Available online at www.sciencedirect.com
SCIENCE DIRECT®

Pattern Recognition 38 (2005) 1209–1223

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Supervised parametric and non-parametric classification of chromosome images

M.P. Sampat[a], A.C. Bovik[b], J.K. Aggarwal[c,*], K.R. Castleman[d]

[a]*Department of Biomedical Engineering, Laboratory for Image and Video Engineering, The University of Texas at Austin, TX 78712, USA*
[b]*Department of Electrical and Computer Engineering, Laboratory for Image and Video Engineering, The University of Texas at Austin, TX 78712, USA*
[c]*Department of Electrical and Computer Engineering, Computer & Vision Research Center, The University of Texas at Austin, TX 78712, USA*
[d]*Advanced Digital Imaging Research, LLC, League City, Texas 77573, USA*

Received 10 October 2003; accepted 7 September 2004

Abstract

This paper describes a fully automatic chromosome classification algorithm for Multiplex Fluorescence In Situ Hybridization (M-FISH) images using supervised parametric and non-parametric techniques. M-FISH is a recently developed chromosome imaging method in which each chromosome is labelled with 5 fluors (dyes) and a DNA stain. The classification problem is modelled as a 25-class 6-feature pixel-by-pixel classification task. The 25 classes are the 24 types of human chromosomes and the background, while the six features correspond to the brightness of the dyes at each pixel. Maximum likelihood estimation, nearest neighbor and k-nearest neighbor methods are implemented for the classification. The highest classification accuracy is achieved with the k-nearest neighbor method and $k = 7$ is an optimal value for this classification task.

Keywords: M-FISH; Nearest neighbor; k-nearest neighbor; Maximum likelihood estimation; Karyotyping

1. Introduction

Cytogenetics is the study of the genetic makeup of cells. Chromosomes are structures that contain the genetic information of cells. Images of chromosomes taken during cell division contain valuable information about the well being of an individual. Chromosome images are useful for diagnosing genetic disorders and for studying cancer. Thus the analysis of chromosomes is an important procedure in cytogenetic studies.

There are 46 human chromosomes which consist of 22 pairs of similar, homologous chromosomes, and two sex-determinative chromosomes. Thus there are 24 types, or classes, of chromosomes. The process of assigning the chromosomes to the different classes is known as Karyotyping [1].

Images of chromosomes are analyzed by cytogeneticists to obtain vital information about the health of an individual. However, manual examination of these images is a laborious and time-consuming process and requires skilled lab technicians [2]. Many successful attempts have been made to automate parts of the chromosome image analysis procedure [13–16]. One of the first steps in chromosome analysis is automated karyotyping.

Images of chromosomes may be obtained using a number of specimen preparation methods. One such method is

[*] Corresponding author. Tel.: +1 512 471 1369; fax: +1 512 471 5532.
E-mail address: aggarwaljk@mail.utexas.edu (J.K. Aggarwal).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2004.09.010

Table 1
M-FISH fluor labelling table

Chromosome	Spectrum Aqua	Spectrum Green	Spectrum Gold	Spectrum Red	Far Red
1	0	0	1	0	0
2	0	0	0	1	0
3	1	0	0	0	0
4	0	1	0	1	1
5	0	0	1	0	1
6	0	1	0	0	0
7	0	0	0	0	1
8	0	0	0	1	1
9	0	0	1	1	0
10	1	0	1	0	0
11	1	0	0	1	0
12	0	1	1	0	0
13	1	1	0	0	0
14	0	1	1	1	0
15	1	0	1	1	0
16	0	1	0	0	1
17	0	1	0	1	0
18	0	0	1	1	1
19	0	1	1	0	1
20	1	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
X	1	0	0	0	1
Y	1	0	1	0	1

The first column represents the chromosome number. Names of the five different fluors are shown in the first row. A 1 indicates that a particular chromosome is labelled by the fluor and a 0 indicates that the chromosome is not labelled by the fluor. Thus each chromosome is labelled by a specific combination of dyes.

Multiplex Fluorescence In Situ Hybridization (M-FISH) [3,4] which is a recently developed chromosome imaging technique. The goal of the research described in this paper is the automated classification of chromosome images that have been obtained by M-FISH.

The first paper on the M-FISH technique was published in 1996 by Speicher et al. [3] and it revolutionized chromosome imaging. In this technique chromosomes are labelled with 5 fluors (dyes) and a fluorescent DNA stain called DAPI (4′,6-Diamidino-2-phenylindole).

DAPI attaches to DNA and thus labels all chromosomes. The fluors attach to specific sequences of DNA. With M-FISH a unique combination of fluors is assigned to each chromosome type. That is, each class of chromosomes absorbs a different combination of fluors [3]. Thus M-FISH is based on a combinatorial labelling strategy. This strategy provides an easy way to label chromosomes in a multiplex fashion, as each fluor is either present(1) or absent(0) [3,5]. Also, at least five distinguishable fluors are needed for combinatorial labelling to uniquely identify all 24 chromosome types as the number of useful combinations of N fluors is $2^N - 1$ [3,5].

The central idea in M-FISH is that each chromosome is labelled by a unique combination of the five fluors. Several such sets of fluors have been developed for M-FISH imaging. One such set of five fluors and the corresponding fluor labelling table is shown in Table 1 [6]. The fluor labelling table enumerates the different combinations of the fluors used to label each chromosome type.

Though in theory the fluor absorption is described as binary, this is not the case in practice for real M-FISH data sets [7].

M-FISH images are captured with a fluorescent microscope. Multiple optical filters are used to view each of the fluorescent fluors. Each of the fluors is visible in one of the spectral channels. Thus a set of M-FISH images can be viewed as a multi-spectral set. An M-FISH data set consists of six images where each image is the response of the chromosome to a particular fluor. A typical M-FISH data set is shown in Fig. 1. Figs. 1(a)–(e) are the images of the responses of the five fluors which are Spectrum Aqua, Far Red, Spectrum Green, Spectrum Red and Spectrum Gold, respectively [6]. Fig. 1(f) shows the response of the DNA stain DAPI. DAPI attaches to DNA and thus all chromosomes are seen in this image.

Semi-automated image analysis of M-FISH data was done by Speicher et al. [5] in 1996. This basically consisted of segmentation, thresholding and classification stages. The DAPI

