M-FISH Chromosome Images Classification by Watershed Based Segmentation Approach

LIJIYA A, SREEJINI K.S, V.K.GOVINDAN
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY, CALICUT, KERALA
INDIA

***Abstract*— Karyotyping is a technique used to display and study the human chromosomes for detecting abnormalities, genetic disorders or defects. M-FISH (Multiplex Fluorescent *In-Situ* Hybridization) provides color karyotyping. In this paper, naïve Bayes classification of M-FISH chromosome images based on watershed based chromosome segmentation is presented. It is observed that the classification of the watershed regions by using the naive Bayes classifier works better than pixel by pixel classification. By adding the feature, standard deviation along with mean of each region, improved classification accuracy was observed. The approach was tested on a database and found to provide an accuracy of 73%.**

Keywords- M-FISH, chromosome, segmentation, karyotyping, watershed transform, Bayes classifier.

I. INTRODUCTION

Cytogenetics is the branch of genetics, deals with the study of the structure and function of the cell, especially chromosomes. Experts can predict genetic disorders or possible abnormalities that may occur in the future generations, by examining the chromosome images. These images are the sources of important information about the health of human beings. Tjio and Levan [1] discovered that the number of human chromosomes is 46 in 1956 and in 1960; the Denver group classification standard was established. In the past many researchers have attempted to automate human chromosome analysis and have produced results though not comparable to manual classification. Many software packages are available for Karyotyping. Automating chromosome classification is the first step in automating the karyotyping process. Cells for chromosome analysis are mostly taken from amniotic fluid or blood samples. *Multiplex or Multi-color Fluorescence In-Situ Hybridization (M-FISH)* is a recently developed chromosome imaging technique for the visualization of chromosome aberrations.

II. BACKGROUND

A normal human cell has 46 chromosomes: 44 autosomes and two sex chromosomes (XX: Female or XY: Male. Chromosomes, the coiled strands of deoxyribonucleic acid (DNA), appear inside the nucleus during cell division (mitosis). Chromosomes exist as a pair, one from each parent.

Chromosomal aberrations can be categorized into numerical and structural aberrations. Numerical aberrations occur due to unusual number of chromosomes. Structural aberrations can be due to translocations, insertions and deletions. Translocation means the rearrangement of a chromosome in which a segment is moved from one location to another, or within the same chromosome. Deletion: a segment of a chromosome can be deleted from a chromosome. Insertion: a segment of a chromosome can be inserted into another chromosome [2].

Karyotype, is the term used to display chromosomes of a cell for diagnostic purpose. In this configuration, the chromosomes are ordered by length from the largest (chromosome 1) to the smallest (chromosome 22), followed by the sex chromosomes. Karyotype images are used in clinical test, to determine if all the chromosomes appear normal and are present in the correct number, since the abnormal cells may have an excess or a deficit of chromosomes. Earlier, chromosomes were classified into only seven groups based on the length and the position of the centromere, called Denver Classification. Manual karyotyping is a very expensive and time-consuming task and needs more trained personnel and is done by visually examining, manually locating, classifying and evaluating all chromosomes.

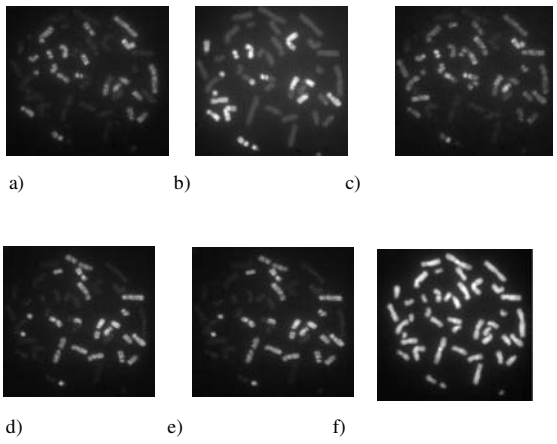

Fig.1. Five channel M-FISH image data. (a) Aqua fluor. (b) Red fluor (c) Far red fluor. (d) Green fluor. (e) Gold fluor. (f) DAPI image

Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering
Editor In Chief Dr. R. K. Singh. ISBN: 978-981-07-1847-3

To develop a karyotype, a cell is photographed under a light microscope during the metaphase stage of cell division. Different staining techniques are used, that allow us to analyze different kinds of abnormalities. A particularly useful cytogenetic technique for the analysis of aberrations is the M-FISH which provides color images.

