CLASSIFICATION OF MULTICOLOR FLUORESCENCE IN SITU HYBRIDIZATION IMAGES USING GAUSSIAN MIXTURE MODELS

AMIN FAZEL **REZA DERAKHSHANI** **YU-PING WANG**

Dept. of CS and Electrical Engineering
School of Computing and Engineering
University of Missouri at Kansas City
Kansas City, Missouri

ABSTRACT

This paper introduces a fully automatic chromosome classification algorithm for Multicolor or Multiplex Fluorescence In-Situ Hybridization (M-FISH) images using Gaussian mixture model technique. M-FISH is a recently developed cellular imaging method for rapid detection of chromosomal abnormalities, where each chromosome is labeled with 5 dyes and counterstained with DAPI florescence stain. The problem is modeled as a 24-class 6-feature pixel-by-pixel classification task. Features are composed of dyed-image pixel brightness. By conducting experiments on ADIR M-FISH database, we demonstrate that our proposed Gaussian mixture model based approach yields significantly better results over the Bayesian classifiers which assume a single Gaussian probability distribution for M-FISH data.

I. INTRODUCTION

Chromosomes are made of DNA molecules which carry information necessary for development and function of an individual. Each normal human cell has 46 chromosomes. Close analysis of chromosome images yield information such as the chromosomal abnormalities which are important in diagnosis of cancers or genetic disorders (Verman and Babu 1995). Historically, chromosomal abnormalities have been examined using Giemsa banding approach. However, there are marker chromosomes or structural abnormalities that cannot be deciphered by G-banding. To mitigate this problem, in the mid 90s Multicolor or Multiplex Fluorescence In Situ Hybridization, or M-FISH, was developed for rapid and high resolution chromosomal analysis. M-FISH images are captured using a fluorescent microscope. In this method, chromosomes are labeled with multiple dyes or colors. These dyes are chromosome-specific so that each type of chromosome appears in a different color (Speicher 1996)(Castleman 1996). M-FISH uses a five color dye template to produce a multi-spectral image. A sixth flouorescent DNA stain called DAPI is also utilized to display all the chromosome images. Thus an M-FISH image set consists of six images. The gray levels of each image, also known as color karyotyping, can be used as features for chromosomal classification.

In this paper we describe a statistical method for classification of M-FISH images called Gaussian mixture models, or GMM. The inputs are M-FISH chromosome images, and the target classes consist of 22 asexual chromosomes (autosomes) and two sex chromosomes. The rest of this paper is organized as follows. In sections II and III the GMM model and a method for its application to M-FISH classification are explained. Section IV presents our experimental results on the well-established ADIR database, and section V concludes this paper.

II. PIXEL-BY-PIXEL GMM CLASSIFIER

In this study, M-FISH chromosome labeling problem is considered as a pixel-by-pixel statistical classification task with six image channels as features. Each pixel is assigned a 6-tuple feature vector whose elements are the intensities of the corresponding color channel of the M-FISH image. Given a manually pre-classified set of images, a class label is assigned to each pixel (training data feature extraction).

The suggested Gaussian Mixture Model (GMM) is a weighted sum of Gaussian probability density functions (PDF), which is capable of providing a more flexible and thus possibly a more accurate model of the M-FISH image data. The GMM PDF is defined as

$$p(x;\theta) = \sum_{c=1}^{C} \alpha_c G(x, \mu_c, \Sigma_c) \tag{1}$$

where μ is the mean vector, Σ is the covariance matrix, α_c is the weight of component c, and G is a Gaussian PDF defined as:

$$G(\mathbf{x}, \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^D \,|\Sigma|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \tag{2}$$

In theory, with a sufficient number of components, any continuous density function can be approximated with arbitrarily high accuracy.

Maximum Likelihood Estimation

One way of constructing a classifier is through conditional probability density functions. In such a case, initial class model selection can be performed using training data. However, adjustment of model parameters requires some measure of goodness, i.e., how well a proposed class probability distribution fits the observed data, which will be described next.