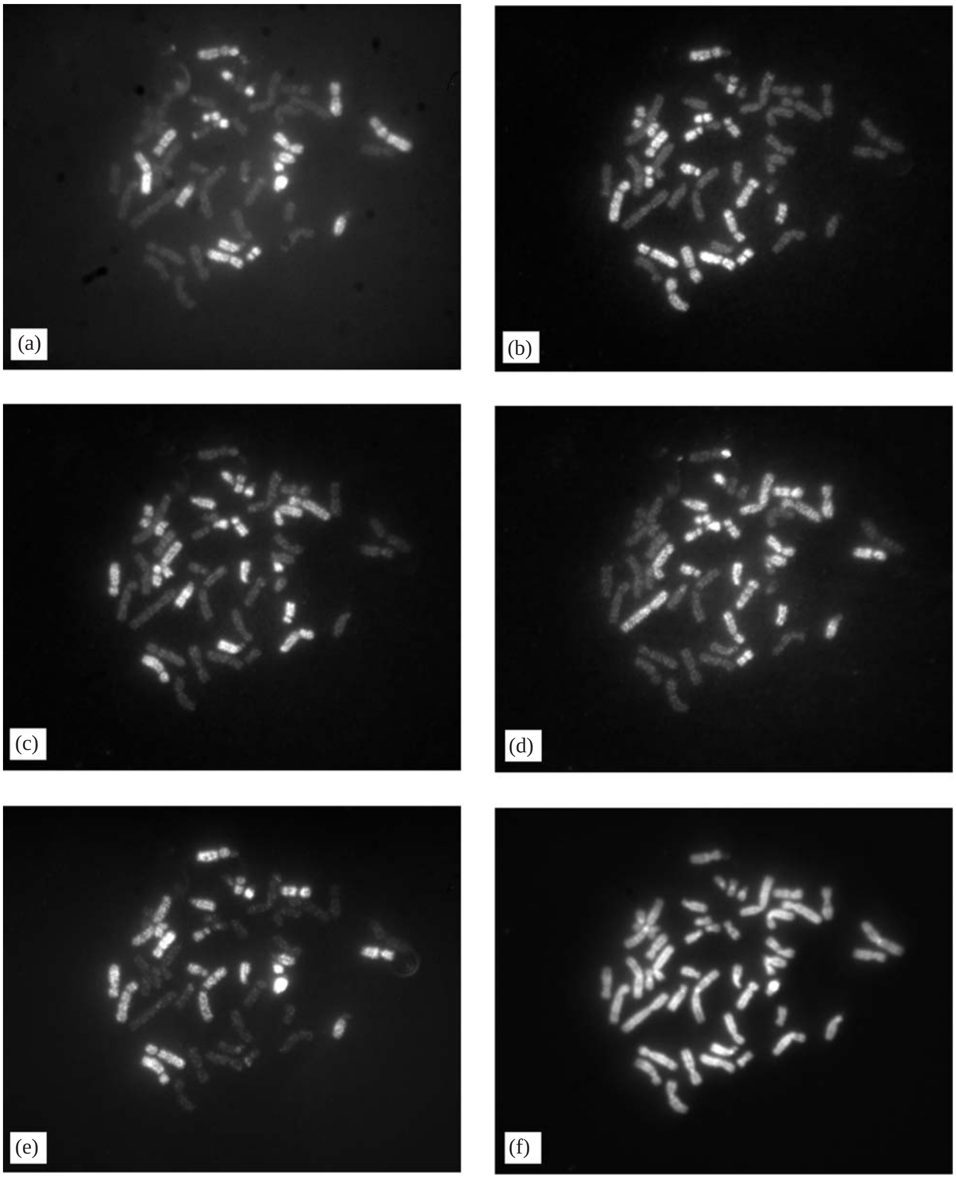

Fig. 1. A set of M-FISH Images. Each image corresponds to the response of a particular fluor. The DAPI stain labels all chromosomes: (a) Fluor: Spectrum Aqua, (b) Fluor: Far Red, (c) Fluor: Spectrum Green, (d) Fluor: Spectrum Red, (e) Fluor: Spectrum Gold, (f) DNA Stain: DAPI.

channel was used to create a mask to segment the chromosomes from the background. This mask and a threshold were applied to each M-FISH image to detect the presence or absence of a fluor at each pixel. Each pixel was then classified by comparing the combined response of the fluors at that pixel to the combinations in a fluor labelling table.

The image analysis was fully automated by Elis et al. [8] in 1998. They modelled the task as a 5-feature 24-class pattern recognition problem and performed adaptive spectral analysis for classification. This consisted of spectral calibration and adaptive region-oriented classification. During the calibration step an optimal vector to represent each class was found by minimizing an energy term. These vectors were called adaptive spectral feature vectors. In the classification step the image was subdivided into various polygons using Voronoi tessellation. The closest adaptive spectral feature vector (spectral class) for each region was computed. These were then classified using an iterative region-growing algorithm. Regions with color vectors best approximating the adaptive spectral feature vectors were

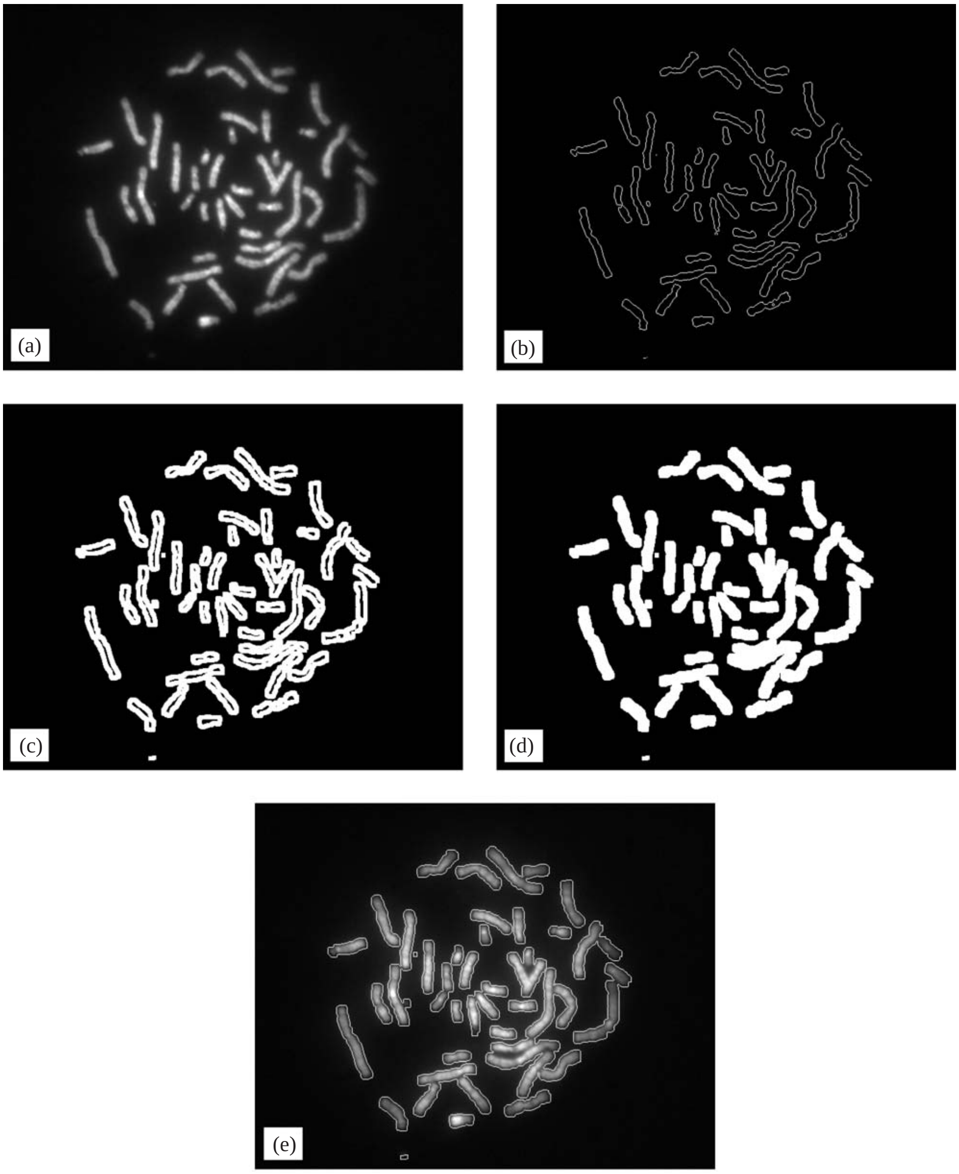

Fig. 2. Selection of testing pixels for classification: (a) original DAPI image, (b) edges detected, (c) edges after dilation, (d) edges filled, (e) boundaries detected from Figure 2(d) overlaid on the original image.

used as the starting points for the region-growing process. Two regions were merged if they belonged to the same class and the merged region was assigned the class of the start region. They claim that pixel-by-pixel classification would produce noisy results and thus did not perform pixel-by-pixel classification [8].

Saracoglu et al. [9] modelled the problem similarly. Their algorithm consisted of three steps, image tessellation, clustering and classification. The image was tessellated into regions with similar properties with a region-growing algorithm. Then an average color vector was computed for each region. For each of the classes, one start vector was selected (from the set of color vectors) such that it was the closest vector to the theoretically optimal color class vector. These 24 start vectors were then used as starting points for a *k*-means clustering algorithm. Each cluster was then classified by comparing its centroid with the theoretical color class vectors. However, none of these papers reported the classification accuracies of their methods over various M-FISH image sets.