There are two types of multicolor FISH imaging systems: M-FISH, developed by Speicher et al. [3] and 'spectral karyotyping' (SKY), developed by Schrock et al. [4] which uses an interferometer. M-FISH images are captured with a fluorescent microscope with multiple optical filters.

M-FISH uses five color dyes that attach to various chromosomes differently to produce a multispectral image, and a sixth dye called DAPI (4 in, 6–diamidino–2–phenylindole) that attaches to all chromosomes to produce a gray-scale image. M-FISH images are captured with a fluorescent microscope with multiple optical filters. Each of the fluors is visible in one of the spectral channels in a way that an M-FISH image consists of six images, and each image is the response of the chromosome to the particular fluor. Thus at least five distinguishable fluors are needed for combinatorial labeling to uniquely identify all 24 chromosome types as the number of useful combinations of N fluors is $2^N - 1$ [3]. An M-FISH image is shown in fig. 1.

M-FISH imaging technique has several advantages [11]:

- The chromosome classification is simplified.
- Subtle chromosomal aberrations are detected.
- It can be used for the identification of small genetic markers that remain elusive after banding

The present work proposes a classification approach using mean and standard deviation of M-FISH image segments obtained by applying watershed segmentation. The rest of the paper is organized as follows: Section III presents a brief review of some of the major existing work in the literature. Image segmentation and classification processes are given in Section IV. The comparative results obtained on standard database for the proposed approach and existing approaches are presented in Section V, and the paper is finally concluded in Section VI.

III. LITERATURE SURVEY

Since the introduction of M-FISH technology, many attempts have been made to automate the human chromosome analysis. The first M-FISH based attempt was by Speicher [3] in 1990. The steps involved are segmentation, thresholding and classification. This method is simple and fast when considering only the pixel classification time and does not require generation of a training data set.

Automatic pixel by pixel classification algorithm for M-FISH images was presented [5]. They approach the pixel classification as a 25 class 6 feature pattern recognition problem and classified using Bayes classifier. The classifier uses small number of non-overlapping images.

Different supervised parametric and non-parametric classification methods, i.e., k-NN, NN, MLE for pixel-by-pixel classification of M-FISH images were proposed [7]. This method does not handle overlapping images and used only a small number of testing images.

A method for joint segmentation-classification of chromosome M-FISH images was presented in [6]. They introduced a probabilistic model of M-FISH chromosomes, which allows for simultaneous segmentation and classification. Steps used are background / foreground separation, connected component labeling, pixel classification, majority filtering, small segment classification which eliminates all remaining small segments and rejoining the over-segmented chromosomes.

Use of pre-processing of the images including background correction and six-channel color compensation method was performed to reduce the noise and the variations were described [8]. They performed joint segmentation and classification of MFISH chromosome images using the 6-feature, 25-class maximum-likelihood classifier. This work does not handle overlapping/ touching chromosomes and used small testing images.

An unsupervised classification method based on fuzzy logic classification and a prior adjusted reclassification was introduced in [9]. The steps involved are foreground-background separation, fuzzy logic classification, and prior adjusted reclassification. It requires spectral information, obtained from color table and does not require training. High average accuracy is achieved, however only a small number of testing images were used.

A watershed based segmentation method for multispectral chromosome image classification is presented in [10]. The first step is the computation of the gradient magnitude of the grayscale DAPI channel. The watershed transform is applied in the next step and a large number of primitive homogeneous regions (over-segmentation) are produced. A binary mask of the DAPI channel is computed in order to further reduce unwanted areas. Finally, for each area a 5 feature vector is computed, each feature representing the average intensity value of each channel. Each segmented region is then classified using a Bayes classifier. Overall accuracy of 89 % is achieved, however only a small number of non-overlapping testing images were used.

