5. A new **postprocessing** method [12], that effectively corrects misclassified pixels while keeping the translocated pixels intact, is developed. By utilizing chromosome size and the likelihoods of pixel membership, the most likely class that an unknown chromosome belongs to is identified, and then pixels are reclassified with an increased prior for the most likely class. With an increased prior probability for a class, misclassified pixels are effectively corrected while translocated pixels remain intact.

1.3 Organization

The rest of the dissertation is structured as follows. In Chapter 2, historical, biological, and medical aspects about chromosomes and their analysis methods are described. In Chapter 3, several normalization methods including image registration, background correction, color compensation, and a new normalization method that uses the expectation maximization algorithm are discussed. Various supervised and unsupervised pixel classification methods are discussed in the Chapter 4, and a new unsupervised nonparametric classification method for M-FISH images are described in the same chapter. An automatic foreground and background segmentation method is also described in Chapter 4. In Chapter 5, a new decomposition method for overlapping and touching M-FISH chromosome images is described. Finally Chapter 6 concludes the dissertation.

Chapter 2

Background

2.1 Chromosomes

Chromosomes, the coiled strands of deoxyribonucleic acid (DNA), appear inside the nucleus during cell division (mitosis). Three billion base pairs (A, C, T, and G) that make up human DNA are organized into twenty four chromosome types: 22 autosomes and the X and Y sex chromosomes. Since chromosomes exist as a pair (one from each parent), there are 46 chromosomes (44 autosomes plus XX or XY) in almost every cell in the body. Mitotic cell division undergoes four phases: prophase, metaphase, anaphase, and telophase. During cell division, two strands of DNA double helix, which have complementary sequences to each other, are separated and make complementary strands of themselves. Then a new strand and a complementary old strand combine, resulting in two pairs of two strands of double helix (interphase). Those pairs of strands are compactly coiled forming chromosomes (in prophase), and chromosomes line up on the equatorial plane (metaphase). A typical chromosome structure is illustrated in Fig. 2.1. The position of the centromere varies from one chromosome to the next, and chromosomes are categorized by its position: metacentric - centromere placed in the middle, submetacentric - centromere placed between the middle and the terminal, acrocentric - centromere placed

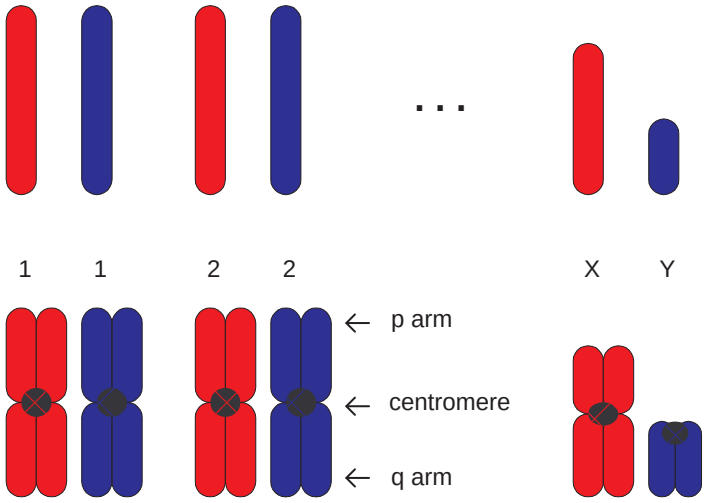


Figure 2.1: Illustration of cell division and chromosome structure. Top row: chromosomes before cell division. Bottom row: chromosomes are duplicated for cell division. Different colors indicate that each copy of a chromosome (a homologous) is inherited from a parent.

nearly at the terminal, and telocentric - centromere placed truly at the terminal (this does not occur in humans). Along with size, the centromere position is an important feature of identifying chromosomes. As the cell divides into two, each chromosome segregates at the centromere and the two halves move to each daughter cell (anaphase). The division completes in telophase.

2.1.1 Chromosomal Aberrations

The human genome (complete set of DNA) is estimated to contain 20,000-25,000 genes, which encode instructions on how to make proteins. Chromosomes are comprised of 50 million to 250 million base pairs. Thus each chromosome contains many genes. Any structural or numerical abnormality

of chromosomes corresponds to a genetic disease. Over 4000 known diseases are linked to genetic abnormalities, and these diseases are currently a major cause of infant mortality [15].

Chromosomal aberrations occur in various forms, and can be categorized into: numerical and structural aberrations.

Numerical aberrations can occur when one of two daughter cells receives both chromosomes of a pair and the other daughter cell receives none during cell division due to, for example, a malfunction of the spindle apparatus. If a chromosome is missing from a pair, it is called **monosomy**. If an extra chromosome is present, it is called a **trisomy**. A common autosomal trisomy in humans is Down's syndrome (trisomy 21). Approximately one in every 800 births is affected with Down's syndrome [15]. Other trisomies include trisomy 8, trisomy 9, trisomy 18, trisomy 22, Triplo-X syndrome (47,XXX), and Klinefelter's syndrome (47,XXY). All other autosomal trisomies are lethal in the embryonic or fetal stage. A common monosomy is Turner's syndrome (monosomy X). Approximately 1 in 2000 female births is affected with Turner's syndrome [16]. Women with this syndrome are usually short in height and are infertile.

Structural aberrations include translocations, insertions, deletions, and inversions. **Translocation** is a rearrangement of a chromosome in which a segment is moved from one location to another, either within the same chromosome or to another chromosome. A segment of a chromosome also can be deleted (**deletion**) from a chromosome, and inserted into another chromo-

some (**insertion**). When a segment including the centromere is inverted, it is called a **paracentric inversion**, and inversion of another region is called a **pericentric inversion**.

Like other genetic diseases, cancer also results from gene mutations. However, with a few exceptions, cancer is not an inherited genetic disease. Cancer is defined as uncontrolled growth of mutated cells (malignant tumor). Cancer is the second leading cause of death in America after heart disease. Half the cancers occur in three organs: lung (28%), colon (13%), and breast (9%). It is estimated that approximately 90% of all cancers are due to environmental factors [15].

Complex chromosomal aberrations are commonly found in cancerous cells. In order to study which genes are responsible for, and to reliably diagnose the disease, many karyotyping techniques have been developed. Using older staining methods, chromosomes could not be uniquely identified. After a staining method that visualized the chromosome banding patterns became available, all chromosomes were uniquely identified, and even their segments could be analyzed, with limitations. The resolving power of chromosomal abnormalities was greatly enhanced as the techniques of molecular cytogenetics evolved. The following section describes the brief historical background of karyotyping techniques.

Group	Chromosome number
A	1 2 3
B	4 5
C	6 7 X 8 9 10 11 12
D	13 14 15
E	16 17 18
F	19 20
G	21 22 Y

Table 2.1: Nomenclature and chromosome classification [1].

2.1.2 Brief History of Karyotyping

Scientists believed the right chromosome number for human was 48 until Tjio and Levan in 1956 found that it was 46 [16]. In 1960, a number of investigators in human cytogenetics met in Denver, Colorado, to propose standard system of chromosome nomenclature of human mitotic chromosomes. The karyotyping techniques available before 1971 did not reveal the internal structure of chromosomes, and thus chromosomes were classified into only seven groups based on the length and the position of the centromere. The nomenclature and the classification are shown in Table 2.1.

In 1968, T. Caspersson and his colleagues in Stockholm reported that certain fluorescent derivatives of quinacrine bind differentially to different parts of chromosomes [17]. This discovery lead to development of Q-banding staining techniques. Giemsa banding (G-banding), developed in 1971, created a unique pattern of bands on each chromosome, making it possible to identify every chromosome and even segments of chromosomes. Since then, Giemsa banding has been used as a standard technique in pre- and postnatal diagnostics.

However, when structural abnormalities such as a translocation, insertion, deletion, or inversion are present, G-banding alone cannot reliably decipher the information.

In the late 1980's, fluorescence in situ hybridization (FISH), the technique which uses fluorescent molecules to paint genes or whole chromosomes, was developed [18]. FISH provides visualization and analysis of multiple genes and chromosomes simultaneously in a sample using either combinatorial or ratio-labeling methods. Ever since, FISH has greatly grown in popularity for chromosome or gene identification and analysis.

In 1996, Speicher *et al.* [7] and Schröck *et al.* [6] introduced systems utilizing multicolor fluorescence in-situ hybridization technique (M-FISH), which is a combinatorial labeling technique developed for the simultaneous analysis of all human chromosomes. To be able to distinguish 24 human chromosomes simultaneously, a minimum of 5 fluorophores are required. Each chromosome is stained with a unique combination of fluorophores, and thus every chromosome is uniquely and simultaneously identified. A sixth fluorophore, DAPI, is counterstained to all chromosomes [9]. In particular, a system developed by Speicher *et al.* is called multiplex-FISH (M-FISH) and a system developed by Schröck *et al.* is called spectral karyotyping (SKY). Both systems have been proven to be useful for the characterization of complex chromosomal rearrangements in cancer cells and for detecting cryptic translocations.

2.1.3 Multicolor FISH Karyotyping systems: M-FISH and SKY

Currently, there are two types of multicolor FISH imaging systems: a system developed by Speicher *et al.*, so called 'multiplex fluorescence in situ hybridization' (M-FISH), which uses a set of optical filters, and a system developed by Schröck *et al.*, so called 'spectral karyotyping' (SKY), which uses an interferometer. In a set of optical filter based system (M-FISH), six images per metaphase spread of corresponding fluorophores are captured using optical bandpass filters. In contrast, the spectral imaging system (SKY) uses a Sagnac interferometer to record the emission spectra of the chromosomes and assigns each pixel a pseudocolor based on its spectrum. Applied Imaging and Metasystems are the major companies that produce M-FISH systems. Applied Spectral Imaging (ASI) is the only producer of the SKY system. Each system has advantages. While the SKY system requires only a single exposure, the M-FISH system takes 6 images with a change of optical filters. However usually SKY needs a longer exposure time than M-FISH. It is easy and inexpensive to add a set of optical filters to an existing microscope to perform the M-FISH analysis whereas the SKY camera is complex and expensive. In addition, the M-FISH system provides an opportunity to users to examine each spectral image in order to verify the classification results. Neither the classification accuracy nor the resolution limit of both systems has been well established. In my research, images from the optical filter based system were used.

2.2 Conventional Karyotyping

Conventional karyotyping uses Giemsa-staining on chromosomes which produces bright and dark patterns (Figure 2.2). The banding patterns are unique to the each type of chromosomes. These bandings are called G-banding. G-banding based chromosome analysis is a standard method for pre and postnatal chromosome analysis, which is routinely performed in clinical laboratories. The relevant image processing, feature extraction, feature normalization, and classification methods have been studied in the past for over 30 years [17, 19–24]. The size, shape, centromere position, and banding patterns are commonly used as features for the chromosome classifications. The general procedure of the classification is to first reduce the noise in the images, enhance the chromosomes [25, 26], segment chromosomes from the background automatically or semi-automatically [2, 3, 27], extract features [28–31], classify chromosomes, and display them in a specific format called karyogram. The classification is usually performed after the chromosome segmentation. Once chromosomes are segmented, the medial axis [32] of chromosomes are found to straighten the chromosome and thus extract the banding patterns [29]. The banding pattern based technique is effective and accurate in determining if the cell line is aneuploid (condition of having abnormal number of chromosomes). However, it is almost impossible to perform the analysis of complex structural abnormalities based on the banding patterns. With recent advances in molecular cytogenetic techniques, a variety of FISH assays have been developed. Since the introduction of M-FISH, it has significantly gained in popularity as

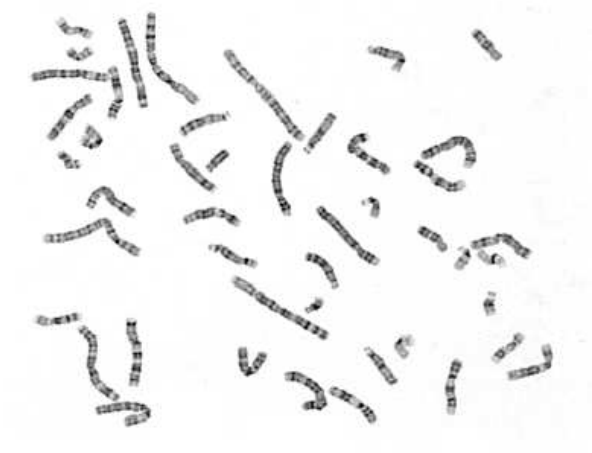
an alternative for chromosome analysis.

2.3 Multichannel/Multispectral Image Classification

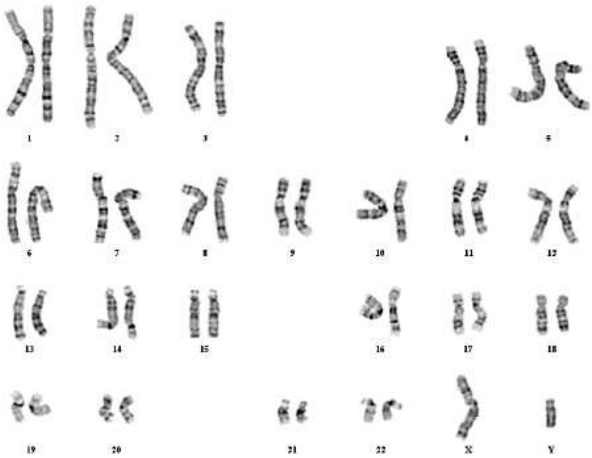
Since M-FISH images are multispectral images, it is worthwhile to investigate segmentation methods in other multispectral images produced by different modalities. Multispectral or multidimensional images are common in medical imaging modalities such as MRI and in geoscience remote sensing. Numerous image segmentation methods based on pixel classification have been developed for various multispectral images.

Depending on the modalities, N number of images are captured, where N ranges from 2 to more than a hundred. Imaging modalities for remote sensing are satellite, multiple band sensors operated from a spaceborne or an airborne platform such as landsat seven-band Thematic mapper (TM), four-band Multispectral Scanner (MSS), and three-band Satellite Pour l'Observation de la Terra (SPOT). In MRI imaging, spin-lattice relaxation time (T1), spin-spin relaxation time (T2), and proton density (PD) are used to capture differences on chemical shifts. The purpose of image segmentation in remote sensing is to study minerals, crops, urban structure and growth, etc. In medical images, different tissue sections are segmented to study anatomy and functions of organs or to diagnose diseases.

Most segmentation approaches in both areas avoid supervised methods due to the expensive process of collecting ground truth. In unsupervised segmentation of multispectral images, virtually all classification techniques



(a) G-banded chromosome image



(b) Karyogram

Figure 2.2: Conventional G-banded chromosome image and its karyogram

currently known have been tried such as Markov random field model based method [33], neural network based methods [34–38], hierarchical clustering and fuzzy classification [39], and k-means clustering [40]. In [40], k-means clustering was used to group the unlabeled data. From the clustered data the class parameters were extracted, and pixels were classified using a supervised method.

Support vector machine (SVM) based approaches have been tried when the number of labeled data points are small, where SVM subsequently refines the support vector (SV) by actively querying for the most ambiguous point in the data [41]. A subspace projection method such as an orthogonal subspace projection (OSP) was also tried in remote sensing for hyperspectral images [42]. While quite useful, OSP has a limitation that the number of spectral bands should be larger than the number of classes. To overcome the limitation of OSP, a Kalman filter based linear mixing approach was developed [43].

Shackelford *et al.* [44] tried combining fuzzy pixel-based and object-based methods for the classification of high-resolution multispectral data over urban areas, and they showed that the fuzzy classifier performed better than the maximum-likelihood classifier, and using the object based approach, correct classification rates increased further, ranging from 76% to 99% depending on the objects. Raghu *et al.* [45] used textural features as well as the spectral information. Each spectral image was filtered by Gabor wavelets and the resulting images concatenated to form new feature vectors.

Different classifiers will produce different accuracies depending on the

feature distribution and the amount of variances between image sets. Most of methods only focus on applying different classification methods to obtain high accuracy instead of preprocessing the signal in a proper way to reduce noise in the features and variances between image sets.

2.4 Previous M-FISH Pixel Classification Methods

The first M-FISH system described in the literature was introduced by Speicher *et al.* [5] in 1996. Their classification method was based on the binary combinations of fluorophore intensities at each pixel. Binary values were obtained after thresholding each channel. This method is simple and fast (considering only the pixel classification time, excluding the time involved in manual corrections of the segmentation map), and does not require generation of a training data set. Their approach demonstrated the usefulness of the M-FISH technique.

In 1998 Eils *et al.* [46] introduced a method called the adaptive region-oriented approach. An image was initially divided into forty Voronoi polygons, and the polygons were subdivided iteratively until all polygons satisfied a homogeneity criteria. Neighboring polygons were merged if they were closer than a threshold distance in feature domain (six dimensional space).

Recently, a supervised 6-feature, 25-class maximum likelihood classification method was introduced in [47–49]. Twenty five classes included twenty four chromosome types and the background. Class distributions were assumed to be normal, and the class parameters were extracted from the training set,

a subset of ADIR's M-FISH database (available at `http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml`). By classifying every pixel in the image including both background and chromosome pixels, chromosomes were successfully segmented from the background. The pixel classification accuracy of this method was about 90% on a small number of images (the list of images and rates are shown in Table 4.3 in Section 4.5).

Schwartzkopf *et al.* [8] developed a joint pixel classification and segmentation method which can handle overlapping and touching chromosomes, using a maximum likelihood framework. After chromosome pixels were classified using a 6-feature, 24-class maximum likelihood classification method, touching and overlapping chromosomes were separated into single chromosomes by maximizing the likelihood of pixel membership and chromosome. While separating overlapping and touching chromosomes, misclassified pixels were corrected resulting in an increased classification accuracy from the initial pixel classification. The initial pixel classification accuracy significantly varied depending on the images, ranging from 20% to 90%. The mean pixel classification accuracy was 68% with a standard deviation of 17.5% [4]. The pixel classification error rate decreased by nearly 50% after using the joint segmentation and classification method [8].

Choi *et al.* [49] have emphasized the importance of feature normalization, and performed background correction and color compensation in order to reduce the background elevation and channel crosstalk. A detailed description of M-FISH image color compensation can be found in [11]. Wang and

Castleman [50] also performed background correction as a normalization step, and reported that after testing on five images, the pixel classification accuracy increased on average from 83% to 91% (the list of images and rates are shown in Table 4.3 in Section 4.5).

2.5 ADIR M-FISH Database

All our research was conducted using the ADIR M-FISH database. The database contains M-FISH images of 203 metaphase spreads from 33 slides. Applied Spectral Imaging, PSI (predecessor to ADIR), and Vysis are the three probe sets that were used for the specimen preparation. Three sets of image file formats are available: PSI format (requires PSI's software to read), PNG format, and JPEG format. Each image is accompanied by ground truth, except for 17 images that are marked as extreme (EX). The set of PNG format images were used in our experiment, and a total of 185 images were tested (85 images for Vysis, 71 images for ASI, and 29 images for PSI). There are 86 Vysis probe images but V1301XY and V1304XY are the same (only V1301XY was used). In the ground truth image, background pixels are assigned value 0, pixels in overlapped region are assigned value 255, and chromosome pixels are assigned a value from 1 to 24 depending on chromosome type. In the case of a translocation, the whole chromosome is labeled as the class which makes up most of the chromosome. The dimension of the images is $647 \times 517 \times 6$ for all images except for two images, V261054 and V270659, whose dimension is $768 \times 568 \times 6$.

Chapter 3

Normalization Methods for M-FISH Images

3.1 Introduction

In order to achieve a high accuracy in pattern recognition, selection and extraction of good features is the most important design factor. Different classifiers may produce different accuracies, but the accuracy is fundamentally bounded by the sample distribution in the feature space. Thus, feature normalization is also a crucial part of classification after feature selection. In particular, when features are obtained independently, the normalization must be performed in order to reduce the intra-variance of the feature distribution among different images. In M-FISH, each channel is captured independently, and each channel has a different integration time due to different signal strengths of fluorophores. As the relative intensity values across the six channels are used as features, intensity variations should be normalized prior to the pixel classification.

In this chapter, normalization methods for M-FISH images are described, which include image registration, background correction, color compensation, and expectation maximization normalization. The expectation maximization normalization significantly increased the pixel classification ac-

curacy.

3.2 Motivation

In M-FISH, 6 fluorophores are combinatorially used to discriminate 24 chromosome types. The color map of the Vysis probe set is shown in Table 3.2. According to the color map, chromosome 1, for example, is stained with DAPI and spectrum Gold dyes. Ideally chromosome 1 should be observed only in the DAPI and Gold channels and should not be visible in the other channels. However, due to the overlap of excitation and emission spectra and the broad sensitivity of image sensors, the obtained images contain a certain amount of crosstalk between the color channels. This phenomenon is called color spread [51]. Thus all chromosomes are visible on all channels with different intensity levels (see Fig. 1.1). Furthermore, each fluorophore has a different sensitivity to the excitation wavelength. Thus some fluorophores require a short integration time while others require a long exposure time. Especially Aqua and Gold dyes require long exposure times in order to visualize the hybridized chromosomes. An example of integration times is [DAPI, Aqua, Green, Gold, Red, Far Red] = [0.14, 6, 0.76, 6, 2.96, 1.4] seconds. When a pixel belongs to chromosome 1, the obtained intensity values are expected to have a pattern of [High, Low, Low, High, Low, Low] for Vysis probes. Unfortunately this pattern can be easily broken when each channel is independently acquired. A long exposure time amplifies the leaked intensity, and in some cases it can be higher than the chromosome intensities on other channels at the same pixel

location. The different DC offset levels of each channel of the imaging device (e.g., three channel color CCD) and non-flat background elevation also bias the signal intensity upward. Furthermore, chromosomes appearing in one spectral channel exhibit different intensity levels: some are darker or brighter than others, partially because of the non-flat background, but more substantially because of the different fluorophore sensitivities for different chromosomes. Examples of real pixel values of chromosome 1 across multiple images are shown in Table 3.1. As shown in the table, there is no obvious pattern in the feature values for the preceding reasons.

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 will yield low accuracy.

As long as k classes are grouped separately in the feature space even if the feature distribution differs from image to image, pixels can be accurately classified without estimating class parameters using unsupervised-nonparametric clustering methods such as k-means clustering or fuzzy k-means clustering. However, when the number of classes is not fixed (e.g. chromosome images: the number of chromosome classes differs by gender or diseases), finding the right number of classes after the clustering by cluster validation adds complexity and may cause inaccuracy. Pixels can be clustered into a maximum number of classes (24 clusters for M-FISH data) using these methods, and clusters that are closer than a threshold should be merged. The threshold

Images	Location (x,y)	Spectrum					
		DAPI	Aqua	Green	Gold	Red	Far Red
v1301xy	243,172	79	75	56	79	52	51
v1312xy	352,194	54	75	54	101	48	50
v1401xy	251,314	176	75	60	44	50	27

Table 3.1: Pixel values of chromosome 1. Even though chromosome 1 is stained with DAPI and Gold, there is no obvious pattern in feature values because of channel crosstalk and independent integration time per channel.

will be again data dependent, which will be different for different images.

Therefore, regardless of the choice of classifiers the variations of the feature distribution should be minimized in order to obtain overall high accuracy in pixel classification.

3.3 Image Registration

The misalignment of spectral images is an inevitable phenomenon due to the fundamental optical properties of the microscopic imaging system. When it occurs, it adversely affects the classification accuracy on pixels, especially on the edges of chromosomes. However we found only four images in the database that exhibited a noticeable misalignment, which are not from the optical properties but from some other unknown source of errors (possibly errors from the software or hardware of the image capturing devices). While image registration of multichannel or multitemporal images is an active research area [52, 53], we do not find a need for an image registration algorithm for M-FISH images. Therefore in this section we will briefly discuss the fun-

Chromosome	Spectrum					
	DAPI	Aqua	Green	Gold	Red	Far Red
1	x			x		
2	x				x	
3	x	x				
4	x		x		x	
5	x			x		x
6	x		x			
7	x					x
8	x				x	x
9	x			x	x	
10	x	x		x		x
11	x	x			x	
12	x		x	x		
13	x	x	x			
14	x		x	x	x	
15	x	x		x	x	
16	x		x			x
17	x		x		x	x
18	x			x	x	x
19	x		x	x		x
20	x	x			x	x
21	x	x	x	x		
22	x	x	x		x	
X	x	x				x
Y	x	x		x		

Table 3.2: Chromosome Labeling Chart of Vysis M-FISH Probe

damental cause of misalignment of M-FISH images, and present an example of misaligned M-FISH images and its effect of classification accuracy due to misalignment.

