Journal of Intelligent & Fuzzy Systems 27 (2014) 3111–3122
DOI:10.3233/IFS-141268
IOS Press

A genetic type-2 fuzzy C-means clustering approach to M-FISH segmentation

Dzung Dinh Nguyen[a], Long Thanh Ngo[a,*] and Junzo Watada[b]
[a]*Department of Information Systems, Le Quy Don Technical University, Hanoi, Vietnam*
[b]*School of Information, Production and Systems, Waseda University, KitaKyushu, Japan*

Abstract. Multiplex Fluorescent In Situ Hybridization (M-FISH) is a multi-channel chromosome image generating technique that allows colors of the human chromosomes to be distinguished. In this technique, all chromosomes are labelled with 5 fluors and a fluorescent DNA stain called DAPI (4 in, 6-Diamidino-2-phenylindole) that attaches to DNA and labels all chromosomes. Therefore, a M-FISH image consists of 6 images, and each image is the response of the chromosome to a particular fluor. In this paper, we propose a genetic interval type-2 fuzzy c-means (GIT2FCM) algorithm, which is developed and applied to the segmentation and classification of M-FISH images. Chromosome pixels from the DAPI channel are segmented by GIT2FCM into two clusters, and these chromosome pixels are used as a mask for the remaining five channels. Then, the GIT2FCM algorithm is applied to classify the chromosome pixels into 24 classes, which correspond to the 22 pairs of homologous chromosomes and two sexual chromosomes. The experiments performed using the M-FISH dataset show the advantages of the proposed algorithm.

Keywords: Type-2 fuzzy C-neans clustering, genetic algorithms, MFISH, image segmentation

1. Introduction

Multiplex Fluorescent In Situ Hybridization (M-FISH) is an important technology that was developed for chromosome analysis. Typically, M-FISH is used to generate a multi-channel chromosome image. This image allows us to distinguish the color of the human chromosomes that contain the genetic information. Almost all human cells include 22 pairs of homologous chromosomes and two sexual chromosomes (XX: female and XY: male). For a normal cell, chromosomes are classified into 24 classes [1, 2]. If each class is represented by a different color, then a geneticist using color chromosome images can easily determine which parts of the chromosomes are lost or rearranged and apply this information to the study of cancer and disorder genetics.

Many algorithms have been proposed for the automatic classification of chromosomes. These algorithms were developed in two main directions: pixel-by-pixel [10] and region-based classification [8, 9]. However, neither the pixel-by-pixel classification algorithms nor the region-based methods have sufficient accuracy (less than 90%) for clinical use [8–10]. The main reason for this low accuracy is the uncertainty or noise, which always exists in the M-FISH image.

Thus, we propose a genetic interval type-2 Fuzzy C-Means clustering algorithm (GIT2FCM), which is a combination of the genetic algorithm (GA) and interval type-2 Fuzzy C-Means (IT2FCM), has shown the advantages in handling uncertainty. While the original IT2FCM faces the difficulties of initializing the centroid of the clusters and determining the number of clusters, GIT2FCM uses the GA to produce better centroids and find the optimal number of clusters. However,

*Corresponding author. Long Thanh Ngo, Department of Information Systems, Le Quy Don Technical University, No. 236 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam. Tel.: +84 914 364 064; Fax: +84 43836 0897; E-mails: ngotlong@mta.edu.vn (Long Thanh Ngo), dinhdung1082@gmail.com (Dzung Dinh Nguyen), junzow@osb.att.ne.jp (Junzo Watada).

the number of clusters in the M-FISH problem is predetermined: it is 2 for the chromosome segmentation and 24 for the chromosome classification. Additionally, a validity index for image segmentation is used as a fitness function for GA to obtain better results.

The GIT2FCM algorithm consists of two steps. In the first step, we randomly initialize the population of the GA. In the second step, a GA is used to adjust the centroid of clusters based on a validity index, which is determined by IT2FCM. Applying this algorithm to the M-FISH problem, the chromosome pixels are segmented into 2 clusters by the GIT2FCM algorithm, and the chromosome pixels of the DAPI channel segmentation result are used as a mask for the remaining five channels. Then, the GIT2FCM algorithm is used to classify the chromosome pixels in the five channels into 24 classes, which correspond to 22 pairs of homologous chromosomes and two sexual chromosomes.

Experiments are completed for the M-FISH images and the results of the classification are compared with the other results.

The paper is organized as follows: Section II reviews related studies on M-FISH segmentation and classification. Section III provides background on interval type-2 fuzzy sets, IT2FCM clustering and the genetic algorithm. Section IV proposes the genetic interval type-2 Fuzzy C-Means clustering. Section V describes M-FISH classification based on the proposed algorithm. Section VI offers some experimental results, and Section VII concludes the paper.

2. Literature review

Researchers have proposed many methods for automating the process of karyotyping. We briefly review some of the major works in this section.

The first M-FISH [3] was a semi-automated analysis system. In this approach, a mask is created from the DAPI channel and a threshold is applied to each pixel in this mask to detect the absence of that pixel in the five remaining fluors.

Almost all of the classification methods used by researchers were developed in one of two directions: pixel-by-pixel classification [10] or region-based classification [8, 9]. Pixel-by-pixel methods either classify each pixel of the M-FISH image or create a binary mask of the DAPI image using edge detection algorithms and then classify each pixel of this mask. In region-based methods, the regions are obtained by decomposing the images that are classified.

Schwartzkopf et al. [12] developed new supervised methods for automatic chromosome identification that exploit the multi-spectral information in M-FISH chromosome images and jointly perform the chromosome segmentation and classification.

Unsupervised and semi-unsupervised classification methods that correct misclassification, are discussed in [6, 19, 21]. Choi et al. [19] proposed a novel unsupervised classification method based on fuzzy logic classification and a prior adjusted reclassification using the chromosome boundaries. The initial classification results can be improved significantly after the prior adjusted reclassification while keeping the translocations intact. A new segmentation method that combines spectral and edge information was also presented. This method provides segmentation accuracy in more than 98% of cases, on average [21] also presented a new unsupervised, non-parametric classification method for M-FISH images that uses the feature normalization method to reduce the difference in feature distributions among images using the expectation maximization (EM) algorithm. This classifier is as accurate as the maximum-likelihood classifier, whose accuracy was also significantly improved after the EM normalization. Karvelis et al. [6] presented a semi-unsupervised method for the M-FISH chromosome image classification. First, the separation of foreground and background is performed by using an automated thresholding procedure. Then, these features are normalized. Second, the K-Means algorithm was applied to cluster the chromosome pixels into the 24 chromosome classes. Although this algorithm does not require a training, it produces a high average accuracy. However, the tests of the algorithm only used a small number of images.

