Joint Segmentation and Classification of M-FISH Chromosome Images

Hyohoon Choi[1], Kenneth R. Castleman[2], Alan C. Bovik[3]
[1]Department of Biomedical Engineering, University of Texas at Austin, TX USA
[2]Advanced Digital Imaging Research, League City, TX USA
[3]Department of Electrical and Computer Engineering, University of Texas at Austin, TX USA

Abstract—**Automatic segmentation and classification of M-FISH chromosome images are jointly performed using a six-feature, 25-class maximum-likelihood classifier. Preprocessing of the images including background correction and six-channel color compensation method are introduced. A feature transformation method, spherical coordinate transformation, is introduced. High correct classification results are obtained.**

Keywords—**Chromosome, Classification, Maximum-likelihood, M-FISH, Segmentation**

I. Introduction

Mulitiplex in-situ hybridization (M-FISH) is a combinatorial labeling technique used for chromosome analysis. To be able to distinguish 24 human chromosomes – 22 somatic chromosomes, and X and Y sex chromosomes, 5 fluorophores are used. An extra fluorophore, DAPI, is counter stained to all chromosomes. Thus six images of corresponding fluorophores are captured per metaphase spread. M-FISH has been utilized and proven to be useful for clinical cytogenetics and cancer research. However, currently available systems still exhibit misclassifications of multiple pixel regions due to number of factors including non-homogeneity of staining, variations of intensity levels within and between image sets, and emission spectra overlaps between fluorophores. Current methods of classifying chromosome pixels involve a manual or semi-automatic image segmentation of the DAPI channel. A maximum-likelihood approach was previously studied without appropriate preprocessing of the images [5], and assumed perfect segmentation for excluding background pixels for classification using the information obtained from the ground truth. To be fully automatic, we suggest a Bayesian rule based statistical classification method that simultaneously performs the segmentation and classification of M-FISH images. We have constructed a six-feature, 25-class maximum-likelihood classifier. The 25 classes are 24 chromosomes plus the background, and the six features are the six color channel intensities. The classification can be done either parametrically or non-parametrically. Both have merits and demerits in terms of classification speed and image processing complexities associated with it. However, it is natural to approach the problem with a supervised parametric method when the characteristics of classes can be well studied.

The maximum-likelihood classifier requires training and testing of the classifier. Since pixel intensities of the images are used as features, noise and variations between sets of images affect the classification. Thus, the preprocessing of images was performed to reduce the noise and the variations. One of the variations comes from the uneven background surface. Background surface is brighter near the chromosomes due to the flair from the chromosomes and has different DC offsets for different channels. This problem should be corrected in cases where the absolute intensity is of importance. Spectral overlap between six color channels introduces another type of noise. Due to the spectral overlap, intensity residuals appear on the channels where chromosomes are not stained. This phenomenon is referred as color spread. The color spread for M-FISH images may eventually affect the classification result. Color spread can be corrected if the color spread matrix is found, which contains information about how much of a specific color spread to the other colors. In M-FISH case, the color spread matrix is a 6×6 matrix, and we have computed a color spread matrix automatically by means of optimization from the measured images. The random white noise can be reduced by median or lowpass filtering. The normalization to reduce the intensity variations between the same channel images of different spreads is one of the crucial parts for a parametric classifier. It is almost true that the distributions of images preserve a similar pattern of bimodal Gaussian with the variations of mean values of the Gaussians. The means of Gaussians between the same channel images were approximately aligned by linear histogram stretching to a fixed range. Misclassifications usually occur at regions where chromosomes overlap, around chromosome perimeters where intensity is weaker, and pixels of centromeres and telomeres where the staining is weak. To compensate the uneven hybridization and to make a chromosome surface more homogeneous and at the same time to preserve the sharp edge of the chromosome, the effect of anisotropic filtering was studied. The intensity variability among images causes misclassification when it is not well adjusted. Thus instead of directly utilizing the pixel intensities as features, a feature transformation method, spherical coordinate transformation, was utilizes. This method utilizes the angle information of 6 dimensional sample values. Angles can be more robust to the intensity variations. Once the noise and image variations are reduced, images are ready to be trained and tested.