3.3.1 Source of Misalignment

A basic principle of optics states that the focal length changes depending on wavelength [54]. Thus when wavelength changes, the in-focus plane of an object changes, resulting in axial chromatic aberration. Also the magnification is inversely proportional to the focal length. Thus depending on the wavelength, the magnification also changes, which results in lateral chromatic aberration. Both aberrations can be found, even in the best currently available objectives. Therefore, when multiple emission wavelengths are used to image the same object, as in fluorescence imaging, chromatic aberration is inevitable. Furthermore, mechanical vibration induced by filter changes may cause misalignment. Classification accuracy on pixels near the edges of chromosomes are affected when misalignment occurs.

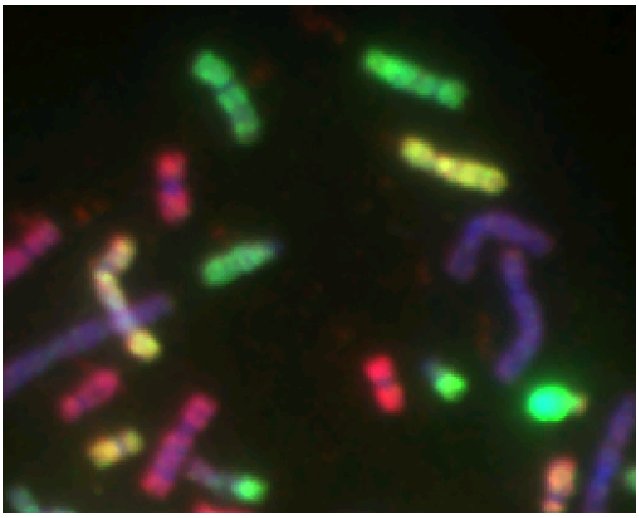
Misalignment due to chromatic aberration should be consistent throughout all images. Variations in the amount of misalignment should be due to other factors such as any mechanical vibrations. Fig. 3.1 shows an arbitrarily selected image from the database. Three channels are displayed as a color image: (DAPI in blue, Gold in green, and Far red in red channel). DAPI and Far red are selected since they have the farthest distance in wavelength, and the corner of an image is shown since the lateral chromatic aberration becomes

more severe as the distance gets further from the optical axis. As the figure shows, the amount of misalignment is negligible (even not noticeable). In fact, all the images we have observed in the database had a negligible amount of misalignment. The misclassifications on the edges of chromosomes commonly occur because intensities on those pixels are weak, making less certain of their memberships, and also when a chromosome appears larger due to blooming on one channel than on another channel, the non-overlapping area around the edge of the chromosome is misclassified. The intensity profiles at a chromosome in Fig. 3.1 (a) are shown in Fig. 3.1 (b). Channel 2, 4, and 6 are colored as R, G, and B. As the profiles show, there is no particular shift of one channel compared to the other, and as mentioned earlier chromosome widths appear differently.

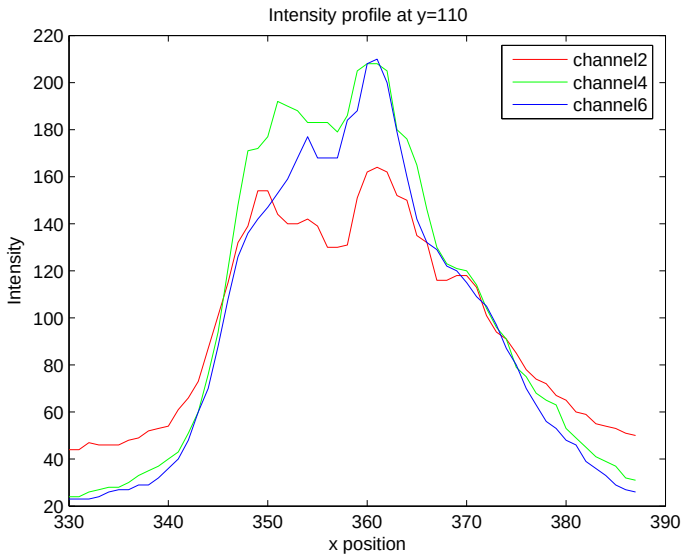
3.3.2 Image Registration of M-FISH Images

Only four images were identified as having severe misalignment (list is shown in Table 4.4 on 94), but the source of misalignment on those images is none of above. Channels were misaligned by simple translations of different amounts.

Fig. 3.2 shows images of before and after image registration. The DAPI channel was used as the reference and all other channels were registered to DAPI. Since the channels are misaligned by translation only, image registration is a rather a simple problem: a whole channel needs to move relative to the DAPI channel to locate the best matching place. Image registration was



(a) R: Far red, G: Gold, B: DAPI



(b) R: Aqua, G: Gold, B: Far red

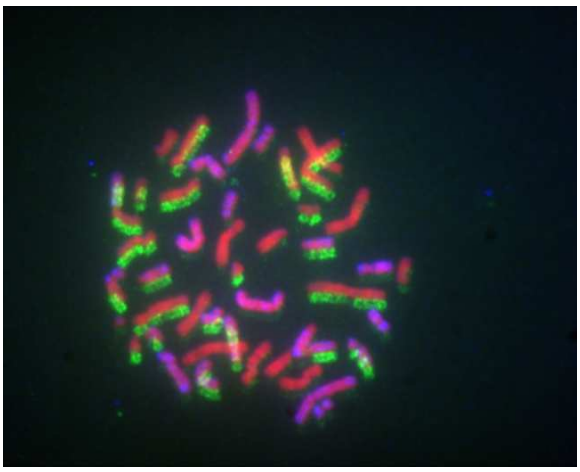
Figure 3.1: Chromatic aberration among different channels is negligible in M-FISH images.

performed using a recently developed metric called SSIM [55], which effectively measures the distance of the qualities of two images. SSIM is 1 only when two images are identical. As the degree of misalignment of two image increases, the quality between two images degrades. Since the amount of degradation in quality measured by SSIM highly correlates with human perception, SSIM values do not respond sensitively to small translations. However it serves our purpose of finding the amount of translation between two images. Fig. 3.3 shows the SSIM values between DAPI and Gold, where Gold was moved 25×25 pixels about the center of DAPI channel.

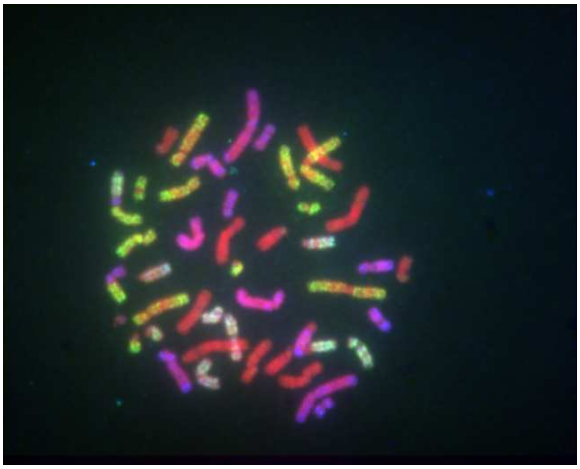
As a conclusion, image misalignment due to chromatic aberrations and mechanical vibrations in M-FISH was negligible in my study. However when misalignment should happen due to other reasons, it can severely affect the pixel classification accuracy. Misalignment due to translations can be corrected using SSIM.

3.4 Background Correction and Color compensation

In this section, a signal model for M-FISH images is introduced in order to recover the true signal based on the model [11]. Note that the true signal may not be exactly the same as the real true signal, but instead it means the ideal signal that we want to obtain after the signal processing. The signal model and its processing is described as follows.



(a) Channel 1, 4, and 5 before registration



(b) Registration of (a)

Figure 3.2: Image registration before and after. Channels 1, 4, and 5 are displayed as a color image. The misalignments of channel 4 and 6 are corrected, and as a result the classification accuracy increased from 43.45% to 66.81%.

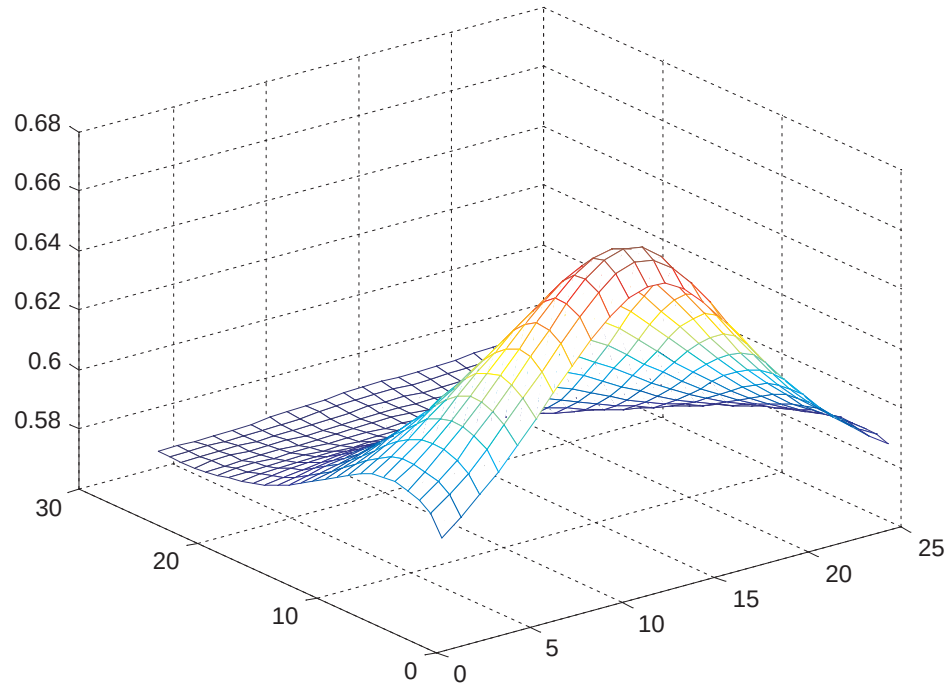


Figure 3.3: SSIM values were computed between DAPI and Gold to find the amount of translation. Two images are registered where the SSIM value is the maximum.

3.4.1 Signal Model

Castleman [51] has modeled the observed signal at a pixel, $\mathbf{y}$, as

$$\mathbf{y} = \mathbf{ECx} + \mathbf{b} \tag{3.1}$$

where $\mathbf{C}$ is the $N \times N$ color spread matrix that specifies how the colors are spread among the channels, $\mathbf{x}$ is the $N \times 1$ vector of true fluorophore intensities, $\mathbf{b}$ is the $N \times 1$ vector of black-level offsets of the imaging sensors (e.g. three channel color CCD or monochrome CCD), and $\mathbf{E}$ is the $N \times N$ diagonal matrix of exposure times of each channel. This model assumes that the gray levels are linear with brightness of the fluorophores.

Then the true signal $\mathbf{x}$ can be found, given $\mathbf{y}$, $\mathbf{E}$, $\mathbf{C}$, and $\mathbf{b}$ by

$$\mathbf{x} = \mathbf{C}^{-1}\mathbf{E}^{-1}\{\mathbf{y} - \mathbf{b}\}. \tag{3.2}$$

Normally, $\mathbf{y}$, $\mathbf{E}$, and $\mathbf{b}$ are given but the color spread matrix is not. Without the color spread matrix, the true signal cannot be recovered. When the specimens are uniquely stained, estimating the color spread matrix is relatively simple. As an example, suppose three biological objects are uniquely stained with three fluorophores as shown in Table 3.3. Then eq. (3.1) can be written

$$\begin{bmatrix} y_{ri} \\ y_{gi} \\ y_{bi} \end{bmatrix} = \begin{bmatrix} E_r & 0 & 0 \\ 0 & E_g & 0 \\ 0 & 0 & E_b \end{bmatrix} \begin{bmatrix} C_{rr} & C_{gr} & C_{br} \\ C_{rg} & C_{gg} & C_{bg} \\ C_{rb} & C_{gb} & C_{bb} \end{bmatrix} \begin{bmatrix} x_{ri} \\ x_{gi} \\ x_{bi} \end{bmatrix} + \begin{bmatrix} b_r \\ b_g \\ b_b \end{bmatrix} \tag{3.3}$$

The color spread matrix, which is dependent on the filter sets, fluorophores, and imaging sensors, can be found from three $\mathbf{y}$ vectors from three specimens. In this example, the intensities of the observed signal $\mathbf{y}_{i \in R,G,orB}$ are $[80,\ 15,\ 10]^T$ for red dye, $[5,\ 80,\ 30]^T$ for green dye, and $[10, 10, 160]^T$ for blue dye respectively, and $[E_r, E_g, E_b] = [1,\ 1,\ 2]$ and $\mathbf{b} = [0,\ 0,\ 0]^T$. R, G, or B is a set of indices of the specimen stained with red, green, or blue dye respectively. Knowing that the intensities of $\mathbf{y}_i$ are originated only from the intensity of red, green, or blue dye, the true pixel values are found by $\mathbf{x}_{i \in R} = [y_r + y_g + y_b/2,\ 0,\ 0]^T = [100,\ 0,\ 0]^T$, similarly $\mathbf{x}_{i \in G} = [0,\ 100,\ 0]^T$, and $\mathbf{x}_{i \in B} = [0,\ 0,\ 100]^T$. Here y_b is divided by 2 because of the integration time. After plugging $\mathbf{x}_i$ into (3.3), the nine unknowns of $\mathbf{C}$ can be found by solving nine linear equations from $\mathbf{E}^{-1}\mathbf{y}_i = \mathbf{C}\mathbf{x}_i$. Simply calculating the intensity ratios of $\mathbf{y}_i$ is the solution in this case. Thus, the color spread matrix in this example is

$$\mathbf{C} = \begin{bmatrix} 0.8 & 0.05 & 0.1 \\ 0.15 & 0.8 & 0.1 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}$$

The first column of the color spread matrix tells that 15% and 5% of the red intensity is spread to the green and blue channels respectively. The inverse matrix of the color spread matrix, called the color compensation matrix, corrects these color spreadings and recovers the true signal intensities. Once the color spread matrix is computed, it can be used for all other images that are captured using the same optical system and fluorophores.

When the specimens are combinatorially stained, the estimate of the true intensities from the observed signal cannot be done in the same way as

		Fluor spectra		
		R	G	B
	1	x	0	0
Objects	2	0	x	0
	3	0	0	x

Table 3.3: Color map: object 1 is stained with red dye, object 2 is stained with green dye, and object 3 is stained with blue dye

when uniquely stained specimens are available. In M-FISH, 6 fluorophores are combinatorially used to discriminate 24 chromosome types. According to the color map shown in Table 3.2, chromosome 1, for example, is stained with DAPI and spectrum gold dyes. In the following sections, the computation of the color spread matrix $\mathbf{C}$ from only the observed signal $\mathbf{y}$ and the exposure times, $\mathbf{E}$ is explained.

We have modeled the measured signal $\mathbf{y}$ of the M-FISH images as

$$\mathbf{y} = \mathbf{E}\{\mathbf{C}\mathbf{x} + \mathbf{b}\} + \mathbf{n} \tag{3.4}$$

where $\mathbf{y}$ is the 6×1 vector of the observed signal at a pixel, $\mathbf{x}$ is the 6×1 vector of the true signal, $\mathbf{C}$ is the 6×6 color spread matrix, $\mathbf{b}$ includes the DC-offset of the CCD and various factors that cause background (non-chromosome area) intensity elevation, $\mathbf{n}$ is the noise of the imaging device such as white noise and shot noise, and $\mathbf{E}$ is the 6×6 diagonal matrix of exposure times. The difference between Castleman's model and our model is that we assume that as the exposure time increases, the background intensity also increases linearly.

Six channels of the M-FISH image are first median filtered with a 3×3

kernel in order to eliminate the shot noise from **n**, and then lowpass filtered with a 3×3 kernel to remove the high frequencies which are mostly dominated by the white noise. Thus, the term **n** is minimized from eq. (3.4).

3.4.2 Background Correction

The background intensity, **b**, is mostly affected by the auto-fluorescence of the slide, the DC offset of the CCD, unattached free fluorescent molecules, the intensity of the defocused objects from out of depth of field, etc. Also, regions having a high density of objects usually have an elevated background intensity relative to regions without objects because of the flair effect. All these factors contribute to the non-flat intensity distribution of the background.

A two-dimensional cubic surface was estimated from the background pixels in order to approximate and remove **b**. The surface that has the minimum mean square error relative to the background pixels is the estimated two-dimensional cubic surface [51].

The background pixels for each channel are found by a k-means clustering method ($k = 2$), in which the threshold is found while iteratively regrouping pixels into two classes until the class means converge. Given a grayscale image I, the intensity distribution is assumed to be a mixture of two gaussians: $p(y|\omega_1) \sim N(\mu_1, \sigma_1)$, $p(y|\omega_2) \sim N(\mu_2, \sigma_2)$, and further assumed that $\sigma_1 = \sigma_2$. D is a set $\{y|y \in I\}$ of n unlabeled samples drawn independently from the mixture density

$$p(y) = p(y|\omega_1)P(\omega_1) + p(y|\omega_2)P(\omega_2)$$

The decision boundary that partitions D into two groups, D_1 and D_2, is computed by minimizing the sum-of-squared error

$$J = \sum_{i=1}^{2} \sum_{y \in D_i} \|y - \mu_i\|^2 .$$

μ_1 and μ_2 that minimize J are found iteratively using

$$T = \mu_1 P(\omega_1) + \mu_2 P(\omega_2) \tag{3.5}$$

where, T is the decision boundary, μ_1 and μ_2 are the class means, and $P(\omega_1)$ and $P(\omega_2)$ are the prior probabilities ($P(\omega_1) + P(\omega_2) = 1$). Given the initial estimates of μ_1 and μ_2, the initial T is found. Using the initial T, the new means are found. This minimization process is repeated until the class means or T converges. $min\{y\}$ and $max\{y\}$ are chosen as the initial values for μ_1 and μ_2 respectively. The samples in D_1, pixels below the threshold T, represent the background pixels.

Given the background pixels, the two-dimensional cubic surface is estimated as follows. The function for a two-dimensional cubic surface is

$$f(x,y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2 + c_6 x^2 y + c_7 x y^2 + c_8 x^3 + c_9 y^3 \tag{3.6}$$

where $c_0 \sim c_9$ are the coefficients that determine the surface shape, x and y are the coordinates, and $f(x,y)$ is the intensity value at (x,y). The ten coefficients are estimated from the given N background pixels by solving $\mathbf{f} = \mathbf{Bc}$, where $\mathbf{f}$ is a $N \times 1$ column vector containing intensity values, $\mathbf{B}$ is a $N \times 10$ matrix containing x and y coordinates, and $\mathbf{c}$ is a 10×1 column vector containing the

ten unknowns:

$$\mathbf{f} = \begin{bmatrix} f(x_1, y_1) \\ \vdots \\ f(x_N, y_N) \end{bmatrix} \mathbf{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_9 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & x_1 & y_1 & x_1y_1 & x_1^2 & y_1^2 & x_1^2y_1 & x_1y_1^2 & x_1^3 & y_1^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & y_N & x_Ny_N & x_N^2 & y_N^2 & x_N^2y_N & x_Ny_N^2 & x_N^3 & y_N^3 \end{bmatrix}$$

The least squares solution for $\mathbf{c}$ that minimizes the sum of the mean square error ($\mathbf{E} = \mathbf{f} - \mathbf{Bc}$) is determined by

$$\mathbf{c} = \left[\mathbf{B}^T\mathbf{B}\right]^{-1}\left[\mathbf{B}^T\mathbf{f}\right].$$

Using the coefficients $\mathbf{c}$ and eq. 3.6, a two-dimensional cubic surface that best fits the given background pixels is obtained. Finally, the surface is then subtracted from each channel of the image removing the above mentioned noises in $\mathbf{b}$ from eq. 3.4.

3.4.3 Color Compensation

After the background correction, the signal model becomes

$$\mathbf{y} = \mathbf{ECx}. \tag{3.7}$$

The formation of this signal can be viewed as in Figure 3.4. Six original signals of $\mathbf{x}$ are linearly mixed by the spread matrix $\mathbf{C}$. The observed signal before the exposure times are applied can be written $\mathbf{E}^{-1}\mathbf{Y}$. The goal is to solve for $\mathbf{X}$ and $\mathbf{C}$ given the observation $\mathbf{E}^{-1}\mathbf{Y}$. Finding the linear mixing matrix $\mathbf{C}$ and the original signal $\mathbf{X}$ from the observed signal is a problem similar to the

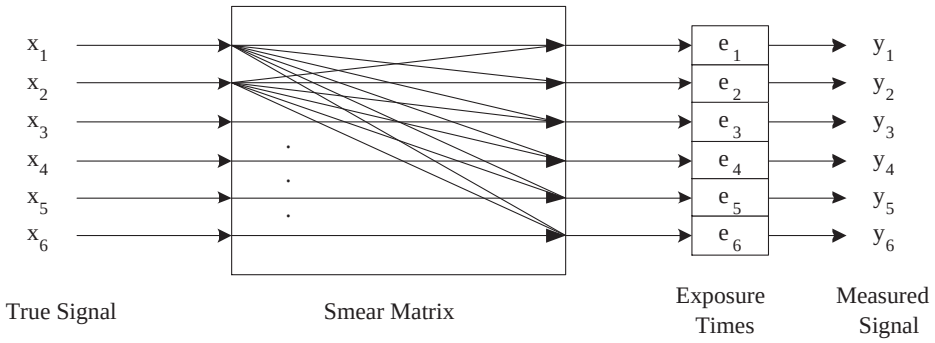


Figure 3.4: Signal formation of M-FISH images. The measured signal $\mathbf{Y} = \mathbf{ECX}$. $\mathbf{x}_c$ is a true color channel and $\mathbf{y}_c$ is a captured color channel corresponding to a specific fluorophore. The matrices $\mathbf{C}$ and $\mathbf{X}$ are unknown.

cocktail-party problem, where there are N speakers and N recording devices, and the N recorded signals are weighted sum of the N true signals. The recently developed technique called Independent Component Analysis (ICA) has been used quite successfully to estimate the mixing channel parameters, $\mathbf{C}$, from the mixed signal $\mathbf{Y}$ based on the assumption that $\mathbf{x}_i[n]$ are statistically independent of each other at every index n. ICA can also be used to separate the M-FISH mixed signal $\mathbf{Y}$ into 6 different statistically independent signals. However, the $\mathbf{X}$ that ICA estimates for M-FISH is not the same as the true signals since the combinatorial labeling causes dependencies among the true signals.

Let $\mathbf{y}_{i\in 1}$ be an observed signal that belongs to chromosome 1. A realistic set of pixel values of $\mathbf{y}_{i\in 1}$ may be $[170, 65, 45, 189, 70, 76]^T$. From $\mathbf{y}_{i\in 1}$ and the color table, we know that at least four values in $\mathbf{x}_{i\in 1}$ should be zero, i.e. $\mathbf{x}_{i\in 1} = [x_1(1), 0, 0, x_1(4), 0, 0]^T$. Values for $\mathbf{x}$ should be zero where the

staining is not present, and values remain unknown where fluorophores are present. Repeating the same exercise for all twenty-four chromosomes, we get 76 unknowns for the true signal, $\mathbf{X}$, where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{24}\}$, a collection of true samples from twenty four classes. All 36 values in the color spread matrix are unknown. Thus, the total number of unknowns are 112. Realistically, we may not find a unique solution given only the observed signal, but we can derive an optimal solution utilizing as much information as possible about the signal.

As shown in eq. (3.7), a pixel value from object 1 is $\mathbf{y}_1 = \mathbf{ECx}_1$. However, in practice, even after the careful noise removal, $\mathbf{y}_1$ will differ from $\mathbf{ECx}_1$: $\mathbf{y}_1 = \mathbf{ECx}_1 + \epsilon$, where ϵ may include factors not considered in our system model such as non-uniform hybridization inside of each chromosome, unremoved noise after median and lowpass filtering and background subtraction, and saturation of the pixels. Twenty four $\mathbf{y}$ vectors are needed to form the linear equations and to solve for 112 unknowns. $\mathbf{Y}$ is a collection of the observed samples from twenty four classes $\omega_{i\in(1,\ldots,24)}$, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{24}\}$, and $\mathbf{y}_i$ is a sample drawn from a normal distribution, i.e. $p(\mathbf{y}|\omega_i) \sim N(\boldsymbol{\mu}_{\mathbf{y}_i}, \Sigma_{\mathbf{y}_i})$. Instead of selecting an arbitrary sample from each class, $\boldsymbol{\mu}_{\mathbf{y}_i}$ are used in eq. 3.7 to compute the maximum likely color spread matrix $\mathbf{C}$ given the data (training images).

