Automatic Locating the Centromere on Human Chromosome Pictures

M. Moradi
Electrical and Computer Engineering Department, Faculty of Engineering,
University of Tehran, Tehran, Iran
moradi@iranbme.net

S. K .Setarehdan
Electrical and Computer Engineering Department, Faculty of Engineering,
University of Tehran, Tehran, Iran
ksetareh@ut.ac.ir

S.R Ghaffari
Cancer Institute, School of Medicine, University of Tehran, Iran
saeed@ghaffari.org

Abstract

Many genetic disorders or possible abnormalities that may occur in the future generations can be predicted through analyzing the shape and morphological characteristics of the chromosomes. Karyotype (a systemized array of the human chromosomes obtained from a single cell either by drawing or by photography using a light microscope [1]) is often used for this purpose. To make a Karyotype it is necessary to identify each one of the 24 chromosomes (22 autosomal and a pair of sex chromosomes) from the microscopic images. The first step to automate this process is then to define the morphological and band pattern based features for each chromosome

An important class of morphological features includes those defined with respect to the location of the chromosome's centromere (part of the chromosome that divides it to the long and short arms). Therefore, localization of centromere is an initial step in designing an automatic karyotyping system.

In this paper, an effective algorithm for chromosome image processing and automatic centromere locating is presented. The procedure is based on the calculation and analyzing the vertical and horizontal projection vectors of the binary image of the chromosome. The binary image is obtained using the thresholding of the input image after histogram modification and analyzing. When applied to the real chromosome images supplied by the Cytogenetic Laboratory of the Cancer Institute of the Imam hospital in Tehran, an average accuracy of 96% for Centromere locating is achieved.

1. Introduction

Karyotyping is a standard tool in medicine. In addition to well known genetic abnormalities like aneuploidy (improper number of chromosomes) or chromosomes with some missing parts, some of the fatal pathological conditions like lukemia are also correlated with chromosome defects [1].

Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems (CBMS'03)
1063-7125/03 $17.00 © 2003 **IEEE**

Karyotyping consists of the identification, classification and presentation of the 23 pairs of the chromosomes in a single picture. This process, which is usually done manually by a human expert is a difficult and time consuming task. In conventional karyotyping, giesma banded cells are photographed under a light microscope (an example picture is shown in figure 1) during the metaphase stage (one of the four stages of the cell division namely: prophase, metaphase, anaphase and telophase). In metaphase, the chromatin is condensed inside the chromosomes making them to be easily observed with a light microscope. A band is defined as part of the chromosome which is clearly distinguishable from its adjacent segments by its darker or brighter appearance. The chromosomes are visualized as consisting of a continuous series of these bright and dark bands. Each of 24 chromosomes has a specific band pattern.

The result of karyotyping process for figure 1, which is done manually by a cytogeneticist is shown in figure 2. Two stages of this process are segmentation and classification of the chromosomes. Automatic classification of chromosomes has been a well studied problem in the last 3 decades [8,9,10]. Natural complexity of the problem is caused by various unpredictable appearances of the chromosomes due to non-rigid nature of them.

Figure 1. G-banded chromosomes as seen under a microscope

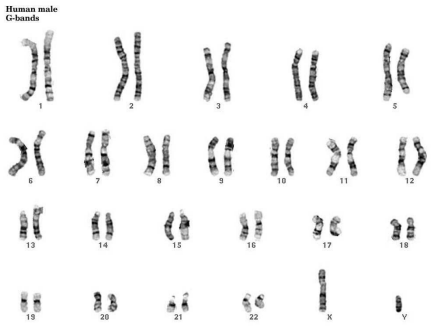

Figure 2. Karyotype of a male

Features used in chromosome classification generally fall into two main categories of the *geometrical* features and the *band pattern based* features [3]. Length of the chromosome and the centromeric index (CI) are the most important geometrical features. CI is the ratio of the length of the short arm of the chromosome to its long arm. These two arms are separated from each other in a point called centromere. From morphological point of view, the centromere is located in the narrowest part of the chromosome along its longitudinal direction.

Based on the location of the centromere along the chromosomes, there are three classes defined for them. In some chromosomes, which are called *metacentric*, the centromere is located in the middle of the long axis of the chromosome and the two arms are almost of the same length (therefore CI ≅1). Chromosomes number 1, 3, 16, 19, and 20 are metacentric. A chromosome is called *acrocentric* when the centromere divides the chromosome into two arms of unequal length. Chromosomes number 13, 14, 15, 21 and 22 belong to this class. In the last class of the chromosomes named *telocentric* the short arm is very small and the centromere is located near to one of the two ends of the

chromosome. From these explanations it is clear that CI is an important feature for classification of the chromosomes.

Different studies were reported for automatic localization of the centromere in the past. The most recent works are based on the technique known as the medial axis transformation (MAT) [3,4]. The main drawback of the MAT based algorithms is the computation cost. For an image of size n×n, the computation complexity is of orders $O(n^2)$ up to $O(n^3)$ depending on the method used for MAT calculation [5].

In comparison, the complexity of the effective algorithm described in this article is of order O(2n) and dose not use the computationally expensive method of MAT. The results we achieved are among the best in the literature.

The rest of the paper is organized as follows: Section 2 describes the chromosome image data set used in this research. In section 3 the block diagram of the algorithm together with the function of each block is explained. Section 4 demonstrates the results of the application of the algorithm to real chromosome images. Finally section 5 concludes the paper.

2. Data

The images used in this study are produced in Cytogenetic Laboratory of Cancer Institute, Imam hospital, Tehran. The images are acquired by conventional photography using a light microscope (Leitz, ortholux) with a magnification factor of 100X Chromosomes were segmented by an expert in the Cytogenetic Laboratory and then scanned by a scanner (Microtek, ScanPlus 6) with a resolution of 300 dpi. The grayscale resolution of the resulting digitized pictures was set to 256 levels.

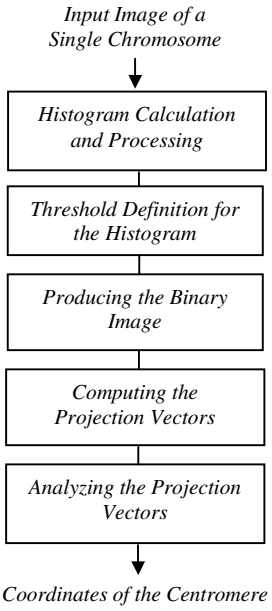

Figure 3. The proposed automatic algorithm

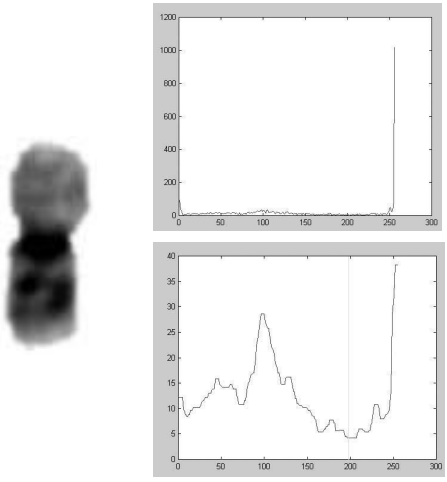

Figure 4. A typical chromosome number 16 and its original (upper plot) and filtered (lower plot) histograms

3. Processing Method

Figure 3 demonstrates a block diagram of the proposed automatic centromere locating system. In the following sections we will describe the function of each block in more detailes.

3.1. Producing the Binary Image

Since we are looking for a morphological feature of chromosome, the grayscale information of the image is not of interest at this stage. Therefore, producing a binary image will make the rest of the process easier. The chromosome images are typically bi-modal [6].

In other words, such an image includes an object over a uni-color background. Histogram of such image usually includes two peaks, one of which corresponds to the background and the other to the object. The preferred threshold separating the object and background-related peaks in the histogram of such an image is usually the gray level representing the global minimum located between the two peaks. For a robust identification of the two major peaks and the global minimum between them, it is necessary to filter out the small variations in the histogram. For this purpose, a first order Savitzky-Golay FIR smoothing filter with a window width of 5 is found to be effective. Savitzky-Golay filters are low-pass filters useful for smoothing data. This time-domain method of smoothing is based on least squares polynomial fitting across a moving window within the data. The method was originally designed to preserve the higher moments within time-domain spectral peak data [11]. Since the background related peak is of many orders of the object related peak in magnitude, a median filter with a window size of 9 is also applied to the histogram to make the peaks visually comparable and easier to locate automatically. Figure 4 shows a typical chromosome number 16 and its original and filtered histograms.

Due to the usually white background of the images, the background related peak in the histogram is always located close to the gray level 255. However, a little bit more care is needed for the determination of the object related peak. After a statistical analysis of the images in the data set, the following two rules are found to be valid for all cases: a) if the mean grayscale of an image is greater than 200 then the object related peak is located between 0 and 220; b) however, for darker images (mean grayscale less than 200), the object related peak is located between 0 and 150.

By locating the two peaks representing the background and the object in the histogram, the gray level coincident with the global minimum between the two peaks is defined next. Using this value as a threshold, all the pixels with gray level below this threshold are set to 1 (white) and the remaining pixels to 0 (black) producing the binary format of the input image.

3.2. Computing the projection vectors

In order to calculate the horizontal projection vector, simply the pixel values of each row are summed up in the binary image. Considering that the binary image includes only 1s (white pixels) and 0s (dark pixels), therefore, each element of the horizontal projection vector is equal to the number of white pixels (1s) in the corresponding row. Similarly, to calculate the vertical projection vector, the pixel values of each column are summed up in the binary image.

Theoretically, these two orthogonal projection vectors contain all the morphological information of the chromosome and can be used for binary image reconstruction and/or for feature extraction. For example they can be used to calculate the extent of the chromosome (i.e. where the chromosome begins and ends on the image plane) or they can be used for locating the centromere of the chromosome, which is explained next.

3.3. Locating the centromere

As it was explained before, from the morphological point of view, the centromere is the narrowest part of the chromosome in its longitudinal direction. This produces a global minimum in the central part of the horizontal projection vector of the chromosome, due to the small number of the white pixels (1s) in the horizontal direction at this region (see Figure 5). Therefore, centromere locating is just locating this global minimum in the central region of the horizontal projection vector. Figure 5 shows the horizontal projection vector of the typical chromosome number 16 of figure 4 with its global minimum in the central region is marked with a circle. This corresponds to the centromere location which is marked with a white line over the chromosome itself.

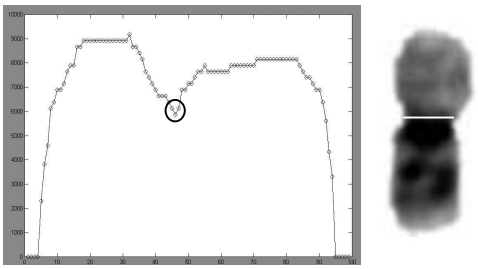

Figure 5. The horizontal projection vector with its global minimum point in the central region is marked with a circle. This corresponds to the centromere location which is marked with a white line over the chromosome itself**.**

4. Results

The proposed automatic centromere locating algorithm is now applied to 87 randomly selected chromosome images from the data set. For comparison purposes, our cytogeneticist colleague was also asked to manually mark the centromere of the chromosomes for all images. The automatically defined centromere locations were then compared to those identified by the cytogeneticist and the results are summarized in Table I. In this comparison the Euclidean distance between the two results is considered as the absolute error of the automatic algorithm. More over, each error value is normalized with the length of the chromosome, which is also determined by the expert. Table I demonstrates the mean and the standard deviation of the absolute and normalized errors calculated over the entire data set.

5. Conclusions

A simple, yet very effective algorithm for automatically locating the centromere in a microscopic image of a human chromosome was presented. Centromere locating is important for feature extraction and classification of the chromosomes, which is a

necessary step towards automatic Karyotyping. The algorithm is based on the calculation and analyzing the vertical and horizontal projection vectors of the binary image of the chromosome. The binary image is obtained by applying a threshold on the input image after histogram modification and analyzing.

TABLE 1. Results of the comparison of the automatically defined centromere locations to those manually identified by the expert observer**.**

Mean value of the absolute error	4.3 (pixel)
Standard deviation of the absolute error	3.8
Mean value of the normalized error	0.041
Standard deviation of the normalized error	0.03

The algorithm was applied to 87 real chromosome images supplied by the Cytogenetic Laboratory of the Cancer Institute of the Imam hospital in Tehran. The mean normalized error (Euclidean distance between the reference and automatically extracted centromere locations normalized by the chromosome length) is about 4%, which is very small and the accuracy is satisfactory.

The main restriction to the algorithm is the highly bent chromosomes in the image. The bending must be less than 90 degree to make the algorithm work successfully.

Acknowledgement

We are grateful to the Cytogenetic Laboratory of the Cancer Institute of Imam hospital in Tehran. The authors would like to express their gratitude to Ms F. Farzanfar for her great help in providing and manual analyzing the images.

References

[1]http://www.pathology.washington.edu/Cytogallery/cytogallery.html

[2] Ozy Sjahptra, James M. Keller, "Evolution of a Fuzzy Rule-Based System for Automatic Chromosome Recognition", IEEE International Fuzzy Systems Conference Proceedings, Seoul Korea, pp: 129-134, Aug 1999

[3] Jong Man Cho, "Chromosome Classification Using Back propagation Neural Networks", IEEE Eng in Medicine and Biology, pp: 28-33, Jan/Feb 2000

[4] Boaz Lerner, "Toward a Completely Automatic Neural Network Based Human Chromosome Analysis", IEEE Trans. On Systems, Man, and Cybernetics-Part B: Cybernetics, vol. 28, pp: 544-552, Aug. 1998

[5] Jing Fu Jenq, Sartaj Sahni, "Serial and Parallel Algorithms for the Medial Axis Transformation", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 14, No 12, pp: 1218-1224, Dec. 1992

[6]D.Vernon, *Machine Vision*, Prentice-Hall, 1991, pp: 49 - 51, 86 - 89.

[7]http://www.dai.ed.ac.uk/HIPR2/threshld.htm

[8] Castleman, K.R. and Melnyk J. H., "Automated System for Chromosome Analysis – Final Report", JPL Document No. 5040-30, Jet Propulsion Laboratory, Pasadena, California, 1976,

[9] Piper, J., and Granum, E., "On Fully Automatic Feature Measurement for Banded Chromosome Classification," Cytometry **1**0:242-255, 1989.

[10] A. Carothers and J. Piper, "Computer-Aided Classification of Human Chromosomes: A Review," Statistics and Computing, vol. 4, no. 3, pp. 161-171, 1994.

[11] Numerical Recipes Software Group, The Art of Scientific Computing, Cambridge University Press.1992