Assume that there is a set of independent samples X = {x_1,…, x_n} drawn from a single distribution described by a probability density function $p(x; \theta)$, where θ is the PDF parameter list. The likelihood function

$$L(X;\theta) = \prod_{n=1}^{N} p(x_n;\theta) \tag{3}$$

yields the likelihood of the data X given a distribution described by parameters θ. The goal is to find a set of distribution-defining parameters $\hat{\theta}$ that maximizes the data likelihood

$$\hat{\theta} = \operatorname{argmax} \mathrm{L}(\mathrm{X};\theta) \tag{4}$$

Equivalently, we can maximize the logarithm of L, $\ln L(X; \theta) = \Sigma \ln p(X; \theta)$, which is called the log-likelihood function and is easier to compute.

EM Estimation

Expectation Maximization (EM) algorithm is an iterative method for calculating the maximum likelihood distribution parameters (Dempster 1977). It is usually used when an analytical solution is not feasible, such as the case for Gaussian mixtures with unknown means and covariances. The following is a brief description of the EM algorithm.

Given a training set with known and unknown features, denote all the known features by X and unknown features by Y. The expectation step of EM finds the following

$$Q(\theta;\theta^i) \equiv E_Y[\ln L(X,Y;\theta \mid X;\theta^i] \quad (5)$$

where θ^i is a previous estimate of PDF parameters, and θ describes an updated estimate. L is the likelihood function given in (4). Q yields the likelihood of the data, including the unknown feature Y, given the current estimate of the distribution as described by θ^i. The maximization step is to maximize $Q(\theta; \theta^i)$ with respect to

$$\theta^{i+1} \leftarrow \arg\max Q(\theta;\theta^i) \quad (6)$$

Log-likelihoods can be processed using the aforementioned EM algorithm. One needs to repeat the above steps until a convergence criterion for the log-likelihood, such as a limit in the incremental change of parameters, is met (Duda 2001). For EM covariance components, we used K-means clustering and covariances of the entire data set. Weights were set to one.

The known data X is interpreted as incomplete data. The missing part Y is the knowledge of which component produced each sample X_n. For each X_n there is a binary vector $y_n = \{y_{n,1},\ldots, y_{n,c}\}$, where $y_{n,c} = 1$ if the sample n is produced by component c, and zero otherwise. The complete log-likelihood can then be written as

$$\ln L(X,Y;\theta) = \sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c} \ln(\alpha_c p(x_n \mid c;\theta)) \quad (7)$$

The expectation step is to compute the conditional expectation of the complete data log-likelihood, i.e. the Q-function, given X and the current estimate θ^i of the parameters. Since the complete data log-likelihood $\ln L(X, Y; \theta)$ is a linear function of the missing Y, the conditional expectation $\ln L(X, Y; \theta)$ can be derived and substituted into $W \equiv E[Y \mid X, \theta]$. Therefore

$$Q(\theta,\theta^i) \equiv E[\ln L(X,Y;\theta) \mid X,\theta^i] = \ln L(X,W;\theta) \quad (8)$$

where the elements of W are defined as

$$w_{n,c} \equiv E[y_{n,c} \mid X,\theta^i] = \Pr[y_{n,c} = 1 \mid x_n,\theta^i] \quad (9)$$

The probability can be calculated using the Bayes law

$$w_{n,c} = \frac{p_c^i\, p(x_n \mid c;\theta^i)}{\sum_{j=1}^{c} \alpha_j^i\, p(x_n \mid j;\theta^i)} \quad (10)$$

where p^i_c is the a priori probability of θ^i estimate and $w_{n,c}$ is the a posteriori probability when $y_{n,c} = 1$ given x_n. In other words, $w_{n,c}$ is the probability that x_n is produced by component c (Figueiredo and Jain 2002).

The maximization step is used next to estimate distribution parameters for the c-component mixture of Gaussians with arbitrary covariances such that

$$p_c^{i+1} = \tfrac{1}{N}\sum_{n=1}^{N} w_{n,c} \quad (11)$$

$$\mu_c^{i+1} = \frac{\sum_{n=1}^{N} x_n w_{n,c}}{\sum_{n=1}^{N} w_{n,c}} \quad (12)$$

$$\Sigma_c^{i+1} = \frac{\sum_{n=1}^{N} w_{n,c}(x_n - \mu_c^{i+1})(x_n - \mu_c^{i+1})^T}{\sum_{n=1}^{N} w_{n,c}} \quad (13)$$

The above new estimates are used to find θ^{i+1}. If a convergence criterion for equations (12) or (13) is not satisfied, the iteration continues with $i \leftarrow i + 1$; and equations (11) through (13) are re-evaluated with new estimates (Duda 2001).

