Multiple Feature Fuzzy c-means Clustering Algorithm for Segmentation of Microarray Images

J. Harikiran[1], P.V. Lakshmi[2], R. Kiran Kumar[3]
[1,2]Department of IT, GIT, GITAM University, India
[3]Department of CS, Krishna University, India

Article Info

Article history:

Received Apr 15, 2015
Revised Jun 3, 2015
Accepted Jun 20, 2015

Keyword:

Empirical Mode Decomposition
Image Processing
Image Segmentation
Microarray

ABSTRACT

Microarray technology allows the simultaneous monitoring of thousands of genes. Based on the gene expression measurements, microarray technology have proven powerful in gene expression profiling for discovering new types of diseases and for predicting the type of a disease. Gridding, segmentation and intensity extraction are the three important steps in microarray image analysis. Clustering algorithms have been used for microarray image segmentation with an advantage that they are not restricted to a particular shape and size for the spots. Instead of using single feature clustering algorithm, this paper presents multiple feature clustering algorithm with three features for each pixel such as pixel intensity, distance from the center of the spot and median of surrounding pixels. In all the traditional clustering algorithms, number of clusters and initial centroids are randomly selected and often specified by the user. In this paper, a new algorithm based on empirical mode decomposition algorithm for the histogram of the input image will generate the number of clusters and initial centroids required for clustering. It overcomes the shortage of random initialization in traditional clustering and achieves high computational speed by reducing the number of iterations. The experimental results show that multiple feature Fuzzy C-means has segmented the microarray image more accurately than other algorithms.

Corresponding Author:

Jonnadula Harikiran,
Departement of Information Technology,
GITAM Institute of Technology,
GITAM University, Visakhapatnam.
Email: jhari.kiran@gmail.com

1. INTRODUCTION

Microarrays widely recognized as the next revolution in molecular biology that enable scientists to monitor the expression levels of thousands of genes in parallel [1]. A microarray is a collection of blocks, each of which contains a number of rows and columns of spots. Each of the spot contains multiple copies of single DNA sequence [2]. The intensity of each spot indicates the expression level of the particular gene [3]. The processing of the microarray images [4] usually consists of the following three steps: (i) gridding, which is the process of segmenting the microarray image into compartments, each compartment having only one spot and background (ii) Segmentation, which is the process of segmenting each compartment into one spot and its background area (iii) Intensity extraction, which calculates red and green foreground intensity pairs and background intensities [5].

In digital image segmentation applications, clustering technique is used to segment regions of interest and to detect borders of objects in an image. Clustering algorithms are based on the similarity or dissimilarity index between pairs of pixels. It is an iterative process which is terminated when all clusters contain similar data. In order to segment the image, the location of each spot must be identified through gridding process. An automatic gridding method by using the horizontal and vertical profile signal of the image presented in [6] is used to perform the image gridding. The algorithm can satisfy the requirements of microarray image segmentation. In the clustering algorithms, parameters such as cluster number and initial centroid positions are chosen randomly and often specified by the user. Instead of randomly initializing the parameters in the clustering algorithms, the ECNC (Estimation of Centroids and Number of Clusters) algorithm using Empirical Mode Decomposition on the histogram of input image will automatically determine the cluster centers and the number of clusters in the image. Using ECNC algorithm as a preliminary stage with clustering algorithms reduces the number of iterations for segmentation and costs less execution time. This algorithm is an extended version for the Hill climbing algorithm presented in [17] for estimation of clustering parameters and works even the image contains low level noise.

Many microarray image segmentation approaches have been proposed in literature. Fixed circle segmentation [7], Adaptive circle Segmentation Technique [8], Seeded region growing methods [9] and clustering algorithms [10] are the methods that deal with microarray image segmentation problem. This paper mainly focuses on clustering algorithms. These algorithms have the advantages that they are not restricted to a particular spot size and shape, does not require an initial state of pixels and no need of post processing. These algorithms have been developed based on the information about the intensities of the pixels only (one feature). But in the microarray image segmentation problem, not only the pixel intensity, but also the distance of pixel from the center of the spot and median of intensity of a certain number of surrounding pixels influences the result of clustering. In this paper, multiple feature fuzzy c-means clustering algorithm is proposed, which utilizes more than one feature. The qualitative and quantitative results show that multiple feature fuzzy C-means clustering algorithm has segmented the image better than other clustering algorithms.