Another approach [11] uses multichannel watershed based segmentation method to decompose the image into a set of homogeneous regions. Classification is performed using region based Bayes classifier and merging. This makes the detection of unhybridized regions simpler. The overall accuracy of 82.4% is achieved. The amount of misclassifications raised by the approach can be reduced by the use of region based classifier and vector median filtering described in [12]. The overall improvement of 9.99% can be achieved.

A semi unsupervised method for M-FISH chromosome image classification is presented in paper [13]. They used an automated threshold selection method in order to extract the pixels which belong to chromosomes. For each segmented pixel, the approach extracts the intensities and normalizes the features using Expectation Normalization algorithm. K-means

clustering is employed to cluster the chromosome pixels. Since the K- means algorithm suffers from the initial position of the clusters, they used emission information for each chromosome class in order to initialize the cluster centers. Overall classification accuracy of 72.48 % is obtained.

IV. MATERIALS AND METHODS

A. Image Segmentation

The Separation of each chromosome from the metaphase image is the major operation carried out in this stage. Basic steps involved are the following:

a) Removal of cells from the DAPI image

Before segmenting chromosomes from the initial image, *the cells are* removed based on the size and circularity. Since DAPI stains all chromosomes, this image is the best one for this process.

b) Minima Selection

Direct application of watershed algorithm leads to over-segmentation due to noise and other local irregularities. Solution is to reduce the number of irrelevant minima. In this work, over-segmentation is controlled by specifying minima.

c) Applying Watershed transform

Watershed transform is applied in the next step which results in tessellation of the image in to different regions. This is a region-based segmentation approach, originally proposed by Digabel and Lantuejoul [14].

The idea of watershed comes from the field of geography. The immersion approach [15] of watershed computation algorithm is used here. A grey scale image can be considered as a topographic surface, different gradient values correspond to different heights. In a topographic surface, watersheds are the lines dividing two catchment basins, each basins corresponds to each local minimum. If we punch a hole in each local minimum and immerse this surface in water, the regions in the image will start filling up with water. Immersion will starts from the points of minimum grey value. When water level in two or more adjacent basins will start merging, dams are built in order to prevent this merging. The flooding process will continue up to the stage at which only the top of dam is visible above the water line [20].

Advantageous of watershed algorithm are the following

- The watershed lines produced are always connected (it divides the image into set of connected pixels) and complete (assigns every pixel to one of the regions).
- The watershed lines correspond to obvious contours of the image.

d) Binary Mask Creation

Segmentation errors are present even after doing the above steps due to uneven hybridization. In order to avoid these errors, binary mask is created from DAPI image after cell removal and superimposing watershed regions on it. Binary mask is created by Otsu's thresholding method [16]. Basic operation behind superimposing is, logical AND operation of watershed lines and blob removed DAPI image.

e) Computation of mean and standard deviation of each segmented Region

Mean and standard deviation of each segmented area is computed. For each segmented area, the intensities of the pixels belonging to that region are then replaced with mean intensity of that region. The present work employs the mean and standard deviation of intensity values of each segmented region for classification.

B. Feature Extraction and Classification

a) Feature Extraction

This stage classifies each segmented area after performing the segmentation. A feature vector is computed from each segmented areas of an image in the M-FISH set.

b) Classification

The segments are classified using naïve Byes classifier. A naive Bayes classifier is a simple probabilistic classifier based on Bayes theorem with strong (naive) independence assumptions. Our goal is to classify the 46 chromosomes in to 22 pairs of similar chromosomes and 2 sex chromosomes (C = 24).

Let $x \in R^d$ denotes the feature vector computed from each segmented area; here d = 10. $P(c_i)$ denotes the prior probability that a feature vector belongs to class c_i where i = 1, 2,….24. $p(x|c_i)$ denotes the class conditional probability distribution function and $P(c_i|x)$ be the posterior probability that the feature vector x belongs to class c_i, given the feature vector x .

By using Bayes theorem,

$$P(c_i|x) = \frac{p(x|c_i)P(c_i)}{\sum_{i=1}^{24} p(x|c_i)P(c_i)} \quad (1)$$

The general multivariate Gaussian density function [18] in d dimension is given by

$$p(x|c_i) = \frac{1}{(2\pi)^{d/2}|\sum_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i)\right) \quad (2)$$

where x is the d-dimensional feature vector from five channels and μ_i is the mean vector of each class c_i , $\sum_i$ is the d x d covariance matrix of the class c_i , and $|\sum_i|$ and $\sum_i^{-1}$ are the determinant and inverse. Also $(x - \mu_i)^t$ denotes the transpose of $(x - \mu_i)$.

