MINIMUM ENTROPY SEGMENTATION APPLIED TO MULTI-SPECTRAL CHROMOSOME IMAGES

Wade Schwartzkopf, Brian L. Evans, and Alan C. Bovik

Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX 78712-1084
{wade,bevans,bovik}@ece.utexas.edu

ABSTRACT

In the early 1990s, the state-of-the-art in commercial chromosome image acquisition was grayscale. Automated chromosome classification was based on the grayscale image and boundary information obtained during segmentation. Multi-spectral image acquisition was developed in 1990 and commercialized in the mid-1990s. One acquisition method, multiplex fluorescence in-situ hybridization (M-FISH), uses five color dyes. We propose a segmentation algorithm for M-FISH images that minimizes the entropy of classified pixels within possible chromosomes. This method is shown to correctly decompose even difficult clusters of touching and overlapping chromosomes. Finally, an example image is given to illustrate the algorithm.

1. INTRODUCTION

Chromosomes are the cell structures that contain genetic information. When chromosomes are photographed, the images contain much information about the health of an individual. The images are useful for diagnosing disorders and studying various diseases. In the past, it has been necessary for laboratory technicians to examine these images visually to collect the useful information contained in these images. However, since many images often have to be inspected and since visual inspection is time consuming and expensive, many attempts have been made to automate chromosome image analysis. For example, automated segmentation algorithms for grayscale chromosome images have been able to correctly decompose about 80-90% of touching and overlapping chromosomes [1, 2, 3]. These automated procedures rely on chromosome shape and texture.

In the 1990's, new techniques were developed to dye chromosomes with multiple colors so that each chromosome class appears to be a distinct color. This makes analysis of chromosome images easier, not only for human inspection, but also for computer analysis. This research focuses on one such dying technique, known as M-FISH (multiplex fluorescence in-situ hybridization). This work takes advantage of the multi-spectral information in M-FISH images to improve past methods of computer segmentation and analysis of chromosome images.

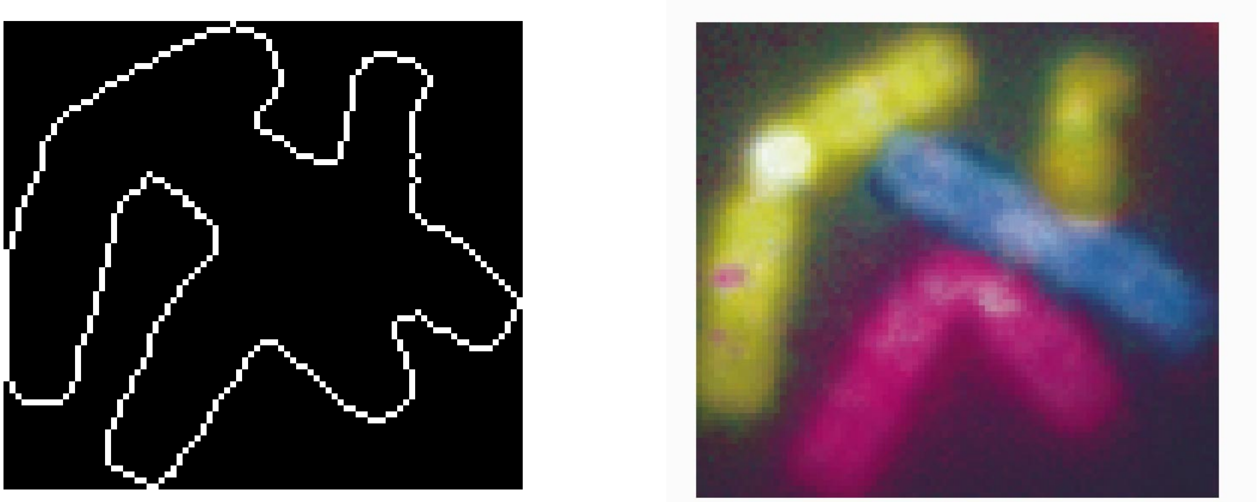
2. MULTI-SPECTRAL IMAGES

A new way to image chromosomes came about with the invention of chromosome painting [4], combinatorial labeling [5] and ratio labeling [6]. These techniques made use of fluorophores (dyes) that attach to a single class of chromosomes, parts of chromosomes, or specific sequences of DNA. Using these techniques, one could create a combination of fluorophores such that each class of chromosomes absorbed a different combination of these fluorophores [7, 8, 9]. Therefore, each chromosome class would appear to be a different color and would be visually distinguishable from all of the other classes.