The paper is organized as follows: Section 2 presents Empirical Mode Decomposition, Section 3 presents ECNC Algorithm, Section 4 presents fuzzy c-means clustering algorithm, Section 5 presents multiple feature clustering algorithm, Section 6 presents Experimental results and finally Section 7 report conclusions.

2. EMPIRICAL MODE DECOMPOSITION

The Empirical Mode Decomposition (EMD) proposed by Norden Huang [11], was a technique for analyzing nonlinear and non-stationary signals. It serves as an alternative to methods such as wavelet analysis and short-time Fourier transform. It decomposes any complicated signal into a finite and often small number of Intrinsic Mode Functions (IMF). The IMF is symmetric with respect to local zero mean and satisfies the following two conditions.

1. The number of extrema and the number of zero crossings must either be equal or differ by one.
2. At any point, the mean value of the envelope defined by local maxima and local minima is zero, indicating the function is locally symmetric.

The decomposition method in EMD is called Shifting Process [13]. The shifting process of the 1-dimensional signal can be adapted as follows.

1. Let Ioriginal be the original signal to be decomposed. Let j=1 (index number of IMF), Initially, I= Ioriginal.
2. Identify the local maxima and local minima points in I.
3. By using interpolation, create the upper envelope Eup of local maxima and the lower envelope Elw of local minima.
4. Compute the mean of the upper envelope and lower envelope.
   Emean= [Eup + Elw]/2
5. Iimf = I- Emean.
6. Repeat steps 2-5 until Iimf can be considered as an IMF.
7. IMF(j)= Iimf, j=j+1, I = I- Iimf ,
8. Repeat steps 2-7 until, the standard deviation of two consecutive IMFs is less than a predefined threshold or the number of extrema in I is less than two.

The first few IMFs obtained from EMD contain the high frequency components which correspond to salient features in original image and the residue represents low frequency component in the image. The original image can be recovered by inverse EMD as follows:

$$I = RES + \sum_{j} IMF(j) \qquad (1)$$

3. ESTIMATION OF CENTROIDS AND NUMBER OF CLUSTERS (ECNC)

1. Let h(k) be the histogram for the input image I with k=0 ,…., G and G being the maximum intensity value in the image
2. Divide the histogram h(k) into IMFs using empirical mode decomposition. The first IMF carries the histogram noise, irregularities and sharp details of the histogram, while the last IMF and residue describe the trend of the histogram. On the other hand, the intermediate IMFs describe the initial histogram with simple and uniform pulses.
3. Consider the summation of intermediate IMFs as follows:

$$\text{IINT} = \sum_{j=2}^{n-1} IMF_j \tag{2}$$

   where n is the number of IMFs.

4. Determine all local minima in IINT.

$$I* = \left\{ \min_{0 \le T \le G} I_{INT}\left(T\right) \right\} \tag{3}$$

   I* is the vector carrying all local minima. All those local minima could express image clusters, but most of them are very close to each other and some of them lie too high to be a cluster. So, truncate the local minima to the important ones that could express an image cluster.

5. The truncation process is carried out in two steps. In the first step, the algorithm truncates all local minima that have a value larger than the threshold, where threshold is equal to average of the values of local minima. The truncation step is expressed as follows:

$$thr = \frac{1}{2N_{I*}} \sum_{I_i^* \in I^*} I_i^* \tag{4}$$

   where N_{I*} is the number of local minima belonging to $I*$ and I_i^* is the local minima belonging to vector $I*$.

$$I^t = \{I_i^*\} \text{, if } I_i^* < \text{thr and } I_i^* \in I* \tag{5}$$

   Where I^t consists of all local minima which are less than the estimated threshold thr.

6. In the second truncation step, the algorithm applies an iterative procedure that calculates the number of image pixels belonging to each candidate image cluster and prunes the cluster with smallest number of image pixels (less than 2 percent of total number of image pixels). The pruned candidate clusters are merged with their closest image clusters.
7. The number of elements in final vector I^t represents the number of clusters denoted by NC.
8. Determine the local maxima in IINT.

$$I_M * = \left\{ \min_{0 \le T \le G} I_{INT}\left(T\right) \right\} \tag{6}$$

   IM* is the vector carrying all local maxima.