Various pre-processing methods such as image registration, dimension reduction and background flattening are discussed in [15, 16]. Wang et al. [15] used a Bayesian classifier for multi-spectral pixel classification with a multi-resolution registration algorithm based on wavelets and spline approximations and indicated that the proposed registration technique leads to an increased pixel classification rate, which in turn translate into improved accuracy in identifying subtle DNA rearrangements. Choi et al. [16] presented an automated method for the segmentation and classification of multi-spectral chromosome images and proposed pre-processing the images using background correction, the six-channel color compensation method and a feature transformation method, i.e., spherical coordinate transformation. Additionally, color compensation techniques for multichannel fluorescence images

whose specimens are combinatorially stained, which are useful for improving the accuracy of karyotyping, are discussed in [24].

Karvelis et al. [7, 9] proposed a region-based watershed segmentation method applied to the DAPI channel for multi-spectral chromosome image classification, and a new method for the multichannel image segmentation and region classification with the marker-controlled watershed transform. The region Bayes classification method which focuses on region classification, is used. The classifier was trained and tested on non-overlapping chromosome images, and an overall accuracy of 89% was achieved. The superiority of the proposed method over methods that use pixel-by-pixel classification was demonstrated. However, only a small number of non-overlapping testing images were used. In their extended further work [8], the segmentation was based on the multichannel watershed transform to define regions that have similar spatial and spectral characteristics. Then, a Bayes classifier, task-specific on region classification, was applied. The proposed method achieved substantially better results than did the other tested methods at a lower computational cost. The combination of the multichannel segmentation and the region-based classification was found to have a better overall classification accuracy than were the pixel-by-pixel approaches.

A support vector machines (SVM) classifier with a multichannel watershed transform to perform M-FISH karyotyping was described in [23]. The method has been tested on images from normal cells, showing a 10.16% improvement in classification accuracy over the Bayesian classifier.

Hua et al. [17] presented embedded M-FISH image coding (EMIC), where the foreground objects/chromosomes and the background objects/images are coded separately. First, critically sampled integer wavelet transforms were applied to both the foreground and the background. Object-based bit-plane coding was used to compress each object and generate separate embedded bit streams that allow continuous lossy-to-lossless compression of the foreground and the background.

Later, Cao et al. [10] presented an adaptive fuzzy c-means algorithm (FCM), which can be used to detect chromosomal abnormalities for cancer and genetic disease diagnosis, and applied the algorithm to the segmentation and classification of M-FISH images. Adaptive FCM was performed using a gain field, which models and corrects any intensity inhomogeneities caused by a microscope imaging system, chromosomes or uneven DNA hybridization. In addition to directly simulating the homogeneously distributed intensities over the image, the gain field regulates centroids of each intensity cluster. This algorithm provides the lowest segmentation error, and the classification error is smaller than that of the traditional FCM and AFCM methods.

Another work of Cao et al. [11] proposed the development of a sparse representation-based classification (SRC) algorithm based on L1-norm minimization for classifying chromosomes from M-FISH images. The algorithm presents a lower classification error than do other pixel-wise M-FISH image classifiers, such as FCM and AFCM and three different sparse representation methods, i.e., the homotopy method, orthogonal matching pursuit (OMP), and least angle regression (LARS).

Overlapping and touching chromosomes remain a problem in pixel-by-pixel classification. Many studies have attempted to resolve this issue [12–14]. Choi et al. [14] proposed a method that evaluates multiple hypotheses based on geometric information, pixel classification results, and chromosome sizes, and a hypothesis that has a maximum-likelihood chosen as the best decomposition of a given cluster. Approximately 90% accuracy was obtained for two or three chromosome clusters, which include approximately 95% of all clusters with two or more chromosomes. This approach exhibits lower computational complexity than does the minimum entropy approach [12, 13], and the good results have been reported. Barrutia et al. [20] presented a new method, non-negative matrix factorization (NMF), which blindly estimates the spectral contributions and corrects for the overlap.

In light of this brief review , we can conclude that region-based classification approaches are more accurate and have lower computation times compared to pixel-by-pixel approaches. Many studies have only been tested on small sets of selected images. For practical purposes, the systems must be able to provide high accuracy on a large data set. Thus, a study is needed to improve the performance on large data sets which high accuracy so that such automated systems are acceptable for commercial karyotyping.

3. Preliminaries

In this section, we briefly introduce the type-2 fuzzy sets, the interval type 2 Fuzzy C-Means algorithm and the genetic algorithm. Details can be found in [30] and [26].

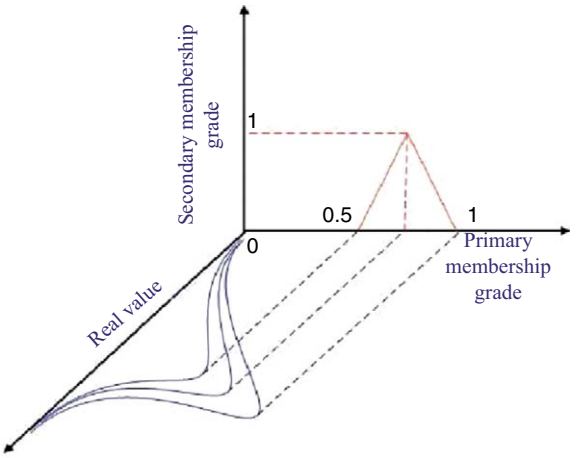


Fig. 1. Explanation of a type-2 fuzzy set.

3.1. Type-2 fuzzy sets

Definition 1. A type-2 fuzzy set, denoted by $\tilde{A}$, is characterized by a type-2 membership function $\mu_{\tilde{A}}(x, u)$ where $x \in X$ and $u \in J_x \subseteq [0, 1]$, i.e.,

$$\tilde{A} = \{((x, u), \mu_{\tilde{A}}(x, u)) | \forall x \in X, \forall u \in J_x \subseteq [0, 1]\} \quad (1)$$

or

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{A}}(x, u))/(x, u),\ J_x \subseteq [0, 1] \quad (2)$$

where $0 \leq \mu_{\tilde{A}}(x, u) \leq 1$.