In our work, 10 sets of images from a slide were divided into two groups, five of them for training and the rest for testing. Training and testing were performed with several different settings: 1. No pre-processing, 2. Background and color correction, median filtering, and normalization, 3.

Background and color correction, median and anisotropic filtering, and normalization, and 4. Background and color correction, median filtering, normalization, and spherical coordinate transformation of the features.

Overall the joint maximum-likelihood segmentation and classification method yields a high performance of correct classification and segmentation.

II. Methodology

1. Pre-processing

A. Background Correction

The background intensity near chromosome cluster area is usually more elevated than that of areas far away from the chromosome cluster mainly because of the flair effects of the chromosomes. This undesired non-flat intensity distribution of background hinders further processing of the images, and eventually affects the classification. The background surface contains the auto-fluorescence, DC offset, flairs from the chromosomes, and other noise that contribute to the background intensity. The two dimensional cubic surface was estimated by pixels from the estimated background area [2]. The surface that has the minimum mean square error with the observed background pixel values was the estimated two-dimensional cubic surface. Then the cubic surface was subtracted from the image. Thus it removes the above mentioned noises. The background area was estimated by thresholding the DAPI channel. The thresholding was performed automatically using the iterative bimodal threshold method with a higher prior given to the lower intensity level so that the chromosomes are safely excluded (Fig. 1B). The iterative bimodal threshold method works similar to K-means clustering, where K = 2. It clusters pixels into two groups iteratively until the means for the two groups do not change, and the decision boundary between two classes is the threshold. The priors for the two clusters can be adjusted depending on the distributions of the clusters. Thus this thresholding method works well when the actual intensity distribution is bimodal. The intensity distribution of M-FISH chromosome images is bimodal with higher *a priori* probability for the background, and the distributions are consistent for all images. For the training, the true classification maps were used to estimate the background area. However, the map includes only the chromosomes and excludes high intensity noise contents such as cells. Thus an automatic thresholded image was included in the map by taking the binary OR operation between them. Note that the precise estimation of background area is not necessary to calculate the general structure of the background cubic surface. The iterative bimodal threshold method with a high prior to the background is sufficient to estimate the background area. For testing, the thresholded image was dilated by a 3 × 3 structuring element. After background was corrected, the spectral overlap between color channels was corrected.

B. Color Compensation

Chromosomes are stained with multiple combinations of fluorophores in M-FISH. Ideally if a chromosome was stained with only two fluorophores, for example, then those two corresponding channels should light up whereas the other channels display zeros. Due to the overlapping spectra of filters of the acquisition system and the overlapping of emission spectra of florescent dyes, color spreading between the spectral channels commonly occurs. The color spreading can be corrected by a linear transformation [2]. The color spread matrix, which describes quantitative ratios of the spectral overlaps between channels, has to be computed. The color spread matrix can be found with a unique solution when each object is stained with a unique fluorophore. However, when no single object is stained with a unique fluorophore, such as in M-FISH, the color spread matrix may not be a unique solution. Thus we have computed the optimal solution by minimizing the mean-square error. Let the measured signal, Y, be a vector of 6 × 1, which has the spectral overlaps and some background elevation due to auto-fluorescence and DC offset. Then $Y = CX + b$. C is the 6 × 6 color spread matrix. b will be removed after the background correction. Thus the true pixel value can be estimated as $X = C^{-1}Y$. The inverse color spread matrix C^{-1} can be found by minimizing the differences between X values and the $C^{-1}Y$ values. In our case, we have 24 objects with 6 fluorophores. There are 144 equations with 104 unknowns after assigning zeros for Xs where chromosomes are not stained. However, this method is sensitive to the initial guess values. Instead when X values are pre-assigned with for example *0*s and *255*s for no fluorophores and fluorophores respectively, the number of unknowns becomes 36. X values per class were assigned in our method as

$$X_k = \frac{Y_k}{\sum Y_k} \sum Y_i + Y_k \quad X_i = 0 \tag{1}$$

where i is the index of no fluorophores and k is the index of fluorophores. This method converges to a solution and finds the color spread matrix C.