9. Thresholding: Find the peaks (local maxima) whose value is higher than one percent of the maximum peak in h(k).
10. Remove the peaks which are very close. This is done by checking the difference between the grey levels of the two individual peaks. If the difference is less than 20, then the peak with lowest value is removed.
11. Neighboring pixels that lead to the same peak are grouped together.
12. The values of the identified peaks represent the initial centroids of the input image.

End

4. FUZZY C-MEANS CLUSTERING ALGORITHM

The Fuzzy C-means [12] is an unsupervised clustering algorithm. The main idea of introducing fuzzy concept in the Fuzzy C-means algorithm is that an object can belong simultaneously to more than one class and does so by varying degrees called memberships. It distributes the membership values in a normalized fashion. It does not require prior knowledge about the data to be classified. It can be used with any number of features and number of classes. The fuzzy C-means is an iterative method which tries to separate the set of data into a number of compact clusters. It improves the partition performance and reveals the classification of objects more reasonable. The predefined parameters such as number of clusters and initial clustering centers are provided by ECNC algorithm. The Fuzzy C-means algorithm is summarized as follows:

Algorithm Fuzzy C-Means (x, N, c, m)

Begin

1. Initialize the membership matrix uij is a value in (0,1) and the fuzziness parameter m (m=2). The sum of all membership values of a pixel belonging to clusters should satisfy the constraint expressed in the following.

$$\sum_{j=1}^{c} U_{ij} = 1 \tag{7}$$

for all i= 1,2,…….N, where c is the number of clusters and N is the number of pixels in microarray image

2. Compute the centroid values for each cluster cj. Each pixel should have a degree of membership to those designated clusters. So the goal is to find the membership values of pixels belonging to each cluster. The algorithm is an iterative optimization that minimizes the cost function defined as follows:

$$F= \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^{m} \| x_j - c_i \|^2 \tag{8}$$

where u_{ij} represents the membership of pixel xj in the ith cluster and m is the fuzziness parameter.

3. Compute the updated membership values uij belonging to clusters for each pixel and cluster centroids according to the given formula.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{2/(m-1)}},$$

$$\text{and}$$

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}. \tag{9}$$

4. Repeat steps 2-3 until the cost function is minimized.

End.

5. MULTIPLE FEATURE CLUSTERING

The clustering algorithms used for microarray image segmentation are based on the information about the intensities of the pixels only. But in microarray image segmentation, the position of the pixel and median value of surrounding pixels also influences the result of clustering and subsequently that leads to segmentation. Based on this observation, multiple feature clustering algorithm is developed for segmentation of microarray images. To apply fuzzy c-means clustering algorithm on a single spot, we take all the pixels that are contained in the spot are, which is obtained after gridding process, and create a dataset D = {x_1, x_2,

x_3, x_4, x_5,……,x_n}, where $x_i = [\, x_i^{(1)}, x_i^{(2)}, x_i^{(3)}]$ is a three dimensional vector that represents the ith pixel in the spot region. We use three features, defined as follows
$x_i^{(1)}$: Represents the pixel intensity value.
$x_i^{(2)}$:Represents the distance from pixel to the center of the spot region.

The spot center is calculated as follows:

1. Apply edge detection to the spot region image using canny method.
2. Perform flood-fill operation on the edge image using imfill method.
3. Obtain label matrix that contain labels for the 8-conneted objects using bwlabel function.
4. Calculate the centroid of each labeled region (connected component) using regionprops method.

$x_i^{(3)}$:Represents the median of the intensity of surrounding pixels.
For each pixel in the spot region, once the features are obtained forming the dataset D, then the fuzzy c-means clustering algorithm is applied. The centroids and number of clusters in the dataset are calculated using ECNC algorithm.