At each value of x, i.e., $x = x'$, the 2-D plane whose axes are u and $\mu_{\tilde{A}}(x', u)$ is called a *vertical slice* of $\mu_{\tilde{A}}(x, u)$. A *secondary membership function* is a vertical slice of $\mu_{\tilde{A}}(x, u)$, that is, for $x \in X$ and $\forall u \in J_{x'} \subseteq [0, 1]$, $\mu_{\tilde{A}}(x = x', u)$ is written in the following form:

$$\mu_{\tilde{A}}(x = x', u) = \int_{u \in J_{x'}} f_{x'}(u)/u,\ J_{x'} \subseteq [0, 1] \quad (3)$$

where $0 \leq f_{x'}(u) \leq 1$.

Type-2 fuzzy sets are called interval type-2 fuzzy sets if the secondary membership function takes the form: $f_{x'}(u) = 1\ \forall u \in J_x$, i.e., an interval type-2 fuzzy set is defined as follows:

Definition 1. An *interval type-2 fuzzy set* $\tilde{A}$ is characterized by an interval type-2 membership function $\mu_{\tilde{A}}(x, u) = 1$ where $x \in X$ and $u \in J_x \subseteq [0, 1]$, i.e.,

$$\tilde{A} = \{((x, u), 1) | \forall x \in X, \forall u \in J_x \subseteq [0, 1]\} \quad (4)$$

The uncertainty of $\tilde{A}$, denoted by FOU, is the union of the primary functions, i.e., $FOU(\tilde{A}) = \bigcup_{x \in X} J_x$. The upper and lower bounds of the membership function (UMF/LMF) are denoted by $\overline{\mu}_{\tilde{A}}(x)$ and $\underline{\mu}_{\tilde{A}}(x)$, of $\tilde{A}$.

3.2. Interval type-2 fuzzy clustering algorithm

Interval type-2 fuzzy c-means (IT2FCM) is an extension of FCM clustering in which two fuzzification coefficients, m_1, m_2, are used to form FOU, corresponding to the upper and lower values of membership. See [26]. The use of fuzzifiers gives rise to different objective functions to be minimized:

$$\begin{cases} J_{m_1}(U, v) = \sum_{k=1}^{N} \sum_{i=1}^{C} (u_{ik})^{m_1} d_{ik}^2 \\ J_{m_2}(U, v) = \sum_{k=1}^{N} \sum_{i=1}^{C} (u_{ik})^{m_2} d_{ik}^2 \end{cases} \quad (5)$$

in which $d_{ik} = \| x_k - v_i \|$ is the Euclidean distance between the pattern x_k and the centroid v_i, C is the number of clusters, N is the number of patterns and $x_k, v_i \in R^M$.

The upper and lower degrees of membership $\overline{u}_{ik}$ and $\underline{u}_{ik}$ are similar to those produced by the FCM algorithm, but they are formed by involving two fuzzification coefficients m_1, m_2 ($m_1 < m_2$) as follows:

$$\overline{u}_{ik} = \begin{cases} \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)^{2/(m_1-1)}} \ if \ \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)} < \dfrac{1}{C} \\ \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)^{2/(m_2-1)}} \ if \ \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)} \geq \dfrac{1}{C} \end{cases} \quad (6)$$

$$\underline{u}_{ik} = \begin{cases} \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)^{2/(m_1-1)}} \ if \ \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)} \geq \dfrac{1}{C} \\ \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)^{2/(m_2-1)}} \ if \ \dfrac{1}{\sum_{j=1}^{C} \left(d_{ik}/d_{jk}\right)} < \dfrac{1}{C} \end{cases} \quad (7)$$

in which $i = 1, ..., C$ and $k = 1, ..., N$.

Because each pattern takes the membership interval bounded by the upper $\overline{u}$ and the lower $\underline{u}$, each centroid of the cluster is represented by the interval between v^L

and v^R. The centroids are computed in the same way as FCM, as follows:

$$v_i = \frac{\sum_{k=1}^{N}(u_{ik})^m x_k}{\sum_{k=1}^{N}(u_{ik})^m} \tag{8}$$

in which $i = 1, ..., C$ and m is arbitrary and is usually given the value 2.

After obtaining v_i^R, v_i^L,the centroid of clusters are defined as follows:

$$v_i = (v_i^R + v_i^L)/2 \tag{9}$$

for membership grades

$$u_i(x_k) = (u_i^R(x_k) + u_i^L(x_k))/2,\ j = 1, ..., C \tag{10}$$

in which

$$u_i^L = \sum_{l=1}^{M} u_{il}/M,\ u_{il} = \begin{cases} \overline{u}_i(x_k) \text{ if } x_{il} \text{ uses } \overline{u}_i(x_k) \text{ for } v_i^L \\ \underline{u}_i(x_k) \qquad otherwise \end{cases} \tag{11}$$

$$u_i^R = \sum_{l=1}^{M} u_{il}/M,\ u_{il} = \begin{cases} \overline{u}_i(x_k) \text{ if } x_{il} \text{ uses } \overline{u}_i(x_k) \text{ for } v_i^R \\ \underline{u}_i(x_k) \qquad otherwise \end{cases} \tag{12}$$

Defuzzification is made for IT2FCM as if $u_i(x_k) > u_j(x_k)$ for $j = 1, ..., C$ and $i \neq j$, then x_k is assigned to cluster i.

3.3. Genetic algorithm

The genetic algorithm (GA) [42] is an artificial system based on the principle of natural selection. As a stochastic algorithm, GA is a robust and powerful optimization method for solving problems with a large search space which are not easily solved by exhaustive methods. In particular, each potential solution is seen as an individual and as appropriate encryption called a chromosome. Genetic algorithms simulate evolution on a population of chromosomes to find a solution to the problem.

In GA applications, the parameters are encoded in chromosomes. A chromosome is encoded with binary, integer or real numbers in an artificial genetic algorithm. Usually, a basic GA consists of three operators: selection, crossover, and mutation [41].

3.3.1. Selection

In our method, the Roulette wheel selection is used and briefly described as follows:

Calculate the adaptability of the population

$$F = \sum_{i=1}^{NP} f_i \tag{13}$$

where NP is the number of individuals in the population and f_i is the fitness of the *ith* individual in the population.

Calculate the probability of an individual:

$$p_i = \frac{f_i}{F} \tag{14}$$

Calculate cumulative probability (q_i) for the *ith* individual:

$$q_i = \sum_{j=1}^{i} p_i \tag{15}$$

In selection process, roulette wheel spins NP times equal to the population size. Each time an individual is selected for a new population. In fact, this step can be done each time as follows:

1. Generate a random number r from the range [0, 1].
2. If $r < p_1$, then select the first individual, otherwise select the *ith* individual such that $q_{(i-1)} < r < q_i$.