III. METHODOLOGY

We used the supervised classification method described in Section II for detection of human chromosomes from M-FISH images. As mentioned earlier, the 46 human chromosomes consist of 22 pairs of homologous autosomes (sex-neutral chromosomes with similar evolutionary origins but different functions), and 2 sex-determinative chromosomes, constituting to 24 types (classes). We considered image background and chromosomes' overlaps as two further classes. Since an M-FISH set consists of six image channels, each pixel was represented by six features with each feature being the gray-scale value of the corresponding color channel. Therefore, a 6-feature, 26 class GMM classifier was used to perform a pixel by pixel classification. The data was obtained through our collaboration with Advanced Digital Imaging Research (ADIR) (http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml). An example of an M-FISH image set is shown in Fig 1.

The ADIR database has manually segmented and classified images, which can be used for training and testing purposes. An example of the class-map is shown in Fig 2 (a). We used these manually processed images as ground truth to determine the accuracy of our classification technique. We trained and tested our GMM classifier on six channel M-FISH image data sets.

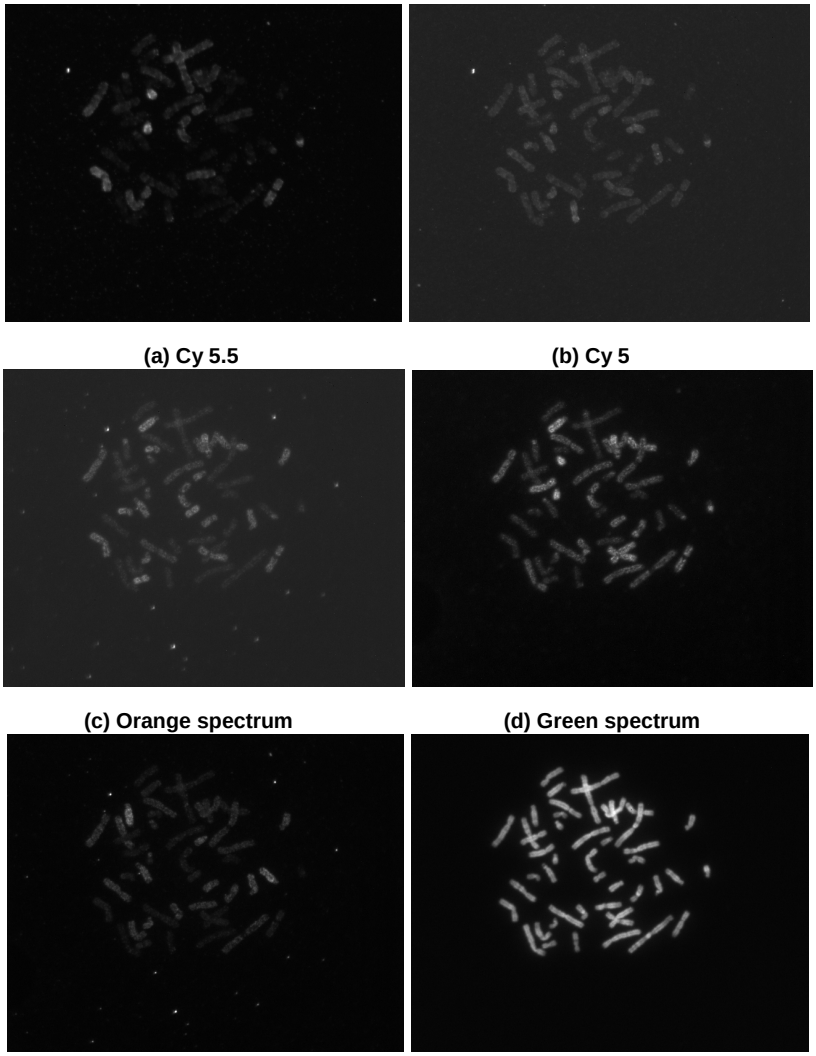
IV. CLASSIFICATION RESULTS

A total of six M-FISH sets of images were classified. Each set has 333,465 pixels. A class-map was generated for each classification output. A pseudo-coloring scheme was used to represent each chromosome class in an image.