For each class, we need to calculate $p(c_i|x)$, the class to which a feature x belongs, is decided by Bayes decision rule.

$$\text{Decide } c_i \text{ , if } P(c_i|x) > P(c_j|x), \text{ for all } j \neq i. \quad (3)$$

Computed prior class probabilities from training samples are,

$$P(c_l) = \frac{(\text{no. of regions belongs to class } c_l)}{\sum_{k=1}^{24} (\text{no. of regions belongs to class } c_k)} \quad (4)$$

Here, classification is done by using mean with standard deviation of the image under test.

c) Neighbor Region Merging

In this stage, neighboring regions belonging to the same class are merged. Adjacent regions can be found by using region adjacency graph. If regions are adjacent then those regions are connected in graph and they must have a common boundary.

V. RESULTS

A. Dataset

Dataset [17] consist of 200 Multispectral images of size 517 X 645 pixels. Each M-FISH set contains five monospectral images recorded at different wavelengths. DAPI channel images are also included. There is no annotation for 17 images, that are "difficult to karyotype" images even by experienced cytogeneticist, due to tightly packed nature of chromosomes and are marked as extreme (EX). For specimen preparation, Applied Spectral Imaging, PSI, Vysis are the probs used. Each M- FISH image set has its "ground truth" image except for EX images. In ground truth image, background pixel values are zero, pixels in the overlapped regions values are 255, and chromosome pixel values are from 1 to 24 depending on chromosome type. In case of translocations, chromosomes are labeled with the class which makes up the most of the chromosome. The images for training and testing are used for this method from this dataset.

B. Classification Accuracy

TABLE I. CLASSIFICATION ACCURACY

#image	*Classification accuracy of various approaches*		
	Proposed: mean &std. deviation	*mean [10]*	*pixel-by-pixel [6]*
1	76.86	78.86	71.67
2	74.30	71.66	61.82
3	75.66	75.07	69.74
4	73.88	66.46	44.59
5	67.86	67.18	85.51
Average	73.71	71.84	66.66

Tables 1 show the comparison of classification accuracy obtained with proposed method, mean only method [10] and pixel by pixel classification method [6]. Here, for all the methods, the same images are used for training and testing. Proposed method obtained the average classification accuracy of 73.71%. For all of the methods, classification accuracy can be improved by preprocessing methods [19, 2]. Classification accuracy, is defined as

$$\text{Chromo. Class Acc} = \frac{\text{chromosome pixels correctly classified}}{\text{total no. of pixels}} \quad (5)$$

C. Classification Map

Actual ground truth and classmap generated for one M-FISH dataset tested in our dataset is shown in Fig 2. Actual ground truth is given in dataset.

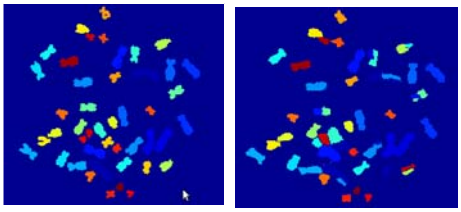

Fig.2 *a) Ground Truth* *b) Classmap Generated*

VI. CONCLUSION AND DISCUSSION

The paper presents M-FISH chromosome classification using watershed segmentation with naïve Bayes classifier. Use of watershed based segmentation, with mechanism for preventing over-segmentation, gives better performance in classification accuracies. The Bayes classification on watershed segmented chromosome regions works for all probes and the results are better than pixel by pixel classification, which always produces noisy results. As the classification is done on the watershed regions, the computational time needed is also much less than the pixels by pixel approach. Further improvements in classification accuracy may be achieved through image registration, background correction and color compensation techniques [19]. In some cases, manually corrected ground truth is also required to get correct classification [2]. Future work is to extend this method to include larger set of images.

REFERENCES

[1] J.H. Tjio and A. Levan, The chromosome number in man, *Hereditas* 42 (1956), 1-6.