3.3.2. Crossover

The purpose of the crossover operation is to create two new individuals from two existing individuals selected randomly from the current population. Typical crossover operations are one-point crossover, two-point crossover, cycle crossover and uniform crossover. In this research, the one-point crossover is used with a fixed crossover probability of μ_c.

For the one-point crossover, two individuals are randomly selected from the population. Assuming the length of an individual to be m, this process randomly selects a point between 1 and $m - 1$ and swaps the content of the two individuals beyond the crossover point to obtain the offspring. For example, two individuals $x = (x_1...x_m)$ and $y = (y_i, ..., y_m)$, with the crossover point k, will be two individuals: $x' = (x_1...x_k, y_{k+1}...y_m)$ and $y' = (y_1...y_k, x_{k+1}...x_m)$.

A crossover between a pair of chromosomes is affected by the crossover probability.

1. Generate a random number r from the range [0, 1].
2. If $r < \mu_c$, then do crossover operator.

3.3.3. Mutation

After crossover, for each gene of each individual chromosome we generate a random number $r[0, 1]$. If $r < \mu_m$, the gene is mutated where $\mu_m \in [0, 1]$ is a fixed probability. All values of a specific gene may be randomly changed. In this paper, a mutation operator with uniform distribution is used as follows:

If x_k gene of chromosome $x = (x_1 ... x_m)$ is mutated, we get a new $x^{'}$ is: if $i = k$ then $x_i^{'} = x_i + \sigma *$ x_i *with* $x_i > 0$ or $x_i^{'} = x_i + \sigma$ *with* $x_i = 0$ else $x_i^{'} = x_i$

4. Genetic type-2 FCM clustering

The objective of the general clustering algorithms is to minimize the objective function. Hence, the objective function may be different for different problems. This paper proposes genetic type-2 fuzzy c-means clustering, which combines type-2 fuzzy c-means clustering method, a good method for handling uncertainty, with a genetic algorithm to find the optimal solution for the objective function.

To apply the genetic algorithm to the clustering problems, we need to model the problems with the chromosomes and evaluate these chromosomes. Here, each chromosome is a centroid of a cluster and the length of the chromosome, K, is equivalent to the number of clusters in the clustering problem. K is in the range $[K_{min}, K_{max}]$, where K_{min} is usually set to 2 and K_{max} is the maximum number cluster centroids, which describes the maximum number of chromosomes for the human species.

Therefore, K_{max} must be selected according to experience. Without assigning the number of clusters in advance, a variable string length is used. Invalid (non-existing) clusters are represented by a negative integer, specifically "-1". The values of the chromosomes are changed in an iterative process to determine the correct number of clusters (the number of valid units in the chromosomes) and the actual cluster centroids for a given clustering problem.

However, the M-FISH problem has the previously determined number of cluster which is 24 for the M-FISH classification. Therefore, we do not need to find the optimal number of clusters. Thus, the length of the chromosome K is unchanged. We will have $K = K_{min} = K_{max}$, and we do not randomly initialize the cluster centroids with "-1".

4.1. Population initialization

In this study, a chromosome is encoded with a unit that represents a potential centroid, and the population size is NP. For population initialization, all values of the chromosomes are chosen randomly from the data space meaning that each chromosome is encoded as the cluster centroid results, which are randomly selected from the input data space.

If a chromosome belongs to the so-called parent generation then its size is K and it is a potential solution of the IT2FCM algorithm. The size of the population, NP, is selected in the experiment.

4.2. The Turi's validity index

In [5, 40] the Turi's validity index is introduced to evaluate the quality of the image clustering problems and proves to be more effective than do the previous used validity indexes such as the Davies-Bouldin index, DunneâĂŹs index and the PC index. Therefore, for image clustering applications, this paper uses this index as the objective function of the clustering algorithm; refer to [5, 40].

The Turi's validity index is given by the following function:

$$V_T = \alpha \times \frac{intra}{inter} \tag{16}$$

where the term *intra* of V_T is the average of the distances to each pixel x within the cluster C_i from centroid z_i, as defined in the following:

$$intra = \frac{1}{N} \sum_{i=1}^{K} \sum_{x \in C_i} (d_i)^2 \tag{17}$$

in which $d_i = ||x - z_i||$

The *inter* is the minimum distance between the cluster centroids and is defined as follows:

$$inter = min(\| v_i - v_j \|^2) \tag{18}$$

where α is a weighting factor, given as

$$\alpha = c \times N(2, 1) + 1 \tag{19}$$

where c is a user-specified parameter and $N(2, 1)$ is a normal distribution function with a mean of 2 and a standard deviation of 1.

$$N(2, 1) = \frac{1}{\sqrt{2\pi \times 1^2}} e^{-\frac{(k-2)^2}{2 \times 1^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(k-2)^2}{2}} \tag{20}$$

where k is the number of clusters (details in [40]).

This validity measure serves the dual purpose of minimizing cluster spread and maximizing the separation of the clusters. However, this value is influenced by the geometry of the cluster centroids.

4.3. The fitness of chromosomes

Because the GA is used to find the optimal solution or minimize the objective function, we need to determine the fitness of chromosomes. In this research, Turi's validity index is used as the objective function because it has the proven ability to obtain high clustering quality for the image processing problem [40]. The chromosomes are encoded as the cluster centroid results. The fitness for chromosomes is computed in the following three steps.

Step 1: The pixel dataset is clustered according to the centroid encoded in the focal chromosome such that each pattern $x_i, i = 1, ..., N$ is assigned to cluster with centroid v_j, $j = 1, ..., K$ according to equations (6), (7) and (10) in IT2-FCM.

Step 2: This step adjusts the values of the cluster centroids encoded in the chromosome, replacing these values with the mean points of the respective clusters. In IT2FCM, the new center v_i^* for the cluster C_i is given as

$$v_i^* = \frac{v_i^R + v_i^L}{2} \tag{21}$$

Step 3: The validity index V_T is calculated using (16). Because the goal of the problem is to achieve a proper clustering by minimizing V_T, the fitness value for chromosome j is defined as $1/V_T^j$, which is equivalent to the clustering with the smallest inner-cluster scatter and the largest cluster separation.

Therefore, the fitness of chromosomes is calculated as follows:

$$f = \frac{1}{V_T} \tag{22}$$

4.4. Genetic type-2 fuzzy C-Means clustering algorithm

The performance of this algorithm is given by a sequence of steps:

The proposed algorithm performs the iterative processes from steps 2 to 4 until a stopping criterion is met. In every generation cycle, the fittest chromosome is preserved until the last generation. Thus, on termination, this chromosome gives us the best solution encountered during the search.

