# MAXIMUM-LIKELIHOOD DECOMPOSITION OF OVERLAPPING AND TOUCHING M-FISH CHROMOSOMES USING GEOMETRY, SIZE AND COLOR INFORMATION

*Hyohoon Choi*∗*, Alan C. Bovik*†*, Kenneth R. Castleman*‡

∗ Department of Biomedical Engineering, University of Texas, Austin, Texas
† Department of Electrical and Computer Engineering, University of Texas, Austin, Texas
‡ Advanced Digital Imaging Research, LLC., 2450 South Shore Blvd., Suite 305, League City, Texas
Email: ∗ hyohoon_choi@mail.utexas.edu, † bovik@ece.utexas.edu, ‡ castleman@adires.com

## ABSTRACT

Since the birth of chromosome analysis by the aid of computers, building a fully automated chromosome analysis system has been the ultimate goal. Along with many other challenges, automating chromosome classification and segmentation has been one of the major challenges especially due to overlapping and touching chromosomes. In this paper we present a novel decomposition method for overlapping and touching chromosomes in M-FISH images. To overcome the limited success of previous decomposition methods that use partial information about a chromosome cluster, we have incorporated more knowledge about the clusters into a maximum-likelihood frame work. The proposed method evaluates multiple hypotheses based on geometric information, pixel classification results, and chromosome sizes, and a hypothesis that has a maximum-likelihood is chosen as the best decomposition of a given cluster. About 90% of accuracy was obtained for two or three chromosome clusters, which consist about 95% of all clusters with two or more chromosomes.

## I. INTRODUCTION

Multicolor fluorescence *in-situ* hybridization is a combinatorial labeling technique developed for the analysis of human chromosomes [1], [2]. To be able to distinguish 24 human chromosomes (22 somatic chromosomes and X and Y sex chromosomes), a minimum of 5 fluorophores are used. Each chromosome is stained with a unique combination of fluorophores so that every chromosome is uniquely identified. An extra fluorophore, DAPI (4'6-diamidino-2-phenyl indole dihydrocloride), is counterstained to all chromosomes. M-FISH provides color karyotyping (visualization of chromosomes in a specific format) by assigning a pseudocolor to each pixel based on the spectral combination, and thus allows simultaneous analysis of numerical and structural abnormalities of whole human chromosomes.

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. There have been numerous segmentation (decomposition) methods developed to automate the analysis process in conventional banded chromosome images. Among them, some methods only handle touching cases and some handle both cases with a limited success. Most of the methods utilize only the geometry information of chromosome clusters such as curvature, skeleton, and convex hulls [3], [4]. 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 that 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. Schwartzkopf et al. [5] 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 are achieved for touching cases and less reliable results for overlapping cases. When only the colors are used, touchings or overlaps of the same kinds of chromosomes can not be segmented, and the segmentation accuracy heavily relies on the initial pixel classification accuracy. Thus the both information, geometry and pixel classification results, have to be merged in order to achieve better segmentation results.

In this paper, 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 hyptheses are evaluated and the most likely

hypothesis is chosen as the correct decomposition.

## II. METHODS

### II-A. Foreground-background segmentation and pixel classification

M-FISH images have six channels. Each channel contains the intensity of a corresponding fluorophore. Since each chromosome is uniquely stained, an intensity combination across 6 channels is unique for each chromosome. Chromosomes are reliably segmented automatically from the background by utilizing the spectral, signal intensity, and edge information. 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. 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. Laplacian of Gaussian (LoG) edge detection 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. *k*-means clustering, adaptive thresholding, LoG edge detection, and a global thresholding methods are combined to achieve a final segmentation result. A composite threshold image is obtained after voting among those 4 methods. For example, a pixel becomes foreground when a majority (3 out of 4) are foreground.

After the chromosome segmentation, only the chromosome pixels are classified using an unsupervised classification method called fuzzy-logic classifier [6].