We calculated the overall accuracy of our approach by comparing the aforementioned class-map to the ground-truth provided in the database. Table I shows the classification accuracy obtained for each M-FISH set, as well as a comparison with the naive Bayesian classifier presented in (Wang and Castleman 2005). Fig 2 shows the class-map along with the results of classifiers using one and two mixtures GMMs for M-FISH A0502XY.

V. CONCLUSIONS

This paper introduced a classification method for M-FISH images. Our GMM pixel-by-pixel classification of six input features with 26 output classes provided an overall classification accuracy of 89.18%, with a noticeable improvement over the Bayesian classifier reported in (Wang and Castleman 2005). The promising results of this method can eventually translate into improved accuracy of M-FISH technique for cancer and other relevant disease diagnosis.



(a) Cy 5.5 (b) Cy 5

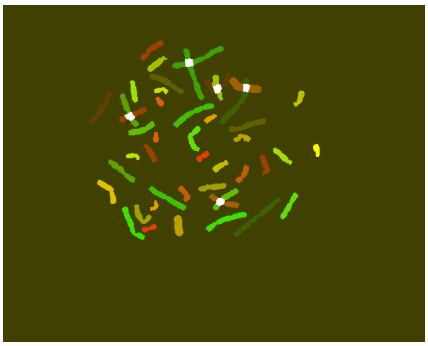
(c) Orange spectrum (d) Green spectrum

(e) Texas Red (f) DAPI

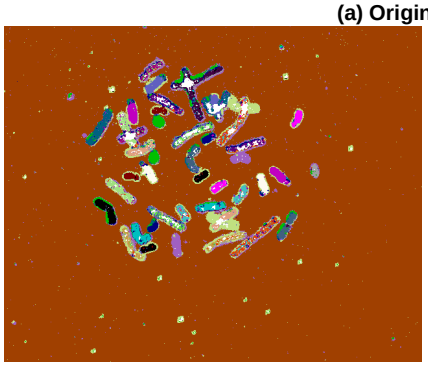
Figure 1: A set of M-FISH images for A0502XY

TABLE I
Pixel-by-Pixel Classification Rate Percentage for Sample M-Fish Images

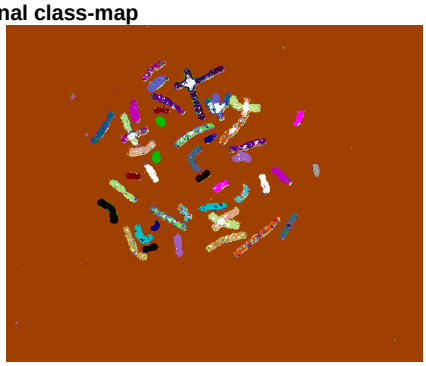
Training/Test Image Set	Bayesian	GMM
V1301	49.2528	85.4887
V1303	70.2981	88.5952
V1304	47.9901	85.3376
V1306	87.7719	90.2275
V1308	56.1503	93.4524
V1309	50.7460	91.9689
Average	**60.3682**	**89.1784**



(a) Original class-map



(b) Detected using one-mixture GMM



(c) Detected using two-mixtures GMM

Figure 2: Classification results for M-FISH image A0502XY

REFERENCES

Castleman, K. R., Riopka, T. P., and Wu, Q., 1996, "FISH image analysis," *IEEE Engineering in Medicine and Biology*, Vol. 15, pp. 67-75.

Dempster A., Larid N., and Rubin D., 1997, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, Vol. 39, pp. 1–38.

Duda, R. O., Hart, P. E., and Stork, D. G. ,2001, *Pattern Classification*, John Wiley & Sons Inc., New York.

Figueiredo, M. A. T., and Jain, A. K., 2002, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 24, pp. 381–396.

Speicher, M. R., Ballard, S. G., and Ward, D. C., 1996, "Karyotyping human chromosomes by combinatorial multi-fluor FISH," *Nature Genetics*, Vol. 12, pp. 368–375.

Verma, R. S., and Babu, A., 1995, *Human Chromosomes: Principles and Techniques*, McGraw-Hill Inc., New York.

Wang, Y., and Castleman, K., 2005, "Automated Registration of Multi-Color Fluorescence In Situ Hybridization (M-FISH) Images for Improving Color Karyotyping," *Cytometry*, Part A, 64A.