Connected Components Labeling and Extraction Based Interphase Removal from Chromosome Images

Sivaramakrishnan Rajaraman[1] and Arun Chokkalingam[2]

[1]*Department of Biomedical Engineering,*
SSN College Of Engineering, Tamil Nadu, India
[2]*Department of Electronics and Communication Engineering, RMKCET,*
Tamil Nadu, India
sivaramakrishnanr@ssn.edu.in,carunece@gmail.com

Abstract

This paper proposes a novel method of connected component labeling and extraction that segments and removes dirt and interphase cells from the chromosome images. This technique was tested with a standard clinical database of human karyotype and excellent segmentation results were achieved. The process involves identifying the various connected components in the input image and assigning labels to create a Label Matrix using the Color Map for these connected components. The connected components or objects that have fewer than the predefined amount of pixels are removed from the image that produces another image where the chromosome except the dirt, stain and interphase cells are uniquely identified and removed. The segmented image is subtracted from the original image, leaving behind only the chromosomes with no dirt, stain and interphase cells, facilitating accurate karyotyping procedures. This technique is extremely helpful when the unwanted object to be segmented and removed shares common intensity levels with the desired information where traditional threshold based procedure will fail to accomplish precise segmentation results.

Keywords: *karyotype, chromosomes, interphase cells, connected components, connected component labeling and extraction, Label Matrix, Color Map, segmentation*

1. Introduction

Human body cells contain 22 pairs of chromosomes called Autosomes and 2 sex chromosomes, X and Y, for the female and male respectively. The pictorial arrangement of these chromosomes present in a cell in accordance to the International System for Cytogenetic Nomenclature (ISCN) [3] is called a Karyotype. Identifying individual chromosomes and grouping them precisely is of extreme importance for the geneticists that help in identifying and resolving genetic abnormalities.

Karyotyping refers to the process of identifying and arranging these chromosomes, aiding genetic diagnosis [2]. A karyotype is the resulting display of ordered pair of chromosomes as shown in Figure 1.

Usually photomicrographs contain unnecessary objects including dirt, stain, and an undivided mass of condensed chromosomes called interphase cells in addition to the chromosomes of interest. The presence of these limiting factors in a photomicrograph poses serious threat at accurate classification of chromosomes. Interphase cells share common intensity levels with the chromosomes and traditional threshold techniques fail to remove these interphase cells where a vital part of the chromosomes of interest are also lost. An algorithm based on Connected Components Labeling and Extraction [6] is proposed herewith that segments and removes dirt and interphase cells from the chromosome images, facilitating accurate, automatic karyotyping procedures. The Connected Components or objects having fewer than the predefined amount of pixels, arrived by a Trial and Error Method, are removed from the image, resulting in another image where the chromosomes except the interphase cells are uniquely identified and removed. The segmented image is subtracted from the original image, leaving behind the chromosomes of interest without any dirt and interphase cells, facilitating accurate karyotyping procedures [5]. Karyotyping refers to the process of identifying and arranging these chromosomes, aiding genetic diagnosis [2]. A Karyotype is the resulting display of ordered pair of chromosomes as shown in Figure 2.

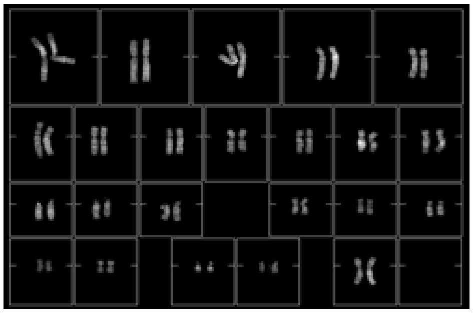


Figure 1. Human Karyotype

Usually photomicrographs contain unnecessary objects including dirt, stain, and an undivided mass of condensed chromosomes called Interphase cells in addition to the chromosomes of interest. Figure 2 shows the presence of interphase cells along with the desired chromosomes. The presence of these limiting factors in a photomicrograph poses serious threat to accurate classification of chromosomes. Interphase cells share common intensity levels with the chromosomes and traditional threshold techniques fail to remove these interphase cells where a vital part of the chromosomes of interest are also removed. An algorithm based on Connected Components Labeling and Extraction [6] is proposed herewith that segments and removes dirt and interphase cells from the chromosomal images, facilitating accurate, automatic karyotyping. The Connected Components or objects having fewer than the defined amount of pixels, arrived by a Trial and Error Method, are removed from the image, resulting in another image where the chromosomes except the interphase cells are uniquely identified and removed. The segmented image is subtracted from the original image, leaving behind the chromosomes of interest without any dirt and interphase cells, facilitating accurate karyotyping procedures [5].

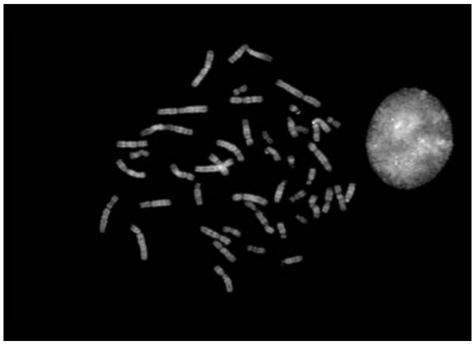
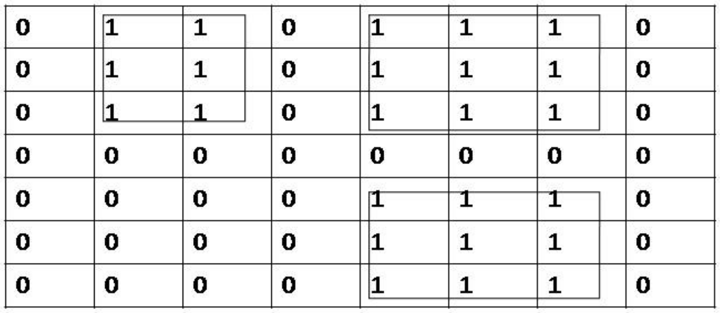


Figure 2. Presence of an Interphase Cell in a Chromosome Image

2. Methodology

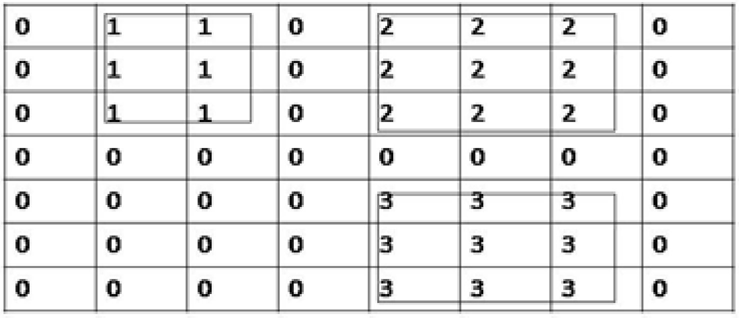
Database of individual chromosomes and karyotypes were obtained from the standard clinical database, formulated by Grisan, *et al.*, [1], at the BIOIMLAB, Laboratory of Biomedical Imaging, Department of Information Engineering, University of Padova, Italy [4]. The Dataset consists of 162 PAL-resolution, Q-banded Pro-metaphase chromosomes. An expert cytologist has performed manual karyotyping and an Optical Microscope has been used to acquire the images. The chromosome images are monochrome, with 8 bits/pixel PAL – resolution. The input image is converted to a binary image. The Global Threshold of the image is computed that converts an intensity image to a binary image. The range of Global Threshold lies in [0, 1] and is normalized. Otsu's method is employed in determining the global threshold level that selects the threshold value in order to minimize the inter-class variance of the thresholded black and white pixels [10]. Connected Components Labeling and Extraction Method is employed to remove the interphase cells. Connected Components form a connected group of pixels in a given image. For instance, the binary image shown in Figure 3 has three connected components.



0	1	1	0	1	1	1	0
0	1	1	0	1	1	1	0
0	1	1	0	1	1	1	0
0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	0
0	0	0	0	1	1	1	0
0	0	0	0	1	1	1	0

Figure 3. Image with 3-connected Components

Connected Components Labeling is the process of identifying the connected components in an image and assigning each one, a unique label, creating a Label Matrix, as shown in Figure 4 [9]. The input image is of the unsigned integer type and non-sparse. The output binary image is logical and has a value of 1 for all white pixels in the input image with luminance value greater than the global threshold level and a value of 0 for all the other black pixels. The Effective Metric (EM) is calculated as an additional parameter at the time of computation of the global threshold [9]. The metric gives an indication of the effectiveness of thresholding the input image and the value lies in the range [0, 1].



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

Figure 4. Labeled Connected Components

The number of connected components in an input image is calculated using the label Matrix. The lower bound is attained by the images with a single value for the gray level, and the upper bound is achieved by binary-valued images. The type of connectivity, size of the image, number of connected objects, and the list of pixels belonging to each connected component is identified [9]. The desired connectivity for the connected components was proposed to have the following scalar values as mentioned in Table 1.

Table 1. Type of Connectivity and Inference

Value	Connectivity	Inference
4	2-D	4 – Connected Neighborhood
6	2-D	6 – Connected Neighborhood
8	3-D	8 – Connected Neighborhood
18	3-D	18 – Connected Neighborhood
26	3-D	26 – Connected Neighborhood

The connected neighborhood is symmetrical about its center element [8]. The single-valued elements define the neighbor locations relative to the connected component. The Label Matrix is constructed to visualize the connected components. This label Matrix is displayed as a pseudo-colored image. The pseudo-colored image is created in a way that the label identifying each object in the Label Matrix is mapped to a different color in the associated 'color map' matrix [9]. The size of the label Matrix depends on the size of the input image and that of the structure of the connected components. An object in the input image is made up of pixels labeled '1' and the second object is made up of pixels labeled '2' and the process continues until all the objects are labeled. Connected Components Labeling algorithm describes the color map, background color, and the method of mapping the objects in the label Matrix to the colors in the color map [9]. The algorithm facilitates selecting individual objects in a given image where we need to specify the pixels in the input image and the algorithm returns the image that accommodates only that region of interest from the input image containing one of the specified pixels [8]. The algorithm also computes various parameters from the input image that facilitates extraction process. Area of the input image is calculated as a measure of the size of the foreground of the image [7]. The algorithm estimates the area of all the pixels in the input image by summing the area of individual pixels. The area of an individual pixel is calculated by observing its 2*2 neighborhood [11]. The neighborhood of the pixel thus plays a major role in determining the structural parameters of the Region Of Interest (ROI). Six different parameters can be obtained, each representing a different area as shown in Table 2.

Each pixel is a part of the four different 2*2 neighborhoods. Thus a single 'ON' pixel, surrounded by 'OFF' pixels has a total area of '1' [11].

There are two basic kinds of measurements made on the input regions that include Pixel Value Measurements and Shape Measurement. These measurements help identify ROI to the best accuracy and excellent segmentation results can be achieved. The list of parameters calculated is tabulated in Table 3. Pixel based measurement parameters are tabulated in Table 4.

Table 2. Area and Number of 'ON' Pixels

Area	Number of 'ON' Pixels
0	0
1/4	1
1/2	2 (adjacent)
3/4	2 (diagonal)
7/8	3
1	4

Table 3. Pixel and Shape based Parameters

Pixel Based	Shape Based
Maximum Intensity, Minimum Intensity, Weighted Centroid, Mean Intensity, Pixel Values	Area, Euler Number, Perimeter, Centroid, Solidity, Eccentricity, Major Axis Length, Minor Axis Length, Orientation

Table 4. Pixel based Measurement Parameters

Parameter	Inference
Maximum Intensity	Pixel with the greatest intensity in the ROI
Minimum Intensity	Pixel with the least intensity in the ROI
Mean Intensity	Mean of the intensity values in the ROI
Weighted Centroid	Specifies ROI center based on location and intensity values
Pixel values	P*1 vector, 'P' is the number of pixels in the ROI.

3. Connected Components Extraction

The Connected Components or objects that have fewer than ‘P’ pixels are removed from the input binary image in Figure 5 that produces another binary image as in Figure 6.

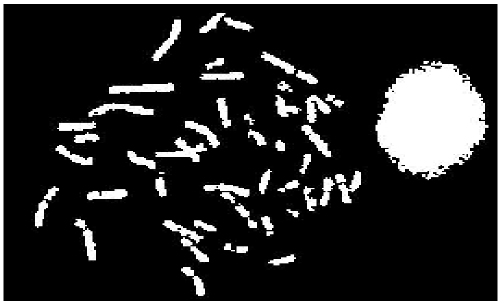


Figure 5. Input Image

The input image containing the chromosomes and the interphase cells are read into the algorithm and the number of pixels ‘P’ is calculated for each of the object in the input image. The desired connectivity can be chosen, allowing the algorithm to be more flexible and user-friendly. Components having fewer than ‘P’ pixels are removed. The value of ‘P’ is chosen based on a Trial and Error method to find the maximum segmentation efficiency that can be achieved with the algorithm. Segmentation Efficiency can be defined as the ratio of the number of true extractions to the total number of inputs, where the input image is a logical array of dimensions [576 768] and the output image will contain only the interphase cells and is logical. The input image is a logical array of dimensions [576 768] and the output image will contain only the interphase cells and is logical.

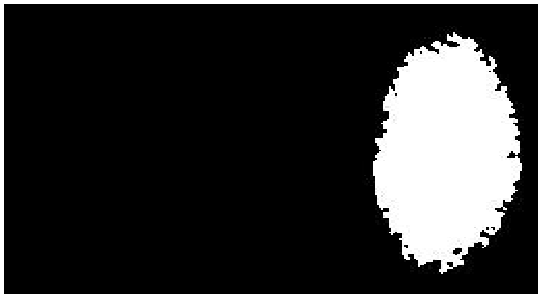


Figure 6. Extracted Connected Components having more than ‘P’ Pixels

4. Results and Discussions

A total of 172 chromosome images are taken and applied as input to this algorithm. Table 5, Table 6, Table 7, Table 8 and Table 9 shows the number of correct and incorrect extraction of interphases based on the chosen value of the number of ‘P’ pixels in the connected components. The Segmentation Efficiency shows a good degree of variation between the numbers of ‘P’ pixels chosen for the connected components.

Table 5. Number of ‘P’ Pixels = 1500

Number of Images	P=1500		
	True Extraction	False Extraction	SE (%)
172	140	32	81.39

Table 6. Number of ‘P’ Pixels = 2000

Number of Images	P=2000		
	True Extraction	False Extraction	SE (%)
172	148	24	86.05

Table 7. Number of ‘P’ Pixels = 2500

Number of Images	P=2500		
	True Extraction	False Extraction	SE (%)
172	154	18	89.53

Table 8. Number of ‘P’ Pixels = 3000

Number of Images	P=3000		
	True Extraction	False Extraction	SE (%)
172	162	10	94.18

Table 9. Number of ‘P’ Pixels = 3500

Number of Images	P=3500		
	True Extraction	False Extraction	SE (%)
172	171	01	99.42

A Segmentation Efficiency of 99.42 was achieved with the value of P=3500 for the chromosome images that almost removes all the dirt and interphase cells from the photomicrographs of chromosomes. The variation in the value of ‘P’ accounts to the possible occlusion between the chromosomes that forms more connected components than usual. Another parameter called ‘Segmentation Accuracy’ (SA) was also determined, defined as the ratio of the area of the segmented ROI using the Connected Components labeling and Extraction Algorithm to the area of the manually segmented ROI.

Table 10 summarizes the Segmentation Accuracy of interphases from a given sample input of 10 chromosome images, segmented using the Connected Components labeling and Extraction algorithm.

Table 10. Segmentation Accuracy (SA)

Name of the Interphase	Segmentation Accuracy (SA) in (%)
Interphase 1.bmp	98.42
Interphase 2.bmp	99.17
Interphase 3.bmp	98.78
Interphase 4.bmp	99.24
Interphase 5.bmp	98.89
Interphase 6.bmp	99.12
Interphase 7.bmp	98.56
Interphase 8.bmp	99.43
Interphase 9.bmp	98.87
Interphase 10.bmp	99.18

Increasing the value of ‘P’ beyond P=3500 did not yield any significant change in the Segmentation Accuracy. The interphase cells, dirt, stain, and other unwanted objects were thus effectively removed using this algorithm as shown in Figure 7.

The mean value of Segmentation Accuracy (SA) using the Connected Components Labeling and Extraction algorithm was found to be 98. 246 for an input set of 172 chromosome images where the interphase cells are uniquely identified and removed.

The segmented image is subtracted from the original image, leaving behind the chromosomes with no dirt and interphase cells, facilitating efficient karyotyping procedures.

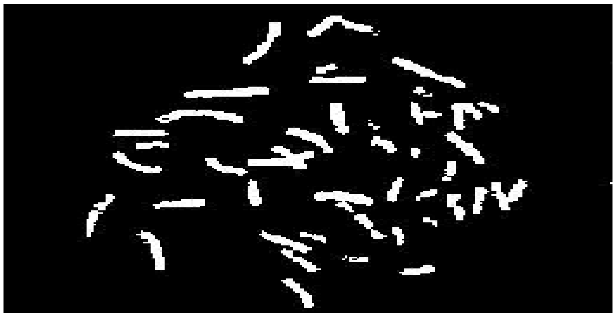


Figure 7. Chromosome Image with Interphase Cells and Dirt Removed

5. Conclusion

Connected Components Labeling and Extraction Method of segmentation and removal is extremely helpful when the unwanted object to be segmented and removed shares common intensity levels with the desired information where traditional threshold based procedures like adaptive thresholding will fail to accomplish precise segmentation results. Segmentation methods like Active Contours and Region Growing can only identify the ROI but will not remove the identified interphase cells. Initial placement of contour for active contours and initial placement of seed for region growing consumes a lot of time where the Connected Components Labeling and Extraction method can yield the most accurate results.

6. Future Enhancements

Interphase cells, dirt, stains and other unwanted objects can be successfully segmented and removed with the proposed technique. The methodology can equally be applied to the recent advancements in karyotype imaging that includes but not restricted to Fluorescence in-situ Hybridization (FISH) and Comparative Genomic Hybridization (CGH) chromosome images and the accuracy can be tested this sentence).

References

[1] E. Grisan, E. Poletti and A. Ruggeri, “Automatic segmentation and disentangling of chromosome in Q-band prometaphase images”, IEEE Trans Inf Technol Biomed., vol. 13, no. 4, **(2009)**, pp. 575-581.

[2] S. W. Katz and A. D. Brink, “Segmentation of chromosome images” Proceedings of the IEEE South African Symposium on Communication, Networking and Broadcasting, **(1993)** August 6; South Africa.

[3] L. G. Shaffer, M. L. Slovak and L. J. Campbell, “An International System for HumanCytogenetic Nomenclature”, ISCN 2009, S. K. Basel (ed.), (**2009**).

[4] Laboratory of Biomedical Imaging, University of Padova, Italy, http://bioimlab.dei.unipd.it.

[5] S. G. Vaidyanathan, CIHCI, vol. 1, no. 1, (**2008**), pp. 237-245.

[6] W. -L. Chan and C. -M. Pun, (Eds.), “Robust Character Recognition using Connected-Component extraction”, Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing, **(2011)** October 14-16; Dalian, China.

[7] R. C. Gonzalez, R. E. Woods and S. L. Eddins, (Eds.), “Digital Image Processing using Matlab”, Prentice Hall, **(2003)**.

[8] Y. Fan, S. Yu and H. Zhao, (Eds.), “A novel line based connected component labeling algorithm”, 3rd IEEE International Conference on Computer Science and Information Technology, **(2010)** July 9-11; Chengdu, China.

[9] http://www.mathworks.com/help/images/labeling-and-measuring-objects-in-a-binary-image.html.

[10] http://www.mathworks.com/help/images/ref/graythresh.html.

[11] http://www.mathworks.com/help/images/ref/bwarea.html.

Authors

Sivaramakrishnan R

Sivaramakrishnan R graduated in the discipline of Electronics and Communication Engineering from Madurai Kamaraj University and completed his Post Graduation in Medical Electronics from Anna University, Chennai, India. He has more than 11 years of teaching experience to his credit in diverse areas of electronics, communication, biomedical and medical electronics. His work on "Early Detection of Cancer using Fuzzy Based Photoplethysmography" has received the prestigious AT&T BELL LABORATORIES AWARD from USA, with a cash grant. He has published various papers at graduate and post graduate levels in national and international journals and conferences. His research areas include image processing, EEG Signal Analysis, and Brain Computer Interface analysis.

Arun C

Professor Arun C received his B.E. Degree in Electronics and Communication Engineering from Shanmuga College of Engineering, India and completed his Post Graduation in Applied Electronics at PSNA College of Engineering and Technology, Dindigul, India. He has completed his Ph.D. under the Information and Communication Discipline in 2010. He has published various papers at graduate, post graduate and doctorate levels in national and international journals and conferences. His research areas include VLSI Signal Processing, Image Processing, and Digital Communication.