To handle the uncertainty of the M-FISH image, the IT2FCM algorithm is used to evaluate the population in step 2. After calculating the fitness of each chromosome of the given population, the best chromosome is compared to the best chromosome of the previous generation (iteration). The best selected chromosome is the result of the problem with the proper number of clusters.

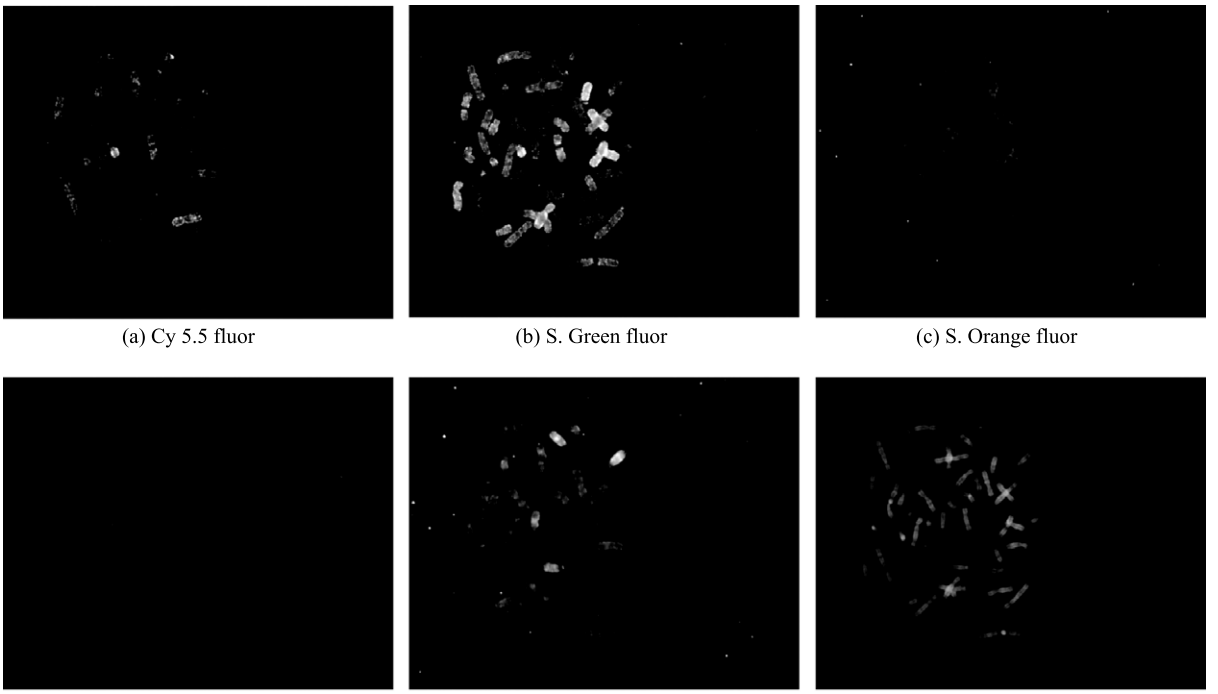
Algorithm 1. Genetic Type-2 FCM

1. Initializing a population of NP chromosomes which each chromosome contains the cluster centroids are randomly selected from the input image.
2. Evaluate on the population by the fitness function (22).
3. Perform GA operators such as: Selection, Crossover and Mutation.
4. Reinsert the new individual chromosomes into the population.
5. Termination criterion: the predetermined number of iterations is achieved or the difference between these two best fitness values lies below a predefined threshold.
6. Output: A chromosome which has the best fitness. It means that the solution with the the cluster centroids is the output of this algorithm.

5. M-FISH classification

M-FISH is a multi-channel chromosome image generated technique that allows the color of the human chromosomes to be distinguished. By analyzing the color images, a geneticist can easily determine which parts are lost or rearranged in the chromosomes, and use this information in the study of cancer and disorders genetics. M-FISH images were taken using a fluorescence microscope with optical filters. Each of the fluors can be observed in one of the spectral channels. An M-FISH image consists of six images, and each image is the response of the chromosome to the particular fluor, as shown in Fig. 2.

In the M-FISH classification method, we can use the pixel values in all of 6-image sets of image channels for classification, but this requires a training stage [35]. However, in the M-FISH technique, all chromosomes are labelled with 5 fluors and a fluorescent DNA stain called DAPI (4 in,6-diamidino-2-phenylindole) that attaches to DNA and labels all chromosomes. DAPI is usually used to create binary masks to classify chromosomes [6, 10, 19]. Thus, in our work, a binary mask of the DAPI image is first created, and the chromosome pixels in the five remain channels are then classified through this mask.



(a) Cy 5.5 fluor (b) S. Green fluor (c) S. Orange fluor

(d) Cy 5 fluor (e) Texas Red fluor (f) DNA DAPI stain

Fig. 2. Six-channels of M-FISH image data

With the M-FISH image segmentation problem, our mission is to use information from the 6-channel image to classify the chromosomes into 24 classes, as described above. Because the original DAPI image contains nuclei and debris along with the chromosomes, we must remove the noise based on the size and density before segmentation. Figure 3 shows the image before and after noise removal.

Based on the proposed algorithm (GIT2FCM), we performed M-FISH classification in two main stages:

Stage 1: the DAPI channel is segmented into two clusters (background and chromosomes). Segmentation is implemented with the following parameters set for GIT2FCM: the value of the parameter C, for the validity index referred from [40], is set to 25; the number of iterations G is the terminating condition; the size of the population is NP; the GA selection method is Roulette Wheel; crossover rate $\mu_c = 0.9$ and mutation rate $\mu_m = 0.01$. By implementing many experiments, the values are set with stable results of $G = 20$ and $NP = 30$.

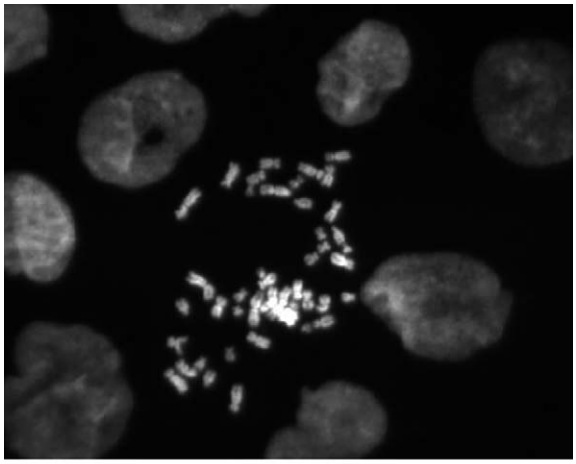
Although the GIT2FCM exhibits the ability to find the optimal number of clusters, the M-FISH problem is applied with a determined number of clusters, 2 for the segmentation stage (one for background and one for the chromosome pixels) and 24 for the classification stage. Thus, in this stage, the chromosome length is the number of clusters, i.e., $K = K_{min} = Kmax = 2$.