An image of each fluorophore can be obtained by employing appropriate optical filters. This way, each pixel could be represented as a vector, where each element represents the intensity of the response to one fluorophore. Instead of obtaining a grayscale image by traditional chromosome imaging techniques, such as Giemsa banding [10], a multi-spectral image could now obtained in which the spectral composition at each point reveals the combination of fluorophores and, thus, the chromosome class of the matter at that point. Using this combinatorial labeling, known as M-FISH, one can determine the chromosome class at every pixel.

Such an imaging technique has a couple of obvious advantages. First, the task of chromosome classification is greatly simplified. Instead of determining and then comparing the chromosome lengths, centromere positions, and banding patterns, one only has to look at the spectral information within that chromosome. The second advantage is that it is possible to detect smaller translocations and rearrangements than were discernible with grayscale chromosome banding patterns only [11].

B. L. Evans was supported by a US National Science Foundation CAREER Award under grant MIP-9702707. W. Schwartzkopf was supported by a grant from Advanced Digital Imaging Research (http://www.adires.com).



a) Boundary of cluster

b) Multi-spectral information in cluster

Figure 1: Comparison of two types of cluster information

3. CHROMOSOME SEGMENTATION

Many previous chromosome segmentation methods [1] begin with thresholding, or adaptive thresholding, followed by labeling of connected components. Then these methods examine the connected components to determine which of them are single chromosomes and which are clusters of multiple chromosomes. Finally, the clusters are divided by choosing cut points on the boundary of the cluster, which are the points at which the boundaries of two different chromosomes meet. For the case of two touching chromosomes, two points must be found, which denote a line that separates the two chromosomes. For the case of two overlapping chromosomes, four cut points must be found in order to create a polygon that defines the area of overlap. Once the proper cut points are discovered, the touching or overlapping chromosomes can then be decomposed by straight cut lines between the points [3] or best-fit cubic curves [2].

Traditional chromosome segmentation methods use shape information from the boundary of the chromosomes to detect and decompose clusters. Cut points are then found by examining the shape of boundary of the cluster [2, 3, 12]. Occasionally, grayscale information from inside the chromosome clusters is also used. One popular method is “valley searching” [13] where a minimum cost algorithm attempts to locate low gray-value valleys running through the cluster to locate separation between the chromosomes.

With M-FISH images, a new source of information is available for segmentation. If one observes the example in Figure 1, it is not immediately clear, by looking only at the boundary of the cluster, what the proper segmentation of the cluster is. It is not apparent, even to many trained observers, whether there is an overlap involved or even how many chromosomes are included in this cluster. However, the M-FISH multi-spectral information makes it more clear what the proper segmentation should be since each chromosome has its own color.

To use the multi-spectral information available in M-FISH, it is necessary to introduce an objective function that uses this multi-spectral information to evaluate possible cut lines. It would then be possible to examine all possible cut lines, or some reasonable subset of them, and pick the pair that maximizes (or minimizes) this function.