In this paper we propose new algorithms for pixel-by-pixel classification of M-FISH images and show that this

Table 2
Overall chromosome classification accuracy for the different methods without majority filtering

Test set	MLE	NN	k-NN ($k=5$)	k-NN ($k=7$)	k-NN ($k=9$)
A	86.2870	87.6290	88.6620	88.7460	88.8040
B	88.3080	90.8400	92.2720	92.6190	92.8190
C	72.3810	85.9460	87.6780	88.0970	88.3080
D	68.0510	82.9520	85.3300	85.8610	86.1830
E	86.5690	84.5900	85.8430	85.9970	85.9990

All results in percentages.

Table 3
Overall classification accuracy for the different methods without majority filtering

Test set	MLE	NN	k-NN ($k=5$)	k-NN ($k=7$)	k-NN ($k=9$)
A	97.3970	97.7030	97.7700	97.7610	97.7710
B	98.2480	98.5350	98.5630	98.5720	98.5790
C	97.1210	98.0890	98.1380	98.1500	98.1630
D	96.3540	97.6560	97.8240	97.8580	97.8860
E	97.8780	98.2680	98.3180	98.3220	98.3220

All results in percentages.

Table 4
Overall chromosome classification accuracy for the different methods with majority filtering

Test set	MLE	NN	k-NN ($k=5$)	k-NN ($k=7$)	k-NN ($k=9$)
A	90.0180	90.9640	91.2200	91.1500	91.1270
B	90.9570	93.4560	94.2710	94.4070	94.4690
C	74.5680	89.8400	90.5340	90.7760	90.8470
D	70.7780	87.7670	88.8210	89.0610	89.1680
E	88.4740	86.4730	87.0830	87.2130	87.1190

All results in percentages.

Table 5
Overall classification accuracy for the different methods with majority filtering

Test set	MLE	NN	k-NN ($k=5$)	k-NN ($k=7$)	k-NN ($k=9$)
A	97.7660	98.0410	98.0130	97.9880	97.9900
B	98.4090	98.6960	98.6770	98.6710	98.6700
C	97.3190	98.3910	98.3490	98.3470	98.3500
D	96.6040	98.0640	98.0940	98.1000	98.1100
E	98.0650	98.4360	98.4220	98.4220	98.4120

All results in percentages.

methodology gives good results. In these algorithms we use all six images of the M-FISH data set and we include the background as a new class. Thus we have modelled the problem as a 6-feature 25-class pattern recognition task. We report the classification accuracies of the method over various M-FISH data sets.

The rest of the paper is organized as follows. Section 2 describes the different classification techniques. The methodology and the data sets used are described in Section 3. The results are presented in Section 4. Finally, Section 5 presents the conclusion.

2. Classification techniques

This section gives a brief review of the different supervised parametric and non-parametric classification techniques that are used in this paper. The aim of these techniques is to classify samples into one of N different

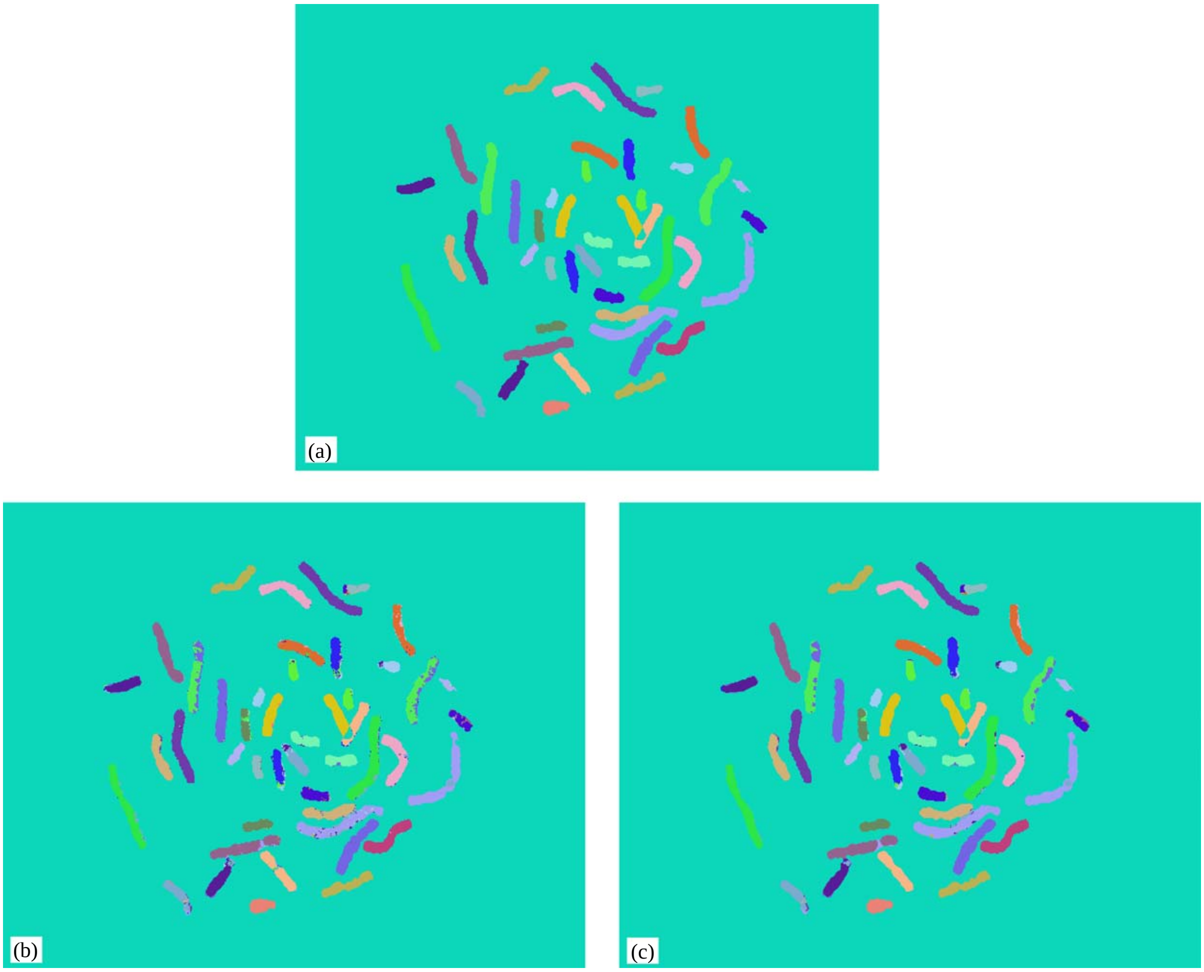

Fig. 3. Classification results for M-FISH Image Set A: (a) original class-map, (b) classified class-map before majority filtering, (c) classified class-map after majority filtering.

classes based on features that describe the sample. Let w_i for $i = 1, \ldots, N$ denote the N classes. If we measure d features for each sample then each sample is described by a d-dimensional *feature vector*. Let x denote such a feature vector. A classifier is first trained on a given labelled set of training samples. A given test sample is then assigned to a particular class by the classifier. The details of the different classifiers are described below [10].

2.1. Supervised parametric method

The supervised parametric method used is maximum likelihood estimation. Let $P(w_i)$ denote the a priori probability that a sample belongs to class w_i where $i = 1, \ldots, N$.

Let $p(x|w_i)$ denote the class-conditional probability density function. It represents the probability distribution function for a feature vector x given that x belongs to class w_i. Let $P(w_i|x)$ be the a posteriori probability, which is the probability that the sample belongs to class w_i given the feature vector x. Given $P(w_i)$ and $p(x|w_i)$, the a posteriori probability for a sample represented by the feature vector x is given by the Bayes formula [10].