C. Filtering and Normalization

The median filtering, which effectively removes the shot noise and reduces the additive Gaussian noise, was applied on the color compensated images. The anisotropic diffusion filter, which diffuses intensities more where the gradient is relatively small and diffuses less where the gradient is large, was followed to reduce the intensity variation inside the chromosome and to keep the edges sharp. Depending on the integration time of image acquisition and the intensity of fluorophores the chromosome brightness can be different among channels or spreads. This can lead to the misclassification of a whole chromosome to another chromosome. Thus the image intensity was normalized from

0 to 1 so that all the images display approximately the same intensity.

2. Spherical Coordinate Transformation

The intensity variation between spreads is one of the factors that leads to misclassifications. The normalization processes try to reduce the intensity variations, but may not be perfect. Instead of applying intensities as features directly, the angles of 6 feature vectors may be more robust to the intensity variation. Assuming that the samples are Gaussian distributed, samples of a class form a cluster on the 6-dimensional space with a mean and a variance. This cluster also can be viewed as being distributed inside a cone starting from the origin toward to the cluster. If the variance perpendicular to the mean vector is small i.e. the angle of the cone is fairly narrow and the overlaps between the cones are small, then the angles of the samples will have more discrimination power between classes. Thus, the following spherical coordinate transformation was applied converting the 6-dimensional Cartesian coordinates $(x, y, z, \xi, \psi, \zeta)$ to $(\alpha, \beta, \chi, \delta, \varepsilon, R)$ as

$$\alpha = \cos^{-1}(y / \sqrt{(x^2 + y^2)}) \quad (3)$$

$$\beta = \cos^{-1}(z / \sqrt{(x^2 + y^2 + z^2)}) \quad (4)$$

$$\chi = \cos^{-1}(\xi / \sqrt{(x^2 + y^2 + z^2 + \xi^2)}) \quad (5)$$

$$\delta = \cos^{-1}(\psi / \sqrt{(x^2 + y^2 + z^2 + \xi^2 + \psi^2)}) \quad (6)$$

$$\varepsilon = \cos^{-1}(\zeta / \sqrt{(x^2 + y^2 + z^2 + \xi^2 + \psi^2 + \zeta^2)}) \quad (7)$$

$$R = \sqrt{(x^2 + y^2 + z^2 + \xi^2 + \psi^2 + \zeta^2)} \quad (8)$$

(3)~(7) are the angles and (8) is the length of the vector. All new 6 features are used. The classification results are shown in Table 1.

3. Maximum-likelihood Classification

The M-FISH database available on ADIR's website was used for this study [4]. The database contains 200 images with ground truth. 10 sets from the same slide were used for the study. 5 sets of images were used for training and 5 sets of images were used for the testing. Both training and testing image sets were preprocessed in the same manner. After preprocessing the images, class parameters of 25 mean vectors and 25 covariance matrixes were calculated from the samples of 6 features. The multivariate normal density in d dimensions is expressed as

$$p(x \mid \omega_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_i)^t \Sigma_i^{-1} (x - \mu_i) \right], \quad (2)$$

where **x** is a $d \times 1$ sample vector, μ is the $d \times 1$ mean vector, Σ_i is the $d \times d$ covariance matrix, and $|\Sigma_i|$ and Σ_i^{-1} are its determinant and inverse, respectively [3]. Each mean vector and covariance matrix characterize each class.