The matrix $\mathbf{Y}$ is

$$\mathbf{Y}^T = \begin{bmatrix} \mu_{\mathbf{y}_1} = \frac{1}{P_1}\sum_{k=1}^{P_1}\mathbf{y}_{1,k} \\ \mu_{\mathbf{y}_2} = \frac{1}{P_2}\sum_{k=1}^{P_2}\mathbf{y}_{2,k} \\ \mu_{\mathbf{y}_3} = \frac{1}{P_3}\sum_{k=1}^{P_3}\mathbf{y}_{3,k} \\ \vdots \\ \mu_{\mathbf{y}_{24}} = \frac{1}{P_{24}}\sum_{k=1}^{P_{24}}\mathbf{y}_{24,k} \end{bmatrix}$$

where $P_{i\in 1,2,3,\cdots,24}$ are the number of pixels that belong to each chromosome i, and $\mathbf{y}_{ik}$ is the k^{th} pixel value in chromosome i, $\mathbf{y}_{ik} = [y_{ik}(1), y_{ik}(2), y_{ik}(3), \cdots, y_{ik}(6)]^T$. If ϵ becomes negligible in $\mathbf{Y}$, $\mathbf{Y}$ can be expressed

$$\mathbf{Y} = \mathbf{ECX}. \tag{3.8}$$

The matrices $\mathbf{C}$ and $\mathbf{X}$ contain the unknowns. In order to form a system of linear equations, eq. (3.8) is written as

$$\mathbf{C}^{-1}\mathbf{E}^{-1}\mathbf{Y} - \mathbf{X} = 0. \tag{3.9}$$

The solution for eq. (3.9) should satisfy the following constraints.

1. The solution should satisfy eq. (3.9).

2. We assume that the intensity of all chromosomes stained with a particular dye should be the same in the original signal. For example, there are 10 chromosomes that are stained with green dye (see Table 3.2), and the mean intensity of each chromosome should be the same. This assumes that all objects have the same hybridization sensitivity to the same fluorophore. However, if there are differences in the sensitivity and

information is not given, then our assumption will give the best estimate. If the information is given, then the sensitivity ratios should be and can be incorporated into the equations.

3. The intensity between the input and output signals should be preserved, i.e. the sum along each column of $\mathbf{E}^{-1}\mathbf{Y}$ should be the same as the sum along each column of $\mathbf{X}$, $\sum \mathbf{E}^{-1}\mathbf{Y}_i = \sum \mathbf{X}_i$.

4. To satisfy the constraint (3), each column of the color spread matrix should sum to 1.

Using these constraints, a nonhomogeneous linear system of 244 equations is formed as $\mathbf{Au}=\mathbf{b}$. The solution for 36 unknowns of $\mathbf{C}$ and 76 unknowns of $\mathbf{X}$ that optimally satisfy the equations is found. $\mathbf{A}$ is the 244×112 coefficient matrix, $\mathbf{u}$ is a column vector of the 112 unknowns, and $\mathbf{b}$ is a column vector of 214 zeros and 30 non-zero values. Among 30 non-zero values in $\mathbf{b}$, 24 values are sums of intensities of each chromosome across spectra and 6 values are 1s, representing sums of each column of the spread matrix. The optimal solution that gives the minimum least squares error of this over-determined problem is computed by $\mathbf{u} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$. In practice, $\mathbf{A}$ will not be error free. The amount of perturbation in $\mathbf{A}$ is directly related to the level of noise in $\mathbf{Y}$. Thus, pixels should be carefully selected, and especially saturated pixels should be avoided since saturation is a nonlinear phenomenon. In actual computation, QR decomposition of $\mathbf{A}$ is used to find the solution and avoid the calculation of $\mathbf{A}^T\mathbf{A}$, since $\mathbf{A}^T\mathbf{A}$ is strongly influenced by round off error [56].

QR decomposition is a matrix factorization method that factorizes $\mathbf{A}$ as $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal matrix ($\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is an upper triangular matrix. The solution is then found by backsubstitution from $\mathbf{Ru} = \mathbf{Q}^T\mathbf{b}$.

Once the color spread matrix $\mathbf{C}$ is found, it can be applied to other images to correct color spreading. An image I is a set of $\{\mathbf{y}_j | \mathbf{y}_j \in I\}$, where $j \in (1, \ldots, N)$ and N is the number of pixels in an image. An image without the channel crosstalk is computed by $\mathbf{x}_j = \mathbf{C}^{-1}\mathbf{E}^{-1}\mathbf{y}_j$ for all j. To account for the fluorophore sensitivities, exposure times $\mathbf{E}$ can be multiplied to $\mathbf{x}_j$.

The color spread matrix can only be estimated when the number of specimen is equal to or larger than the number of fluorophores.

3.5 Results of Color Compensation of M-FISH Images

In this section, an example of calculating the color spread matrix is shown, and results of color compensated M-FISH images and are shown. In addition we quantitatively show the improvement in image quality after the color compensation using mean squared error (MSE) and the structural similarity index (SSIM), a recently developed metric that has been shown to significantly surpass the MSE as a means for quantifying structural similarities between two images [55]. Given two signals $\mathbf{x}$ and $\mathbf{y}$, SSIM includes the luminance, contrast, and structure terms, and it is expressed

$$SSIM(x, y) = \frac{(2\mu_x\mu_y)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.10}$$

	Spectra		
	1	2	3
1	x	x	0
2	x	0	x
3	x	x	x

Objects

	Spectra		
	1	2	3
1	0.8	0.05	0.1
2	0.15	0.8	0.1
3	0.05	0.15	0.8

1	4	0	0
2	0	1	0
3	0	0	2

Table 3.4: Example. Color labeling table (L), color spread matrix (M), and exposure times (R)

where $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, and L is the dynamic range of the pixel values (255 for 8-bit images). For the constants, we used $K_1 = 0.01$ and $K_2 = 0.03$ (refer to [55] for the details). SSIM is calculated within a circular moving window, which moves pixel-by-pixel over the entire image. It is easily shown [55] that $0 \leq SSIM(\mathbf{x}, \mathbf{y}) \leq 1$, where $SSIM(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$.

3.5.1 Example of Computing the Color Spread Matrix

Formulating a linear system of equations from the observed signal $\mathbf{Y}$ is illustrated in this example. Suppose three objects are stained with three fluorophores combinatorially according to the color map shown in Table 3.4, and the color spread matrix of the imaging system combined with those three fluorophores is also defined in Table 3.4. The color map shows that object one is stained with fluorophore 1 and 2, and object three is stained with all three fluorophores. The color spread matrix indicates that, for each fluorophore, twenty percent of the original signal intensity is spread to the other channels.

Let's define the original signal $\mathbf{X}$ as

$$\mathbf{X}^T = \begin{bmatrix} 50 & 200 & 0 \\ 50 & 0 & 100 \\ 50 & 200 & 100 \end{bmatrix}.$$

Rows in $\mathbf{X}^T$ represent objects and columns represent spectra. The observed signal $\mathbf{Y}$ is defined as $\mathbf{Y} = \mathbf{ECX}$. Remember that the matrix $\mathbf{Y}$ is a set of means of each object, i.e.

$$\mathbf{Y}^T = \begin{bmatrix} \mu_{\mathbf{y}_1} = \frac{1}{P_1} \sum_{k=1}^{P_1} \mathbf{y}_{1k} \\ \mu_{\mathbf{y}_2} = \frac{1}{P_2} \sum_{k=1}^{P_2} \mathbf{y}_{2k} \\ \mu_{\mathbf{y}_3} = \frac{1}{P_3} \sum_{k=1}^{P_3} \mathbf{y}_{3k} \end{bmatrix}$$

where $P_{i \in 1,2,3}$ are the number of pixels that belong to each object and $\mathbf{y}_{ik} = [y_{ik}(1), y_{ik}(2), y_{ik}(3)]$. A pixel value from object 1 is $\mathbf{y}_1 = \mathbf{ECx}_1$. Then the matrix of means of the observed signal is

$$\mathbf{Y}^T = \begin{bmatrix} 200 & 167.5 & 65 \\ 200 & 17.5 & 165 \\ 240 & 177.5 & 225 \end{bmatrix}.$$

Now, given $\mathbf{Y}$, $\mathbf{E}$, and the color table, we will estimate the color spread matrix $\mathbf{C}$ and the original signal $\mathbf{X}$. We have nine unknowns for the color spread matrix and seven unknowns for the true signal. The solution for the total of sixteen unknowns can be found by solving the following equation $\mathbf{C}^{-1}\mathbf{E}^{-1}\mathbf{Y} - \mathbf{X} = 0$ with conditions defined in 3.4.3. The equation can be written

$$\begin{bmatrix} u_1 & u_2 & u_3 \\ u_4 & u_5 & u_6 \\ u_7 & u_8 & u_9 \end{bmatrix} \begin{bmatrix} 50 & 50 & 60 \\ 167.5 & 17.5 & 177.5 \\ 32.5 & 82.5 & 112.5 \end{bmatrix} - \begin{bmatrix} u_{10} & u_{12} & u_{14} \\ u_{11} & 0 & u_{15} \\ 0 & u_{13} & u_{16} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (3.11)$$

Eq. (3.11) can be written as a linear system of m equations in 16 unknowns, $\mathbf{Au} = \mathbf{b}$, where $\mathbf{A}$ is the coefficient matrix, $\mathbf{u}$ is the column vector

of the unknowns, and $\mathbf{b}$ is an $m \times 1$ column vector. From eq. (3.11), nine equations are formed. The sums of columns of $\mathbf{E}^{-1}\mathbf{Y}$ should be the same as the sums of columns of $\mathbf{X}$. This gives three equations. The sum of each column of $\mathbf{C}^{-1}$ should be 1. This gives three more equations. Further, values in a row of $\mathbf{X}$ should be the same, yielding four more equations. Thus, a total of 19 equations are formed. A linear system of m equations in n unknowns has a unique solution if the coefficient matrix $\mathbf{A}$ and the augmented matrix $\tilde{\mathbf{A}}$ has the same rank, and the rank equals n. In this example, rank($\mathbf{A}$) = rank($\tilde{\mathbf{A}}$) = 16. The solution is found by the QR decomposition. $\mathbf{u}(1 \cdots 9)$ contains the solution for $\mathbf{C}^{-1}$. Then the estimated color spread matrix is

$$\hat{\mathbf{C}} = \begin{bmatrix} 0.8 & 0.05 & 0.1 \\ 0.15 & 0.8 & 0.1 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}$$

$\mathbf{u}(10 \cdots 16)$ contains the solution for the unknown $\mathbf{X}$ values. The true signal estimated is

$$\hat{\mathbf{X}}^T = \begin{bmatrix} 50 & 200 & 0 \\ 50 & 0 & 100 \\ 50 & 200 & 100 \end{bmatrix}$$

$\hat{\mathbf{C}} = \mathbf{C}$ and $\hat{\mathbf{X}} = \mathbf{X}$. Thus, the MSE between the estimation and the truth is zero. The proposed method finds the unknowns with no error in this example.

3.5.2 Quantification of Image Quality Improvement After Color Compensation

Six-channel synthetic images, representing the ideal color compensation result, are generated using the ground truth from the database in order to quantify the image quality improvement. Non-overlapping pixels are as-

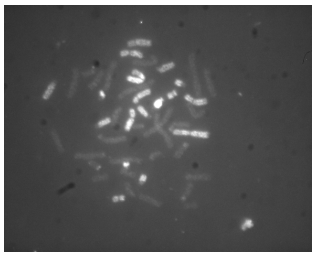
Images	MSE_B	MSE_A	$MSSIM_B$	$MSSIM_A$
v1301xy	2196	534	0.074	0.693
v1303xy	922	184	0.074	0.765
v1309xy	1339	482	0.081	0.779
v1310xy	847	157	0.072	0.800
v1311xy	700	145	0.072	0.784
v1313xy	767	190	0.079	0.800
Average	1128	282	0.075	0.770

Table 3.5: Image quality improvement. Subindex B represents before color compensation, and subindex A represents after color compensation

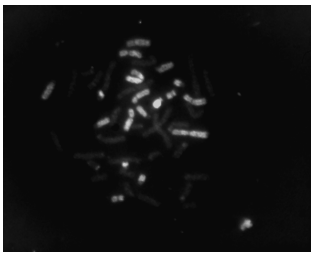
signed 128 and overlapping pixels are assigned 255 on corresponding channels of synthetic images. The MSE and the mean SSIM (MSSIM) are measured from before and after color compensated images against the synthetic images. An average of MSEs and MSSIMs across six channels per image is shown in Table 3.5. MSSIM becomes one when two images are identical. As shown in Table 3.5, the MSE reduced by a factor of 4 and the MSSIM increased by a factor of 10 after the color compensation.

3.5.3 Background Correction and Color Compensation Results

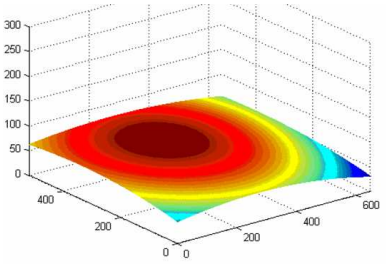
Figure 3.5 shows the result of background correction. The original image in Fig. 3.5 has an elevated background and displays channel crosstalk. The estimated cubic surface of the background is shown in Fig. 3.5(c). The background corrected image shown in Fig. 3.5(b) is obtained after subtracting Fig. 3.5(c) from Fig. 3.5(a). Fig. 3.5(d) is a profile drawn from rows in the middle of Fig. 3.5(a) and Fig. 3.5(b), and it clearly shows that the background elevation is effectively removed.



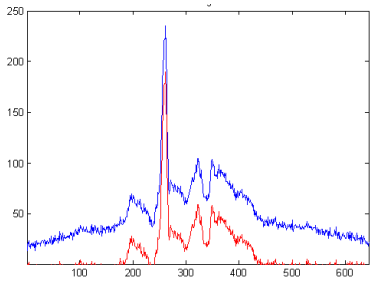
(a) Original aqua channel



(b) Background corrected image

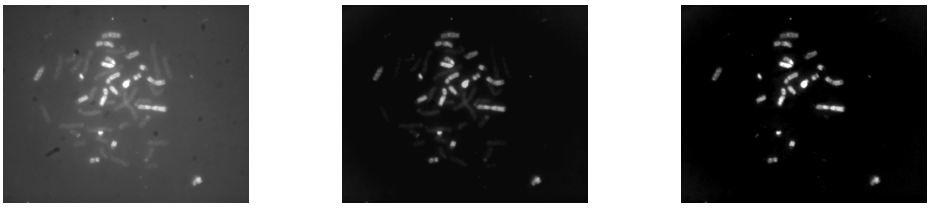


(c) Estimated cubic surface



(d) A profile: Top-before, Bottom-after

Figure 3.5: Background correction. Elevated background intensity is removed after the background correction, but the channel crosstalk still remained.



(a) Original image (b) Background correction (c) Color compensation

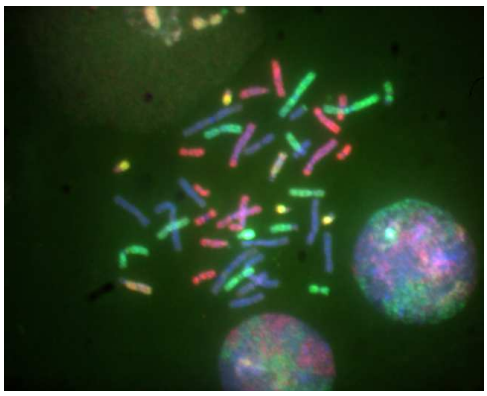
Figure 3.6: Color compensation. The color compensation removed the channel crosstalk effectively. A significant increase in image quality is achieved on image (c).

After correcting the background of five images that are captured from the same slide, pixels from each chromosome class are collected from those images. The means of each class are computed to form the matrix $\mathbf{Y}$. Then the color spread (compensation) matrix is calculated. Fig. 3.6(c) shows the color compensation result, and as shown in the figure, all the crosstalk was effectively removed. A significant improvement in image quality is achieved after the color compensation.

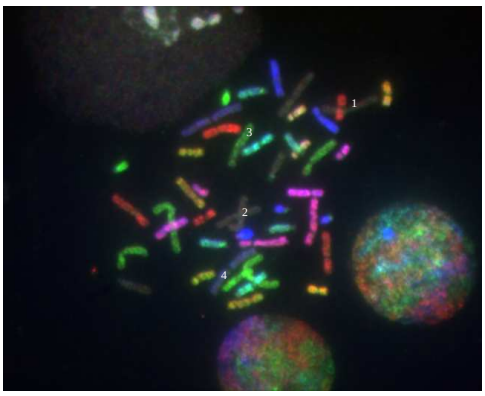
Color compensation is an effective method of improving the quality of M-FISH images by removing the channel crosstalk. Figure 3.7 shows an example of before and after the color compensation: (a) Green, Aqua, and DAPI channels of V1301XY are combined as a color image, (b) Far red, Red, and Gold channels are combined and shown as a color image, (c) and (d) show the result of background correction, (e) and (f) show the color compensation result (Simple scaling has been applied to the color compensated image). As shown in (e) and (f) the quality of the image has been improved significantly by

removing channel crosstalk. Pixel values numbered on (b) are shown in Table 3.6. It is not easy to distinguish which chromosomes are truly hybridized and which are due to crosstalk in Fig. 3.7 (b). Chromosomes marked with number 1 and 2 are due to crosstalk and they are effectively removed in Fig. 3.7 (f).

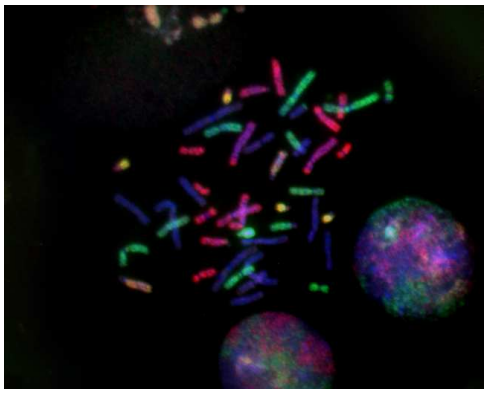
Table 3.6 shows pixel values after the background correction and color compensation. NP, BC, and CC means no processing, background correction, and color compensation respectively. As the values show, the intensity corresponding to the channel crosstalk has been removed effectively. The background correction helps reveal the pattern and color compensation further enhances the pattern as shown in Fig. 3.7 and in Table 3.6. Accordingly, pixel classification accuracy also increased significantly after the background correction (results are shown in Section 4.5). However our experiments on a small number of images showed that color compensating images after the background correction did not improve the overall classification accuracy. This suggests that revealing the pattern helps classification but enhancing the pattern is not enough. The pattern must satisfy certain criteria, which is explained in the following section. One more drawback of color compensation with respect to pixel classification is that the color spread matrix should be recalculated for images with different intensity (feature) distributions. To summarize, the color compensation improves the image quality significantly but may not be a useful preprocessing step for pixel classification.



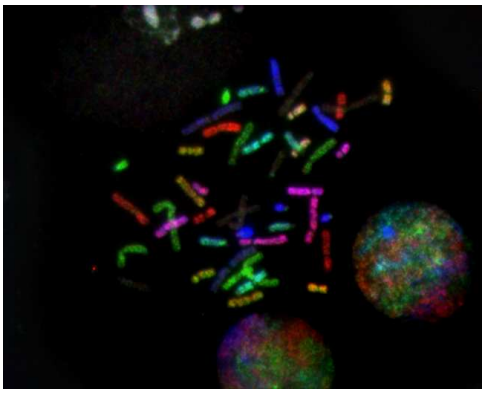
(a) Before color compensation



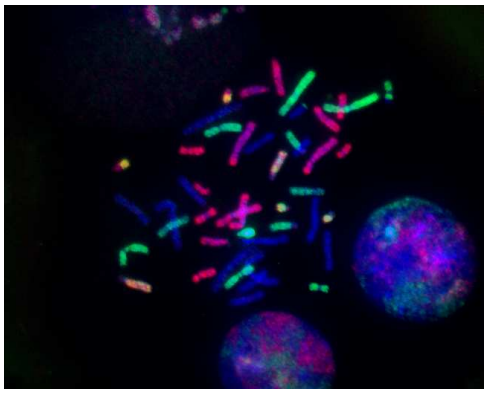
(b) Before color compensation



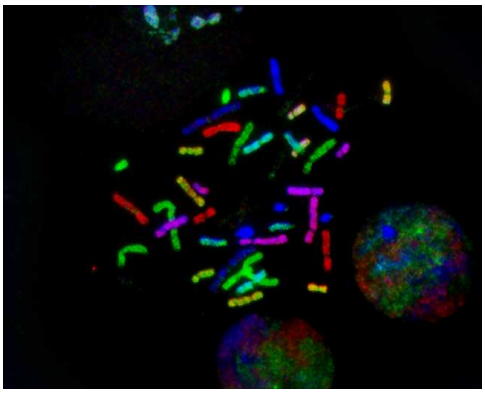
(c) Background correction of (a)



(d) Background correction of (b)



(e) Color compensation of (a)



(f) Color compensation of (b)

Figure 3.7: Color compensation result on image V1301XY.

Pixel	Spectrum						
	Processing	Aqua	Green	Gold	Red	Far Red	Class
1	NP	191	57	38	44	55	3
	BC	133	26	11	17	34	
	CC	196	1	0	0	11	
	Pattern	1	0	0	0	0	
2	NP	137	219	75	84	85	6
	BC	61	173	32	43	50	
	CC	18	208	1	27	16	
	Pattern	0	1	0	0	0	
3	NP	95	154	60	128	53	4
	BC	30	118	27	97	28	
	CC	0	151	0	163	0	
	Pattern	0	1	0	1	0	
4	NP	104	77	120	80	71	1
	BC	29	29	76	37	34	
	CC	0	0	128	0	1	
	Pattern	0	0	1	0	0	

Table 3.6: Pixel values numbered on Fig. 3.7 (b).

3.6 Expectation Maximization Normalization

Even after background correction and color compensation, intensity variations within a chromosome and among chromosomes in a channel and between channels, caused by uneven hybridization in a chromosome and unequal fluorophore sensitivities depending on chromosomes, remain as a source of classification error. Within a channel, chromosomes that are supposed to be bright are expected to have a similar intensity level, but often chromosome intensities differ considerably. The bright chromosomes in one channel are not consistently brighter than other chromosome in other channels where they are supposed to appear. For example, a chromosome labeled with three fluorophores show significantly different intensity levels across those three channels due to unequal fluorophore sensitivities and unequal exposure times. These intensity differences across channels are also inconsistent across images. This inconsistency causes classification errors since the feature vector $\mathbf{y}$ becomes inconsistent. Given an individual feature value, e.g. gray level of 60, it is uncertain whether it comes from a hybridized chromosome or from noise. Only when a feature vector is formed does the relative intensity difference among feature values deliver meaningful information about the pixel membership. The relative intensity difference among feature values in a feature vector is called texture, which is independent of the mean value of the vector. Note that the texture here is defined as the shape of dark and bright pattern in a feature vector. Two feature vectors $\mathbf{y}_1 = [1, 0, 1, 0, 1]$ and $\mathbf{y}_2 = [100, 60, 100, 60, 100]$ have the identical texture, while a third vector $\mathbf{y}_3 = [68, 5, 240, 10,$

210] has a similar pattern as $\mathbf{y}_1$ and $\mathbf{y}_2$ but has a different texture, if we define the texture as $(\mathbf{y}_j - \mu_{\mathbf{y}_j})/\sigma_{\mathbf{y}_j}$.

Suppose that $\mathbf{y}_3 \in \omega_{10}$ (chromosome 10), and the pattern of $\mathbf{y}_3$ is consistent throughout all $\mathbf{y} \in \omega_{10}$, then a supervised classification method should work well without further normalizing the data. Even though background correction significantly reduces the variations in $\mathbf{y}_i \in \omega_i$, for all i, there are pixels misclassified due to the aforementioned variations. Therefore, hybridized chromosomes must have a high intensity level across all spectral channels, and at the same time, unhybridized sections, including intensity due to spectral crosstalk, should have a certain intensity level that is lower than the intensity of hybridized chromosomes across all spectral channels. The normalization process should minimize the difference of the intensity distributions for all images. This can be achieved by normalizing the variables (the features).