$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)}, \tag{1}$$

where $p(x) = \sum_{i=1}^{N} p(x|w_i)P(w_i)$. The formula is applicable for all probability density functions; however, depending on the nature of the data, the normal density function is often used to model the distribution of feature values of a particular class. The general multivariate normal density function in d dimensions is given by:

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \sum\nolimits^{-1} (x-\mu)\right], \tag{2}$$

where x is a d component feature vector, μ is the d component mean vector, Σ is the $d \times d$ covariance matrix, and $|\Sigma|$ and Σ^{-1} are its determinant and inverse, respectively. It

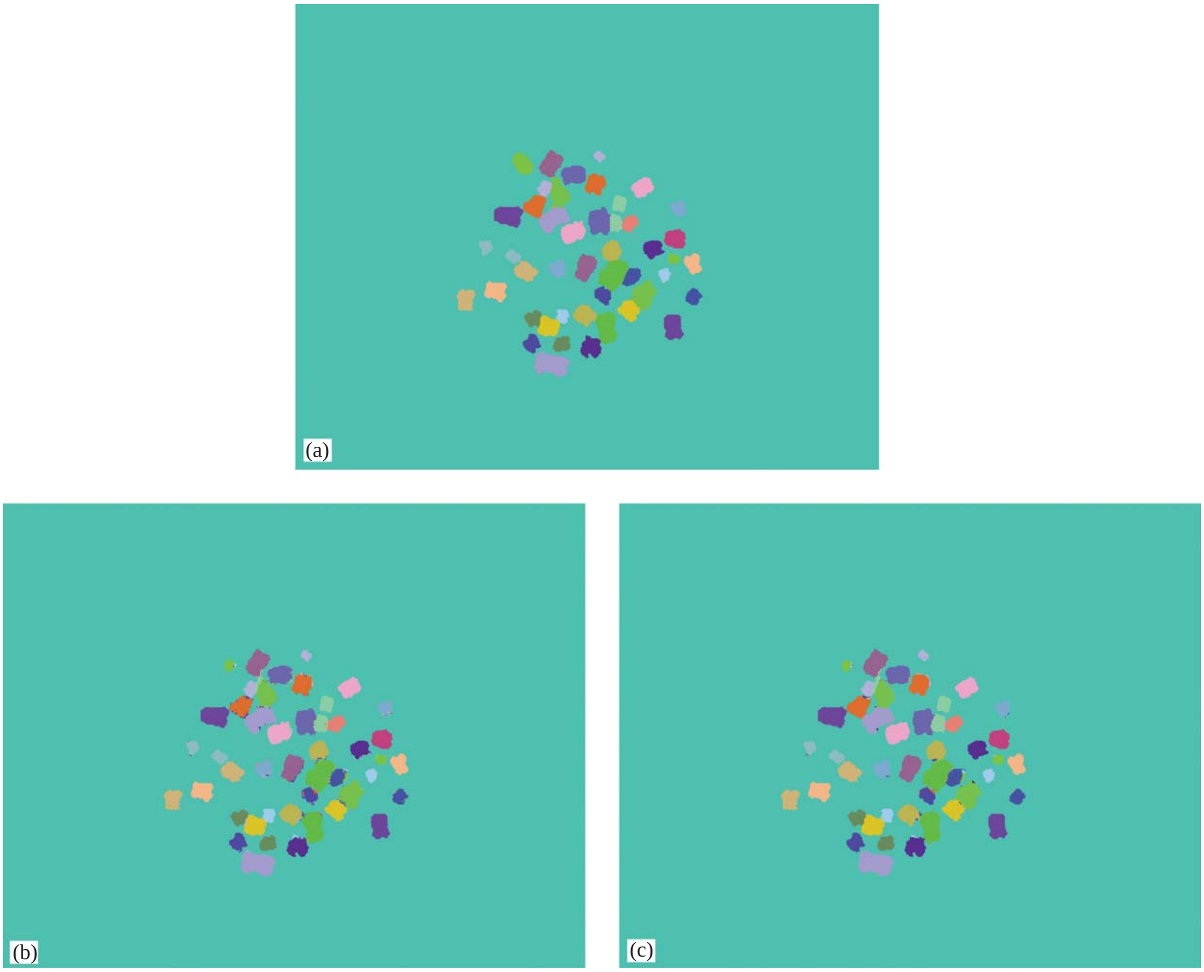

Fig. 4. Classification results for M-FISH Image Set B: (a) original class-map, (b) classified class-map before majority filtering, (c) classified class-map after majority filtering.

is assumed that the density function for each class is a 6-dimensional Gaussian function. The parameters μ and Σ of the probability density function for each class are calculated from the training samples belonging to that class. *Note that the maximum likelihood estimates for μ and Σ of each class are the mean vector and covariance matrix of the training samples of that class*. Any given test sample, described by the feature vector x, can be classified by using the Bayes Decision Rule, which is:

$$\textit{decide } w_i \textit{ if } P(w_i|x) > P(w_j|x)\ \forall j \neq i. \tag{3}$$

2.2. Supervised non-parametric methods

The supervised non-parametric methods selected for classification are the nearest neighbor and the k-nearest neighbor methods. In these methods no assumptions are made about the probability density function for each class. These methods are used because the assumption that the probability density function for each class is a 6-dimensional normal distribution may not necessarily be true, and a classifier may perform better if these assumptions are not made.

2.2.1. Nearest neighbor

Let $T=\{s_1, s_2, \ldots, s_n\}$ denote the set of n-labelled training samples. Each sample is a d-dimensional vector. Let $s_i \in T$ be the training sample nearest to a given test sample t in terms of some metric or distance function. The nearest neighbor rule for classifying t is to assign it to the class to which s_i belongs [10]. The metric we use is the Euclidean distance.

2.2.2. k-nearest neighbor

Let $T=\{s_1, s_2, \ldots, s_n\}$ denote the set of n-labelled training samples. Given a test sample t, let $R=\{r_1, r_2, \ldots, r_k\}$