6. EXPERIMENTAL RESULTS

Qualitative Analysis:

The proposed clustering algorithm is performed on two microarray images drawn from the standard microarray database corresponds to breast category aCGH tumor tissue. Image 1 consists of a total of 38808 pixels and Image 2 consists of 64880 pixels. Gridding is performed on the input images by the method proposed in [13], to segment the image into compartments, where each compartment is having only one spot region and background. The gridding output is shown in Figure 1. Multiple feature clustering algorithm is applied to each compartment for segmenting the foreground and background region. The ECNC algorithm is executed on the histogram of input images for identification of number of clusters and initial centroids which is required for clustering algorithm. The output of the proposed method on a compartment from image 1 and image 2 is shown in Figure 1.

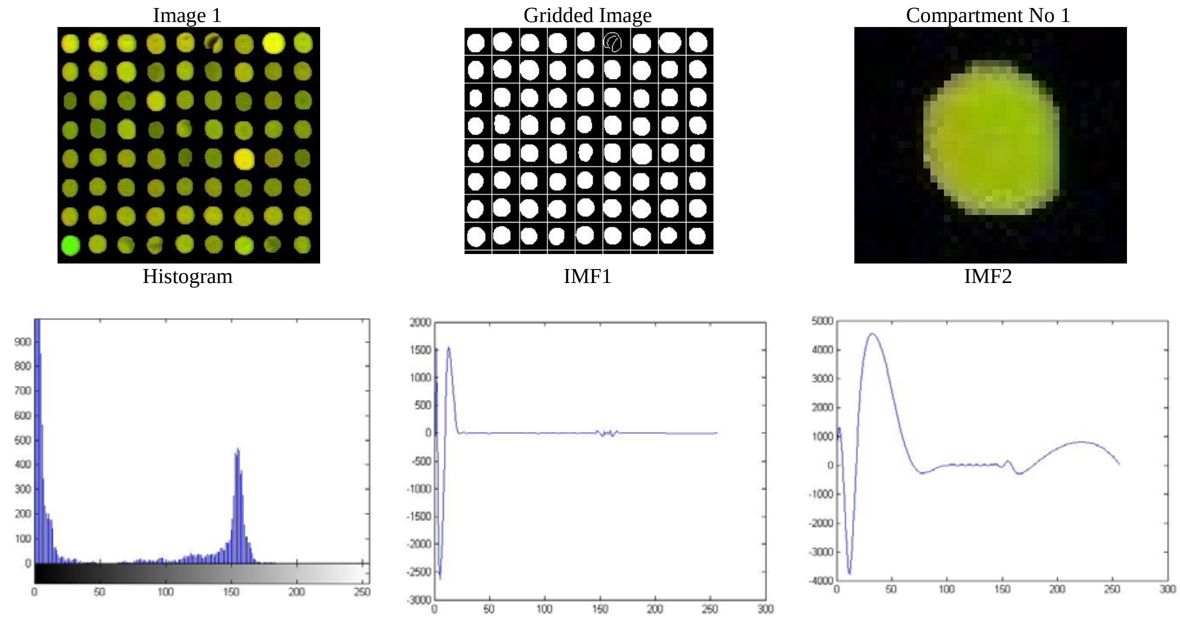


Image 1

Gridded Image

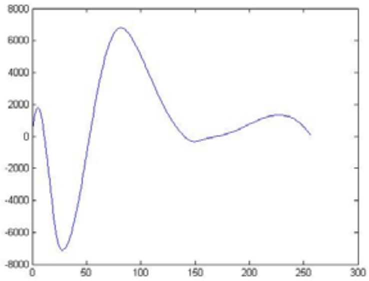
Compartment No 1

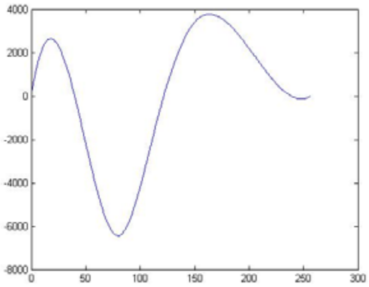
Histogram

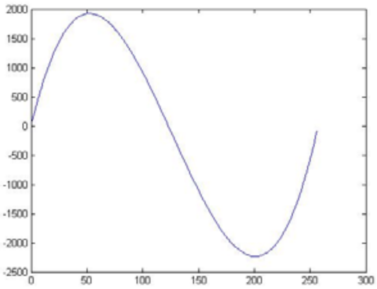
IMF1

IMF2

IMF3 IMF4 IMF5



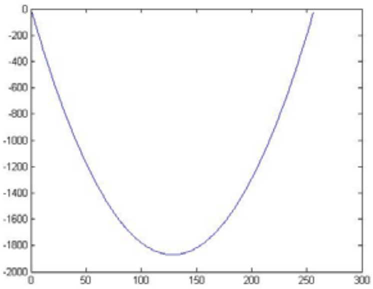


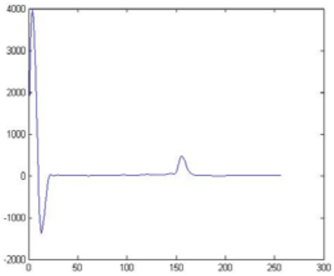


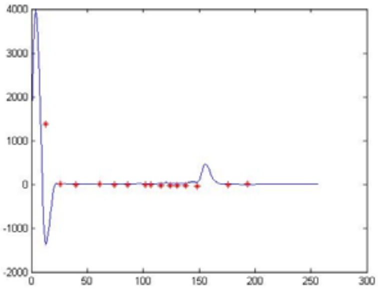
IMF6

Combined IMF

Local Minima



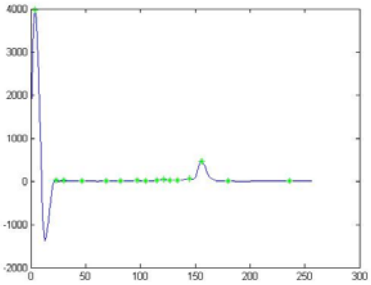




Local Maxima

Centroids

Segmented Image using Multiple features



No of clusters :2
Centroids are 4, 145

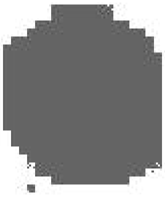
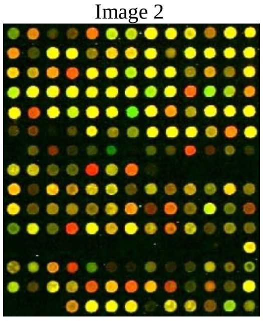
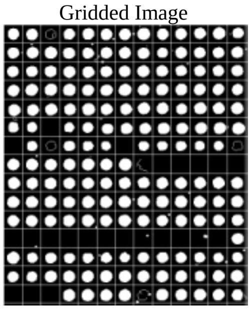