Stage 2: After the segmentation of the DAPI image, a cluster contains the pixels belonging to the chromosomes. Suppose that this cluster contains N pixels. Each pixel $I(x, y)$ in this cluster possesses the intensity values of five M-FISH channels, i.e., $I(x, y) = [x_1, x_2, x_3, x_4, x_5] \in R^5$, where x_i is the intensity value of the ith M-FISH channel at pixel (x, y). If the ranges of the features in each dimension vary considerably, the performance of the classification phase may be affected [3]. Thus, the features are normalized as follows:

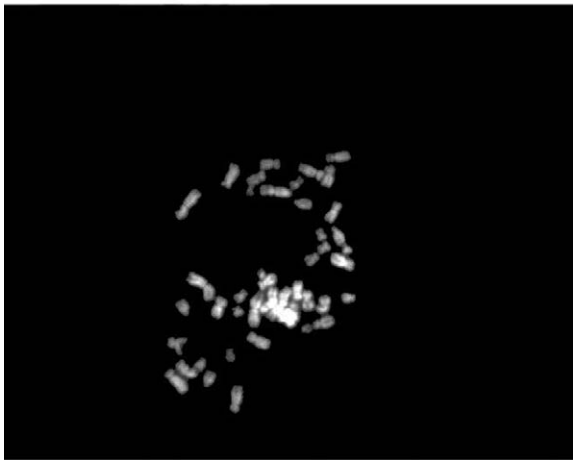
$$y_i = \frac{x_i^j - z^j}{s^j} \tag{23}$$

where x_i^j is the intensity of the ith pixel of channel jth, z^j and s^j are the mean and standard deviation of channel jth, respectively.

$$z^j = \frac{1}{N}\sum_{i=1}^{N}(x_i)^j \tag{24}$$



(a) Original DAPI



(b) DAPI after noise removal

Fig. 3. Noise removal of the DAPI image.

$$s^j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^j - z^j)^2} \tag{25}$$

Now, each pixel in N pixels of the chromosome cluster of the DAPI channel is represented by a feature vector $Y_i = [y_i^1, y_i^2, y_i^3, y_i^4, y_i^5]$, $i = 1, ..., N$, where y_i^j is the normalized intensity of channel jth.

Our purpose is to classify the data set $Y = [Y_i]$, $i = 1, ..., N$ into 24 classes, which correspond to 22 pairs of homologous chromosomes and 2 sexual chromosomes (X and Y) or, more specifically to cluster the data into 24 clusters. The classifications are implemented with the following parameters of the GIT2FCM: the parameter C, the validity index described in [40], is set to 25; the number of iterations G is the terminating condition of the proposed algorithm; the size of the population is NP; the GA selection method is Roulette Wheel; crossover rate $\mu_c = 0.9$ and mutation rate $\mu_m = 0.01$. In the segmentation phase, the values of $G = 20$ and $NP = 50$ are obtained through experiments with stable results. In a similar way, the chromosome length is set to 24 ($K = K_{min} = Kmax = 24$).

The performance of the segmentation and classification are evaluated using the correct detection rate (CR), false detection rate (FR) and classification ratio (CT) [4], which are defined by the following equations:

$$\text{CR} = \frac{\text{pixels correctly segmented}}{\text{total chromosome pixels}} \tag{26}$$

$$\text{FR} = \frac{\text{background pixels segmented as chromosome}}{\text{chromosome pixels}} \tag{27}$$

$$\text{CT} = \frac{\text{chromosome pixels correctly classified}}{\text{total chromosome pixels}} \tag{28}$$

6. Experimental results

The images used for testing are taken from the M-FISH database [46]. This database contains 20 M-FISH image sets and each M-FISH image set consist of 5 mono-spectral images recorded at different wavelengths, DAPI and its "ground truth" image according to ISCN (International System for Human Cytogenetic Nomenclature) for each M-FISH image. The ground truth image used to determine the accuracy of the M-FISH images classification is labeled based on the gray level of each pixel so that each gray level represents a chromosome type; the value of the background pixels is set to zero, and the value of pixels in the overlapping regions is 255.

We performed the image segmentation and calculated CR and FR according to Equation (26) and Equation (27) from 120 images using IT2FCM [26] and the GIT2FCM method.

In Table 1, the results were summarized by Karvelis et al. [8, 9] for the ADIR M-FISH database [45], which contains the test database [46] with 200 M-FISH images. In this work, the CR was 83.59%±9.89% for only 15 non-overlapping M-FISH images and 82%±12% when fusing 183 M-FISH images (excluding 17 images, which were marked as extreme (EX) for "difficult to karyotype", from the set of 200).

The segmentation results of Otsu's method in [25] and the results of IAFCM, AFCM and FCM methods

Table 1
Segmentation results for the GIT2FCM, IT2FCM, IAFCM, AFCM, FCM, Otsu and region [9], and watershed [8] methods

Methods	GIT2FCM	IT2FCM	IAFCM	AFCM	FCM	Otsu	Region	Watershed
CR(%)	92.5 ± 3.5	92.2 ± 6.5	89.5 ± 10.5	96.5 ± 4.6	92.0 ± 8.4	89.1 ± 9.2	83.6 ± 9.9	82 ± 12
FR(%)	1.4 ± 0.6	3.7 ± 2.3	3.6 ± 2.8	20.9 ± 12.9	9.7 ± 7.7	11.7 ± 8.7	NA	NA