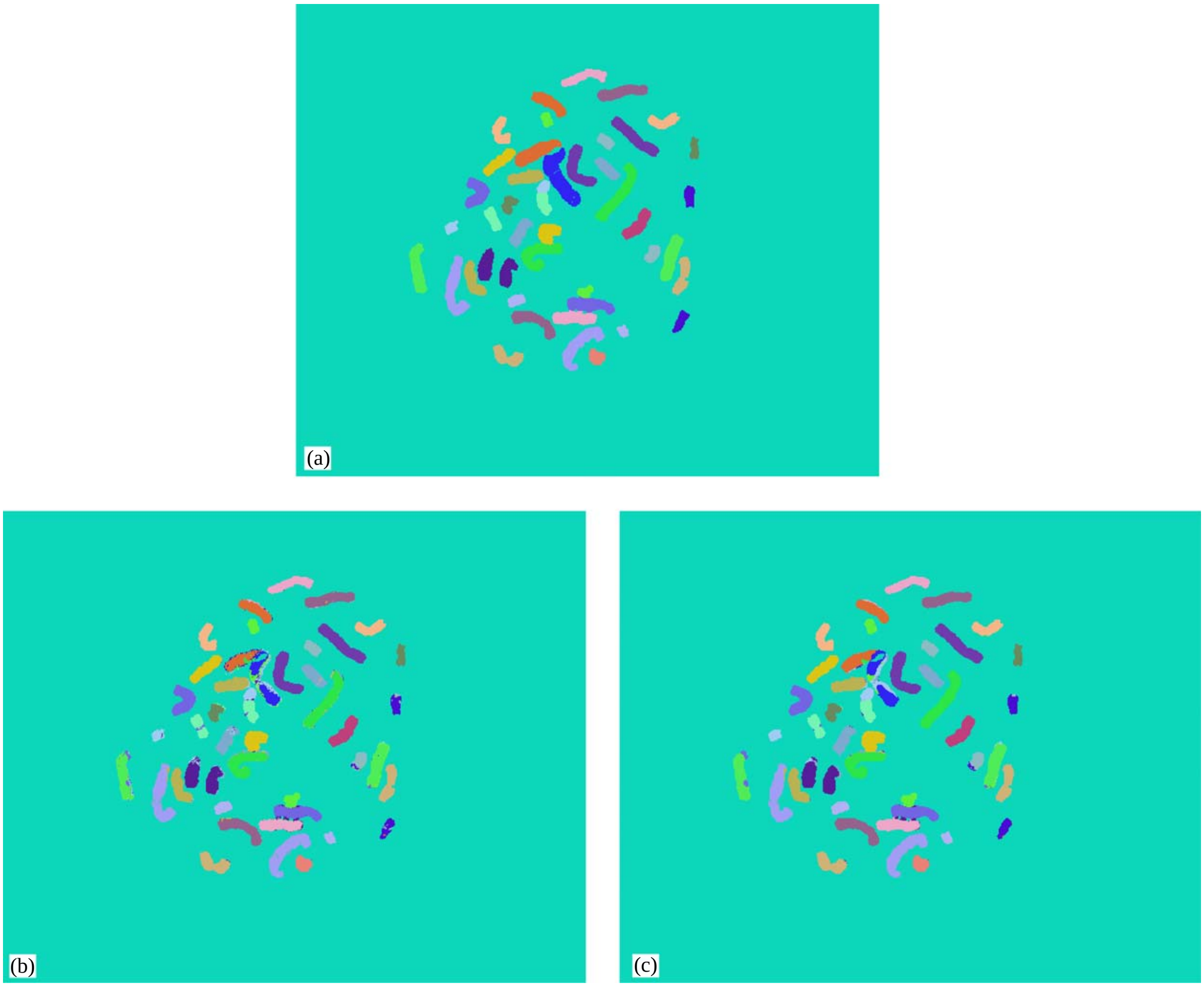

Fig. 5. Classification results for M-FISH Image Set C: (a) original class-map, (b) classified class-map before majority filtering, (c) classified class-map after majority filtering.

be a set of the k-nearest training samples to t in terms of some metric. The k-nearest neighbor rule is to assign the sample t to the class that occurs most frequently among the k-nearest training samples. Again the metric used is the Euclidean distance. The values of k used are 5, 7 and 9 neighbors. If the ranges of the data in each dimension vary considerably, this may affect the performance of the nearest neighbor and k-nearest neighbor drastically. Thus both the training and testing data must be normalized. We used the following method for normalization of the data.

$$y = (x - \mu)/(3 * \sigma), \tag{4}$$

where x is the d-dimensional original data sample, μ is the d-dimensional mean vector of the given training samples, σ is the standard deviation of the training samples, and y is the normalized data sample.

3. Methodology

The supervised parametric and non-parametric methods described in Section 2 were used for classification. For all of the methods, we used the same training and testing samples so that a fair comparison could be made between them. To compare the performance of the two methods, the overall classification accuracy and the chromosome classification accuracy were measured. The chromosome classification accuracy is the accuracy of classifying only those pixels belonging to chromosomes. Since a majority of the pixels are background pixels, the overall pixel classification accuracy mainly reflects segmentation. Thus, it is important to measure the chromosome classification accuracy to get a good idea of the diagnostic performance of the classifier.

The images for training and testing were selected from a public database of M-FISH images. This database is made

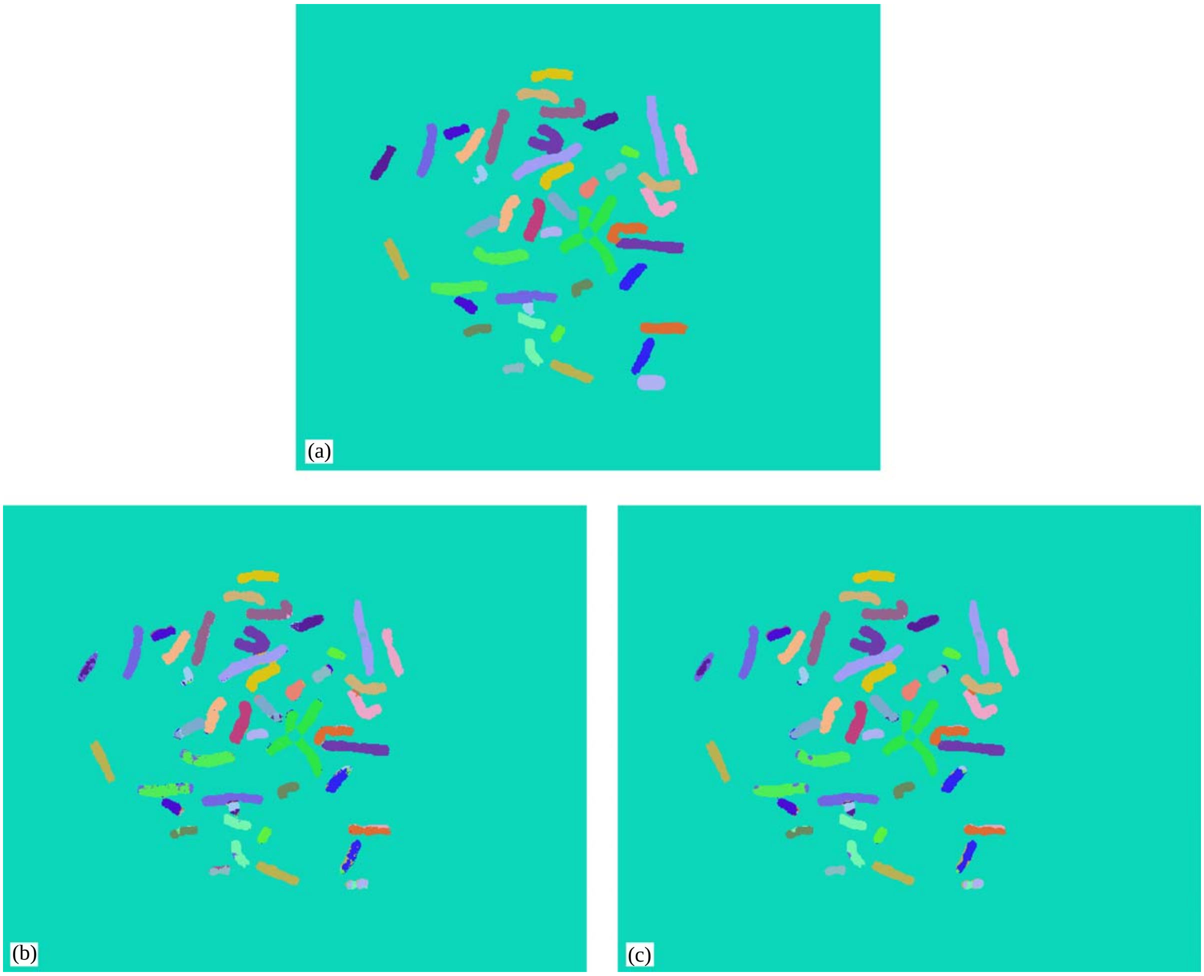

Fig. 6. Classification results for M-FISH Image Set D: (a) original class-map, (b) classified class-map before majority filtering, (c) classified class-map after majority filtering.

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 labelled class-map image in which each pixel is labelled according to the class to which it actually belongs. This image was used to determine the accuracy of the different classification techniques.

For training, pixels belonging to each of the classes were chosen randomly ten times, from one set of M-FISH images. Thus ten different training data sets were created. Pixels from other sets of M-FISH images were chosen for testing. Thus there was no overlap between the training and testing data. Each set of testing data was then classified with respect to each of the training data sets. The classification results (the overall accuracy and the chromosome accuracy) obtained from the ten trials were then averaged to obtain the final classification results for each test set. This was done for each classification method and for every test set. Since 90% or more of the pixels of each M-FISH set were background pixels, only a subset of pixels from each set were selected for testing. The selection of pixels for testing is described in Section 3.1.

3.1. *Selection of pixels for classification*

The goal was to create a binary image(mask) in which the pixels to be selected for testing are labelled “1” whereas the pixels not to be selected are labelled “0”. As mentioned before, the DAPI stain labels all of the chromosomes, and thus the image of the DAPI channel was used for the selection of pixels. This image is shown in Fig. 2(a). First the edges of the chromosomes in the DAPI image were

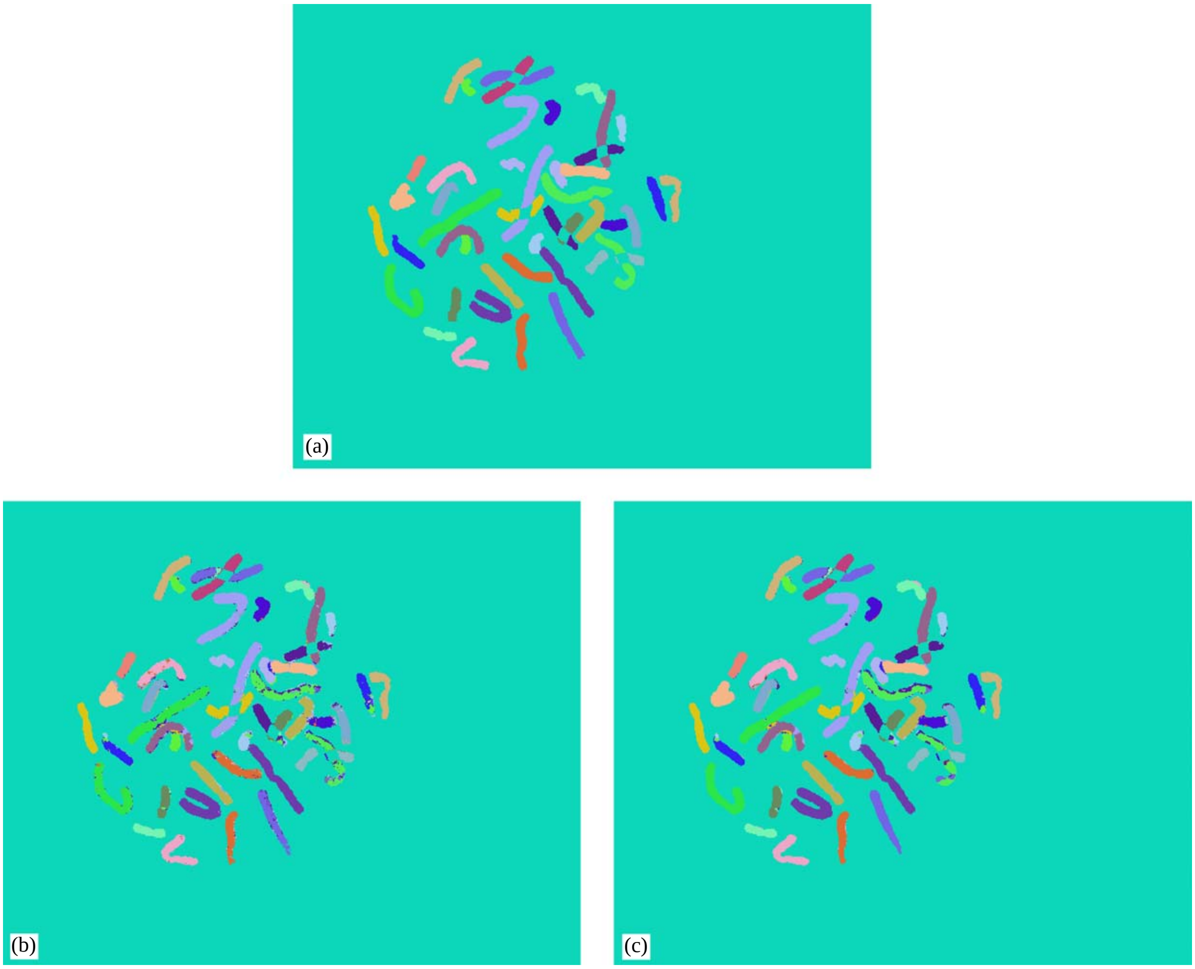

Fig. 7. Classification results for M-FISH Image Set E: (a) original class-map, (b) classified class-map before majority filtering, (c) classified class-map after majority filtering.

detected using the Laplacian of Gaussian edge detector. Fig. 2(b) shows the edges detected. A review of this method appears in Refs. [11,12]. The edge image was then dilated using a morphological operator, as shown in Fig. 2(c). This was done because perfect segmentation of the chromosomes is difficult to achieve and it was seen that some faint pixels belonging to some chromosomes fell outside the edges detected. Dilation ensured that these pixels were also included in the classification stage. Finally all pixels lying inside the edges of the chromosomes were set to 1, and those lying outside were set to 0 to create the mask shown in Fig. 2(d). The boundaries of the objects in Fig. 2(d) were detected and overlaid on the original image in Fig. 2(e).

3.2. Classification and post-processing

The pixels selected by the process described in Section 3.1 were classified by maximum likelihood estimation (MLE), nearest neighbor (NN) and *k*-nearest neighbor ($k=5$, 7 and 9) classifiers. Before training, all pixels were first normalized by the procedure described in Section 2.2.2. All of these classifiers were then trained with the same set of training samples. A class-map for each output was generated. In this image each pixel was labelled according to the class it was classified to.

Isolated pixel classification errors were observed after the classification. To remove these errors, a 5-by-5 majority filter was applied to the classification output. In majority filtering, an *n*-by-*n* window is centered about each pixel in a given image. The value that occurs the maximum number of times among the values lying within the window is determined. This output is placed at the location of the center pixel, that is, the pixel about which the window was centered. This procedure is then repeated for every pixel in the image. Majority filtering significantly improve the classification accuracy.

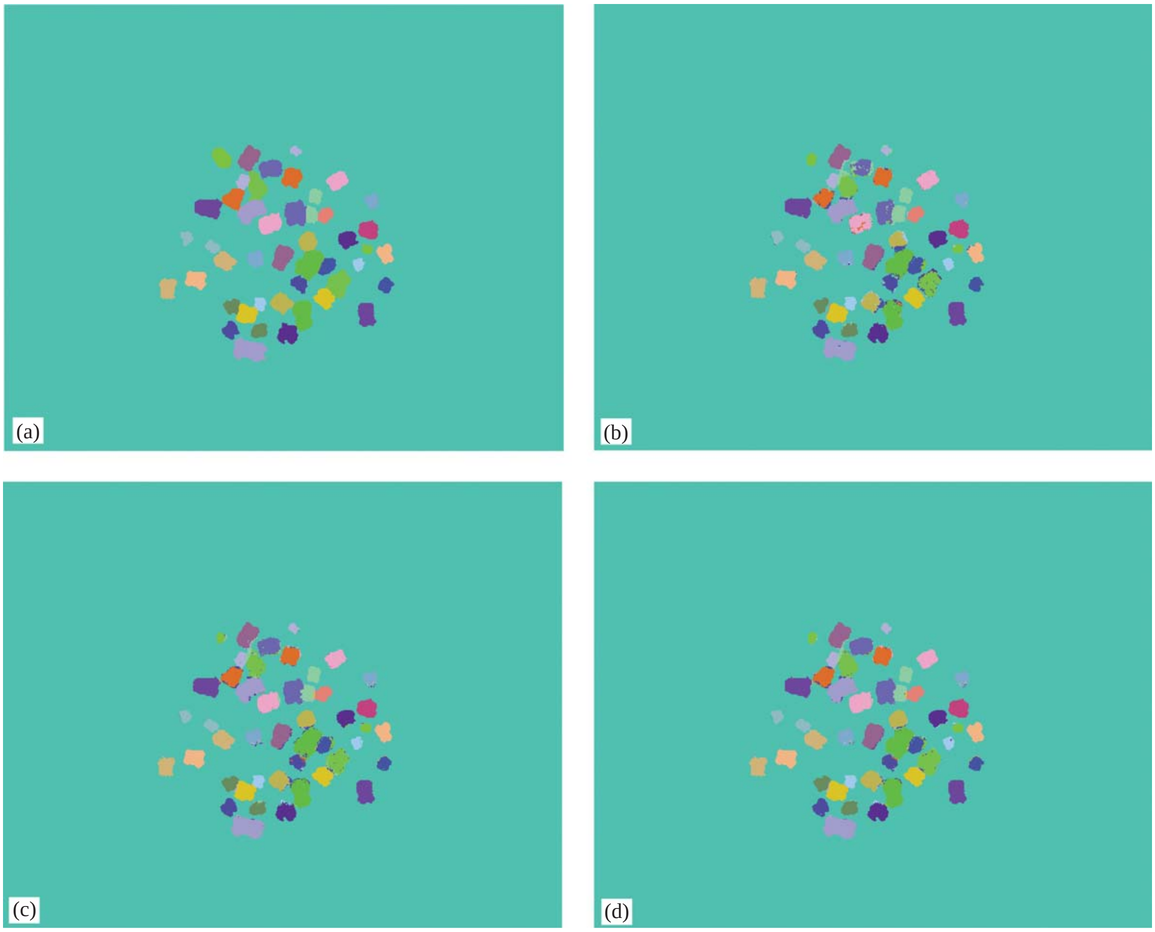

Fig. 8. The different classification results obtained with the MLE, NN and k-NN $(k - 7)$ classifiers, for M-FISH Image Set B: (a) original class-map, (b) output class-map obtained with MLE classifier, (c) output class-map obtained with NN classifier, (d) output class-map obtained k-NN classifier $(k = 7)$.