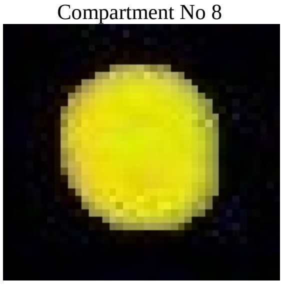
Image 2

Gridded Image

Compartment No 8



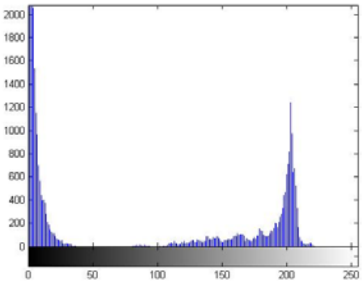


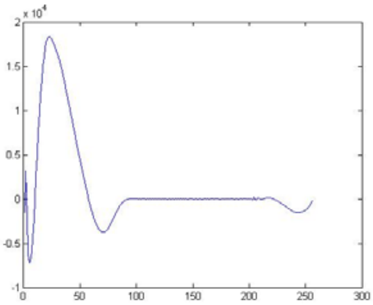


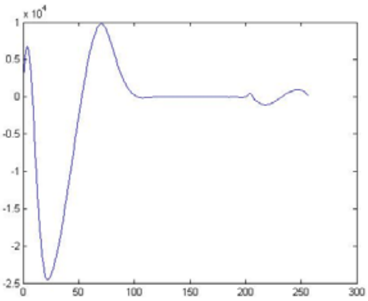
Histogram

IMF1

IMF2



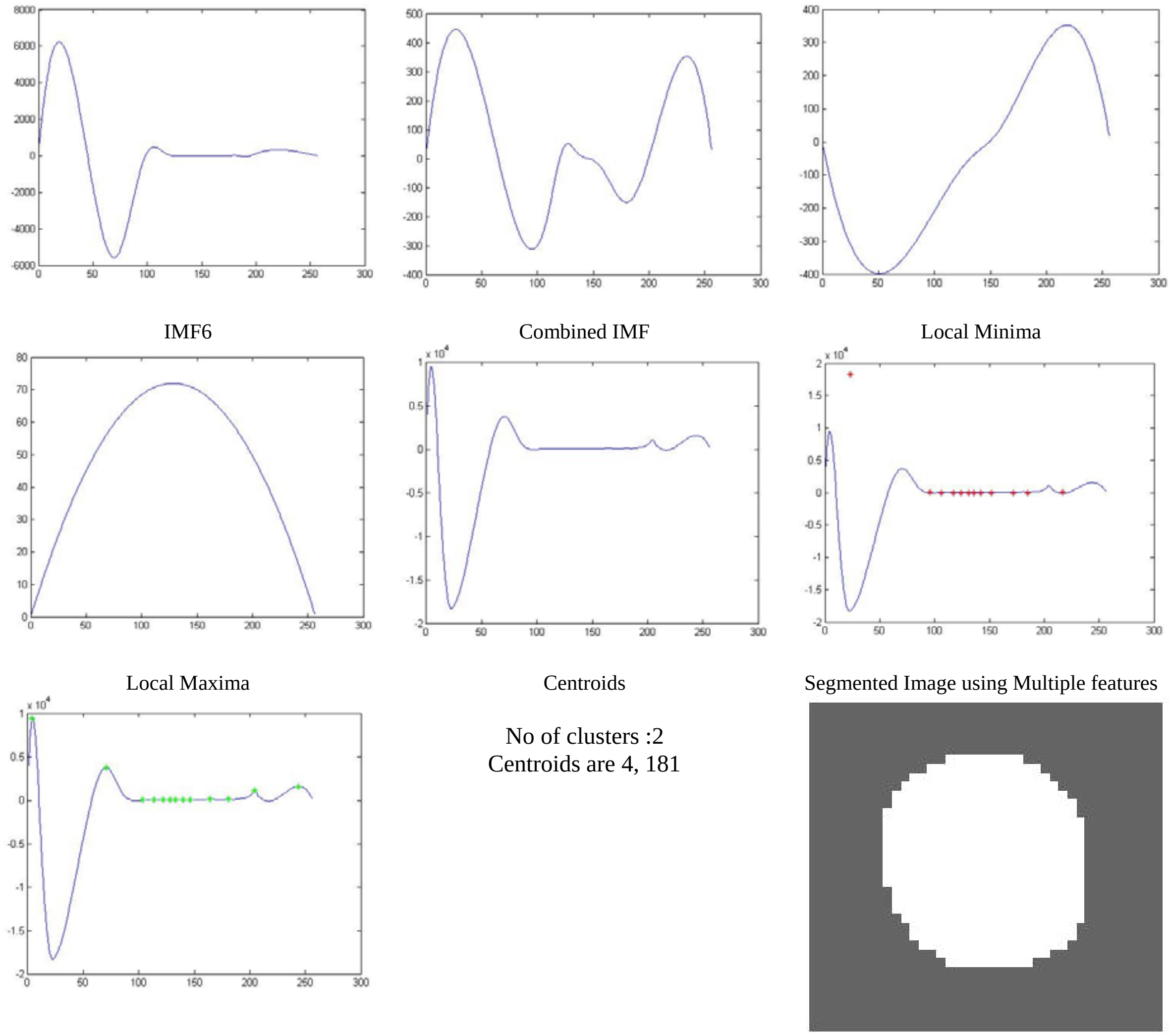




IMF3

IMF4

IMF5



No of clusters :2
Centroids are 4, 181

Figure 1. Segmentation result

Quantitative Analysis:

Quantitative analysis is a numerically oriented procedure to figure out the performance of algorithms without any human error. The Mean Square Error (MSE) [14, 15] is significant metric to validate the quality of image. It measures the square error between pixels of the original and the resultant images. The MSE is mathematically defined as

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i\in c_j} ||v_i - c_j||^2 \qquad (10)$$

Where N is the total number of pixels in an image and xi is the pixel which belongs to the jth cluster. The lower difference between the resultant and the original image reflects that all the data in the region are located near to its centre. Table 1 shows the quantitative evaluations of three clustering algorithms. The results confirm that multiple feature fuzzy c-means algorithm produces the lowest MSE value for segmenting the microarray image. As the initial centroids required for clustering algorithms are determined by ECNC algorithm, the number of iterative steps required for classifying the objects is reduced. While the initial centroids obtained by ECNC are unique, the segmented result is more stable compared with traditional algorithms. Table 2 shows the comparison of iterative steps numbers for clustering algorithms with and without ECNC.