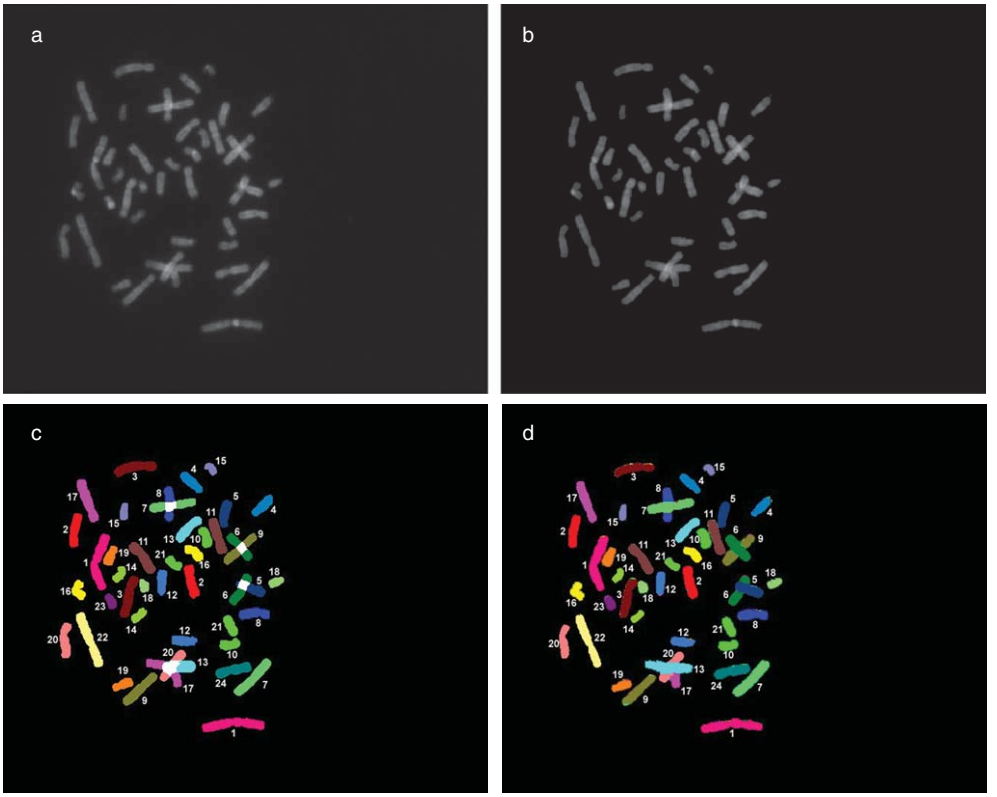


Fig. 4. a) a DAPI channel; b) a DAPI channel after segmentation; c) the ground truth; d) GIT2FCM classification with 24 classes.

in [10], which were tested on the same database, are also listed and have a higher CR than does Karvelis's method. However, with the lowest false detection ratio (FR), our experiments showed that the segmentation results of the proposed GIT2FCM algorithm are better than those of the others. In addition, the chromosome segmentation accuracy (CR) is also higher than that of the other methods, except the AFCM algorithm.

After the segmentation stage, we conducted a classification stage to classify M-FISH images into 24 classes, which correspond to 22 pairs of homologous chromosomes and two sexual chromosomes. A sample classification result is shown in Fig. 4.

In Fig. 4, (a) is for the DAPI channel image, (b) is for the DAPI channel image after segmentation, (c) is for the ground truth image and (d) is for the image classification result GIT2FCM algorithm where each class is shown as a different color.

We calculated the CT values using Equation (28) from 20 M-FISH image classification results with the IT2FCM [26] and GIT2FCM methods. In Equation (28), the number of the correctly classified chromosome pixels is determined in reference to the "ground truth" image. Although the overlapping regions of chromosomes are not labelled in the "ground truth" image and these regions can belong to any overlapping chromosomes, they are considered as the correctly classified chromosome pixels in the classification results.

The classification results from 20 tested cells with 120 M-FISH images are listed in Table 2. In addition to the classification result of the GIT2FCM (the proposed algorithm), the classification results were performed on

Table 2
Classification results for the GIT2FCM, IT2FCM, IAFCM, AFCM, FCM and the k-means methods

Results	GIT2FCM	IT2FCM	IAFCM	AFCM	FCM	K-Mean
Average	90.3	86.3	88.5	84.2	84.1	72.48
Standard deviation	3.6	6.2	5.5	5.6	8.5	7.01

the same database [46] for the IT2FCM [26], IAFCM, AFCM and FCM algorithms [10] and the k-means algorithm [6] are shown in the table for comparison with the proposed algorithm.

In Table 2, the classification results (the CT values) show that the proposed algorithm (GIT2FCM) is better than the others because the former has the highest average classification ratio (CT) and the lowest standard deviation values.

7. Conclusions

This paper presented a clustering algorithm based on a genetic technique that determines a good initial value for the centroid and realized image segmentation based on IT2FCM and Turi's validity index (GIT2FCM) to improve the results. The experimental results obtained using the M-FISH images showed a more accurate chromosome segmentation (CR) value and a lower false detection ratio (FR) than did the "conventional" algorithms, including FCM, IAFCM and k-means. Additionally, the M-FISH classification results have higher average CT values than do the other FCM, such as IAFCM, AFCM and FCM.

The next goal is to perform research related to speeding-up the algorithms for processing large datasets based on the parallel architecture of GPU computing.

References

[1] M.W. Thompson, R.B. McInnes and H.G. Willard, *Genetics in Medicine*, 5th Edition, (2001), WB Saunders Company.

[2] B. Dave and W. Sanger, Role of cytogenetics and molecular cytogenetics in the diagnosis of genetic imbalances, *Seminars in Pediatric Neurology* **14**(1) (2007), 2–6.

[3] M.R. Speicher, S.G. Ballard and D.C. Ward, Karyotyping human chromosomes by combinatorial multi-fluor fish, *Nature Genetics* **12**(4) (1996), 368–375.

[4] E. Schrock, S. du Manoir and T. Veldman, Multicolor spectral karyotyping of human chromosomes, *Science* **273**(5274) (1996), 494–497.

[5] A. Halder and S. Pramanik, An unsupervised dynamic image segmentation using fuzzy hopfield neural network based genetic algorithm, *International Journal of Computer Science Issues* **9** (2012), 525–532.

[6] P. Karvelis, A. Likas and D.I. Fotiadis, Semi unsupervised M-FISH chromosome image classification, *10th IEEE Int Cont on Digital Object Identifier* (2010), 1–4.

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

[8] P.S. Karvelis, A.T. Tzallas, D.I. Fotiadis and I. Georgiou, A multichannel watershed based segmentation method for multispectral chromosome classification, *IEEE T on Medical Image*, **27**(5) (2008), 697–708.

[9] P.S. Karvelis, D.I. Fotiadis and A. Tzall, Region based segmentation and classification of multi-spectral chromosome images, *20th Int Symposium on Digital Object Identifier Computer-Based Medical Systems*, CBMS 09, (2009), 251–256.

[10] H.B. Cao, H.W. Deng and Y.P. Wang, Segmentation of M-FISH images for improved classification of chromosomes with an adaptive fuzzy C-means clustering algorithm, *IEEE T on Fuzzy Systems* **20**(1) (2012), 1–8.