4. Results

Five M-FISH image sets, labelled A to E, were classified using the methods described above. Each set has 333, 465 pixels. From each of these, a subset of pixels was selected for testing by applying the pixel selection algorithm described in Section 3.1. For each set, the average overall classification accuracy and the average chromosome classification accuracy were computed. A class-map was generated for each classification output. A separate color was used to represent each chromosome class in the image. The overall and chromosome accuracies were computed by comparing this class-map to the class-map provided in the database.

Tables 2 and 3 show the chromosome classification accuracy and the overall classification accuracy obtained for each M-FISH set without application of the majority filter. Tables 4 and 5 show the chromosome classification accuracy and the overall classification accuracy obtained after application of the majority filter to the classification result. Majority filtering improves classification accuracy by reducing the number of isolated pixel classification errors. It reduced the average chromosome misclassification rate by 2%.

Fig. 3 shows the classification results for the M-FISH Image Set A. The actual class-map is shown in Fig. 3(a) and the computed class-maps before and after majority filtering are shown in Figs. 3(b) and (c), respectively. Similarly, the results for the other M-FISH image sets (B–E) are shown in Figs. 4–7, respectively. These figures show the results obtained with the k-nearest neighbor method $(k = 7)$. Fig. 8 shows the different classification results for M-FISH Image Set B, obtained with the MLE, NN and k-NN $(k = 7)$ classifiers.

A 25 × 25 confusion matrix for one of the classified outputs is shown in Table 6 . The rows and columns of this table correspond to the actual and predicted classes. The first row and column correspond to the class numbers. In this matrix, class 0 corresponds to the background and thus a maximum

Table 6
The 25 × 25 confusion matrix for M-FISH Image Set B

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
0	312119	13	11	0	3	1	8	4	6	0	0	0	0	10	3	9	3	2	0	6	3	0	301	0	0
1	220	1373	0	0	8	2	0	0	8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	118	0	1361	0	3	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0
3	249	0	0	1058	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	92	0	16	0	1018	0	20	0	0	0	0	0	15	0	0	0	0	0	0	1	0	0	0	0	0
5	101	2	0	0	0	996	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	155	0	0	0	6	0	995	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	206	0	0	0	0	4	0	884	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	49	0	6	0	0	0	0	0	753	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0
9	221	0	9	0	0	0	0	0	0	730	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	192	0	0	0	1	0	0	0	0	0	810	0	0	0	0	0	15	1	0	0	0	0	0	0	0
11	357	0	0	2	0	0	0	0	0	0	0	775	0	0	0	10	0	0	0	0	0	0	2	0	0
12	162	3	0	0	0	0	0	0	0	0	0	0	728	0	11	0	0	0	0	0	0	0	0	0	0
13	85	10	1	0	1	0	0	0	1	0	0	0	0	705	0	2	1	0	0	0	0	0	0	0	0
14	85	1	24	0	17	0	0	0	0	0	4	0	35	0	451	1	0	0	0	0	0	0	0	0	0
15	115	4	0	0	0	0	0	0	0	0	0	0	0	11	0	494	0	0	0	0	0	4	0	0	0
16	201	1	0	0	36	11	61	7	2	0	1	0	0	0	0	0	417	4	0	35	0	0	0	0	0
17	111	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	512	0	1	0	0	0	0	0
18	226	0	2	0	0	21	0	0	62	0	0	0	0	0	0	0	0	0	496	0	4	0	0	0	0
19	115	3	0	0	5	4	2	0	1	0	6	0	20	0	0	0	12	13	0	277	0	0	0	0	0
20	112	0	0	0	0	0	0	0	2	0	0	4	0	0	0	0	0	0	17	0	348	0	0	0	0
21	194	0	0	0	3	0	0	0	0	0	5	0	2	31	12	5	3	0	0	0	0	328	14	0	0
22	148	0	9	0	38	0	0	0	0	0	0	0	7	9	44	2	0	1	0	0	0	0	224	0	0
23	125	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	401	0
24	67	0	0	0	0	0	0	0	0	0	1	0	0	0	0	10	18	0	0	0	0	0	0	0	253

The columns correspond to the actual classes and the rows correspond to the predicted classes. Class 0 corresponds to the background. Class 1 corresponds to chromosome 1, and so on. The first row and columns represent the class numbers.

number of pixels fall in the $(0, 0)$ square. Note that most of the entries of this matrix are zeros.

The non-parametric methods give higher classification accuracies than the parametric method. The *k*-nearest neighbor method outperformed the maximum likelihood and nearest neighbor methods. As the value of *k* was increased, the classification accuracy increased. However, we observe very little improvement in accuracy as *k* was increased from 7 to 9 and beyond. Thus increasing *k* beyond 7 is not beneficial.

5. Conclusion

In this paper we have developed new, fully automated algorithms for pixel-by-pixel classification of M-FISH images and showed that high classification accuracies can be achieved with this methodology. The overall classification accuracy achieved is 98.3% and the overall chromosome classification accuracy achieved is 90.52%.

The classification task is modelled as a 6-feature, 25-class classification problem. Supervised parametric and non-parametric techniques were implemented, and it was found that the Non-Parametric methods performed better than the parametric method. The highest classification accuracy was obtained by the *k*-nearest neighbor method, and $k = 7$ is an optimal value for this classification task. We also showed that post-processing techniques such as majority filtering can help improve the classification accuracy.

6. Summary

This paper describes a fully automatic chromosome classification algorithm for Multiplex Fluorescence In Situ Hybridization (M-FISH) images using supervised parametric and non-parametric techniques.

Chromosomes are structures that contain the genetic information of cells. Images of chromosomes taken during cell division contain valuable information about the well being of an individual. Chromosome images are useful for diagnosing genetic disorders and for studying cancer. Thus the analysis of chromosomes is an important procedure in cytogenetic studies, which deal with the genetic makeup of cells.

There are 46 human chromosomes, which consist of 22 pairs of similar, homologous chromosomes, and two sex-determinative chromosomes. Thus there are 24 types, or classes, of chromosomes. The process of assigning the chromosomes to the different classes is known as karyotyping.

Images of chromosomes are analyzed by cytogeneticists to obtain vital information about the health of an individual. However, manual examination of these images is a laborious and time-consuming process and requires skilled lab technicians. Many successful attempts have been made to automate parts of the chromosome image analysis procedure. One of the first steps in chromosome analysis is automated karyotyping.

Images of chromosomes may be obtained using a number of specimen preparation methods. One such method is Multiplex Fluorescence In Situ Hybridization (M-FISH), which is a recently developed chromosome imaging technique. The goal of the research described in this paper is the automated classification of chromosome images that have been obtained by M-FISH.

In this technique chromosomes are labeled with five fluors (dyes) and a fluorescent DNA stain called DAPI ($4'$,6-Diamidino-2-phenylindole). The fluors attach to specific sequences of DNA. With M-FISH a unique combination of fluors is assigned to each chromosome type. That is, each class of chromosomes absorbs a different combination of fluors. This strategy provides an easy way to label chromosomes in a multiplex fashion, as each fluor is either present or absent.