[2] Hyo Hun Choi, "Automatic Segmentation and Classification of Multiplex-Fluorescence In-Situ Hybridization Chromosome Images", Phd Dissertation, The University of Texas at Austin, 2006.

[3] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Karyotyping human chromosomes by combinatorial multi-fluor FISH," *Nat. Genet.*, vol. 12, pp. 368 – 375, 1996.

[4] E. Schrock, S. duManoir, T. Veldman, B. Schhoell, J. Weinberg, M. FergusonSmith, Y. Ning, D. Ledbetter, I. BarAm, D. Soenksen,Y. Garini, and T. Ried, "Multicolor spectral karyotyping of human chromosomes", Science, vol. 273, no. 5274, pp. 494 - 497, 1996.

[5] M.P. Sampat, A.C. Bovik, J K. Aggarwal, and K.R. Castleman, "Pixel-by-Pixel classification of MFISH images," in Proc. 24th IEEE Ann. Intern. Conf. (EMBS), Houston, 2002, pp. 999-1000.

[6] W.C. Schwartzkopf, A.C. Bovik, and B.L. Evans, "Maximum likelihood techniques for joint segmentation-classification of multispectral chromosome images," IEEE Trans. Med. Imag., vol. 24, pp. 1593-1610, Dec.2005.

[7] M. P. Sampat, A. C. Bovik, J. K. Aggarwal, and K. R. Castleman, "Supervised parametric and non-parametric classification of chromosome images," *Pattern Recognit.*, vol. 38, pp. 1209 – 1223, Aug. 2005.

[8] 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.

[9] Choi, K. R. Castleman and A C. Bovik, "Segmentation and fuzzy logic classification of M-FISH chromosome images," in Proc. IEEE Intern. Conf. on Image Processing (lClP), Atlanta, 2006, pp. 69-72.

[10] P. S. Karvelis, D. I. Fotiadis, M. Syrrou, and I. Georgiou, "A watershed based segmentation method for multispectral chromosome images classification," in *Proc. 28th IEEE Ann. Intern. Conf. (EMBS)*, New York, 2006, pp. 3009–3012.

[11] P Karvelis, A Tzallas, D. Fotiadis, and 1. Georgiou, "A multichannel watershed-based segmentation method for multispectral chromosome classification," IEEE Trans. on Med. Imag., vol. 27, no. 5, pp. 697-708,2008.

[12] P. S. Karvelis, D. I. Fotiadis, D. I. Tsalikakis, and I. Georgiou, "Enhancement of Multichannel Chromosome Classification Using a Region-Based Classifier and Vector Median Filtering," IEEE Trans. Information Technology in Biomedicine., vol. 13, No.4, pp. 561-570, July 2009.

[13] P. S. Karvelis, Aristidis Likas, Dimitrios I Fotiadis, "Semi Unsupervised M-FISH chromosome image classification," in 10th IEEE Intern. Conf on information Technology and Applications in Medicine.(ITAB), 2008, pp. 1 – 4.

[14] Digabel, H., and Lantuéjoul, C. "Iterative algorithms". In Actes du Second Symposium Européen d'Analyse Quantitative des Microstructures es en Sciences des Matériaux , Biologie et Médecine, Caen, 4-7 October 1977 (1978), J.-L. Chermant, Ed., Riederer Verlag, Stuttgart, pp. 85-99.

[15] Vincent, L., and Soille, P. "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations", IEEE Transaction on Pattern Analysis and Machine Intelligence, 13, No.6, pp. 583–598, June 1991.

[16] N. Otsu, "A threshold selection method for gray-levels histograms", IEEE Trans. On Pat. Anal. And Machine Int. ,13 ,pp.583 – 598, 1993.

[17] http://dip4fish.blogspot.com/2011/11/mfish-dataset-available.html

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

[19] Hyohoon choi, Kenneth R. Castleman, Alan C. Bovik, "Color Compensation of Multicolor M-fish Images", In IEEE Transaction on Medical Imaging, volume 28, January 2009.

[20] Rafael C. Gonzalez & Richard E. Woods. Digital Image Processing, 2nd edition. Prentice-Hall, 2002.