An M-FISH image M is composed of six gray scale images $\{I_1, I_2, I_3, I_4, I_5, I_6\}$, each corresponding to a spectral channel. Each gray scale image I_k contains gray scale values y that belong to background $I_b(k)$ and chromosomes $I_c(k)$, i.e. $I_k = \{y | y \in I_b(k) \cup y \in I_c(k)\}$, and $I_c(k) = \{y | y \in \omega_1 \cup y \in \omega_2\}$, where $\omega_1 =$ intensity due to no fluorophore and $\omega_2 =$ intensity due to a fluorophore. The distribution of y in $I_c(k)$ is assumed to be a mixture of two gaussians: $p(y|\omega_1) \sim N(\mu_1, \sigma_1)$ and $p(y|\omega_2) \sim N(\mu_2, \sigma_2)$, and $\mu_1 < \mu_2$. Then $I_c(k)$ is a set of unlabeled samples drawn independently from the mixture density

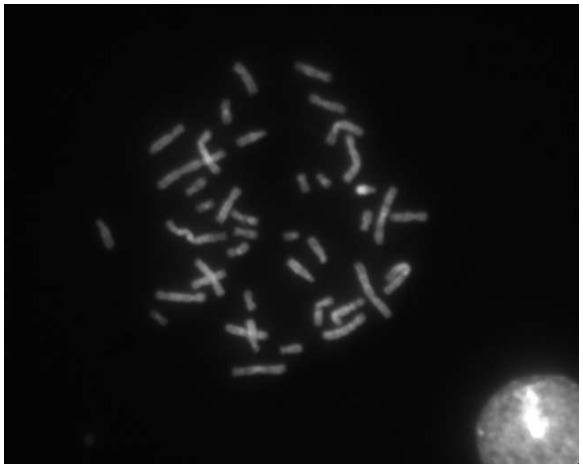
$$p(y) = p(y|\omega_1)P(\omega_1) + p(y|\omega_2)P(\omega_2). \tag{3.12}$$

Since the models are identical for all channels, the channel index k is not specified for y. A parameter vector $\boldsymbol{\theta}$ contains $(\mu_1, \mu_2, \sigma_1, \sigma_2, P(\omega_1), P(\omega_2))$. $P(\omega_1)$ and $P(\omega_2)$ are prior probabilities and also called mixing parameters.

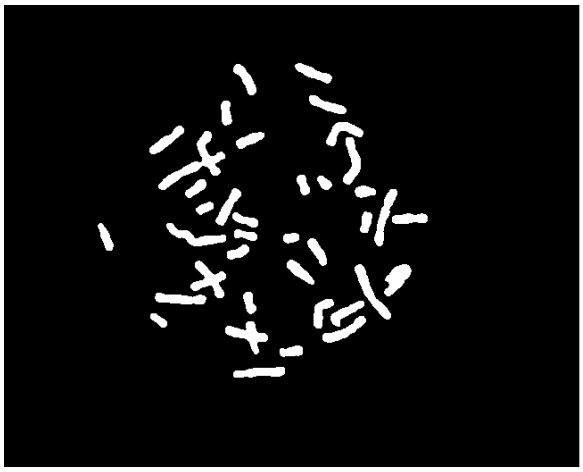
The separation between $I_c(k)$ and $I_b(k)$ can be obtained by a new automatic segmentation method [12], which combines global and local intensity, spectral information, and edge information to segment chromosomes from the background. Cells are also removed, based on their size and circularity (see Fig. 3.8).

Following segmentation, only pixels that fall inside chromosomes are classified. Among the 6 features, the DAPI channel provides information regarding whether a pixel belongs to chromosomes or to background. Since chromosome-background classification (segmentation) is already accomplished, DAPI information becomes redundant when classifying only chromosome pixels. Thus the remaining five features are normalized and used for classification.

Figure. 3.9 shows an example of the mixture density distributions of $I_c(k)$ of an M-FISH image, V1401XX. The black bars in Fig. 3.9 represent the range of gray scale values for a pixel that has a [High Low High Low High] labeling pattern. As one can notice, a significant portion of High values in $I_c(3)$ overlaps with Low values in $I_c(2)$ and $I_c(4)$, resulting in a totally unexpected pattern. This unexpected pattern will result in a low classification accuracy simply because the distributions of this image and the training data (or expected patterns) are different. We want to emphasize that this low classification accuracy comes from the difference in the patterns between



(a) V130740XY DAPI Channel



(b) Segmentation result

Figure 3.8: Segmentation result. Chromosomes are automatically segmented from background by utilizing 6 spectral information, global and local intensity, and edge information. Cells are also removed based on the size and circularity.

the training and the testing data, and the accuracy is less dependent on the fundamental error rate (Bayes error) of the testing data. In other words, the joint distribution of five features of the testing data may have extremely small overlaps (low errors) among classes, but has its own patterns that are different from the training data, which will result in a low classification accuracy.

Figure. 3.10 further illustrates this point. Suppose that each feature has a bimodal distribution, and there are two features describing four classes. The straight lines in the figure are the decision boundaries for the four classes. As one can see, the fundamental error rate of each data set is determined by the distribution of its marginal density functions. As the overlap between the two modes in each feature increases, the error rate of the data increases. Both of the data in Fig. 3.10 seem to have small error rates. However the testing data's classification accuracy will be low because the distribution is considerably different from the distribution of the training data set. The two distributions should be made as similar as possible by the normalization process in order to minimize the classification error.

Given a bimodal marginal density function (eq. (3.12)) and its parameters, the normalization process should cause $y \in \omega_1$ and $y \in \omega_2$ to fall within certain ranges and the decision boundary between ω_1 and ω_2 to lie at a certain point. The parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, P(\omega_1), P(\omega_2))$ are unknown, and the samples are unlabeled. $\boldsymbol{\theta}$ can be found by the maximum-likelihood estimation procedure.

When all parameters are unknown, and if no constraints are placed on

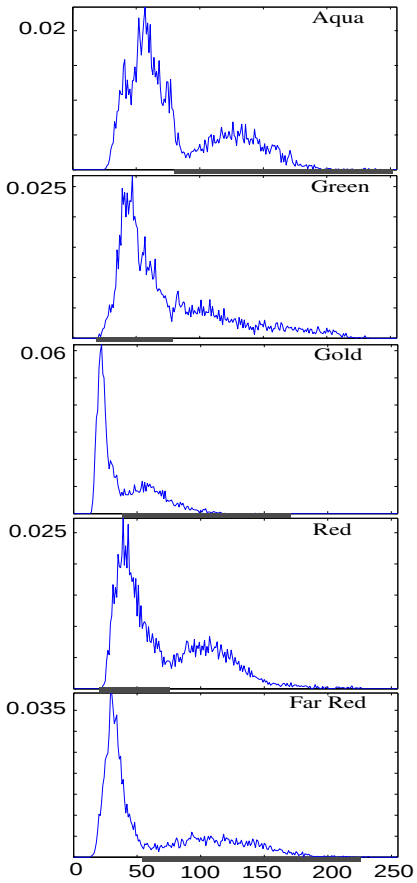
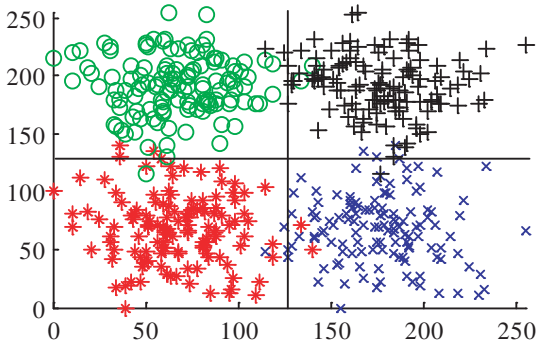
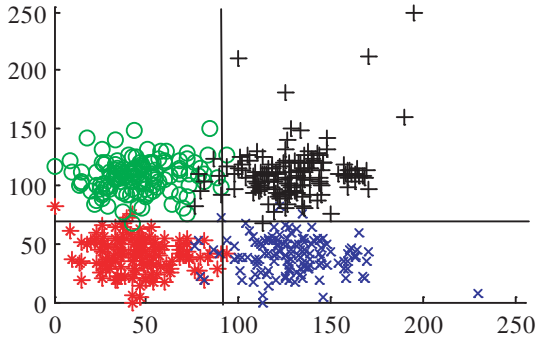


Figure 3.9: The mixture density distribution of $I_c(k)$ of V1401XX.



(a) Training data



(b) Testing data

Figure 3.10: Distributions of training data and testing data. x and y axes are the feature values (thus, a feature vector forms a point in the figure). Each data set has its own fundamental error rate by its own distribution, but the classification accuracy for the testing data will be low because the distributions are different between the two data sets. The distributions should be normalized in order to obtain a high classification accuracy.

the covariance matrix (for multi-dimensional data), the maximum-likelihood principle yields useless singular solutions. However, meaningful solutions can still be obtained if we restrict our attention to the largest of the finite local maxima of the likelihood function, assuming that the likelihood function is well-behaved at such maxima [57]. Then the parameter vectors $\boldsymbol{\theta_i} = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, P(\omega_i))$ can be estimated iteratively using the following equations:

$$\hat{P}(\omega_i) = \frac{1}{N} \sum_{j=1}^{N} \hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}}) \tag{3.13}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{j=1}^{N} \hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}}) \mathbf{y}_j}{\sum_{j=1}^{N} \hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}})} \tag{3.14}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{j=1}^{N} \hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}})(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_i)^T}{\sum_{j=1}^{N} \hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}})} \tag{3.15}$$

where N = number of unlabeled samples drawn independently from the mixture density of c classes, $i = 1, \ldots, c$, and

$$\begin{aligned} \hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}}) &= \frac{p(\mathbf{y}_j | \omega_i, \hat{\boldsymbol{\theta}}_i) \hat{P}(\omega_i)}{\sum_{l=1}^{c} p(\mathbf{y}_j | \omega_l, \hat{\boldsymbol{\theta}}_l) \hat{P}(\omega_l)} \\ &= \frac{|\hat{\boldsymbol{\Sigma}}_i|^{-1/2} exp\left[-\frac{1}{2}(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_j - \hat{\boldsymbol{\mu}}_i)\right] \hat{P}(\omega_i)}{\sum_{l=1}^{c} |\hat{\boldsymbol{\Sigma}}_l|^{-1/2} exp\left[-\frac{1}{2}(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_l)^T \hat{\boldsymbol{\Sigma}}_l^{-1} (\mathbf{y}_j - \hat{\boldsymbol{\mu}}_l)\right] \hat{P}(\omega_l)} \end{aligned} \tag{3.16}$$

Among the various techniques that can be used to obtain a solution, one approach is to use an initial estimate to evaluate (3.16) for $\hat{P}(\omega_i | \mathbf{y}_j, \hat{\boldsymbol{\theta}})$, then use (3.13) - (3.15) to update the estimate [57]. This iterative method is also called expectation-maximization (EM). Since the solution depends on the initial estimates and to obtain fast convergence, a k-means clustering method

is used to estimate the initial parameters. k-means clustering is a simple but popular method of finding the c mean vectors $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_c}$. Given the c initial mean vectors $\boldsymbol{\mu_m}$, the samples are classified to the nearest $\boldsymbol{\mu_m}$. Then by approximating $\hat{P}(\omega_i|\mathbf{y}_j, \hat{\boldsymbol{\theta}})$ in (3.14) as

$$\hat{P}(\omega_i|\mathbf{y}_j, \hat{\boldsymbol{\theta}}) \cong \begin{cases} 1 & \text{if } i = m \\ 0 & \text{otherwise} \end{cases}$$

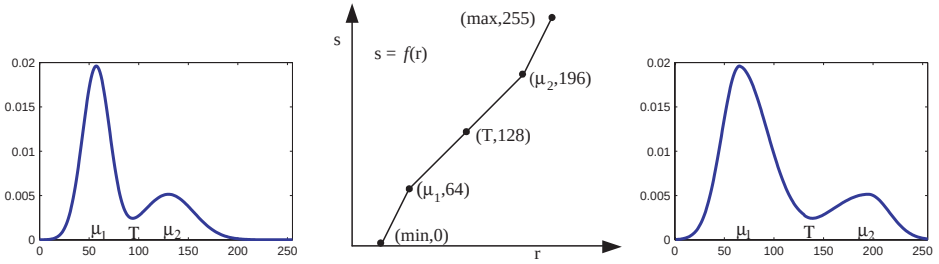
new estimates of the c mean vectors are obtained. The iteration repeats until the means converge. Usually c randomly chosen samples are used as the initial c means. In our case, the minimum and the maximum gray scale values in each channel are used as the initial mean values. Once $\hat{\mu}_1$ and $\hat{\mu}_2$ are found via the k-means clustering, the values $\hat{\sigma}_i^2$ are estimated from the samples classified to ω_1 and ω_2. These means and variances along with equal priors are used as initial estimates for (3.13) - (3.15). Once the parameters are estimated by the EM method, the decision boundary between ω_1 and ω_2 is found by

$$T = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{3.17}$$

where

$$\begin{aligned} A &= \hat{\sigma}_2^2 - \hat{\sigma}_1^2 \\ B &= 2\hat{\sigma}_1^2\hat{\mu_2} - 2\hat{\sigma}_2^2\hat{\mu_1} \\ C &= \hat{\sigma}^2\hat{\mu}_1^2 - \hat{\sigma}_1^2\hat{\mu}_2^2 - 2\hat{\sigma}_2^2 ln(\frac{\hat{\sigma}_2\hat{P}(\omega_1)}{\hat{\sigma}_1\hat{P}(\omega_2)}). \end{aligned}$$

Given the parameter vectors and the decision boundary, the sample distribution is normalized by piece-wise linear transformations as shown in Fig. 3.11.



(a) Before normalization (b) Normalization (c) After normalization

Figure 3.11: A marginal density function in (a) is normalized as in (c) by the piece-wise linear gray level mapping function in (b). The horizontal axes represent gray scale range.

The input intensity r is mapped to the output intensity s by

$$f(r) = \begin{cases} \frac{64}{\mu_1 - min(r)}(r - min(r)) & \text{if } min(r) \leq r < \mu_1 \\ \frac{64}{T - \mu_1}(r - \mu_1) + 64 & \text{if } \mu_1 \leq r < T \\ \frac{64}{\mu_2 - T}(r - T) + 128 & \text{if } T \leq r < \mu_2 \\ \frac{63}{max(r) - \mu_2}(r - \mu_2) + 192 & \text{if } \mu_2 \leq r < max(r) \end{cases} \tag{3.18}$$

where $min(r)$ is the minimum intensity level and $max(r)$ is the maximum intensity level in r.

3.7 Results of EM Normalization

Each data set has its own unique error rate (Bayes error) based on the feature distribution. While the fundamental error rate of each data set is one problem that causes classification error, the significant error comes from having different distributions for different data sets. The EM normalization process is focused on reducing the distribution differences among the different data sets. Thus, the classification accuracy improves significantly after normalization (rates are shown in the following section).

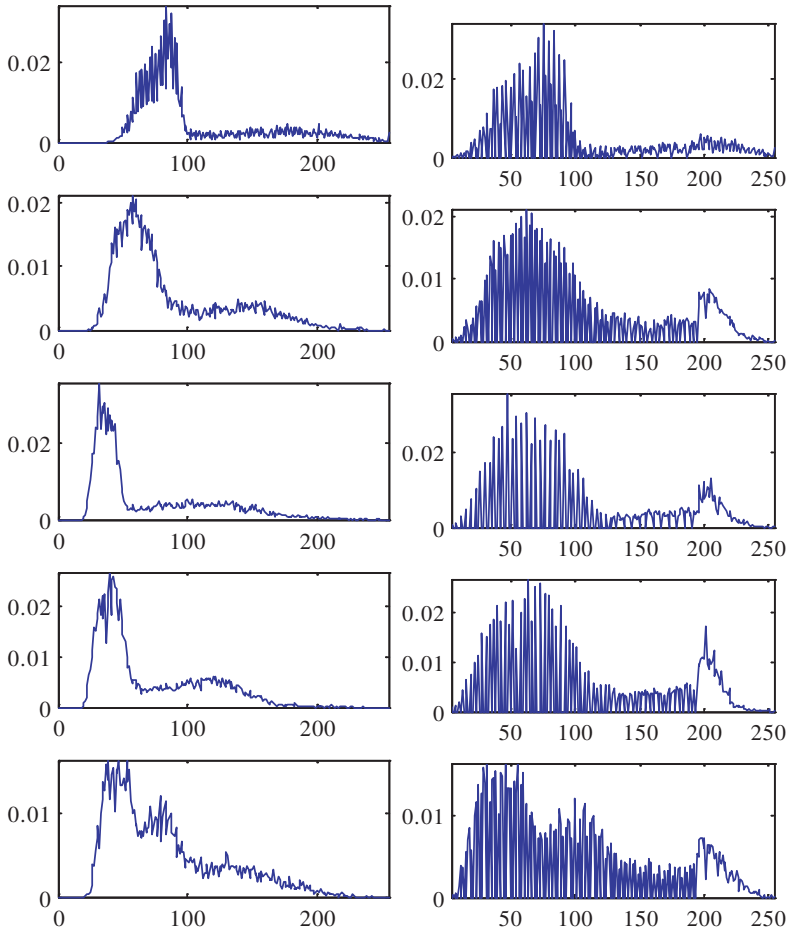
The mixture density parameters for each feature were found by (3.13) - (3.15), and then the decision boundary between the modes was found by (3.17). Given the parameters and the decision boundary T, the features were normalized by (3.18).

In particular, Fig. 3.12 shows the intensity distributions of the features of V290562 before and after EM normalization. As the figure shows, the uncertainty of a gray scale at a channel being High (hybridized) or Low (not hybridized) is removed in the normalized data.

Figure 3.13 shows the gray scale images before and after the EM normalization. As the figure shows, chromosomes that are hybridized have higher intensity levels than the intensities due to non-hybridized chromosomes. This normalization ensures that the patterns become consistent throughout all images.

3.8 Conclusion

We have shown a method of estimating the color spread matrix for M-FISH images. Examples of formulating a linear system of equations in order to estimate the color spread matrix were presented. The color compensated M-FISH images provide superior image quality compared to the unprocessed images. The image quality improvement is also quantitatively shown using SSIM and MSE. MSSIM improved by a factor of 10, and MSE reduced by a factor of four on average after the color compensation. The new technique can be applied easily to any fluorescence microscopy images where specimens are



(a) Before normalization (b) After normalization

Figure 3.12: Feature distribution (normalized histogram) of V1290562 before and after the EM normalization. x axis represents gray scale and y axis represents the normalized frequency of a gray level. The EM normalized images are shown in Fig. 3.13, and the classification result is shown in Fig 4.6.

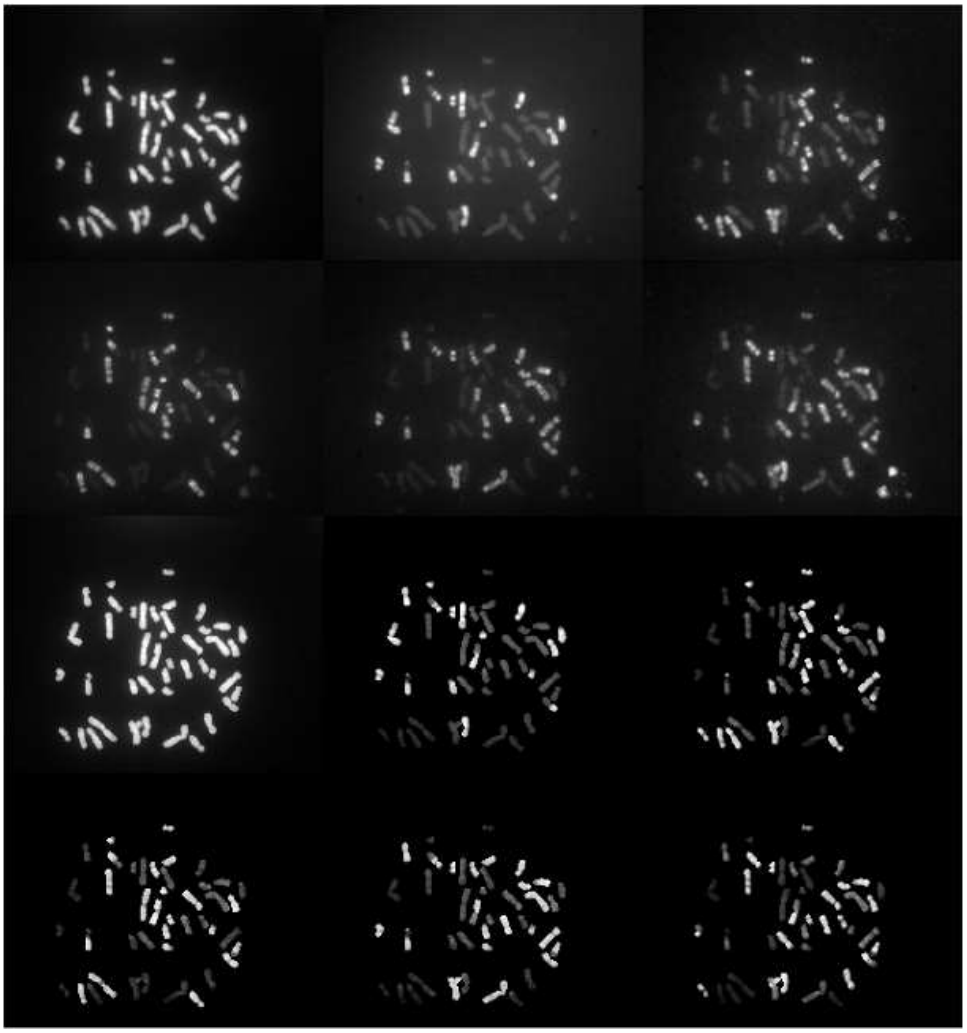


Figure 3.13: EM normalization result of V1290562 before and after the EM normalization.

uniquely or combinatorially stained, and it can be extended to any multichannel images that exhibit a signal formation pattern that is similar to M-FISH images.

We also introduced a new normalization method using the expectation maximization algorithm. Previously the variation in the feature distributions among the different M-FISH images was not emphasized as a source of misclassification. Even if it was recognized, there was no good method of reducing the variation. Assuming the distribution of each feature in the chromosome region is a mixture of two normal density functions, the maximum-likelihood parameters were estimated for the mixture density and each feature was normalized based on the parameters. In the following chapter we show that the overall pixel classification accuracy improved by 40% after EM normalization from 50% (with no proprocessing) to 70% (with EM normalization), and also show that the improvement is statistically significant with no preprocessing and with background correction.

Chapter 4

Pixel Classification Methods For M-FISH Images

In this chapter, various pixel classification methods for M-FISH images are described, which include supervised parametric, supervised nonparametric, and unsupervised nonparametric methods including two new classification methods for M-FISH images that do not require training of a classifier (unsupervised) nor require class parameter estimation (nonparametric).

Given a number of objects, the choice of classifier depends on the knowledge about the samples in the feature domain such as the number of classes, the prior probabilities, the forms for the class-conditional probability density functions, the values for the density functions, and the category labels of the samples.

When the labels are available, we can learn the statistical properties of the samples and design a classifier that utilizes that knowledge. The **maximum-likelihood classifier** is one kind, which estimates the class parameters from the training data and an unknown sample is classified to a class that yields the maximum likelihood of the sample belonging to the class. When the number of samples representing classes is large, the estimation of

the parameters will become close to the true parameters. On the contrary, when the number of training samples is small (e.g. the face recognition problem where only a few images are available for each class), the estimation of the class parameters will be inaccurate or even impossible. For such cases, the **nearest neighbor classifier** or k**-nearest neighbor classifier** is a suitable choice, which assigns samples to the class of the nearest training sample.

If the labels are not available, the class parameters can be estimated using an unsupervised method. The samples can be grouped into a number of classes without estimating the parameters, and this is called an unsupervised nonparametric method. These include k**-means clustering** and **fuzzy** k**-means clustering**. These clustering methods cluster the data into a fixed number of groups regardless of the true number of classes in the data. If the clustering method is the only option, when the number of classes changes depending on a set of data or depending on time, then the right number of classes should be validated. If the labels are not available but the patterns for each class are expected to have ideal prototypes, then the **template matching method** (similar to the nearest neighbor method) or the **fuzzy-logic classifier** can be used.

M-FISH images have six channels. Each channel contains the intensity of a corresponding fluorophore. Since each chromosome is uniquely stained, the intensity combinations across 6 channels are unique for each chromosome type. Previously, we have designed a 6-channel 25-class maximum likelihood classifier [48, 49]. 25 classes include 24 chromosomes plus background. By clas-

sifying every pixel in the image using this maximum-likelihood classifier, both segmentation and classification of chromosomes were achieved simultaneously. The overall accuracy of the segmentation was about 90% on a small number of images using this method. When a portion of the chromosome pixels are classified as background or vice versa, the lost region cannot be recovered without prior knowledge about the chromsome boundaries. Furthermore, classifying every pixel in the images is wasteful since chromosome pixels only occupy less than 10% of the image. Thus, prior to the pixel classification, an accurate segmentation method is desired.

4.1 Foreground-background segmentation

In order to compute reliable boundaries between objects and background, we combined multiple methods that utilize not only spectral information but also edge information. Laplacian of Gaussian (LoG) edge detection performed on the DAPI channel provides nice closed boundaries of chromosomes that correspond well to human perception. However, it also picks up unwanted artifacts from the background. In general, chromosome intensities are brighter than the neighboring background, although the background surface is not globally uniform. When object intensity is brighter than the neighboring pixels, adaptive thresholding is an effective segmentation method. This method effectively separates chromosomes from background. Due to its simplicity and effectiveness, adaptive thresholding is widely used for chromosome image segmentation. However, when a number of pixels in the foreground are

darker than neighboring foreground pixels, adaptive thresholding creates holes inside the chromosome or disconnects the chromosome into pieces. To utilize the spectral information, 6-feature 2-class K-means clustering method is used. This clustering method is preferable to the maximum-likelihood method because it does not require training. It groups six dimensional data into two classes while iteratively regrouping the data points until the class means converge. Its classification results are similar to those of the maximum-likelihood classifier since they both utilize the same information. Adaptive thresholding, LoG edge detection, K-means clustering, and global thresholding methods are combined together to achieve a final segmentation result. A composite binary image is obtained after voting among those 4 methods. For example, a pixel becomes foreground when a majority (3 out of 4) are foreground.

Prior to the segmentation, a non-uniform background was corrected by fitting a cubic surface to the estimated background pixels and subtracting it from each channel [51]. The background pixels for each channel were estimated by a global thresholding method, an iterative clustering method, in which the threshold was found while iteratively grouping pixels into two classes until the class means converge. The decision boundary between the two classes was the threshold (eq. 3.5 on page 37). Pixels below threshold were used for the surface estimation. After the background correction, cells are identified based on the circularity and size measures (the detailed procedure of cell removal is given in the following section). Once the background was corrected and cells are removed from the image, adaptive thresholding, LoG edge detection, and

(a) DAPI channel (b) Ground truth (c) Segmentation Result

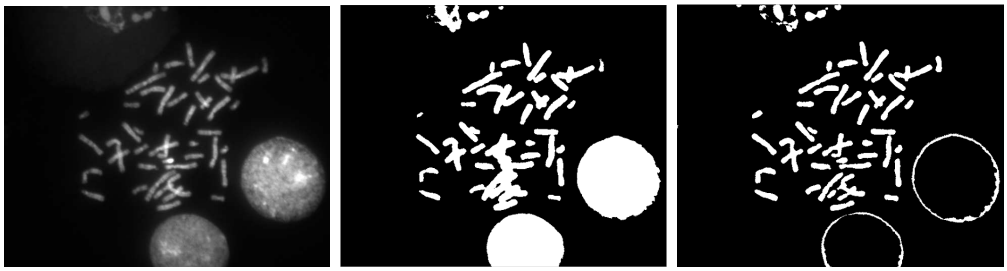
Figure 4.1: Segmentation result. Notice that the cell is effectively removed.

6-feature 2-class K-means clustering were performed. A composite threshold image was created after voting. An example is shown in Fig. 4.1. Fig. 4.1 (b) the ground truth was generated by thresholding the DAPI channel and by manually correcting mistakes. During the manual correction, some chromosomes were mistakenly drawn larger than their proper sizes such as the chromosome indicated by an arrow in Fig. 4.1 (b). Fig. 4.1 (c) agrees well with human perception. The segmentation accuracy was also quantitatively measured by comparing with the ground truth. Among 10 images, the lowest and the highest correct rates were 97.5% and 98.7%, and the average was 98.2%.

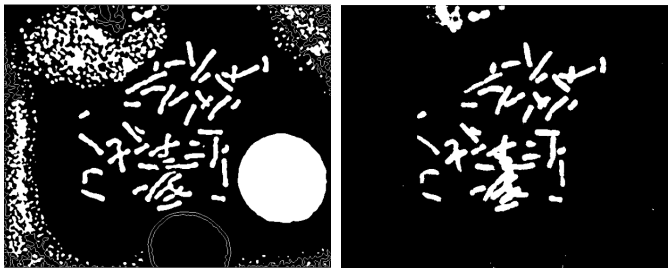
4.1.1 Detailed Procedure for Cell Removal

The cell identification procedure is as follows.

1. All six channels are summed together and scaled to fit 8bit grayscales



(a) DAPI channel (b) Global Threshold (c) Adaptive Threshold



(d) LoG Threshold (e) k-Means Clustering

(f) Ground Truth (g) Segmentation Result

Figure 4.2: Segmentation steps.

since the cells do not always appear in the DAPI channel.

2. This composite image is thresholded using the iterative global thresholding method having the prior of 0.4 for the lower gray scales.

3. The holes inside cells are filled.

4. A morphological open operation with a 5×5 circular structuring element is applied to smooth the boundary. Let's call the resulting image T1.

5. A morphological erode operation with a 51×51 circular structuring element is applied to T1 to remove objects smaller than 25 pixels in width. The structuring element is chosen to ensure that most of the chromosomes are removed and only cells are left. Let's call the resulting image T2.

6. Each blob (found by 8-connectivity) on T2 are examined for the circularity. Let's call a blob T3.

7. T3 is further smoothed by the open operation using a 7×7 circular structuring element, creating T4.

8. The circularity of T4 is measured using $S = 4\pi A/P^2$, where A = area of T4 and P = length of perimeter. The length P is measured using Freeman's chaincode by tracking the boundary, and $P = \alpha \times N_e + \beta \times N_o + \gamma \times N_c$, where α, β, and γ are 0.980, 1.406, -0.091 respectively and N_e = number of even chaincode, N_o = number of odd chaincode,

and N_c = number of corners where the chaincode changes. Let's call the circularity of T4 S1.

9. If S1 is larger than 0.65, the corresponding blob on T1 is examined for its circularity. If its circularity is larger than 0.75 then the blob is identified as a cell.

4.2 Supervised Classification Methods

4.2.1 Maximum-Likelihood Classifier

Unknown samples can be classified by the statistical properties of the samples, if the number of classes, the forms for the class-conditional density functions, the class parameters for the density functions, and the prior probabilities of the classes are known. However in practice, all these parameters and the density functions are not given. Instead, they can be learned from the training data. In chromosome classification, the maximum number of classes is known since the number of chromosome types is fixed. The prior probabilities of the classes can be estimated from the training data or estimated by assuming that an equal number of male and female specimens will be encountered. The parameters for the density functions can be estimated once the forms for the density functions are given. One can visually confirm the forms for the density functions by plotting the sample distributions or using the Parzen window density estimation [57]. However as the dimensionality of the feature space increases, the required number of samples grows exponentially, and visualization is impossible when the dimensionality is more than three. In general, one

can assume the form of the density functions based on knowledge of the data. These forms can be chosen from standard unimodal density functions that best describe the true underlying densities. In M-FISH pixel classification, we assume the sample distributions are normal i.e. $p(\mathbf{x}|\omega) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Given the labeled samples, we estimate the maximum-likelihood values for $\boldsymbol{\theta} = (\boldsymbol{\mu},\boldsymbol{\Sigma})$ for all classes. Suppose that n samples in a training data set D_i are independently drawn from $p(\mathbf{x}|\omega_i)$, where $i = 1, \ldots, c$ and $c =$ number of classes. The likelihood of $\boldsymbol{\theta}_i$ with respect to the set of samples is

$$p(D_i|\boldsymbol{\theta}_i) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}_i). \tag{4.1}$$

The maximum-likelihood estimate of $\boldsymbol{\theta}_i$ is the value $\hat{\boldsymbol{\theta}}_i$ that maximizes $p(D_i|\boldsymbol{\theta}_i)$. The log-likelihood function can be used for the analytical convenience since it is monotonically varying. The estimation is identical for all classes. Thus eq. 4.1 can be written as

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\boldsymbol{\theta}), \tag{4.2}$$

and the maximum-likelihood estimate for $\boldsymbol{\theta}$ can be obtained from a set of equations

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0}. \tag{4.3}$$

The log-likelihood function of the normal density function in one dimension is written as

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \tag{4.4}$$

where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. By solving eq. 4.3, we obtain the following maximum-likelihood solution for μ and σ^2:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

and

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2.$$

With a similar analysis, the maximum-likelihood solution for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the multivariate case are

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

and

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T.$$

This process of estimating the class parameters is called the training of the classifier. Once the classifier is trained, unknown samples can be classified. The likelihood of an unknown sample $\mathbf{x}$ belonging to a class ω_i is written in Bayes formula as

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{4.5}$$

where

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x}|\omega_i)P(\omega_i).$$

The Bayes decision rule (Bayes classifier) assigns an unknown sample $\mathbf{x}$ is to class ω_j if the posterior probability for ω_j is the maximum compared

to all other posterior probabilities. The expression can be written as following without $p(\mathbf{x})$ since it does not affect the decision:

$$p(\mathbf{x}|\omega_j)P(\omega_j) > p(\mathbf{x}|\omega_i)P(\omega_i) \quad \text{for all} \quad i \neq j.$$

In our case, we assumed the equal prior for all chromosomes. Thus the decision rule is solely based on the likelihoods $p(\mathbf{x}|\omega_i)$, i.e. assign $\mathbf{x}$ to class ω_j if

$$p(\mathbf{x}|\omega_j) > p(\mathbf{x}|\omega_i) \quad \text{for all} \quad i \neq j.$$

4.2.2 k Nearest Neighbor Classification

For the maximum-likelihood classifier, we have assumed the underlying density functions to be normal (unimodal). However, in many practical problems the density functions may not be unimodal. In such cases, two approaches can be possible: 1) the multimodal density functions can be modeled as having multiple sub-classes if the forms of the densities can be verified somehow, 2) in cases where the dimensionality prohibits density estimation, a nonparametric method can be used with arbitrary distributions without assuming a form for the density functions.

A popular nonparametric method is the nearest neighbor or the k nearest neighbor method. Given a set of training data, a test sample $\mathbf{x}$ is assigned to a class ω_i when the nearest neighbor of $\mathbf{x}$ in the training data belongs to the class ω_i. The error rate of the nearest neighbor method is greater than the Bayes rate, and never worse than twice the Bayes rate when an unlimited number of training samples is used. A simple extension of the nearest neighbor

method is the k nearest neighbor method, which assign a test sample $\mathbf{x}$ to the most frequent class that its k nearest neighbors belong. As the value of k grows toward infinity, the error rate becomes the Bayes rate. However, in practice the aforementioned is not always true since the number of training samples is limited. In fact, the error rate can even increase as k increases. However it is a useful method when the number of training samples is so small that the class parameters cannot be estimated, or the underlying density functions do not fit a simple unimodal density function.

4.3 Unsupervised Classification Methods

Supervised classification methods, such as the **Bayes classifier** (parametric) and ***k*-nearest neighbor clustering** (nonparametric), require training data. If the number of classes and the form of the class-conditional probability density functions are known, the class parameters can be estimated from the training data, and a parametric classification method can be used. If the number of classes is known but the form of the class-conditional probability functions are unknown, then a nonparametric method such as k-nearest neighbor clustering can be used. In general, collecting and labeling a large set of samples can be extremely costly and even prohibitive for some cases. Fortunately we have a large collection of M-FISH images with ground truth. Thus the use of a supervised method is an adequate approach here. However, in an early stage of investigation regarding the structure of the data based on some features, an unsupervised method is desired since the samples are

unlabeled. Then unsupervised methods can be used to generate the training data set and further to extract useful features. Popular unsupervised methods are **k-means clustering** and **fuzzy k-means clustering**, which group the samples into k clusters whether or not k classes actually exist in the data. These methods can be readily used for normal XX (23 classes) or XY (24 classes) samples where the number of classes is fixed and known. When only the maximum number of classes is known, the use of one of these methods requires the cluster validation to assess the right number of classes by finding the right threshold, which may or may not be feasible depending on the data.

4.3.1 Minimum Distance Classifier

In order to overcome the limitation of these unsupervised methods, we introduce a simple but effective unsupervised and nonparametric classification method for M-FISH images. The concept arises from the fact that a set of samples bound to a particular probe set has an expected intensity pattern for each class. Those fundamental patterns can be used as templates or ideal prototypes for the classes. If our normalization process is effective, then the distance between a normalized sample and its correct class mean (template) should be as small as possible. If that is true, then the minimum-distance classifier [57], without actually training the classifier, can be used to classify pixels, and the classification accuracy can be used to evaluate the effectiveness of the normalization.

The derivation of the minimum-distance classifier is as follows. In

Bayesian decision theory, the minimum-error-rate classification can be achieved by using a set of discriminant functions $g_i(\mathbf{y})$, $i = 1, \ldots, c$, and the classifier assigns $\mathbf{y}$ to class ω_i if

$$g_i(\mathbf{y}) > g_j(\mathbf{y}) \quad \text{for all} \quad j \neq i \tag{4.6}$$

where

$$g_i(\mathbf{y}) = \ln p(\mathbf{y}|\omega_i) + \ln P(\omega_i).$$

If the density functions are multivariate normal in d dimensions, the class-conditional probability density functions are expressed as

$$p(\mathbf{y}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{y} - \boldsymbol{\mu}_i)\right],$$

and then the discriminant functions can be written as

$$g_i(\mathbf{y}) = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{y} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i).$$

If we assume that the features are statistically independent and have the same variance σ^2, then the discriminant functions become

$$g_i(\mathbf{y}) = -\frac{\|\mathbf{y} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

after ignoring the additive constants. After further simplification, we obtain linear discriminant functions whose decision boundary is the hyperplane perpendicular to the line linking the class means. If the prior probabilities are assumed to be the same for all c classes, then the classifier assigns a feature vector $\mathbf{y}$ to the class that yields the minimum Euclidean distance $\|\mathbf{y} - \boldsymbol{\mu}_i\|$.

This classifier is essentially the same as the nearest neighbor classifier. In general, multiple samples are used per class to represent a class in the nearest neighbor method, and after computing the distances to all those samples from an unknown sample $\mathbf{y}$, the sample $\mathbf{y}$ is assigned to the most frequent class among k-nearest neighbors. In the template matching case, only one sample (ideally the class mean) per class is used to represent the class.

In our case, the template patterns are determined by the color table (e.g. Table 3.2). Let a template sample from class ω_1 be $\mathbf{x}_1 = [0,\ 0,\ x,\ 0,\ 0]$, from ω_2 be $\mathbf{x}_2 = [0,\ 0,\ 0,\ x,\ 0]$, and so on, where x can be any positive real number, then the template patterns are defined as

$$\boldsymbol{\mu}_i^t = \frac{\mathbf{x}_i - \mu_{\mathbf{x}_i}}{\sigma_{\mathbf{x}_i}}.$$

After normalization (by EM or background correction), it is important that the samples $\mathbf{y}$ should be further normalized for this classifier, before pixel classification, by $\mathbf{y}' = (\mathbf{y} - \mu_{\mathbf{y}})/\sigma_{\mathbf{y}}$. Thus an unknown sample $\mathbf{y}$ is assigned to ω_i if

$$\left\|\mathbf{y}' - \boldsymbol{\mu}_i^t\right\| < \left\|\mathbf{y}' - \boldsymbol{\mu}_j^t\right\| \quad \text{for all} \quad j \neq i. \tag{4.7}$$

4.3.2 Fuzzy Logic Classifier

A fuzzy-logic classifier is an unsupervised classification method that does not need to assume an underlying distribution, nor does it estimate the distribution. Furthermore, the computational complexity is far less (at least 10 times) than that of the maximum-likelihood classifier, while the classification

accuracy is comparable. It only requires information regarding the labeling of each class (e.g. Table 3.2).

The discriminant functions of the fuzzy logic classifier are formulated as follows:

$$g_i(\mathbf{x}) = \prod_{j=1}^{6} f(x_j) P(\omega_i) \tag{4.8}$$

where i is the class index ($i = 1 \sim 24$), and j is the spectrum index ($j = 1 \sim 6$), $P(\omega_i)$ is the *a priori* probability for class i, and $\mathbf{x}$ is a sample vector.

$$f(x_j) = \begin{cases} x_j & \text{if } T(i,j) = 1 \\ 1 - x_j & \text{if } T(i,j) = 0 \end{cases} \tag{4.9}$$

where T is the color table (e.g. Table 3.2). For example, the discriminant function for class 1 will be (assuming equal priors)

$$g_1(\mathbf{x}) = x_1 \times (1 - x_2) \times (1 - x_3) \times x_4 \times (1 - x_5) \times (1 - x_6) \tag{4.10}$$

A pixel $\mathbf{x}$ belongs to class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$. Only pixels inside foreground are classified using this classifier.

4.4 Postprocessing Methods

4.4.1 Majority and Plurality Filtering

Since many misclassified pixels are surrounded by correctly classified pixels, small local pixel misclassifications can be corrected using neighborhood information. A kernel of a proper size is applied, pixel by pixel, to the initial classification result. In **majority filtering**, a pixel value is replaced with a majority of the pixel values under the kernel, if the majority exists. If a

majority is not found, then the pixel values remains unchanged. Given an $N \times N$ kernel, a majority is the value that occurs more than $N^2/2$ times. In plurality filtering, a pixel value is replaced with the most common value under the kernel. If there is a tie, the pixel value remains unchanged. When the kernel is placed near the boundaries, the background pixels are ignored for the counting. Caution needs to be used when selecting the proper kernel size. However, there is always the danger of removing translocations using these methods. Plurality filtering was used as an intermediate step in the chromosome decomposition process (Chapter 5). An alternative method of correcting misclassifications without removing the translocations is described in the following section.

4.4.2 Prior Adjusted Reclassification

The boundary information is extremely useful when correcting pixel misclassifications. Misclassifications usually occur where chromosomes touch or overlap and near the boundaries of chromosomes. Here we introduce a method that eliminates misclassifications effectively, while preserving translocations, when the boundary information is available.

The majority and plurality filters correct misclassifications regardless of the likelihoods of the pixel being evaluated. The confidence level of a pixel belonging to a class can vary significantly. A pixel may belong to several classes almost equally likely, or only to a particular class with a high likelihood compared to belonging to any other class. Interestingly, we have observed that

when a pixel $\mathbf{x}_1$ belongs to ω_1 but is misclassified as ω_2, the posterior probability difference is small, $P(\omega_2|\mathbf{x}_1) > P(\omega_1|\mathbf{x}_1)$ and $P(\omega_2|\mathbf{x}_1) - P(\omega_1|\mathbf{x}_1) = \epsilon$. The posterior probability is derived from Bayes rule as follows:

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

When $\mathbf{x}_1$ truly belongs to ω_2, the posterior probability difference is usually large: $P(\omega_2|\mathbf{x}_1) >> P(\omega_1|\mathbf{x}_1)$. In the former case, $\mathbf{x}_1$ could be easily reclassified as ω_1 by a small increase in the prior for ω_1. In the later case, a small increase in the prior for ω_1 would not change the classification result. Therefore, the misclassified pixels can be effectively corrected by increasing the prior probability for the correct class. Of course, for this method to work, the right class to increase the prior must be determined for a given chromosome.

A set of pixels that belongs to a boundary B_i is defined as S_i. S_i may contain pixels that belong to multiple classes due to misclassifications or a translocation. Given B_i, there exists the most likely class ω_m among $\{\omega_1, \ldots, \omega_{24}\}$ that S_i belongs to. Given B_i, m is found by the following formula:

$$m = \arg\max_m \left\{ P_s(\omega_m|\mathbf{s}) \sum_{i=1}^{24} P_p(\omega_m|\mathbf{x}_i) P_N(\omega_m) \right\} \tag{4.11}$$

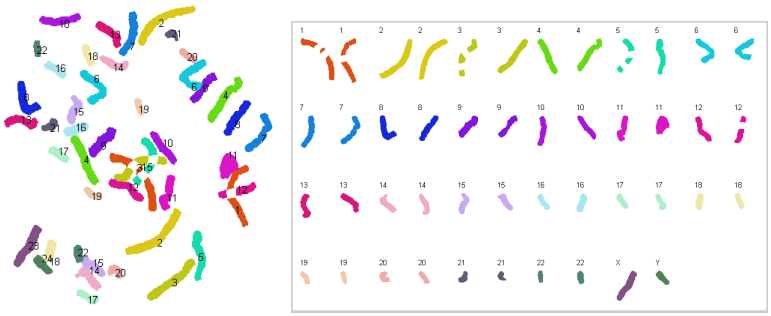
where $P_s(\omega_m|\mathbf{s})$ is the posterior probability given $\mathbf{s}$, $\mathbf{s}$ is the normalized size of B_i, $P_p(\omega_m|\mathbf{x}_i)$ is the posterior probability given a vector that belongs to class ω_i, and $P_N(\omega_m)$ is the normalized number of pixels that belong to ω_m. Three factors are considered in determining the most likely class: the chromosome size, the sum of *a posteriori* probabilities for each class, and the class

population. These three factors are effectively incorporated in order to correct errors. Once ω_m is found, all pixels in B_i are reclassified with a higher prior for ω_m. Note that this method preserves translocations (because those pixels will not be easily reclassified as ω_m, even though the prior for ω_m, a different class from the translocated pixels, is increased) while correcting the misclassifications effectively. An example of prior adjusted misclassification is shown in Fig. 4.3. Pixels are initially classified using the fuzzy-logic classifier (explained in Section 4.3.2).

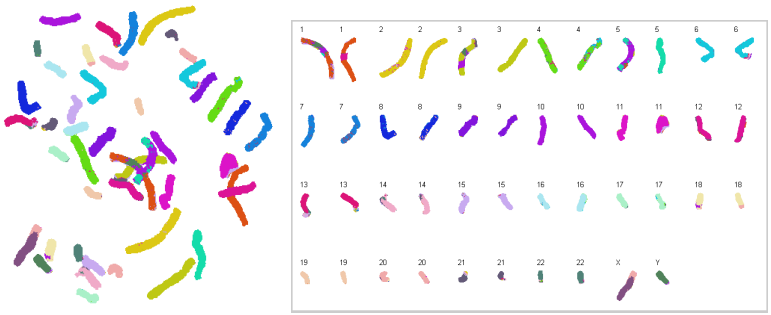
Fig. 4.4 shows another example of the prior adjusted reclassification (image: V240452). The pixel classification accuracy improved from 86.15% on Fig. 4.4 (b) to 93.26% on Fig. 4.4 (c) while preserving the existing translocation. Chromosome 4 has a translocation of chromosome 9. The translocated segmented is not affected by the increased prior on chromosome 4. However, since this chromosome 4 has an added segment from chromosome 9, its size is much closer to that of chromosome 2. Thus the second chromosome 4 was misidentified as chromosome 2. Subsequently, the wrong prior was increased on the second chromosome 4. In such cases, the right prior can be assigned interactively. This problem should be corrected in the future.

4.5 Accuracies of Classification Methods Before and After Normalization

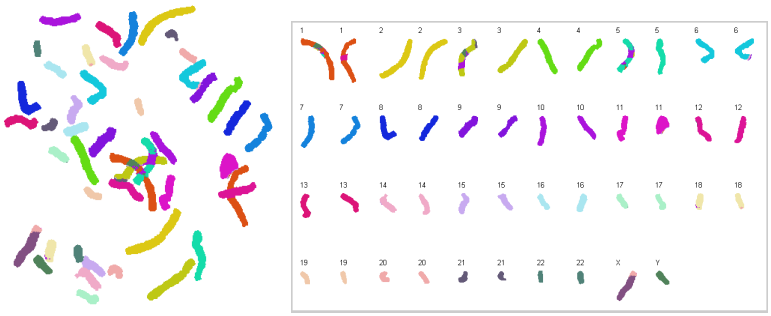
The pixel classifications were performed with three different conditions: no preprocessing, background correction, and EM normalization. Both



(a) Ground truth (b) Karyogram of (a)

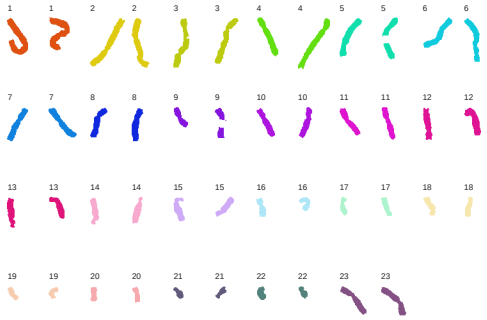


(c) Initial classification (d) Karyogram of (c)

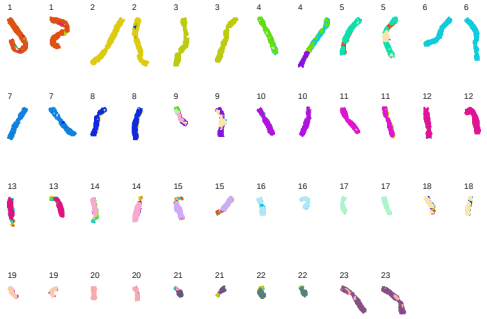


(e) Reclassification (f) Karyogram of (e)

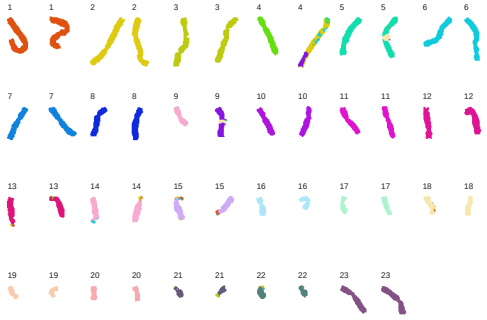
Figure 4.3: Fuzzy logic classification and prior adjusted reclassification



(a) Karyogram of ground truth



(b) Karyogram of initial pixel classification



(c) Karyogram of prior adjusted reclassification

Figure 4.4: Fuzzy logic classification and prior adjusted reclassification. The pixel classification accuracy improved from 86.15% to 93.26%. The chromosome 4 has a translocation of 9. The translocated segmented is not affected by the increased prior on chromosome 4.

Vysis	ASI	PSI
V1301XY	A0101XY	P0801XY
V1302XY	A0102XY	P0802XY
V1303XY	A0103XY	P0803XY
V1305XY	A0104XY	P0804XY
V1306XY	A0105XY	P0805XY
V1801XY	A0201XY	P0808XY
V1802XY	A0202XY	P1102XY
V1803XY	A0205XY	P1103XY
V1805XY		P1104XY

Table 4.1: Training images

unsupervised-nonparametric (the minimum-distance classifier) and supervised-parametric (the maximum-likelihood classifier) methods were used for classification.

Since the maximum-likelihood classifier requires training, a set of images of normal male specimens was selected as training samples for each probe set as shown in Table 4.1. A total of 26 out of 185 images were used for training: 9 out of 85 images for Vysis images, 8 out of 71 images for ASI images, and 9 out of 29 images for PSI images. All 185 images were tested using both classification methods. Eq. 4.6 was used for the maximum-likelihood classifier assuming the distributions were normal and eq. 4.6 was used for the minimum-distance classifier to classify pixels.

As table 4.2 shows, the overall classification accuracy without any normalization was about 50%, which increased significantly after background correction to about 60%, and further improved with EM normalization to about

Methods	Minimum-distance classifier			Maximum-likelihood classifier		
	NP	BC	EM	NP	BC	EM
Accuracy [%]	47.12	60.11	68.70	47.86	62.46	72.72

Table 4.2: The overall classification accuracy. NP = no preprocessing, BC = background correction, and EM = expectation maximization normalization.

70% for both classification methods. EM normalization increased the classification accuracy from 50% to 70%, which is a 40% increase in accuracy.

Table 4.3 shows the classification accuracies of the commonly cited images in previous papers regarding M-FISH pixel classification. Note that the results shown in this paper are the initial pixel classification accuracies without any post-processing to correct obvious misclassifications using such methods as majority filtering, and also note the rates are regarding the chromosome pixels only. Since chromosomes occupy less than 10% of the image, even if the entire pixels in the image are classified, the rates for background and chromosomes should be reported separately. Our results are by far the most accurate compared to the other classification methods, such as the fuzzy logic classifier (unsupervised nonparametric method) [12], fuzzy k-means clustering (supervised nonparametric method) [50], k-nearest neighbors method (supervised nonparametric method) [58], and the maximum-likelihood classifier (supervised parametric method) [49]. All of these classifiers will show an improved classification accuracy after EM normlization.

In order to evaluate the statistical significance of the effect of the EM normalization, bootstrap estimation was used. Given 185 data points for clas-

	Minimum-distance classifier			Maximum-likelihood classifier		
Images	NP	BC	EM	NP	BC	EM
V1301XY	63.77	88.51	90.81			
V1302XY	83.35	92.35	92.99			
V1303XY	81.15	90.13	92.77			
V1305XY	93.01	92.36	94.72			
V1306XY	87.64	89.81	94.66			
V1308XY	77.95	86.63	96.28	84.45	93.89	96.49
V1309XY	56.16	83.58	84.50	70.34	84.55	86.29
V1310XY	80.57	88.00	84.19	86.03	88.50	86.90
V1311XY	94.01	93.48	94.50	90.79	90.99	94.54
V1312XY	87.31	93.44	94.83	95.09	95.25	95.20
V1313XY	89.92	91.69	94.53	93.76	94.82	94.88
Average	81.35	90.00	92.25	86.74	91.33	92.38

Table 4.3: Classification accuracies [%] of the commonly cited images. Images with empty values in the ML method are used as training.

sification accuracies per method, 185 samples were selected at random (from a uniform distribution) iteratively for 1000 times, and at each iteration the mean was calculated. The distribution of the means for each method is shown in Fig. 4.5. The error bars represent the 95th percentile of the means. As the graph shows, the accuracies after the EM normalization are statistically significant. The difference between the two classifiers are not significant except after the EM normalization. However we should mention that 26 images were used for training and their classification results are also included in the ML accuracy. Therefore, it is reasonable to assume that the difference is slightly smaller than 4%, and whether it is statistically significant or not, the difference is marginal between the two classifiers.

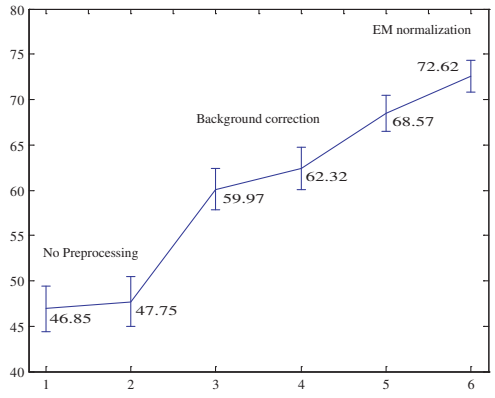


Figure 4.5: Statistical significance of each classification method. The bootstrapping of each method. Left to right: NP_MD, NP_ML, BC_MD, BC_ML, EM_MD, and EM_ML. The error bars are drawn at the 95th percentile.

Fig. 4.6 shows an example of a color coded classification result (its spectral images are shown in Fig 3.13). The classification accuracies using the MD classifier without preprocessing, with background correction, and with EM normalization were 55.34%, 75.64%, and 84.03% respectively. This particular image has 6 translocations but they are unmarked in the ground truth in the database. After carefully examining all six spectral images and manually constructing the new ground truth, the recalculated accuracy was 91.52%. There are 104 images that contain abnormalities, and among them 63 images contain translocations. If the ground truth in the databse were marked with translocations, the true overall classification accuracy may have been slightly higher.

There are many images that give low classification accuracies even after EM normalization. The common factor among those images is that the

(a) Ground truth (b) Corrected ground truth (c) EM_MD

Figure 4.6: Classification result of V290562 (spectral channels are shown in Fig. 3.13).

image quality is poor for various reasons. All 185 images were individually self trained and tested to evaluate the quality of the feature distribution. The mean accuracy was 89.95% with 51.30% as the minimum and 99.00% as the maximum (see Fig. 4.7). Images that gave lower than 85% correct classification rate were identified (also visually confirmed) as bad images. In addition, three images that had higher than 90% rate were added to the bad ones because they had wrong probe labeling. They were labeled as Vysis when, in fact, they were hybridized using the PSI probe. A total of 40 images were identified as bad, and the list is shown in Table 4.4.

In M-FISH, all 6 spectral channels are expected to be perfectly aligned to each other. While that is true for most cases, four images were identified as misaligned. The misalignment can come about from various reasons including mechanical shift during image capture and the use of poor quality lenses with spherical and chromatic aberration. The misalignment in V291562 was a vertical shift of 10 pixels only in channels 4 and 6. The misalignment in P080628,

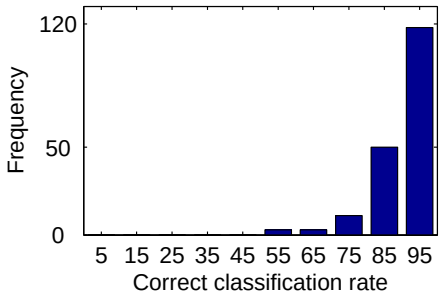


Figure 4.7: Correct classification rate of individually self trained and tested images. Ten bins are used from 0 to 10, 10 to 20, . . . , 90 to 100.

P080729, and P0804XY was found in the DAPI channel only.

It is interesting that the image quality varied depending on the probes. Images with the Vysis probe were captured with a good quality in general. Many ASI probe images displayed a poor quality. In many cases, at least one or more spectral channels occupied only a low intensity range. The cause of it can be either that the hybridization process was done poorly, or the exposure times were not set correctly. Many PSI probe images showed low SNR as the signals do not have the high contrast. However, this comparison may not generalize the quality of the probes, since the image quality will depend on the quality of the specimen preparation and the settings in the microscope. Since images in the database were collected from five different labs, we do not know whether the image quality difference comes from the human error or from the probe difference.

Excluding the bad images, the classification accuracy of the remaining 145 images are shown in Fig. 4.8. The mean accuracies were 51.66%, 52.42%,

File name	Condition	File name	Condition	File name	Condition	File name	Condition
V250253	PQ	A020818	PQ	A0507XY	PQ	P070218	PQ
V260754	CT	A0202XY	PQ	A0604XY	PQ	P080628	PQ, MA(1)
V260856	CT	A0205XY	PQ	A0609XY	PQ	P080729	PQ, MA(1)
V290162	CT	A0206XY	PQ	A0614XY	PQ	P080930	PQ
V290362	CT	A0207XY	PQ	A0621XY	PQ	P0802XY	PQ
V290962	CT	A0209XY	PQ	A200344	PQ	P0804XY	PQ, MA(1)
V291562	MA(4,6)	A0402XY	PQ	A200444	PQ	P0808XY	PQ
A0102XY	PQ	A0403XY	PQ	A200544	PQ	V1701XY	WP
A020402	PQ	A0503XY	PQ	A200644	PQ	V1702XY	WP
A020315	PQ	A0506XY	PQ	P070109	PQ	V1703XY	WP

Table 4.4: List of bad quality images. PQ = poor quality due to either ill-hybridization or wrong exposure times, CT = channel crosstalk, MA = misalignment, WP = wrong probe.

65.40%, 67.58%, 74.49%, and 77.80% for NP_MD, NP_ML, BC_MD, BC_ML, EM_MD, and EM_ML respectively.

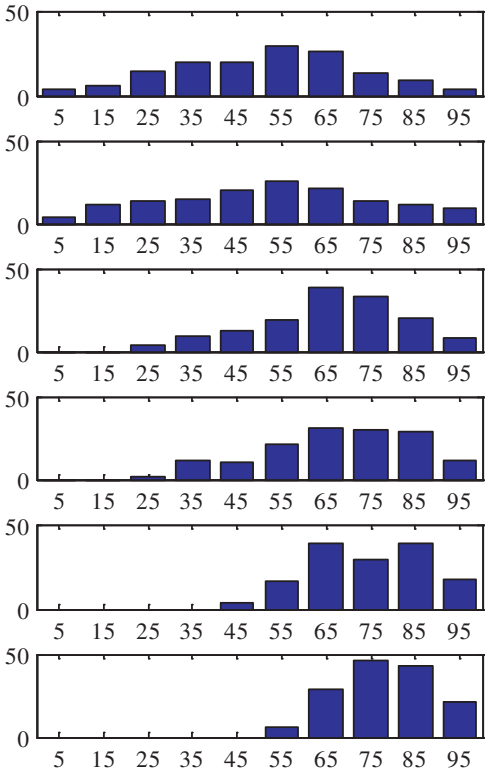


Figure 4.8: Histogram of classification accuracies. x axis represents the classification accuracy [%], y axis represents the frequency. Ten bins are used from 0 to 10, 10 to 20, ..., 90 to 100. Top to bottom: NP_MD, NP_ML, BC_MD, BC_ML, EM_MD, and EM_ML respectively.

Chapter 5

Decomposition of overlapping and touching M-FISH chromosomes

5.1 Introduction

Automatic segmentation of partially occluded and/or touching objects is an extremely challenging task. Chromosome images are subject to the partial occlusion and touching of chromosomes. This is one of the major factors that hinders automating the analysis. There have been numerous segmentation (decomposition) methods developed for conventional banded chromosome images. Among them, some methods only handle touching cases and some handle both cases with limited success. Most of the methods utilize only geometry information of chromosome clusters, such as curvature, skeleton, and convex hulls [2, 3]. The geometry based methods only analyze the boundary shape of a chromosome cluster. Even though the boundary shape contains rich information about the cluster formation, there are many cases where the boundary information itself is not sufficient such as a touching of two chromosomes by their short sides or long sides forming a long chromosome or a thick chromosome. These touching cases can be easily discerned when the pixel memberships are presented by two distinctive colors, as in M-FISH. When the pixel classification accuracy is high, the color information itself may be

sufficient for the chromosome segmentation for most cases. Schwartzkopf *et al.* [8] proposed a maximum likelihood decomposition method using the pixel classification results and chromosome size for M-FISH images. Authors compared their results to that of commercially available software (Cytovision), and reported that much better results were achieved for touching cases and less reliable results for overlapping cases. When only the colors are used, touchings or overlaps of the same type of chromosomes cannot be segmented, and the segmentation accuracy heavily relies on the initial pixel classification accuracy. Thus the both geometry and pixel classification results have to be merged in order to achieve better segmentation results.

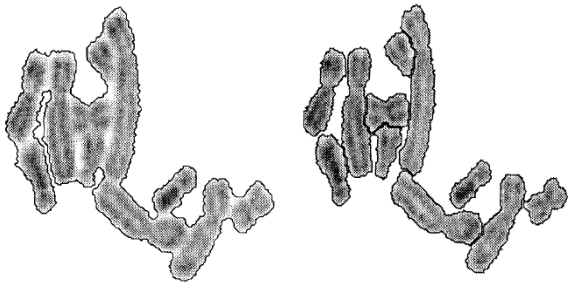
In this chapter, we present a novel decomposition method for overlapping and touching chromosomes that utilizes the geometry of a cluster, pixel classification results, and chromosome sizes. We also introduce basic elements of overlap and touching cases. These basic elements yield hypotheses of possible overlapping and/or touching cases. Given a cluster, multiple hypotheses are evaluated, and the most likely hypothesis is chosen as the correct decomposition.

5.2 Background

5.2.1 G-banded Chromosome Decomposition

Ji [2] had developed a simple but effective method to segment touching chromosomes based on two hypotheses: (a) at points where chromosomes touch, the optical density is relatively low; (b) where chromosomes touch, the

(a) Result 1 (b) Result 2



(c) Result 3

Figure 5.1: Separation results of Ji's method [2].

cluster boundary tends to form an acute angle. Based on these ideas, touching chromosomes were effectively segmented. This algorithm is implemented in one of the current commercially available karyotyping systems. Some of the results are shown in Fig. 5.1. As shown in the figure, overlapping chromosomes are not segmented.

Agam and Dinstein [3] developed a method that can handle both overlapping and touching chromosomes by analyzing the concave points on the boundary. After connecting all concave points as shown in Fig. 5.2, pairs of parallel lines are retained as valid cut lines (Fig. 5.2 right). A polygon that is contracted and bent on one point was fitted to chromosomes, and among all possible combinations, a combination that satisfied the best criteria was

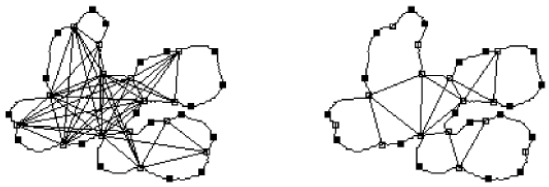


Figure 5.2: Possible separation lines of Agam and Dinstein's method [3]

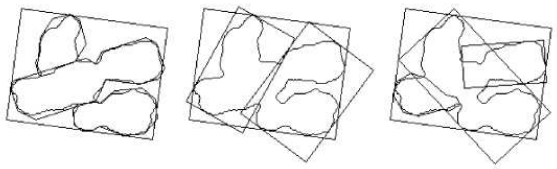


Figure 5.3: Several hypotheses for a cluster of three chromosomes [3].

chosen as the correct separation. Three possible combinations for a cluster of three touching chromosomes are shown in Fig. 5.3. A rectangle was drawn when a chromosome does not satisfy a certain condition for fitting the polygon. The developed method was tested on 25 selected images that were suitable for the analysis. The accuracies for two, three, and more than four chromosome clusters were 88%, 68%, and 63% respectively.

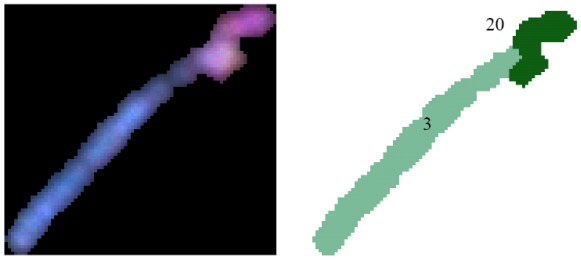
5.2.2 M-FISH Chromosome Decomposition

Since the pixel membership information is available for M-FISH images, Schwartzkopf *et al.* [8] developed a joint pixel classification and segmentation method which can handle overlapping and touching chromosomes for M-FISH images, utilizing the color information in a maximum likelihood framework. After the initial pixel classification using a 6-feature, 24-class

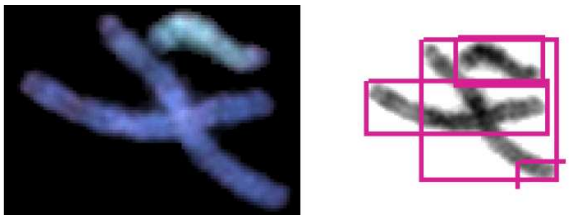
(a) Initial pixel classification (b) Majority filtering (c) Segmentation result

Figure 5.4: Segmentation results of an M-FISH image by Schwartzkopf's method [4].

maximum-likelihood classifier, a 17×17 majority filtering was applied to correct small misclassifications. Touching and overlapping chromosomes were separated into a set that maximizes the overall likelihood with respect to pixel membership and chromosome size. An example result is shown in Fig. 5.4. Figure 5.5 (a) shows that Schwartzkopf's method successfully segmented a touching case that appeared as a long chromosome, which could not be segmented using the commercial Cytovision software. Figure 5.5 (b) shows that Schwartzkopf's method did not work because two overlapping chromosomes belonged to the same class. The separated chromosomes result in an increased pixel classification accuracy since the algorithm corrects misclassifications while merging color blobs. However the merging process is greedy instead of optimal: given a number of blobs, the method joins the pair that yields the maximum likelihood compared to all other pairs, and this may not lead to the correct segmentation.



(a)



(b)

Figure 5.5: (a) Schwartzkopf's method successfully decomposed touching chromosomes, whereas grayscale based method (using Cytovision software) could not since two chromosomes appear as a long chromosome. (b) Grayscale based method could decompose, whereas Schwartzkopf's method could not since two overlapping chromosomes belong to the same class [4].

5.3 Methods

The limitations of the geometry-based and color-based methods can be overcome by incorporating information from both. The new method utilizes the geometry of a cluster, pixel classification results and chromosome sizes. This section describes the details of the implementation.

After chromosome segmentation from the background, only the chromosome pixels are normalized using EM normalization and are classified using an unsupervised nonparametric method, or the minimum-distance classifier, which is described in Section 4.3.1.

5.3.1 Elements of clusters

We define a group of connected pixels by 4-connectivity as a cluster, S_i. Clusters are found after segmenting the chromosomes from background using the segmentation method described in Section 4.1, and eroding the segmentation result with a 3×3 structuring element. The erosion is performed to avoid evaluating simple touching cases where chromosomes are connected by one pixel. Each cluster is dilated back before being evaluated for touching and overlapping.

A cluster can be formed by one chromosome or multiple chromosomes. Whether a cluster is formed by one or multiple chromosomes, every cluster is subjected to evaluation.

We define three sets of basic elements for clusters as follows (see Fig.

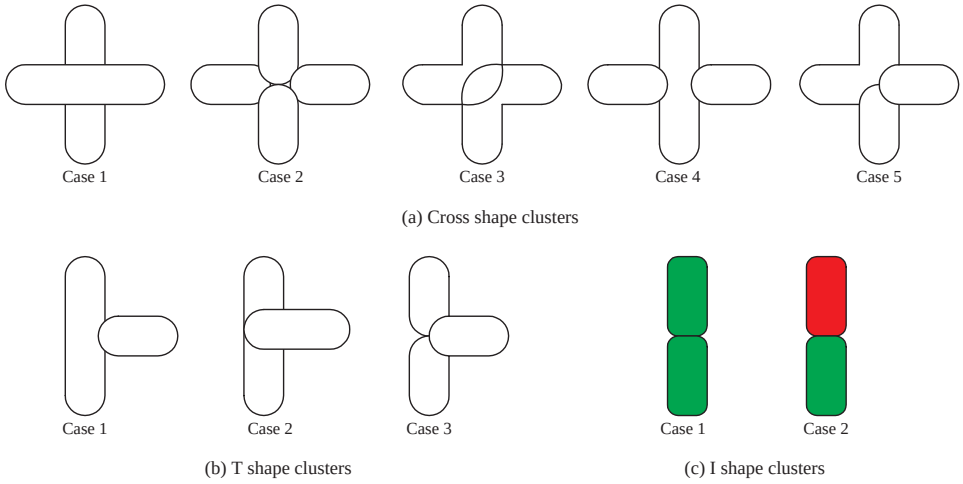
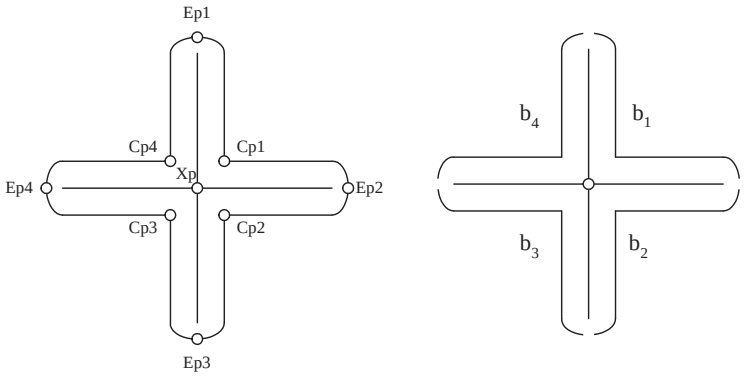


Figure 5.6: Elements of clusters.

5.6)

1. Cross shape cluster
2. T shape cluster
3. I shape cluster

Most of the cluster formations can be decomposed into the basic elements. Given a cluster, we also define the landmarks such as cut points (Cp), cross points (Xp), and end points (Ep) on the skeleton and on the boundary of the cluster as shown in Fig. 5.7. There exists an Xp that is connected to an Ep, and an Ep connects two boundary segments b. Given $\{Ep,Xp\}$ and two boundary segments, the closest points on b's from the Xp are the cut points associated with the Xp. A cluster can have multiple Xp's and each Xp



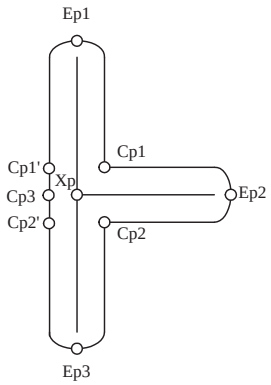
(a) Landmarks (b) Boundary segments

Figure 5.7: Landmarks of a cross shape cluster

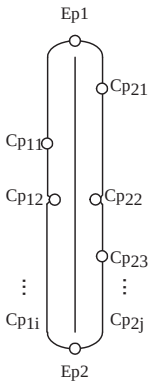
has three or four cut points. Once all the landmarks are found, all possible decompositions are evaluated.

The cross shape cluster has 5 cases: case 1 is an overlap of two chromosomes, case 2 is a touching of four chromosomes, case 3 is a touching of two chromosomes, and case 4 and 5 are touchings of three chromosomes (see Fig. 5.6). Two chromosomes are found by connecting $\{Cp1 - Cp4, Cp2 - Cp3\}$ and $\{Cp1 - Cp2, Cp4 - Cp3\}$ for case 1. Four chromosomes are found by connecting $\{Cp1 - Xp - Cp3, Cp2 - Xp - Cp4\}$ for case 2. Case 3 has two subcases, where two chromosomes are found by connecting $\{Cp1 - Xp - Cp3\}$ for one case, and $\{Cp2 - Xp - Cp4\}$ for another case. The same analogy can be applied to case 4 and 5. Case 4 has two subcases and case 5 has four subcases. In total, there are ten hypotheses to evaluate in the cross case.

The T shape cluster has 3 cases: case 1 is a touching of two chromosomes (three subcases), case 2 is a partial overlap of two chromosomes (three



(a) Landmarks of T shape cluster



(b) Landmarks of I shape cluster

Figure 5.8: Landmarks of clusters

subcases), and case 3 is a touching of three chromosomes at the center. In total there are seven hypotheses to evaluate in T case. Case 1 is evaluated by connecting $\{Cp1 - Cp2\}$, $\{Cp1 - Cp3\}$, or $\{Cp2 - Cp3\}$. Case 2 is evaluated by connecting $\{Cp1 - Cp1', Cp2 - Cp2'\}$ and $\{Cp1' - Cp2', Cp1 - Cp2\}$ (see Fig. 5.8). $Cp1'$ and $Cp2'$ are found by extending lines from $Cp1$ and $Cp2$ with the slope of a line between Xp and $Ep2$. In fact, any shape that has three end points is a T shape cluster (imagine a Y shape cluster). Thus, $Cp1'$ and $Cp2'$ are evaluated from all three arms. Case 3 is evaluated by connecting $\{Cp1 - Xp - Cp2\}$.

Not all of these shapes occur equally likely. Forming a cross shape cluster with four different chromosomes by their short sides will certainly have a lower chance of occurring than the case of two chromosomes crossing each other. We have examined all chromosome clusters in the database. There were

Cluster shape	Case 1	Case 2	Case 3	Case 4	Case 5	Total
Cross shape	99	0	13	5	0	117
T shape	132	32	0			156
I shape						127
Extra						7

Table 5.1: The number of occurrences of the basic shapes.

117 Cross shape clusters: 99 for case 1, 0 for case 2, 13 for case 3, 5 for case 4, and 0 for case 5. There were 156 T shape clusters: 132 for case 1, 32 for case 2, and 0 for case 1. These are tabulated in Table 5.1. These values act as a prior probability for a given hypothesis. Thus, the cases that have zero prior probability is removed from the hypothesis evaluation for this study.

We define a cluster that does not have a cross point as an I shape cluster which may have touchings of the same chromosomes or different chromosomes. The I shape cluster has an arbitrary number of cases. The number of segments are determined by the number of concave points on the boundary. There are two end points in I shape cluster that divide the boundary into two segments. Concave points across each boundary are connected and the minimum number of pairs of which have minimum distances determine the final number of chromosome segments. For example, Fig. 5.8 (b) will have three segments separated by two lines, $\{Cp11 - Cp21\}$ and $\{Cp12 - Cp22\}$. In general there can be i and j number of convex points on each side of boundaries. Given N segments, 2^{N-1} combinations are evaluated. If three segments are found, for example, then there are four possible chromosome formations: {1-2-3}, {1-2, 3}, {1, 2, 3}, and {1, 2-3}, i.e. in words, all three segments form a chro-

mosome, segments 1 and 2 form a chromosome and segment 3 form another chromosome, and so on. In order to account I shape clusters that are formed by different chromosomes but have no obvious concave points, chromosome segments are determined by the pixel classification results (color). An area with a homogeneous color forms a segment. Again, given M segments, $2^{M-1}-1$ combinations are evaluated (evaluation of all segments as one chromosome is considered in concave points based method). Thus, a total of $(2^N+2^M)/2-1$ hypotheses are evaluated for an I shape cluster.

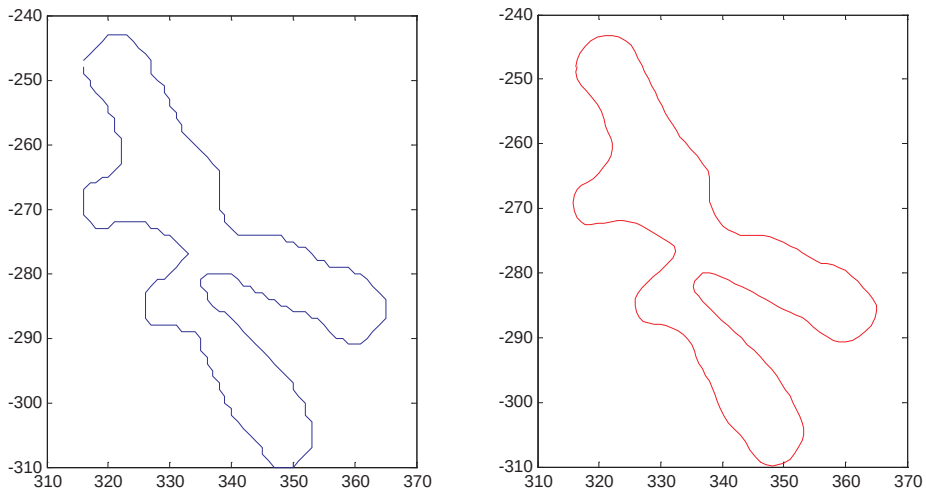
5.3.2 Concave Points Detection

Concave points are found by analyzing the angle changes of the boundary points. Boundaries are first smoothed using a wavelet denoising method. $\mathbf{x}$ and $\mathbf{y}$ are the coordinates of boundary points. $\mathbf{x}$ and $\mathbf{y}$ are independently decomposed up to 3 decomposition levels, and the corresponding wavelet coefficients are soft thresholded at each level. Then the inverse wavelet transform of those coefficients yields the smoothed boundaries made up of non-interger $\mathbf{x}$ and $\mathbf{y}$ values, to avoid sampling grid effects. A result of boundary smoothing is shown in Fig. 5.9.

After the boundary is smoothed, the tangents α at every point are calculated from vectors that connect the current point to the second neighboring point:

$$\alpha(i) = tan^{-1}\left(\frac{\mathbf{y}(i+2)-\mathbf{y}(i)}{\mathbf{x}(i+2)-\mathbf{x}(i)}\right) \quad [rad].$$

The second neighbors are used in the equation to further reduce the effect of



(a) Boundary points (b) Smoothed boundary

Figure 5.9: Boundary is smoothed using a wavelet denoising method.

noise in the boundary. The direction of the boundary tracking can be either clockwise or counterclockwise. Regardless of the tracking direction, a segment of boundary can be either concave or convex depending on the perspective. Depending on whether the inside of an object is on the left or on the right of the tracking direction, the angle changes on the boundary are found by

$$\theta(i) = \begin{cases} 180 + \Delta\alpha(i) & \text{if the inside is on the left} \\ -180 + \Delta\alpha(i) & \text{otherwise} \end{cases} \tag{5.1}$$

where $\Delta\alpha(i) = \alpha(i+1) - \alpha(i)$.

Given the boundary shown in Fig. 5.9, θ, the derivative of the tangent of the boundary is shown in Fig. 5.10. Given θ, the concave regions are found where

$$\begin{cases} \theta < 180 & \text{if the inside is on the left} \\ \theta > -180 & \text{otherwise} \end{cases}$$

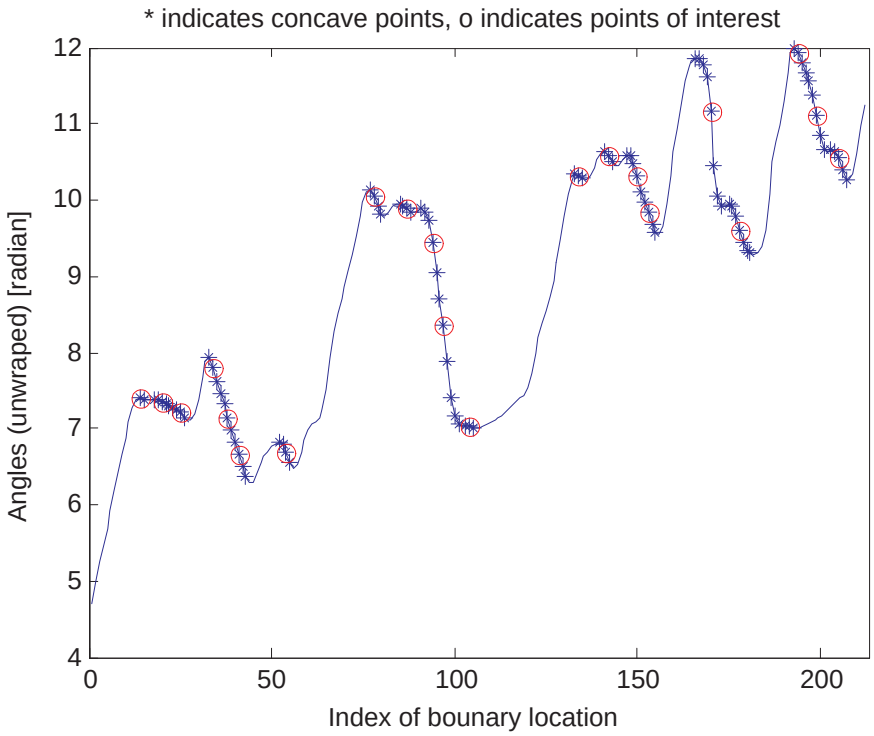
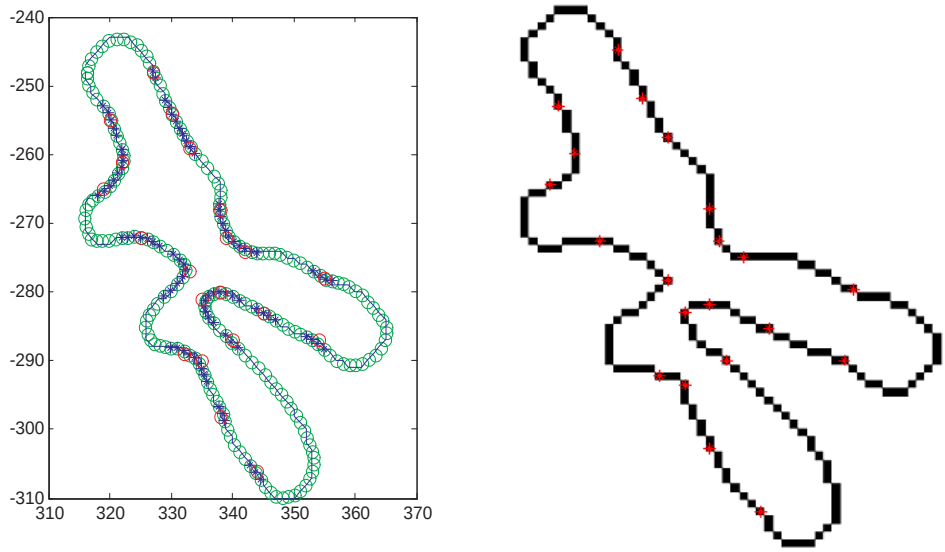


Figure 5.10: The derivative of the tangent of the boundary shown in Fig. 5.9.

The points that are marked by $*$ are concave regions on the boundary, and the points marked by circles, which are the points of interest, are found where the second derivative of θ in concave regions crosses zero from negative to positive. The concave regions are mapped on the downslope of θ, and the convex regions are mapped on the up-slope of θ. Thus using this method, the convex or concave regions can be found simultaneously. The results of concave point finding are shown in Fig. 5.11.

The detected concave points are rotation invariant. Even though an example is shown with an overlap case, the concave point finding is performed



(a) Concave regions and concave points (b) Concave points

Figure 5.11: Concave point detection. (a) Solid line is the original boundary, green circles are the smoothed boundary, $*$s represent the concave regions, and circles show the concave points of the boundary. (b) The concave points detected and are marked with red stars on the original boundary.

only on the I shape clusters.

5.3.3 Evaluation of the hypothesis

Given a cluster, there are a number of hypotheses, and each of them is composed of single or multiple chromosomes. The best possible cut is achieved when a cluster is decomposed into the right number of chromosomes with the right sizes and at the same time maximizing the homogeneity of the pixel memberships within each chromosome. Among N_k hypotheses for a given cluster, a hypothesis that maximizes the following posterior probability is chosen as the most likely decomposition of the cluster:

$$P(S_k|\psi_k) = \frac{p(\psi_k|S_k)P(S_k)}{p(\psi_k)} \tag{5.2}$$

where

S_k = the state of nature of the k^{th} hypothesis,

$\psi_k = (s_i, P(\omega_i))$ = (size, weight), $i = (1, \ldots, N_c(k))$,

$N_c(k)$ = number of chromosomes in the k^{th} hypothesis,

$p(\psi_k|S_k)$ = the likelihood of S_k with respect to ψ_k,

$P(S_k)$ = the prior probability of the k^{th} hypothesis, and

$p(\psi_k)$ = the evidence, $\sum_{k=1}^{N_k} p(\psi_k|S_k)P(S_k)$.

The variable S is described by the parameter vector ψ, which contains the sizes of all chromosomes s_i and their weights $P(\omega_i)$ given a hypothesis, where

ω is the variable for chromosome categories. Note that $P(\omega_i)$ are different from the actual prior probabilities of each chromosome. They are the weights of the segmented chromosomes in a cluster. $P(\omega_i)$ becomes larger as more pixels in chromosome i are classified as ω_i. Also note that chromosome i here means a group of connected pixels being evaluated as a chromosome in a cluster.

The likelihood of a hypothesis is computed by

$$p(\psi_k|S_k) = \prod_{i=1}^{N_c(k)} p(s_i|\omega_i)P(\omega_i) \tag{5.3}$$

where,

$P(\omega_i) = \frac{N_i}{N_{ci}}$,

ω_i = the most popular class in chromosome i,

N_i = the number of pixels belong to ω_i in chromosome i,

N_{ci} = the total number of pixels belong to chromosome i,

$s_i = \frac{N_{ci}}{N_T}$, normalized size of chromosome i,

N_T = total number of chromosome pixels in an image, and

$p(s_i|\omega_i)$ = class-conditional probability density function for chromosome size s_i given class ω_i.

The prior probabilities for hypotheses $P(S_k)$ are computed from the Table 5.1 (e.g. the prior for the T shape, case 1 is 132/156). Among all

hypotheses, the one that has the maximum posterior probability is chosen as the correct decomposition of the cluster: the hypothesis S_k is chosen when

$$P(S_k|\psi_k) > P(S_j|\psi_j) \quad \text{for all} \quad j \neq k \tag{5.4}$$

Since $p(\psi_k)$ in eq. 5.2 is a normalization factor that does not affect the decision in eq. 5.4, it is factored out from eq. 5.4.

For clusters with a combination of multiple shape elements and I shape clusters, the equal priors are used for $P(S_k)$ (for computational simplicity, $P(S_k) = 1$ was used for these cases).

The class-conditional probability density functions for the size are defined as

$$p(s|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(-\frac{1}{2}\left(\frac{s_i - \mu_i}{\sigma_i}\right)^2\right) \tag{5.5}$$

The class parameters, μ_i and σ_i $(1 \leq i \leq 24)$, were initially calculated from 16 images in the database (the list is shown in Table 5.2). A caution needs to be taken when computing the size parameters from the database since the database does not provide information of the pixel memberships where pixels overlap. The pixels belonging to multiple chromosomes should be counted as many times as the number of chromosomes to which they belong. The mean sizes calculated from the database did not exactly match the standard chromosome sizes in order (shown in Table 5.3, note that parameters were estimated from the normal chromosomes), because the segmentation results in the database were not perfect. Also, the variance in size was large: the size probability of a chromosome becomes unreliable whenever the size of a

v1301xy	v1303xy	v1309xy	v1310xy
v1311xy	v1313xy	v1302xy	v1305xy
v1306xy	v1308xy	v1312xy	v1701xy
v1702xy	v1703xy	v1902xy	v2704xy

Table 5.2: Training images used for the size parameter estimation.

chromosome deviates much from its mean size. However, the large variance is somewhat desirable since we are not dealing only with normal chromosomes. The chromosome sizes vary due to structural abnormalities (e.g. deletions, insertions, and translocations) and segmentation errors. Therefore, the mean values were adjusted based on the standard chromosome sizes (obtained from NCBI's website, and shown as chromosome sizes [Mbps] in Table 5.3). Among 24 chromosome types, the normalized sizes of large chromosomes calculated from the M-FISH database correlated well with the standard sizes (not the exact numbers but the relative sizes and their orders). Thus, seven chromosomes (chromosome 1 to 7) were used to calculate the unit megabase pairs per pixel. The mean unit Mbps/pixel was 2.3154 for the images in the database. Using the mean unit Mbps the normalized chromosome sizes were recalculated as shown in column NS_NCBI in Table 5.3. The probability density functions of chromosome sizes are shown in Fig. 5.12. Instead of using the estimated σ_i, which were different for different chromosomes, the equal variance of 5×10^{-5} was used for all chromosomes.

Chromosome	Chromosome length [Mbps]	NS_NCBI	NS_Database
1	245203898	0.0414	0.0417
2	243315028	0.0411	0.0404
3	199411731	0.0337	0.0344
4	191610523	0.0323	0.0326
5	180967295	0.0305	0.0302
6	170740541	0.0288	0.0286
7	158431299	0.0267	0.0267
8	145908738	0.0246	0.0238
9	134505819	0.0227	0.0216
10	135480874	0.0229	0.0217
11	134978784	0.0228	0.022
12	133464434	0.0225	0.0224
13	114151656	0.0193	0.0198
14	105311216	0.0178	0.0179
15	100114055	0.0169	0.0178
16	89995999	0.0152	0.0143
17	81691216	0.0138	0.013
18	77753510	0.0131	0.0135
19	63790860	0.0108	0.01
20	63644868	0.0107	0.0106
21	46976537	0.0079	0.0088
22	49476972	0.0084	0.0093
X	152634166	0.0258	0.026
Y	50961097	0.0086	0.0118

Table 5.3: Normalized mean chromosome sizes. NS_NCBI represents normalized mean chromosome sizes calculated using the known chromosome lengths (obtained from NCBI's website). NS_database represents the normalized mean chromosome sizes calculated from the ADIR M-FISH database.

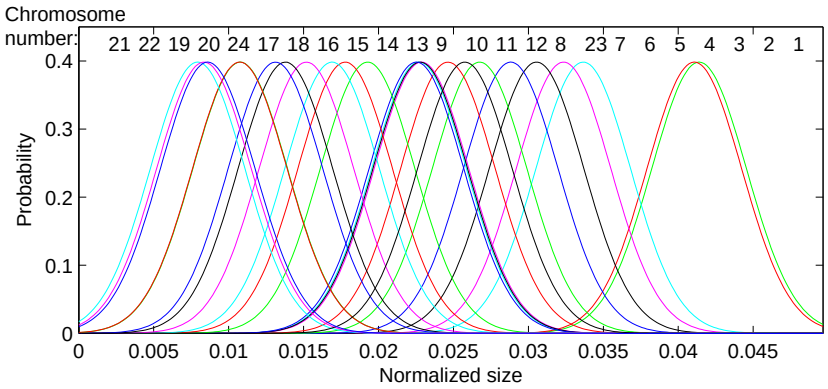


Figure 5.12: Probability density functions of the normalized chromosome sizes.

5.3.4 Decomposition Steps

Given a group of pixels and their class memberships, the landmarks are found on the boundary as shown in Fig. 5.13. Based on the landmarks, the cluster is identified as Cross shape, T-shape, I-shape, or Multiple-shape. If the cluster is defined as either Cross or T shape, then the corresponding subcases are evaluated. Among the subcases, the case that has the maximum-likelihood is chosen as the best separation of the cluster given the shape constraint. Then, the individual chromosomes in the best subcase are evaluated for touchings (I-shape evaluation). An example of the decomposition procedure for a cluster of two chromosomes that belong to the same class crossing each other is shown in Fig. 5.14. The overlaps of the same class chromosomes are decomposed successfully using the developed method. Fig. 5.15 shows the decomposition steps for a T-shape cluster.

If the landmarks of a cluster include more than one cross point, Xp, the cluster is identified as a Multiple-shape. Ignoring all the other Xp's,

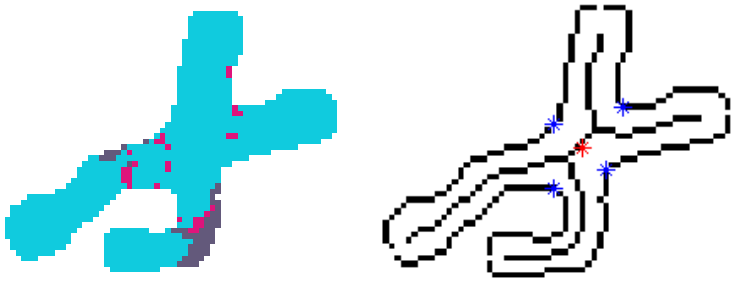


Figure 5.13: Landmarks of a cross shape cluster

the decomposition is evaluated at only one Xp and its associated cutpoints. The initial evaluation results in multiple single chromosomes and multiple clusters. Those multiple clusters are decomposed in the same manner. This process repeats until all Xp's in all clusters are evaluated. Then the set of single chromosomes that yields the maximum-likelihood is chosen as the best decomposition of all the evaluations. Then finally the single chromosomes in the best decomposition are evaluated for the touchings.

5.4 Results

Chromosomes are first segmented from the background using the new segmentation method explained in Section 4.1. Then chromosome pixels are classified using the minimum-distance classifier after the EM normalization [10]. Given a cluster, the landmarks on the boundary and skeleton are computed as shown in Fig 5.16, and the cluster is decomposed into multiple hypotheses and the likelihood of each hypothesis is computed by eq. 5.3. When there are multiple Xp's, hypotheses are evaluated at each Xp consecutively.

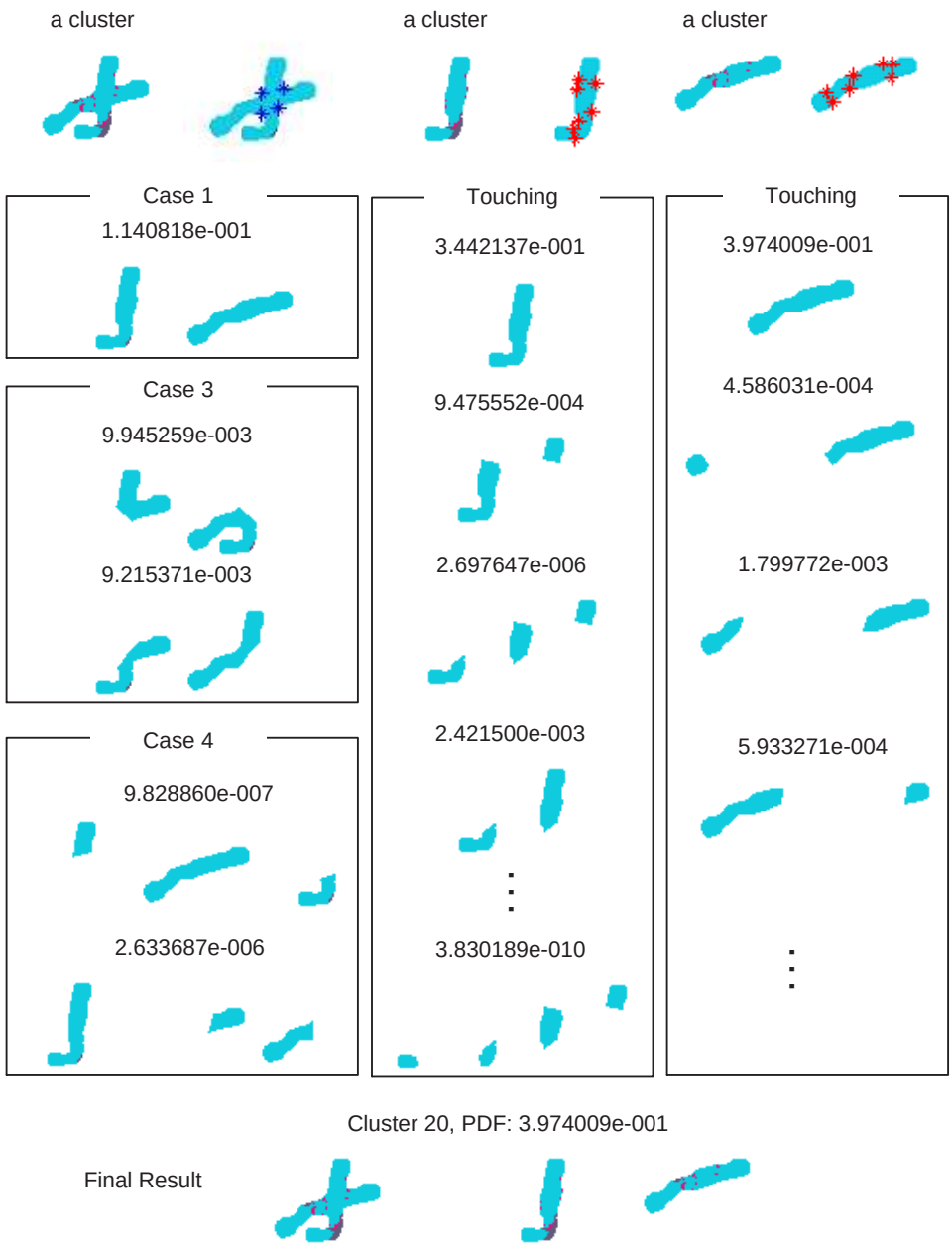


Figure 5.14: Decomposition of a Cross shape cluster. Two chromosomes that belong to the same class crossing each other are successfully decomposed using the developed method.

Figure 5.15: Decomposition of a T-shape cluster. No previous methods could decompose this kind of partial overlaps correctly.