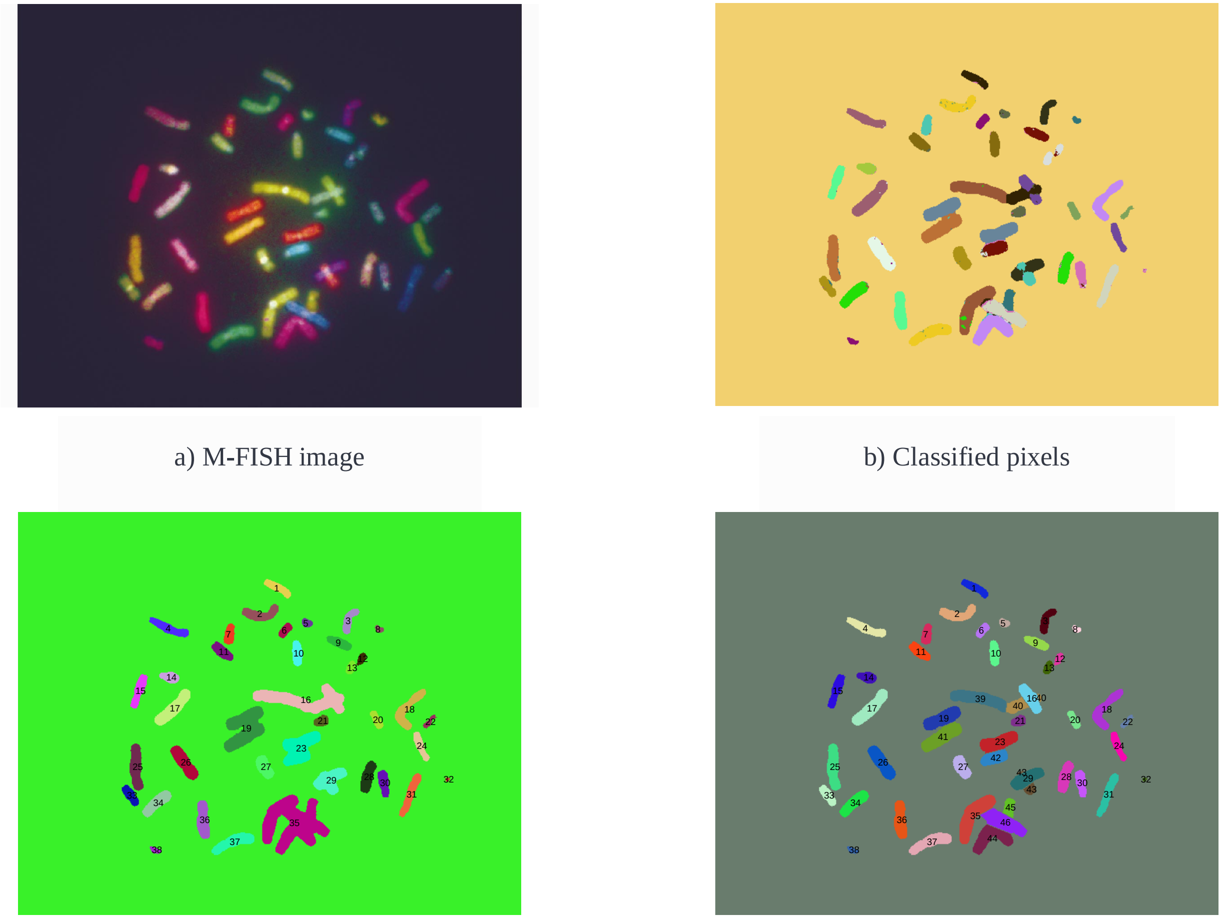
For the objective function, we use a measure of entropy. In particular, we use Shannon’s definition of entropy [14]

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

where n is the number of possible classes. In a perfectly classified and segmented image, the entropy of each segment will be zero, since all the pixels in each segment will be classified into the same class. The larger the number of different classes that are found within a segment, the higher the entropy will be.

The probabilities of the classes are calculated empirically. That is, the probability of any class is the number of pixels of that class in an object divided by the number of pixels in that object.

The following algorithm (outlined in Fig. 3), which segments chromosomes using M-FISH multi-spectral data and entropy, was proposed and tested.



a) M-FISH image

b) Classified pixels

c) Connected components

d) Entropy-segmented image

Figure 2: Example of an M-FISH image segmented with entropy segmentation

4. MINIMUM ENTROPY ALGORITHM

We locate the chromosome material (via thresholding or some other method) an image of a sixth dye, DAPI (4,6-diamidino-2-phenylindole nucleic acid stain), which binds to all chromosomes, and then find the connected components. Next, we label every connected component as a separate object and calculate its entropy. If its entropy is above a given threshold, then we examine that object for possible cut lines.

Initially all boundary points of the object are taken to be possible cut points, where every combination of two boundary points makes a cut line. However, this is a large number of combinations that can be narrowed down. For instance, in order to preserve connectivity, all points in a valid cut line must be contained within the cluster. Furthermore, we can prudently remove a number of other points. Some have suggested using only points of high curvature along the boundary of the cluster [3].

We consider only points whose cut line was contained within the chromosome clusters and had an 8-connected neighbor whose class differed from theirs. We also require that the cut line not result in a division whose size was smaller than a certain threshold. This avoids the removal of single pixels, whose entropy is always zero, as well as other small sections of noise along the boundary.

Candidate cut lines are made by straight lines between all combinations of candidate cut points. For each candidate cut line, we calculate the entropy of the resulting division. If the sum of the resulting entropies is less than the entropy of the original cluster, then the cut line is considered valid. We choose the valid cut line that results in the least total entropy. If no valid cut lines exist, then the object is not divided. If a valid cut line is found, we remove the points in the optimal cut line from the object so that the two divisions are no longer connected. These two divisions are then considered two new objects. We then examine the two new objects and entropies to see if these new objects need to be further divided.

```
1. Separate DAPI image background/foreground.
2. Find connected components of foreground image.
3. Classify pixels in each connected component with
   multi-spectral information.
4. For each connected component,
   A. Find the cut line that gives the lowest entropy
      as measured by (1). If dividing decreases the
      total entropy of the object, then split object.
   B. Continue dividing until no more entropy-
      reducing cut lines can be found.
   C. Match resulting segments within the cluster
      pairwise. If combining the pair decreases the
      total entropy, then recombine segments.
   D. Continue recombining pairs until no entropy-
      reducing segments can be found.
```

Figure 3: Minimum entropy algorithm

Once all entropy-reducing divisions are made, we recombine objects that may have been labeled as separate objects due to overlap. To do this, we combine all objects within a cluster in pairs. If the entropy of any pair is greater than the sum of the entropies of the two individual objects, we label the combination as valid. We choose the combination that yields the lowest entropy. After these two objects are combined, we examine the cluster again, using the same method to look for other objects in it that need to be combined. If no valid combination is found, the algorithm stops and moves to the next cluster.

5. EXAMPLE

Fig. 2 segments an M-FISH image using the entropy method. The original chromosome image is in Fig. 2a. Fig. 2b classifies each pixel in Fig. 2a where color represents the class. Pixel classification was done with Applied Imaging's Powergene M-FISH software [15]. Fig. 2c shows the connected components of the image after thresholding. In this image, several groups of chromosomes (those labeled 16, 19, 23, 29, and 35) are labeled as a single object. Fig. 2d shows the entropy-segmented chromosomes. All touching chromosomes were correctly split. The two overlapped chromosomes (40 and 43) were correctly identified. In these chromosomes, both ends are labeled as one chromosome.

6. CONCLUSION

This paper introduces entropy as a criterion for selecting cut lines to decompose groups of chromosomes that touch and overlap each other. This algorithm uses multi-spectral information in chromosome images for more accurate segmentation. This algorithm was able to decompose clusters of touching and overlapping chromosomes. Furthermore, this algorithm may be applied to other types of multi-spectral images if these images contain touching or overlapping objects with different spectral signatures.

C code for this algorithm is available at

http://www.ece.utexas.edu/~wade/mfish

A public database of 200 hand-segmented M-FISH images is available from Advanced Digital Imaging Research at

http://www.adires.com/projects/mfish_db.shtml

This database contains 200 hand-segmented M-FISH images, or approximately 9000 individual chromosomes.

7. REFERENCES

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

[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] D. Pinkel, T. Straume, and J. W. Gray, “Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization,” *Proc. National Academy of Sciences of the United States of America*, vol. 83, pp. 2934-2938, 1986.

[5] P. M. Nederlof, S. van der Flier, J. Wiegant, A. K. Raap, H. J. Tanke, J. S. Ploem, and M. van der Ploeg, “Multiple fluorescence in situ hybridization,” *Cytometry*, vol. 11, pp. 126-131, 1990.

[6] P. M. Nederlof, S. van der Flier, J. Vrolijk, H. J. Tanke, and A. K. Raap, “Fluorescence Ratio Measurements of Double-labeled Probes for Multiple in Situ Hybridization by Digital Imaging Microscopy,” *Cytometry*, vol. 13, pp. 839-845, 1992.

[7] 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.

[8] M. M. Le Beau, “One FISH, two FISH, red FISH, blue FISH,” *Nature Genetics*, vol. 12, pp. 341-344, 1996.

[9] T. Ried, A. Baldini, T. C. Rand, and D. C. Ward, “Simultaneous visualization of seven different DNA probes by in situ hybridization using combinatorial fluorescence and digital imaging microscopy,” *Proc. National Academy of Sciences of the United States of America*, vol. 89, pp. 1388-1392, 1992.

[10] M. Seabright, “A rapid banding technique for human chromosomes,” *Lancet* ii, pp. 971-2, 1971.

[11] T. Veldman, C. Vignon, E. Schröck, J. D. Rowley, and T. Ried, “Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping,” *Nature Genetics*, vol. 15, pp. 406-410, 1997.

[12] Q. Wu, *Automated Identification of Human Chromosomes as an Exercise in Building Intelligent Image Recognition Systems*, Catholic University of Leuven, Leuven, Belgium, Doctoral Dissertation, 1987.

[13] A. M. Vossepoel, *Analysis of Image Segmentation for Automated Chromosome Identification*, University of Leiden, Leiden, Netherlands, Doctoral Dissertation, 1987.

[14] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Sys. Tech. Journal*, vol. 27, pp. 379-423, 623-656, 1948.

[15] http://www.aicorp.com