[11] H.B. Cao, H.W. Deng and Y.P. Wang, Classification of multicolor fluorescence *in situ* hybridization (M-FISH) images with sparse representation, *IEEE T on NanoBioScience* **11**(2) (2012), 112–118.

[12] W.C. Schwartzkopf, B.L. Evans and A.C. Bovik, Minimum entropy segmentation applied to multi-spectral chromosome images, *Proc IEEE Int Conf on Image Processing, II* (2001), 865–868.

[13] W.C. Schwartzkopf, B.L. Evans and A.C. Bovik, Entropy estimation for segmentation of multi-spectral chromosome images, *IEEE Southwest Symposium on Image Analysis and Interpretation* (2002), 234–238.

[14] H. Choi, A.C. Bovik and K.R. Castleman, Maximum-likelihood decomposition of overlapping and touching M-FISH chromosomes using geometry, size and color information, *the 28th IEEE EMBS Int' Conference* (2006), 3130–3133.

[15] Y.P. Wang, M-FISH image registration and classification, *IEEE Int, Symp Biomedicine Image: Nano to Macro* **1** (2004), 57–60.

[16] H. Choi, K.R. Castleman and A.C. Bovik, Joint segmentation and classification of M-FISH chromosome images, in *Proc 26th IEEE Annu Int Conf (EMBS)* **1** (2004), 1636–1639.

[17] J. Hua, Z. Xiong, Q. Wu and K.R. Castleman, wavelet-based compression of M-FISH images, *IEEE T on Biomedical Engineering* **52**(5) (2005), 890–900.

[18] W.C. Schwartzkopf, A.C. Bovik and B.L. Evans, Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images, *IEEE T Medical Imaging* **24**(12) (2005), 1593–1610.

[19] H. Choi, K.R. Castleman and A.C. Bovik, Segmentation and fuzzy-logic classification of M-FISH chromosome images, *Proc IEEE Int Conf Image Process* (ICIP 2006), 69–72.

[20] A.M. Barrutia, J.G. Munoz, B. Ucar, I.F. Garcia and C.O. Solorzano, Blind spectral unmixing of M-FISH images by non-negative matrix factorization, *the 29th Int' Conference of the IEEE EMBS* (2007), 6247–6250.

[21] H. Choi, A.C. Bovik and K.R. Castleman, Feature normalization via expectation maximization and unsupervised

nonparametric classification For M-FISH chromosome images, *IEEE T on Medical Imaging* **27**(8) (2008), 1107–1119.
[22] Y.P. Wang, Detection of chromosomal abnormalities with multi-color fluorescence *in situ* hybridization (M-FISH) imaging and multi-spectral wavelet analysis, *30th International IEEE EMBS Conference* (2008), 1222–1225.
[23] I. Georgiou, P. Sakaloglou, P.S. Karvelis and D.I. Fotiadis, Enhancement of the classification of multichannel chromosome images using support vector machines, *31st International Conference of the IEEE EMBS* (2009), 3601–3604.
[24] H. Choi, K.R. Castleman and A.C. Bovik, Color compensation of multicolor FISH images, *IEEE T on Medical Imaging* **28**(1) (2009), 129–136.
[25] N. Otsu, A threshold selection method from gray-level histograms, *IEEE T on System, Man, Cybernetic* **9**(1) (1979), 62–66.
[26] C. Hwang and F.C.H. Rhee, Uncertain fuzzy clustering: Interval type-2 fuzzy approach to C-means, *IEEE T on Fuzzy Systems* **15**(1) (2007), 107–120.
[27] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman and A.Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, *IEEE T on Pattern Analysis and Machine Intelligence* **24**(7) (2002), 881–892.
[28] M.C. Hung, J. Wu, J.H. Chang and D.L. Yang, An efficient k-means clustering algorithm using simple partitioning, *J of Info Science and Engineering* **21**(6) (2005), 1157–1177.
[29] K.R. Zalik, An efficient k'-means clustering algorithm, *Pattern Recognition Letters* **29**(9), 1385–1391.
[30] J. Mendel and R. John, (2002), Type-2 fuzzy set made simple, *IEEE T on Fuzzy Systems* **10**(2) (2008), 117–127.
[31] N.N. Karnik and J.M. Mendel, Operations on type-2 fuzzy sets, *Fuzzy Sets and Systems* **122**(2) (2001), 327–348.
[32] M.R. Rezaee, B.P.F. Lelieveldt and J.H.C. Reiber, A new cluster validity index for the fuzzy c-mean, *Pattern Recognition Letter* **19**(3) (1998), 237–246.
[33] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, (1981), New York: Plenum.
[34] J. Ma and B. Cao, The Mahalanobis distance based rival penalized competitive learning algorithm, *Lect Note Compute Science* **3971** (2006), 442–447.
[35] M.P. Sampat, K.R. Castleman and A.C. Bovik, Pixel-by-pixel classification of MFISH images, *Engineering in Medicine and Biology* **2** (2002), 999–1000.
[36] L. Xu, Bayesian Ying-Yang machine, clustering and number of clusters, *Pattern Recognition Letter* **18**(11) (1997), 1167–1178.
[37] D. Steinley and M.J. Brusco, Initialization k-means batch clustering: A critical evaluation of several techniques, *Journal of Classification* **24**(1) (2007), 99–121.
[38] Y.M. Cheug, On rival penalization controlled competitive learning for clustering with automatic cluster number selection, *IEEE T on Knowledge and Data Engineering* **17**(11) (2005), 1583–1588.
[39] S.J. Robert, R. Everson and I. Rezek, Maximum certainty data partitioning, *Pattern Recognition* **33**(5) (2000), 833–839.
[40] R.H. Turi, *Clustering-Based Color Image Segmentation*, PhD Thesis, Monash University (2001), Australia.
[41] S. Haykin, *Neural Networks: A Comprehensive Foundation* (1999), Prentice-Hall, Inc.
[42] J.H. Holland, *Adaptation in Natural and Artificial Systems* (1975), MIT Press.
[43] Y.G. Tang, F.C. Sun and Z.Q. Sun, Improved Validation Index for Fuzzy Clustering, *American Control Conference* (2005), 1120–1125.
[44] W. Wang and Y. Zhang, On fuzzy cluster validity indices, *Fuzzy Sets and Systems* **158**(3) (2007), 2095–2117.
[45] The ADIR M-FISH Image Database (2008), http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml
[46] M-FISH data (2010), http://sites.google.com/site/xiaobaocao006/database-for-download.