Once the class parameters were calculated from the training set, each new image set was introduced to the classifier after performing the same pre-processing as done in training. Each pixel on the 6 channel images is classified using Bayes rule, which is $P(\omega_i \mid x) = \frac{p(x \mid \omega_i) P(\omega_i)}{p(x)}$, where $p(x) = \sum_{i=1}^{c} p(x \mid \omega_i) P(\omega_i)$, $P(\omega_i)$ are the *a priori* probabilities, c is the number of classes. A sample **x** belongs to class i when $P(\omega_i \mid x) > P(\omega_j \mid x) \quad i, j = 1 \ldots c \quad j \neq i$. Since $p(x)$ is just a normalizing factor to make a posteriori probability function sum to unity, the selection rule, $P(\omega_i \mid x) > P(\omega_j \mid x)$, can be simplified as $p(x \mid \omega_i) > p(x \mid \omega_j)$. The classifier defined by this decision rule is a maximum-likelihood classifier. For the simplicity, the *a priori* probabilities are assumed to be the same for all classes. These will be adjusted according to the ratios of class populations in the later studies.

III. Results and Discussion

Fig. 1 shows the results of background thresholding, color compensation, and classification results. The classification results are tabulated on Table 1. The correct classification rate for the background increases as more processes are added. In particular, case 4 yields almost perfect background classification. Case 1, 2, and 3 show relatively the same classification accuracy for the chromosomes, but the classified chromosomes are wider than the ground truth while the edges of chromosomes are misclassified as different chromosomes. In case 4, however, classified chromosomes are thinner than the ground truth and chromosome pixels around edge are classified as background, thus the classification rate is lower for the chromosomes. We cannot claim that the provided karyotyping maps are absolutely correct. In particular pixels around the chromosome perimeter are fuzzy, thus it is difficult to draw the clear cut of the boundary. For this reason, quantitative measures can mislead the interpretation of the results. Thus we have to analyze the results both quantitatively and qualitatively. For example, the classified chromosomes may look thinner than the provided map while inside of the chromosomes are correctly classified. This is the case where qualitative measure tells a low correct classification when the actual classification result is almost perfect quantitatively (see Fig. 1H). The background is effectively corrected (Fig. 1B), and the color spread is correctly compensated (Fig. 1C). The classification results with anisotropic filtering and spherical coordinate transformation (Fig. 1G and 1H) show more uniform classification inside chromosomes. Fig. 1G clearly shows that the anisotropic filtering helps reduce the intensity variations inside chromosomes. The difference between Fig.

1F and 1H is that 1F uses the intensity as features and 1H uses the spherical coordinate transformed features as features. Fig. 1H suggests that the spherical coordinate transformed features are more robust to the intensity variation. The background is almost perfectly classified, thus segmented, in all cases.

Misclassified pixels may be reduced using a feedback loop from the classification result, and incorporating the banding pattern, contextual information, and other information. Adjusting *a priori* probability based on those extra information will lead to less classification error.

TABLE 1. Correct Classification Rates.
BG (Background). CHR (Chromosomes). Case 1: No preprocessing, Case 2: Background correction, color compensation, normalization, Case 3: Background correction, color compensation, anisotropic filtering, normalization, and Case 4: Background correction, color compensation, normalization, spherical coordinate transformation.

Image Set	Correct Classification Rate							
	Case 1		Case 2		Case 3		Case 4	
	BG	CHR	BG	CHR	BG	CHR	BG	CHR
1	87.62	84.53	90.58	79.40	92.63	78.45	95.70	75.95
2	97.98	90.31	97.66	90.69	96.64	89.61	99.46	82.43
3	98.43	86.33	98.28	82.64	97.87	82.03	99.59	77.35
4	98.07	92.89	97.86	89.14	97.58	87.40	99.47	83.95
5	97.70	93.09	97.40	93.73	96.83	93.33	99.40	85.52
Ave.	95.96	89.43	96.36	87.12	96.31	86.16	98.72	81.04

IV. CONCLUSION

In this paper we have introduced a novel classification method for M-FISH chromosome images. M-FISH images are jointly segmented and classified with a six-feature, 25-class maximum-likelihood classifier. The preprocessing methods including the background correction and the six-channel color compensation methods are demonstrated. The spherical coordinate transformation method was introduced as a feature transformation technique. The high correct classification results are obtained. The classification accuracy and robustness of the classification will improve as more information such as spatial and contextual data is utilized.