### II-B. Elements of clusters

We define a group of connected pixels as a cluster. Thus, a cluster may 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. 1)

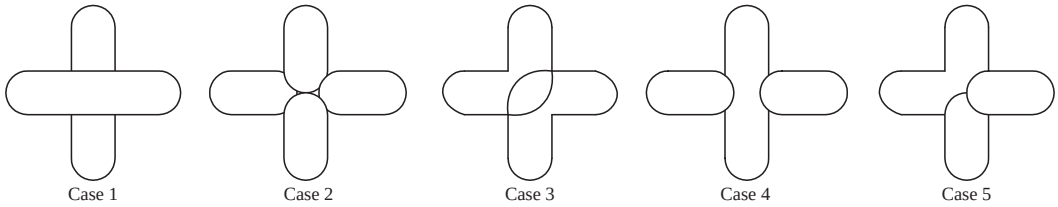
1) Cross shape cluster
2) T shape cluster
3) I shape cluster

Most of clusters are formed by combinations of 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. 2. There exists an $Xp$ that is connected to an $Ep$, and an $Ep$ connects two boundary segments $b$. Given $\{Ep,Xp\}$ and two $b$s, 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 $Xp$ has three or four cut points. Once all the landmarks are found, all possible decompositions are evaluated.

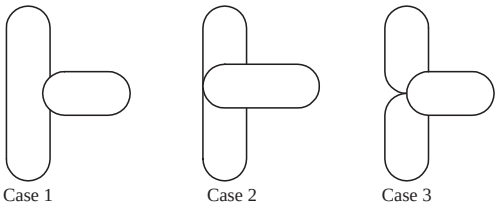
The cross shape cluster has 5 cases: case 1 is an overlap of two chromosomes, case 2 is a touching of four chromosomes, case 3 is a touching of two chromosomes, and case 4 and 5 are touchings of three chromosomes (see Fig. 1). Two chromosomes are found by connecting $\{Cp1-Cp4, Cp2-Cp3\}$ and $\{Cp1-Cp2, Cp4-Cp3\}$ 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 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 subcases), and case 3 is a touching of three chromosomes. In total there are seven hypotheses to evaluate in T case.

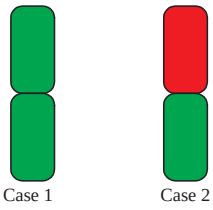
We define a cluster that does not have a cross point as I shape cluster which may have touchings of the same chromosomes or different chromosomes. The I shape cluster has an arbitrary number of cases. The number of segments are determined by the number of concave points on the boundary. There are two 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. 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 chromosome, segments 1 and 2 form a chromosome and segment 3 form another chromosome, and so on. For 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, $2^{M-1}$ combinations are evaluated. Thus, a total of $(2^N + 2^M)/2$ hypotheses are evaluated for an I shape cluster.




(a) Cross shape clusters

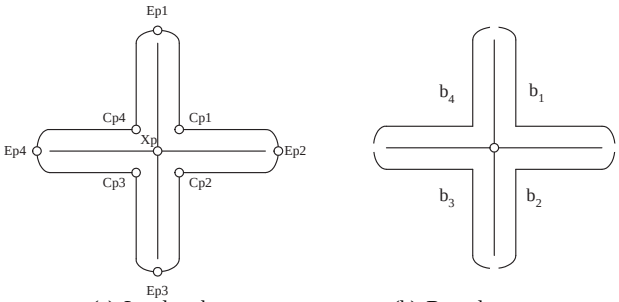



(b) T shape clusters




(c) I shape clusters

**Fig. 1**. Elements of clusters




(a) Landmarks (b) Boundary segments

**Fig. 2**. Landmarks of a cluster - Definition

### II-C. 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 likelihood of a hypothesis is calculated by

$$p_h = \prod_{i=1}^{N_c} p(s_i|\omega_i)P(\omega_i) \qquad (1)$$

where,

$N_c$ = number of chromosomes in a hypothesis
$P(\omega_i)$ = $\frac{N_i}{N_{ci}}$
$\omega_i$ = most popular class in chromosome $i$
$N_i$ = number of pixels belong to $\omega_i$ in chromosome $i$
$N_{ci}$ = 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
$p(s_i|\omega_i)$ = class-conditional probability density function for chromosome size $s_i$ given class $\omega_i$

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

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

The class parameters, $\mu_i$ and $\sigma_i$ ($1 \leq i \leq 24$), are calculated from Advanced Digital Imaging Research's M-FISH image database (available at `http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml`).

$p(s_i|\omega_i)P(\omega_i)$ in eq. 1 is the posterior probability function without the normalization factor in Bayes rule. $P(\omega_i)$ acts as priors for $\omega_i$, and its value is high when most of pixels in chromosome $i$ are classified as $\omega_i$ and vice versa. The posterior probability function gives the likelihood that an unknown chromosome $i$ belonging to $\omega_i$ by its size and classification results. The total likelihood of a hypothesis is a product of the posterior probabilities. Among all hypotheses, the one that has the maximum likelihood is chosen as the correct decomposition of the cluster.

## III. RESULTS

Chromosomes are first segmented from the background using the segmentation method explained in Section II-A. Then chromosome pixels are classified using a fuzzy-logic classifier [6]. Given a cluster, the landmarks on the boundary and skeleton are computed as shown in Fig 3, and the cluster is decomposed into multiple hypotheses and the likelihood of each hypothesis is computed by eq. 1. When there are multiple $Xp$s, hypotheses are evaluated at each $Xp$ consecutively. After decomposing at all $Xp$s, the maximum likely hypothesis is chosen as the best decomposition of the cluster.

| $Ncc$ | $NC$ | $N_{WD}$ | Accuracy [%] |
|---|---|---|---|
| 1 | 428 | 0 | 100 |
| 2 | 47 | 5 | 89 |
| 3 | 9 | 1 | 89 |
| $\geq 4$ | 3 | 1 | 67 |

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

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

**Fig. 3**. Landmarks of a cluster - Real case

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 I. 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.

Fig. 4 shows decomposition results of various clusters.

## IV. CONCLUSION

We have presented a new decomposition method for overlapping and touching M-FISH chromosomes. Previous chromosome decomposition methods utilized partial information of chromosome clusters resulting in a limited success. A cluster was 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. A hypothesis that has a maximum-likelihood is chosen as the best decomposition of a given cluster. About 90% of accuracy was obtained for two or three chromosome clusters, which consist about 95% of all clusters with two or more chromosomes, and 100% accuracy was obtained for clusters with a single chromosome.

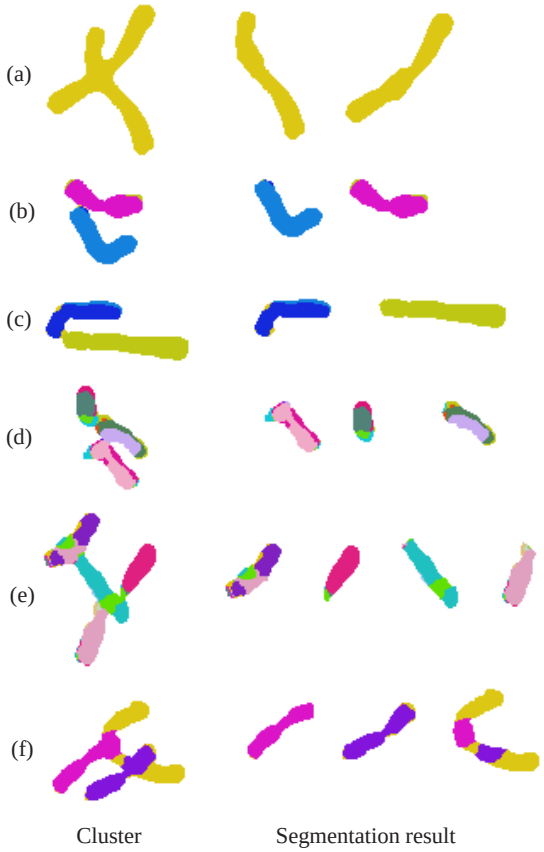



**Fig. 4**. Segmentation results. (a) Cross case, (b) T case, (c) I case, (d) T and I case, (e) Cross and T case, and (f) Cross and T case

## V. REFERENCES


[1] 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.

[2] 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.

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

[5] 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.

[6] H. Choi, K. R. Castleman, and A. C. Bovik, "Segmentation and fuzzy-logic classification of m-fish chromosome images," *International conference on image processing*, October 2006.