M-FISH images are captured with a fluorescent microscope. Multiple optical filters are used to view each of the fluorescent fluors. Each of the fluors is visible in one of the spectral channels. Thus a set of M-FISH images can be viewed as a multi-spectral set. An M-FISH data set consists of six images where each image is the response of the chromosome to a particular fluor.

In this paper we propose new algorithms for pixel-by-pixel classification of M-FISH images and show that this methodology gives good results. In these algorithms we use all six images of the M-FISH data set and we include the background as a new class. The classification problem is modeled as a 25-class, 6-feature pixel-by-pixel classification task. The 25 classes are the 24 types of human chromosomes and the background, while the six features correspond to the brightness of the dyes at each pixel. Maximum likelihood estimation, nearest neighbor and *k*-nearest neighbour methods are implemented for the classification. The highest classification accuracy is achieved with the *k*-nearest neighbor method and $k = 7$ is an optimal value for this classification task.

References

[1] R.S. Verma, A. Babu, Human Chromosomes: Principles and Techniques, second ed., McGraw-Hill, Inc., New York, 1995.

[2] Q. Wu, K. Castleman, Automated chromosome classification using wavelet-based band pattern descriptors, Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems, 2000, pp. 189–194.

[3] M. Speicher, S. Ballard, D. Ward, Karyotyping human chromosomes by combinatorial Multi-Fluor FISH, Nat. Genetics 12 (1996) 368–375.

[4] M. Beau, One FISH, two FISH, red FISH, blue FISH, Nat. Genetics 12 (1996) 341–344.

[5] M. Speicher, S. Ballard, D. Ward, Computer image analysis of combinatorial Multi-Fluor FISH, Bioimaging 4 (1996) 52–64.

[6] W. Schwartzkopf, ADIR M-FISH Image Database, Technical Report, Advanced Digital Imaging Research, August 2000.

[7] W. Schwartzkopf, Maximum likelihood techniques for joint segmentation-classification of multi-spectral chromosome images, Ph.D. Thesis, The University of Texas at Austin, December 2002.

[8] R. Eils, S. Uhrig, K. Saracoglu, K. Satzler, A. Bolzer, I. Petersen, J. Chassery, M. Ganser, M. Speicher, An optimized, fully automated system for fast and accurate identification of chromosomal rearrangements by multiplex-fish (m-fish), Cytogenet. Cell Genet. 82 (1998) 160–171.

[9] K. Saracoglu, J. Brown, L. Kearney, S. Uhrig, J. Azofeifa, C. Fauth, M. Speicher, R. Eils, New concepts to improve resolution and sensitivity of molecular cytogenetic diagnostics by multicolor fluorescence in situ hybridization, Cytometry 44 (2001) 7–15.

[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley-Interscience, San Diego, 2001.

[11] A.C. Bovik, Handbook of Image and Video Engineering, Academic Press, New York, 2000.

[12] K.R. Castleman, Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[13] A. Carothers, J. Piper, Computer-aided classification of human chromosomes: A review, Statist. Comput. 4 (3) (1994) 161–171.

[14] W. Schwartzkopf, B.L. Evans, A. Bovik, Minimum entropy segmentation applied to multi-spectral chromosome images, Proceedings of the IEEE International Conference on Image Processing II, 2001, pp. 865–868.

[15] K. Castleman, Digital imaging and cytogenetics a historical perspective, Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems (2000) 3.

[16] K.R. Castleman, R. Elis, L. Morrison, J. Piper, K. Saracoglu, M.A. Schulze, Classification accuracy in multiple color fluorescence imaging microscopy, Cytometry 41 (2000) 139–147.

About the Author—J.K. AGGARWAL has served on the faculty of The University of Texas at Austin College of Engineering since 1964 and is currently the Cullen Professor of Electrical and Computer Engineering and Director of the Computer and Vision Research Center. His research interests include image processing, computer vision and pattern recognition. A Fellow of IEEE since 1976 and IAPR since 1998, he received the Senior Research Award of the Society of Engineering Education in 1992, and the 1996 Technical Achievement Award of the IEEE Computer Society. He is the recipient of the 2004 King-Sun Fu Prize of the International Association for Pattern Recognition, and the 2005 Leon K. Kirchmayer Award for Graduate Teaching of IEEE.

He is author or editor of seven books and 39 book chapters; author of over 175 journals papers, as well as numerous proceedings papers and technical reports. He served as Chairman of the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence (1987–1989), Director of the NATO Advanced Research Workshop on Multisensor Fusion for Computer Vision, Grenoble, France (1989), Chairman of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1993), and President of the International Association for Pattern Recognition (1992–1994).

About the Author—AL BOVIK is currently the Cullen Trust for Higher Education Endowed Professor at The University of Texas at Austin in the Department of Electrical and Computer Engineering. He is the Director of the Laboratory for Image and Video Engineering (LIVE) in the Center for Perceptual Systems. During the Spring of 1992, he held a visiting position in the Division of Applied Sciences, Harvard University, Cambridge, Massachusetts. His current research interests include digital video, image processing, and computational aspects of biological visual perception. He has published over 350 technical articles in these areas and holds two U.S. patents. He is the editor/author of the *Handbook of Image and Video Processing*, published by Academic Press in April in 2000.

Dr. Bovik was named Distinguished Lecturer of the IEEE Signal Processing Society in 2000, received the IEEE Signal Processing Society Meritorious Service Award in 1998, the IEEE Third Millennium Medal in 2000, the University of Texas Engineering Foundation Halliburton Award in 1991 and is a two-time Honorable Mention winner of the international Pattern Recognition Society Award for Outstanding Contribution (1988 and 1993). He was named a Dean's Fellow in the College of Engineering in the Year 2002. He is a Fellow of the IEEE and has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; Editor-in-Chief, *IEEE Transactions on Image Processing*, 1996–2002; Editorial Board, *The Proceedings of the IEEE*, 1998-present; Series Editor for Morgan and Claypool Publishing Company, 2003-present; and Founding General chairman, *First IEEE International Conference on Image Processing*, held in Austin, Texas, in November, 1994. Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.

About the Author—KEN CASTLEMAN after receiving the BSEE, MSEE, and Ph.D. degrees from the The University of Texas at Austin, Dr. Castleman worked in the Image Processing Laboratory at the NASA Jet Propulsion Laboratory from 1970 through 1985. During that time he taught at CalTech and served as a research fellow at both USC and UCLA.

Since 1985 he has been the President of Advanced Digital Imaging Research, LLC (formerly Perceptive Systems, Inc.) a company which he founded. He has written two college-level textbooks and more than 60 scientific papers on the subject of digital image processing and its applications. He is an Adjunct Professor of Biomedical Engineering at The University of Texas at Austin, a fellow of the American Institute of Medical and Biological Engineering, and a member of the United States Space Foundation's Space Technology Hall of Fame. He serves on a forensic imaging advisory board for the FBI, and he was called in by NASA to assist the analysis of launch imagery from the Challenger accident in 1986 and the Columbia accident in 2003.

About the Author—MEHUL P. SAMPAT, M.S., is a doctoral candidate in Biomedical Engineering at the University of Texas at Austin. He graduated from the University of Mumbai in May 2000 with a bachelor's degree in Biomedical Engineering. He came to the University of Texas at Austin in Fall 2000 and received the Masters degree in Biomedical Engineering in Spring 2002.

Mehul has been a graduate research assistant at the Laboratory for Image and Video Engineering (LIVE) under the supervision of Dr. Alan C. Bovik since Spring 2001 where he worked on developing fully automated classification techniques for M-FISH images for his Masters degree. For his dissertation work, he is working on computer-aided detection of spiculated masses and architectural distortions in mammograms. He is also a member of the Biomedical Informatics Lab (BMIL) at UT.

He was awarded a predoctoral traineeship award through the Department of Defense (DoD) Breast Cancer Research Program. He was one of the 50 students selected to attend the 2004 IEEE International Summer School on Medical Imaging.