REFERENCES

[1] 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.

[2] K. R. Castleman, *Digital Image Processing*. Upper Saddle River, NJ: Prentice-Hall, 1996.

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

[4] M-FISH Database. Available: http://www.adires.com

[5] W. C. Schwartzkopf, "Maximum Likelihood Techniques for Joint Segmentation-Classification of Multi-spectral Chromosome Images," Ph.D. dissertation, Dept. of Elec. and Comp. Eng., University of Texas at Austin, Texas, USA, 2002.

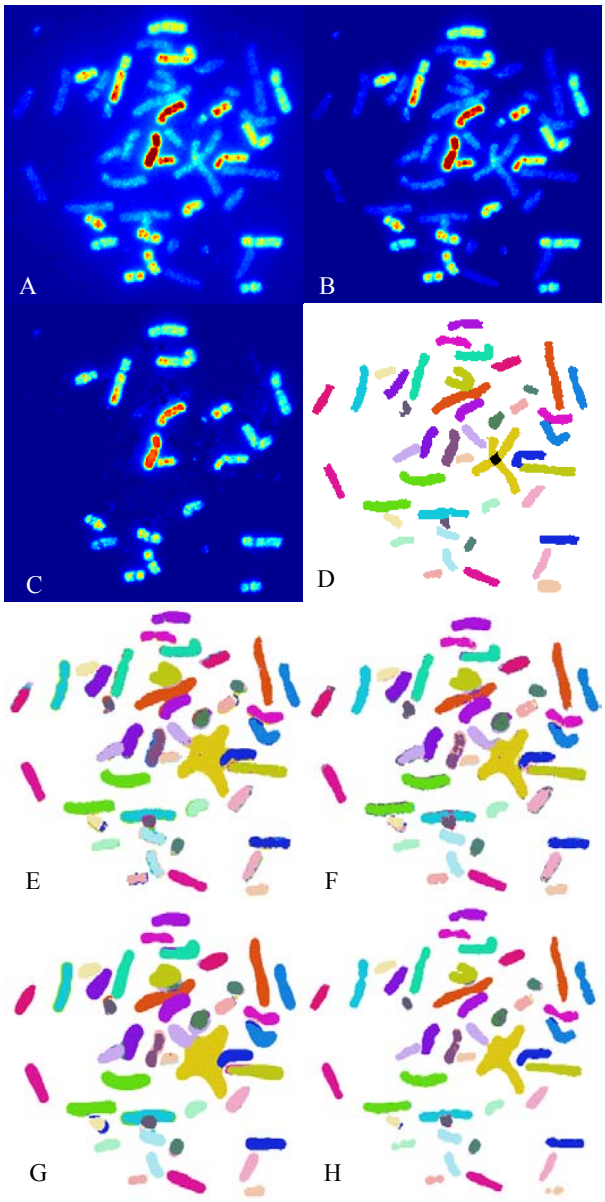

Fig. 1. Preprocessing and classification of a male chromosome image stained with Vysis probe (4 in Table 1). A: Far red channel image before the pre-processing, B: Background correction of A, C: Color compensation of B, D: Karyotype of the chromosome image (a different color is assigned to each class), E: Classification result with no preprocessing, F: Classification result with background and color correction and normalization, G: Classification result with background and color correction, anisotropic filtering, and normalization, and H: Classification result with background and color correction, normalization, and spherical coordinate transformation.