Table 1. MSE values

Method	MSE Values (Compartment No 1) In image 1	MSE Values (Compartment No 8) In image 2
K-means	282.781	346.47
Fuzzy c-means	216.392	228.69
Multiple feature Fuzzy C-means	198.327	186.276

Table 2. Comparison of iterative step numbers

	Clustering algorithm	Iterative steps (without ECNC)	Iterative steps (with ECNC)
(Compartment No 1) In image 1	K-means	10	4
	Fuzzy C-means	14	6
	Multiple feature Fuzzy C-means	17	9

	Clustering algorithm	Iterative steps (without ECNC)	Iterative steps (with ECNC)
(Compartment No 8) In image 2	K-means	11	6
	Fuzzy C-means	16	12
	Multiple feature Fuzzy C-means	19	11

7. CONCLUSION

This paper presents multiple feature fuzzy c-means clustering algorithm for microarray image segmentation. Instead of using single feature i.e., pixel intensity, two other features such as distance of the pixel from the spot center and median value of surrounding pixels are used for segmentation. The qualitative and quantitative analysis done proved that multiple feature Fuzzy C-means has higher segmentation quality than other clustering algorithms with single feature. Clustering algorithm combined with ECNC overcomes the problem of random selection of number of clusters and initialization of centroids. The proposed method reduces the number of iterations for segmentation of microarray image and costs less execution time.

REFERENCES

[1] M. Schena, D. Shalon, Ronald W. Davis and Patrick O. Brown, "*Quantitative Monitoring of gene expression patterns with a complementary DNA microarray*", Science, Oct 20; 270(5235): 467-70.

[2] Wei-Bang Chen, Chengcui Zhang and Wen-Lin Liu, "An Automated Gridding and Segmentation method for cDNA Microarray Image Analysis", *19th IEEE Symposium on Computer-Based Medical Systems*.

[3] Tsung-Han Tsai Chein-Po Yang, Wei-Chi Tsai, Pin-Hua Chen, "Error Reduction on Automatic Segmentation in Microarray Image", *IEEE 2007*.

[4] Eleni Zacharia and Dimitirs Maroulis, "Microarray Image Analysis based on an Evolutionary Approach", *IEEE 2008*.

[5] J. Harikiran, B. Avinash, Dr. P.V. Lakshmi, Dr. R. Kiran Kumar,"Automatic Gridding Method for Microarray Images", *Journal of Applied Theoritical and Information Technology*",Vol 65, No 1, pp 235-241, 2014.

[6] Volkan Uslan, Omur Bucak, "clustering based spot segmentation of microarray cDNA Microarray Images", *International Conference of the IEE EMB, 2010.*

[7] M. Eisen, ScanAlyze User's manual, 1999,

[8] J. Buhler, T. Ideker and D. Haynor, "Dapple: Improved Techniques for Finding spots on DMA Microarray Images", *Tech. Rep. UWTR 2000-08-05*, University of Washington, 2000.

[9] R. Adams and L. Bischof, "Seeded Region Growing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 16,no. 6, pp.641-647, 1994.

[10] J. Harikiran, P.V. Lakshmi, R. Kiran Kumar, "Fast Clustering Algorithms for Segmentation of Microarray Image", *International Journal of Scientific and Engineering Research*, Vol 5, Issue 10, pp. 569-574.

[11] N.E. Huang, Z. Shen, S.R. Long, "The empirical mode decomposition and the Hilbert Spectrum for non-linear and non-stationary time series analysis". *Proc. Roy. Soc, London*.A, Vol. 454, pp. 903-995, 1998

[12] