

Gaussian Mixture Model (GMM)
classification of
Multi-Color Fluorescence In Situ Hybridization (M-FISH)
Images

ECE590B – Biomedical Imaging
Project Report
May 2006

By:
Amin Fazel

Abstract

This report describes a fully automatic chromosome classification algorithm for Mul-tiplex Fluorescence In-Situ Hybridization (M-FISH) images using Gaussian mixture model technique. M-FISH is a recently developed chromosome imaging method in which each chromosome is labeled with 5 dyes and is also counterstained with DAPI. The classification problem is modeled as a 26-class 6-feature pixel-by-pixel classification task. The 26 classes are the 24 types of human chromosomes, the background and chromosome's overlap, while the six features correspond to the brightness of the dyes at each pixel. Experiments conducted on the ADIR M-FISH database demonstrate the superior performance of the proposed approach.

I. INTRODUCTION

Human chromosomes are the body's information carriers. They are formed from DNA molecules and all of the data necessary for an organism's development and maintenance are stored in them. They contain vast amounts of information; in fact, every cell in a normal human being contains 46 chromosomes. Thus by examining of chromosome images one can obtain important information about health of human beings [1]. These images are useful for cancer diagnosis and genetic disorders. In the past these images were examined visually by tedious manual processes of locating, classifying, and evaluating the chromosomes. In the mid-1990s, Multiplex fluorescence in situ hybridization (M-FISH) was developed for chromosome analysis as a new technology. M-FISH images are captured with a fluorescent microscope. This method dyes chromosomes with multiple colors in which each chromosome type appeared as a different color [2,3]. M-FISH uses five color dyes that attach to various chromosomes differently to produce a multi-spectral image, and a sixth dye that attaches to all chromosomes to produce a grayscale image. Thus it is possible to visualize improved method for the location, segmentation and classification of chromosome images by taking advantage of the color information in M-FISH images. An M-FISH set consists of six images, so each pixel in the image can be viewed as a six feature vector.

This report addresses the automatic classification of M-FISH chromosome images into 24 chromosome types (22 autosomes and 2 sex chromosomes) by using the Gaussian mixture models (GMM). In Sections II and III the GMM Classifier and the methodology are discussed respectively. In Section IV the results are presented.

II. GMM PIXEL-BY-PIXEL CLASSIFIER

Since here we have done a pixel-by-pixel classification with six image channels, in order to classify a pixel one must take into account the gray level value of that pixel from all those channels. We will compute a six dimensional feature vector from and assign classes to each pixel. Gaussian mixture model is a weighted sum of Gaussian probability density functions which are referred to as Gaussian components of the mixture model describing a class; therefore represent different sub classes inside one class. The probability density function is defined as a weighted sum of Gaussians

$$P(x; \mathbf{q}) = \sum_{c=1}^C \mathbf{a}_c G(x, \mathbf{m}_c, \Sigma_c) \quad (1)$$

Where \mathbf{m} is the mean vector, Σ is the covariance matrix, \mathbf{a}_c is the weight of the component c , and $G(x, \mathbf{m}_c, \Sigma_c)$ is the Gaussian probability density function which is defined as:

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

It is clear that modeling a pixel class with a GMM rather than a single Gaussian gives a great deal of added flexibility to the model. Indeed, if one is allowed an arbitrary number of components, any continuous density function can be approximated to any desired accuracy.

Maximum Likelihood Estimation

In construction of a classifier the class-conditional probability density functions need to be determined. The initial model selection can be done for example by visualizing the training data, but the adjustment of the model parameters requires some measure of goodness, i.e., how well the distribution fits the observed data. Assume that there is a set of independent samples $X = \{x_1, \dots, x_n\}$ drawn from a single distribution described by a probability density function $p(x; \mathbf{q})$ where \mathbf{q} is the PDF parameter list. The likelihood function

$$L(X; \mathbf{q}) = \prod_{n=1}^N P(x_n; \mathbf{q}) \quad (3)$$

tells the likelihood of the data X given the distribution or, more specifically, given the distribution parameters \mathbf{q} . The goal is to find $\hat{\mathbf{q}}$ that maximizes the likelihood:

$$\hat{\mathbf{q}} = \operatorname{argmax} L(X; \mathbf{q}) \quad (4)$$

Usually this function is not maximized directly but the logarithm

$$\ln L(X; \mathbf{q}) = \sum_{n=1}^N \ln p(x_n; \mathbf{q}) \quad (5)$$

called the log-likelihood function which is analytically easier to handle.

EM Estimation

The expectation maximization (EM) algorithm is an iterative method for calculating maximum likelihood distribution parameter [4]. It can also be used to handle cases where an analytical approach for maximum likelihood estimation is infeasible, such as Gaussian mixtures with unknown and unrestricted covariance matrices and means. Assume that each training sample contains known features and unknown features. Mark all good features of all samples with X and all unknown features of all samples with Y . The expectation step (E-step) for the EM algorithm is to form the function

$$Q(\mathbf{q}; \mathbf{q}^i) \equiv E_Y [\ln L(X, Y; \mathbf{q} | X; \mathbf{q}^i)] \quad (6)$$

where \mathbf{q}^i is the previous estimate for the distribution parameters and \mathbf{q} is the variable for a new estimate describing the (full) distribution. L is the likelihood function. The function calculates the likelihood of the data, including the unknown feature Y marginalized with respect to the current estimate of the distribution described by \mathbf{q}^i . The maximization step (M-step) is to maximize $Q(\mathbf{q}; \mathbf{q}^i)$ with respect to \mathbf{q} and set

$$\mathbf{q}^{i+1} \leftarrow \arg \max Q(\mathbf{q}; \mathbf{q}^i) \quad (7)$$

The steps are repeated until a convergence criterion is met [5] and iterations are stopped when the change in the values falls below a threshold.

The EM algorithm starts from an initial guess for the distribution parameters and the log-likelihood is guaranteed to increase on each iteration until it converges [5]. K-means clustering was used for initializing the component means. Component covariance matrices are initialized to diagonal part of the whole data covariance matrix. Weights are set to equal 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 $y_n = \{y_{n,1}, \dots, y_{n,C}\}$, where $y_{n,c} = 1$, if the sample was produced by the component c , or zero otherwise. The complete data log-likelihood is

$$\ln L(X, Y; \mathbf{q}) = \sum_{n=1}^N \sum_{c=1}^C y_{n,c} \ln(\mathbf{a}_c p(x_n | c; \mathbf{q})) \quad (8)$$

The E-step is to compute the conditional expectation of the complete data log-likelihood, the Q-function, given X and the current estimate \mathbf{q}^i of the parameters. Since the

complete data log-likelihood $\ln L(X, Y; \mathbf{q})$ is linear with respect to the missing Y , the conditional expectation $\ln L(X, Y; \mathbf{q}) = \sum_{n=1}^N \sum_{c=1}^C y_{n,c} \ln(\mathbf{a}_c p(x_n | c; \mathbf{q}))$ has simply to be computed and put it into $\ln L(X, Y; \mathbf{q})$.

Therefore

$$Q(\mathbf{q}, \mathbf{q}^i) \equiv E[\ln L(X, Y; \mathbf{q}) | X, \mathbf{q}^i] = \ln L(X, W; \mathbf{q}) \quad (9)$$

where the elements of W are defined as

$$w_{n,c} \equiv E[y_{n,c} | X, \mathbf{q}^i] = \Pr[y_{n,c} = 1 | x_n, \mathbf{q}^i] \quad (10)$$

The probability can be calculated with the Bayes law

$$w_{n,c} = \frac{p_c^i p(x_n | c; \mathbf{q}^i)}{\sum_{j=1}^c \mathbf{a}_j^i p(x_n | j; \mathbf{q}^i)} \quad (11)$$

where p_c^i is the a priori probability (of estimate \mathbf{q}^i) and $w_{n,c}$ is the a posteriori probability that $y_{n,c} = 1$ after observing x_n . In other words, $w_{n,c}$ is the probability that x_n was produced by component c [6].

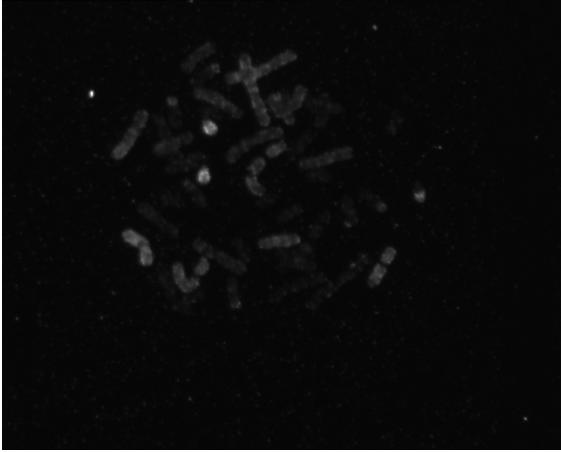
Applying the M-step to the problem of estimating the distribution parameters for C-component Gaussian mixture with arbitrary covariance matrices, the resulting iteration formulas are as follows:

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

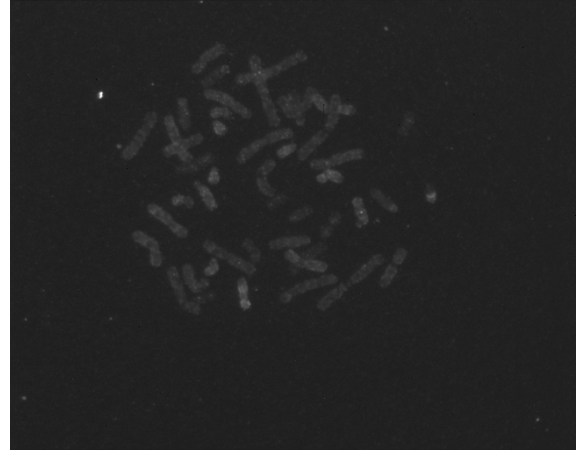
$$\mathbf{m}_c^{i+1} = \frac{\sum_{n=1}^N x_n w_{n,c}}{\sum_{n=1}^N w_{n,c}} \quad (13)$$

$$\Sigma_c^{i+1} = \frac{\sum_{n=1}^N w_{n,c} (x_n - \mathbf{m}_c^{i+1})(x_n - \mathbf{m}_c^{i+1})^T}{\sum_{n=1}^N w_{n,c}} \quad (14)$$

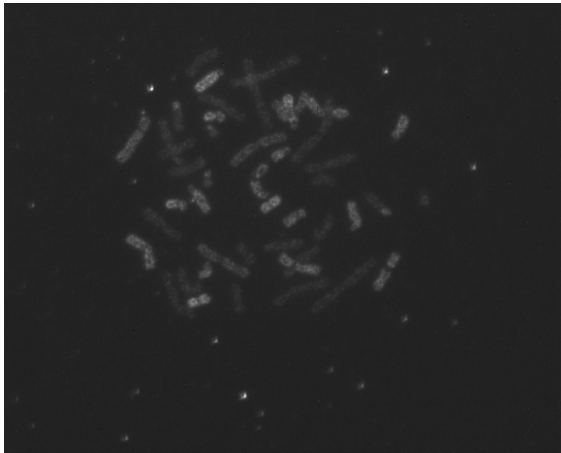
The new estimates are gathered to \mathbf{q}^{i+1} . If the convergence criterion (Eqs. 12 or 13) is not satisfied, $i \leftarrow i+1$ and Eqs. 11–14 are evaluated again with new estimates [5].



(a) Cy 5.5



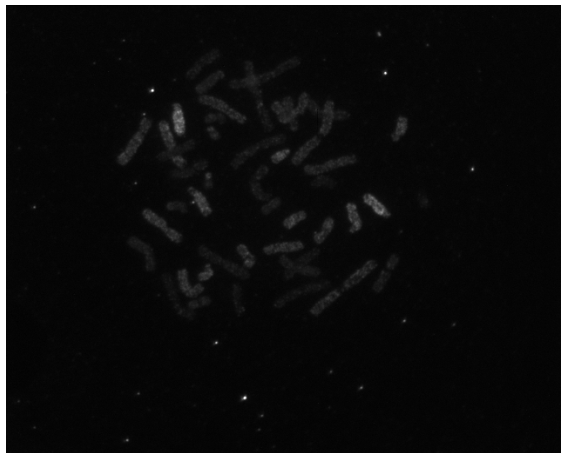
(b) Cy 5



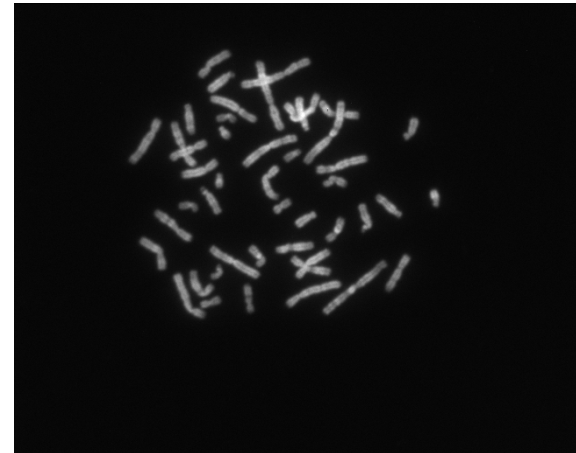
(c) Spectrum Orange



(d) Spectrum Green



(e) Texas Red



(f) DAPI

Figure 1- A set of M-FISH image - A0502XY

III. METHODOLOGY

The supervised methods described in Section 2 were used for classification. Our objective is to classify human chromosomes. The 46 human chromosomes consist of 22 pairs of similar, homologous and 2 sex-determinative chromosomes. These represent 24 classes. The background and chromosome's overlap were included as two classes. Since an M-FISH set consists of 6 image channels, each pixel can be represented by six features, which are the gray-scale values in the six color channels (five labels and DAPI counterstain). These channels is shown in Fig 1. Therefore, a 6-feature 26 class GMM classifier was used to do a pixel by pixel classification. The images for training and testing were selected from a public database of hand segmented M-FISH images. This database is made available online by Advanced Digital Imaging Research and can be accessed at: http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml

For each set of M-FISH images the database also contains a labeled class-map image in which each pixel is labeled according to the class to which it actually belongs. Such a class-map is show in Fig 2 (a). This image was used to determine the accuracy of the classification techniques. The classifier was trained on set of six M-FISH images and also tested on these sets.

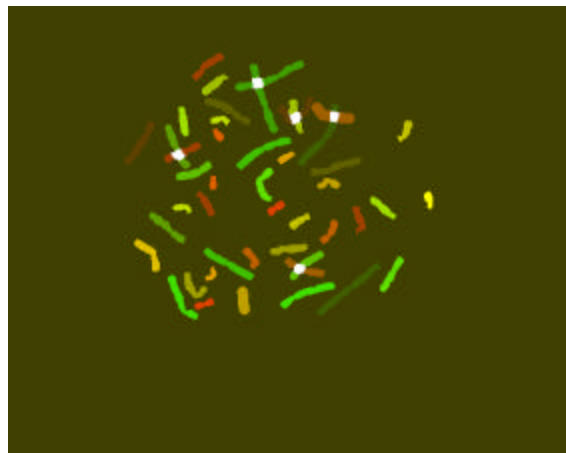
IV. EXPERIMENTAL RESULTS

We have done an experiment using the above classification method to evaluate the performance of this method, using the ADIR M-FISH database. Six M-FISH image sets were classified in which each set has 333,465 pixels. A class-map was generated for each classification output. A separate color was used to represent each chromosome class in the image. The overall accuracy was computed by comparing this class-map to the class-map provided in the database. Tables 1 shows the classification accuracy obtained for each M-FISH set and compared to other classification method which are presented in [7]. An actual class-map and the results of classifier using one and two mixtures in GMM are shown in Fig 2 (a), Fig 2 (b), and Fig 2 (c) for M-FISH A0502XY respectively.

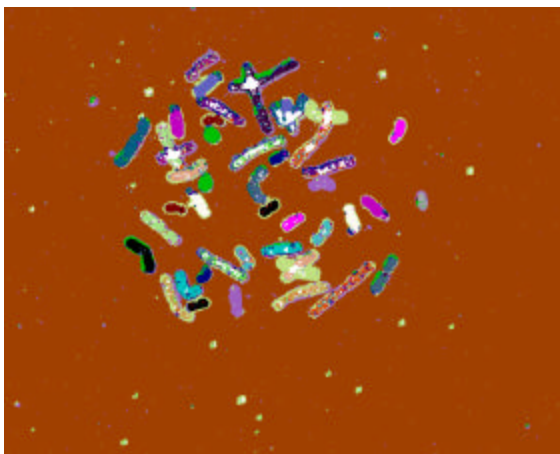
Table 1

Pixel-by-Pixel Classification Rate (%) of M-FISH Image sets

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) Classified class-map using one mixture in GMM (c) Classified class-map using 2 mixtures in GMM

Figure 2- classification result for M-FISH A0502XY

CONCLUSION

This report introduced a classification method based on Gaussian mixture models for multispectral FISH images and showed that the promising classification accuracies can be achieved with this method. The pixel-by-pixel classification task is modeled as 6-feature, 26-class classification problem. The overall classification accuracy achieved is %89.18. This algorithm leads to improve classification rate, and hence improves the accuracy of M-FISH technique for cancer diagnosis and other diseases.

References

- 1 R. S. Verma, A. Babu, Human Chromosomes: Principles and Techniques, 2nd Edition, McGraw-Hill, Inc., 1995.
- 2 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.
- 3 K. R. Castleman, T. P. Riopka, and Q. Wu, "FISH image analysis," *IEEE Engineering in Medicine and Biology*, vol. 15, pp. 67-75, 1996.
- 4 A. Dempster, N. Larid, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, pp. 1–38, 1977.
- 5 R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001.
- 6 M.A.T. Figueiredo and A.K. Jain. "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), pp. 381–396, March 2002.
- 7 Yu-Ping Wang and Ken Castleman, "Automated Registration of Multi-Color Fluorescence In Situ Hybridization (M-FISH) Images for Improving Color Karyotyping," *Cytometry, Part A*, 64A (2), April, 2005.