Ncc	NC	N_{WD}	Accuracy [%]
1	428	0	100
2	47	5	89
3	9	1	89
≥ 4	3	1	67

Table 5.4: Decomposition results. Ncc = number of chromosomes in a cluster, NC = number of clusters, and N_{WD} = number of wrong decomposition

After decomposing at all Xp's, the maximum likely hypothesis is chosen as the best decomposition of the cluster.

We have tested our algorithm on 12 images from ADIR's M-FISH image database. A total of 487 clusters were evaluated. Outstanding results were obtained as shown in Table 5.4. Among 487 clusters, most of them were single chromosomes and they were all correctly identified instead of breaking into multiple chromosomes. Among clusters that have 2 or more chromosomes, about 95% was less than three chromosome cases. About 90% of accuracy was obtained for those cases.

Figure 5.17 and Fig. 5.18 show decomposition results of various clusters, and Fig. 5.19 and Fig. 5.20 show an example of automatic foreground-background segmentation, classification, and chromosome decomposition results.

5.5 Conclusion

We have presented a new decomposition method for overlapping and touching M-FISH chromosomes. Previous chromosome decomposition meth-

(a) T and Cross case (b) Landmarks of the cluster

Figure 5.16: Landmarks of a cluster.

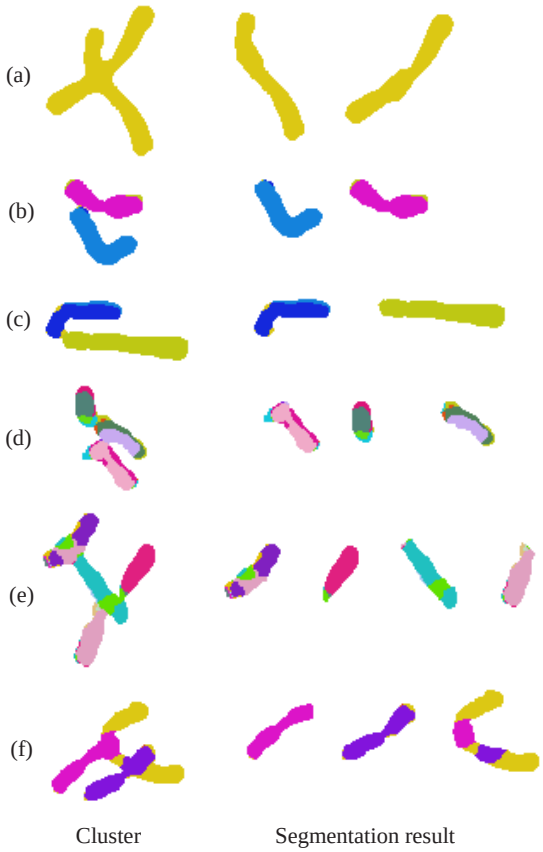
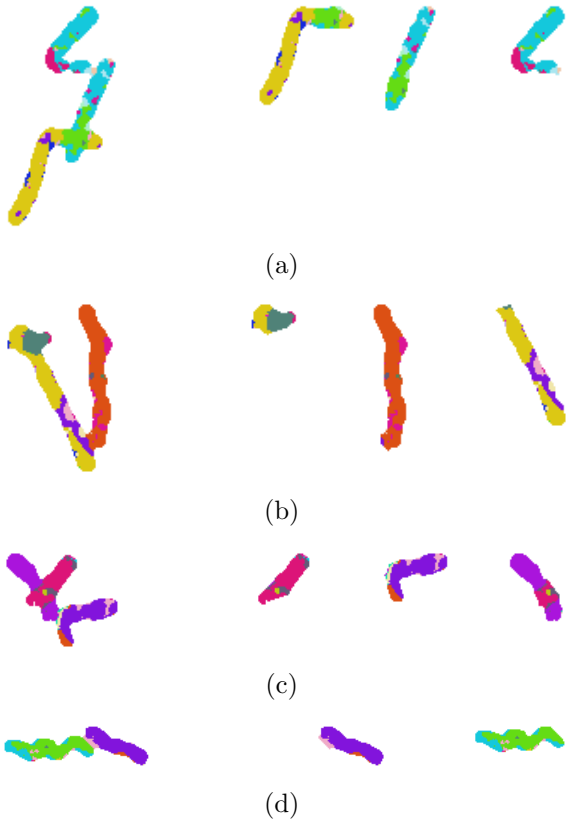


Figure 5.17: Decomposition results. (a) Cross case, (b) T case, (c) I case, (d) T and I case, (e) Cross and T case, and (f) Cross and T case



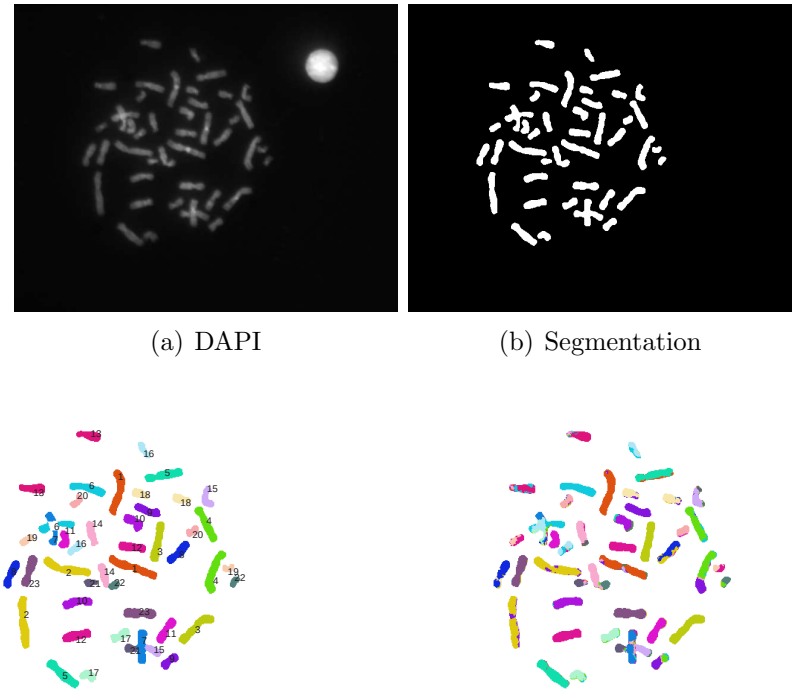
(a)

(b)

(c)

(d)

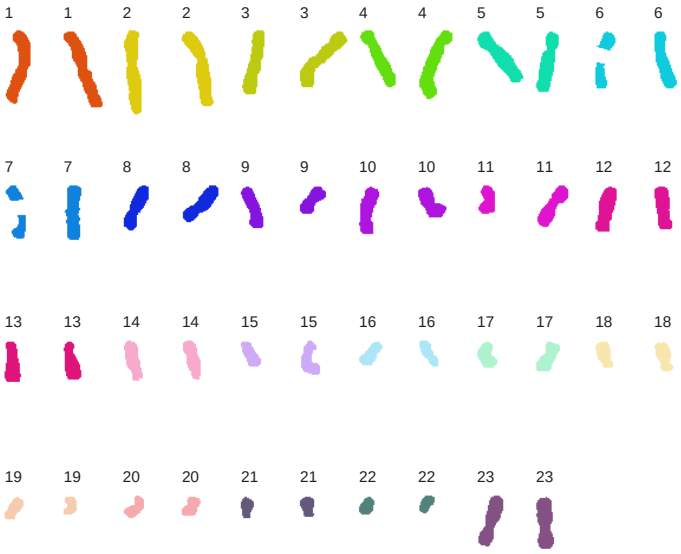
Figure 5.18: More results of chromosome cluster decomposition. The developed decomposition method is robust to misclassification errors.



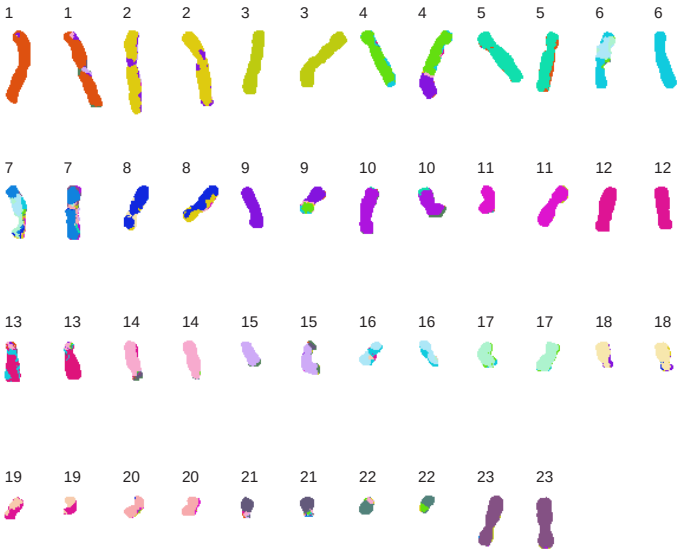
(a) DAPI (b) Segmentation

(c) Ground truth (d) Pixel classification

Figure 5.19: Automatic karyotyping



(a) Karyogram of ground truth



(b) Karyogram

Figure 5.20: Automatic karyotyping, continued from Fig. 5.19. In (b) translocation between 4 and 9 are shown, and overlapping chromosomes are segmented correctly and automatically.

ods utilized partial information of chromosome clusters resulting in limited success. Clusters are better decomposed by incorporating more knowledge. Multiple hypotheses were formed based on color and the geometry defined by the basic elements of a cluster, and then evaluated based on the pixel classification results and chromosome sizes. The hypothesis that has the maximum-likelihood is chosen as the best decomposition of a given cluster. About 90% accuracy was obtained for two and three chromosome clusters, which comprize about 95% of all clusters of two or more chromosomes, and 100% accuracy was obtained for clusters with a single chromosome.

Chapter 6

Conclusion and Future Work

6.1 Conclusions

In this dissertation, methods of improving pixel classification accuracy and automating decomposition of overlapping and touching chromosomes were presented. Using the new feature normalization method, the intra-variance of the feature distribution among different images was reduced, and thus the classification accuracy significantly improved after normalization. A new color compensation method for combinatorially stained FISH images was presented. The color compensation removed the channel crosstalk effectively and improved the image quality significantly. Two new unsupervised nonparametric classification methods for M-FISH images were presented, which are a fuzzy logic classifier and a template matching algorithm. Both methods provide a significant advantage in terms of computation time compared to supervised methods, and their accuracies were comparable to that of a maximum-likelihood classifier. Overlapping and touching chromosomes were effectively decomposed using the developed decomposition method. Given a cluster, a number of hypotheses were formed utilizing the geometry of a cluster, pixel classification results, and chromosome sizes, and a hypothesis that maximized the likelihood function was chosen as the correct decomposition. After chromo-

somes are individually identified, misclassified pixels were effectively corrected while preserving the translocated pixels, using the prior adjusted reclassification method.

6.2 Future Work

Developing a completely automated karyotyping system is near realization. Highly sophisticated and industrially useful algorithms have been developed through this research. Using the developed methods, the amount of human intervention will be significantly reduced: chromosomes are reliably and accurately segmented from the background, pixels are accurately classified, and clusters of overlapping and touching chromosomes are automatically separated. However, there still remains room for improvement.

6.2.1 Pixel classification accuracy

The average self trained and tested accuracy was about 90%, but the average classification accuracy after normalization was about 73%. The accuracies were over 95 % for a set of images that exhibit good quality while bad quality images produced below 30% of accuracy.

Specimen preparation:

This suggests that the most important part of achieving good classification results is to prepare the specimens with a great care. Preparation of specimens should strictly follow the protocols in order to minimize the chemical noise and to ensure the production of high quality signals.

Image quality control:

Image capture should also be performed carefully in order to obtain high quality images. Sharp, in-focus images should be captured for each spectral channel with a proper exposure time that does not saturate the signals. For an automated system, the proper parameters of image capturing process for a microscope can be learned from the good quality images. Images that do not satisfy the quality criteria can be discarded automatically. The quality criteria can include the mean intensity, fitness of a bimodal density function, and variances of each mode in the mixture density function. The DAPI channel should exhibit a bimodal function, representing intensity distributions of background and chromosome pixels. The intensity distribution under the pixels identified as chromosomes should exhibit a bimodal density function in each channel, representing intensity distributions of channel crosstalk and signal from truly hybridized chromosomes.

6.2.2 Automatic chromosome cluster decomposition

While the developed method in this dissertation can decompose many cases of overlapping and touching chromosomes, this part still needs to be improved. As the chromosomes form a cluster, the signal intensities of each chromosome affect the intensity of other chromosomes making the pixel membership less certain. Misclassifications usually occur where chromosomes are close to each other and when they overlap. One can investigate the signal intensities of the presumed misclassified pixels to determine whether the signal

intensity is affected by the nearby chromosomes or due to a true chromosomal aberration. However, there are only 31 possible color combinations in M-FISH when 5 spectra are used, and 24 of them are assigned for chromosomes. Only 7 unused combinations are left to identify a few cases of chromosome overlaps when there are many more overlap cases. As the amount of overlaps and touching increase, the analysis of a particular metaphase spread becomes unreliable or even useless. Therefore, the importance of sample preparation should be emphasized again. It will be ideal if the spreading of chromosomes can be controlled so that chromosomes are spread out with a minimum number of overlaps.

The new developed foreground/background segmentation method for M-FISH images does not produce many touching cases as compared to the technique that Ji [2] used in his touching chromosome segmentation method. However, Ji's method still can be incorporated into our method when evaluating multiple touching cases. The decomposition results will also improve if a shape constraint, such as in [3], is incorporated.

Bibliography

[1] D. G. Harnden, H. P. Klinger, J. T. Jensen, and M. Kaelbling, Eds., *An International System for Human Cytogenetic Nomanclature (1985).* S. Karger, 1985.

[2] J. Liang, "Intelligent splitting in the chromosome domain," *Pattern Recognition*, vol. 22, no. 5, pp. 519–532, 1989.

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

[4] W. Schwartzkopf, "Maximum likelihood techniques for joint segmentation-classification of multi-spectral chromosome images," Ph.D. dissertation, University of Texas at Austin, Department of Electrical and Computer Engineering, Dec. 2002.

[5] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Karyotyping human chromosomes by combinatorial multi-fluor fish," *Nature Genetics*, vol. 12, pp. 368–375, 1996.

[6] E. Schrock, S. du Manoir, T. Veldman, B. Schoell, J. Wienberg, M. A. Ferguson-Smith, Y. Ning, D. H. Ledbetter, I. Bar-Am, D. Soenksen,

Y. Garini, and T. Ried, "Multicolor spectral karyotyping of human chromosomes," *Science*, vol. 273, pp. 494–497, 1996.

[7] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Computer image analysis of combinatorial multi-fluor fish," *Bioimaging*, vol. 4, pp. 52–64, 1996.

[8] 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.

[9] T. Liehr and U. Claussen, "Multicolor-fish approaches for the charaterization of human chromosomes in clinical genetics and tumor cytogenetics," *Current Genomics*, vol. 3, pp. 213–235, 2002.

[10] H. Choi, A. C. Bovik, and K. R. Castleman, "Normalization via expectation maximization and unsupervised nonparametric classification for m-fish chromosome images," *Submitted to, IEEE Transaction on Medical Imaging*, 2006.

[11] H. Choi, K. R. Castleman, and A. C. Bovik, "Color compensation of multi-color fish imags," *Submitted to, IEEE transactions on medical imaging*, 2006.

[12] ——, "Segmentation and fuzzy-logic classification of m-fish chromosome images," *International conference on image processing*, October 2006.

[13] H. Choi, A. C. Bovik, and K. R. Castleman, "Maximum-likelihood decomposition of overlapping and touching m-fish chromosomes using shape, size and color information," *IEEE International Conference of the Engineering in Medicine and Biology Society*, September 2006.

[14] ——, "Automatic karyotyping via maximum-likelihood decomposition of overlapping and touching m-fish chromosomes using geometry, size, and pixel classification," *In preparation*, 2006.

[15] A. J. Vander, J. H. Sherman, and D. S. Luciano, *Human Physiology*, 8th ed. New York: McGraw-Hill, 2001.

[16] H. E. Sutton, *An introduction to human genetics*, 3rd ed. Saunders College, 1980.

[17] T. Caspersson, K. R. Castleman, G. Lomakka, E. S. Modest, A. Moller, R. Nathan, R. J. Wall, and L. Zech, "Automated karyotyping of quincrine mustard stained human chromosomes," *Experimental Cell Research*, vol. 67, pp. 233–235, 1971.

[18] P. Lichter, "Multicolor fishing: what's the catch?" *Trends Genet.*, vol. 13, pp. 475–479, 1997.

[19] K. R. Castleman, "Match recognition in chromosome band structure," *Biomed. Sci. Instrum.*, vol. 4, pp. 256–264, 1968.

[20] K. Paton, "Automatic chromosome identification by the maximum-likelihood method," *Annals of Human Genetics*, vol. 33, pp. 177–184, 1969.

[21] R. S. Ledley, H. A. Lubs, and F. H. Ruddle, "Introduction to chromosome analysis," *Computers in Biology and Medicine*, vol. 2, pp. 107–128, 1972.

[22] C. Lundsteen and J. Piper, *Automation of Cytogenetics*. Berlin: Springer-Verlag, 1989.

[23] M. Tso, P. Kleinschmidt, I. Mitterreiter, and J. Graham, "An efficient transportation algorithm for automatic chromosome karyotyping," *Pattern Recognition Letters*, vol. 12, pp. 117–126, 1991.

[24] A. Carothers and J. Piper, "Computer-aided classification of human chromosomes: a review," *Statistics and Computing*, vol. 4, pp. 161–171, 1994.

[25] Q. Wu and K. R. Castleman, "Wavelet-based enhancement of human chromosome images," *Proceedings of the 20th Annual International Conference of the IREEE in Medicine and Biology Society*, vol. 20, no. 2, pp. 963–966, 1998.

[26] Y.-P. Wang, Q. Wu, K. R. Castleman, and Z. Xiong, "Chromosome image enhancement using multiscale differential operators," *IEEE Transactions on Medical Imaging*, vol. 22, no. 5, pp. 685–693, May 2003.

[27] L. Vanderheydt, F. Dom, A. Oosterlinck, and H. C. D. Berghe, "Two dimensional shape decomposition using fuzzy subset theory applied to automated chromosome analysis," *Pattern Recognition*, vol. 13, no. 2, pp. 147–157, 1981.

[28] J. Piper and E. Granum, "On fully automatic feature measurement for banded chromosome classification," *Cytometry*, vol. 10, pp. 242–255, 1989.

[29] F. C. A. Groen, T. K. Kate, A. W. M. Smeulders, and I. T. Young, "Human chromosome classification based on local band descriptors," *Pattern Recognition Letters*, vol. 9, pp. 211–222, 1989.

[30] A. J. Dennis, K. S. Tang, and S. Zimmerman, "Band features as classification measures for g-banded chromosome analysis," *Comput. Biol. Med.*, vol. 23, no. 2, pp. 115–129, 1993.

[31] Q. Wu and K. R. Castleman, "Automated chromosome classification using wavelet-based band pattern descriptors," *13th IEEE Symposium on Computer-Based Medical Systems*, pp. 189–194, June 2000.

[32] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "Medial axis transform based features and a neural network for human chromosome classification," *Pattern Recognition*, vol. 28, no. 11, pp. 1673–1683, 1995.

[33] A. Sarkar, M. Biswas, B. Kartikeyan, V. Kumar, K. Majumder, and D. Pal, "A mrf model-based segmentation approach to classification for multispectral imagery," *Ieee Transactions On Geoscience And Remote Sensing*, vol. 40, no. 5, pp. 1102–1113, May 2002.

[34] C. Mao, S. Liu, and J. Lin, "Classification of multispectral images through a rough-fuzzy neural network," *Optical Engineering*, vol. 43, no. 1, pp.

103–112, Jan 2004.

[35] J.-S. Lin and S.-H. Liu, "Classification of multispectral images based on a fuzzy-possibilistic neural network," *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 32, no. 4, pp. 499–506, Nov 2002.

[36] W. Reddick, J. Glass, E. Cook, T. Elkin, and R. Deaton, "Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 911–918, Dec 1997.

[37] A. Kumar, S. Basu, and K. Majumdar, "Robust classification of multispectral data using multiple neural networks and fuzzy integral," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 3, pp. 787–790, May 1997.

[38] W. Reddick, J. Glass, E. Cook, T. Elkin, and R. Deaton, "Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 911–918, Dec 1997.

[39] S. Lee and M. M. Crawford, "Multi-channel/multi-sensor image classification using hierarchical clustering and fuzzy classification," *IEEE*, 2000.

[40] D. Koechner, J. Rasure, R. Griffey, and T. Sauer, "Clustering and classification of multispectral magnetic resonance images," *Proceedings of Third*

Annual IEEE Symposium on Computer-Based Medical Systems, pp. 32–37, June 1990.

[41] P. Mitra, B. Shankar, and S. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, Jul 2004.

[42] J. Harsanyi and C. I. Chagn, "Hyperspectral image classification and dimensionality redution: An orthogonal subspace projection approach," *IEEE Transtions on geoscience and remote sensing*, vol. 32, pp. 779–785, May 1994.

[43] C. I. Chang and C. M. Brumbley, "A kalman filtering approach to multispectral image classification and detection fo changes in signature abundance," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 37, no. 1, pp. 257–268, January 1999.

[44] A. Shackelford and C. Davis, "A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 10, pp. 2354–2363, Oct 2003.

[45] P. P. Raghu and B. Yegnanarayana, "Multispectral image classification using gabor filters and stochastic relaxation neural network," *Neural Networks*, vol. 10, no. 3, pp. 561–572, 1997.

[46] R. Eils, S. Uhrig, K. Saracoglu, K. Sätzler, A. Bolzer, I. Petersen, J. M. Chassery, M. Ganser, and M. R. Speicher, "An optimized, fully automated system for fast and accurate identification of chromosomal rearrangements by multiplex-fish (m-fish)," *Cytogenetics and Cell Genetics*, vol. 82, pp. 160–171, 1998.

[47] F. A. Merchant, K. N. Good, H. Choi, and K. R. Castleman, "Automated detection of chromosomal rearrangements in multicolor fluorescence in-situ hybridizatioin images," *Proceedings of the Second Joint Conference of the IEEE EMBS and BMES Conference*, vol. 2, pp. 1074 – 1075, Oct 2002.

[48] M. P. Sampat, K. R. Castleman, and A. C. Bovik, "Pixel-by-pixel classification of m-fish images," *Proceedings of the Second Joint Conference of the IEEE EMBS and BMES Conference*, vol. 2, pp. 999 – 1000, Oct 2002.

[49] H. Choi, K. R. Castleman, and A. C. Bovik, "Joint segmentation and classification of m-fish chromosome images," *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, 2004.

[50] Y. Wang and K. R. Castleman, "Normalization of multicolor fluorescence in situ hybridization (m-fish) images for improving color karyotyping," *Cytometry Part A*, vol. 64A, no. 2, pp. 101–109, 2005.

[51] K. R. Castleman, *Digital Image Processing.* Upper Saddle River, NJ: Prentice-Hall, 1996.

[52] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, October 2003.

[53] J. P. W. Pluim and J. M. Fitzpatrick, "Image registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 11, November 2003.

[54] M. Kozubek and P. Matula, "An efficient algorithm for measurement and correction of chromatic aberrations in fluorescence microscopy," *Journal of Microscopy*, vol. 200, pp. 206–217, 2000.

[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactios on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[56] G. H. Golub and C. F. V. Loan, *Matrix computations*, 3rd ed. Baltimore: Johns Hopkins University Press, 1996.

[57] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Second ed. San Diego: Harcourt Brace Jovanovich, November 2000.

[58] 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, Aug. 2005.

Vita

Hyo Hun (Hyohoon) Choi was born in Korean on March 1 1973. He received the Bachelor of Science in Engineering in Electrical Engineering from the Kangwon National Univerisity, Kangwon, Korea in 1999. He joined the Biomedical Engineering Program at the University of Texas at Austin in Fall 1999 and obtained the Master of Science degree in May 2001 under the supervision of Prof. Jonathan W. Valvano. He worked as a research engineer at Advanced Digital Imaging Research, League city, Texas for a year. He joined the Biomedical Engineering Department at the Univeristy of Texas at Austin in Fall 2002 and worked since then as a graduate research assistant at the Laboratory for Image and Video Engineering (LIVE) under the supervision of Prof. Alan C. Bovik. During the summers of 2003, 2004, 2005, and 2006, he worked as a summer intern at Advanced Digital Imaging Research under the supervision of Dr. Kenneth R. Castleman.

His current research interests include pattern recognition, image and signal processing, computer vision, multi-dimensional data visualization, diagnostic system development, biometrics, and bioinformatics.

Permanent address: 265 3/5 Jangsadong, Sockchoshi, Kangwondo,
217-130, Republic of Korea

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.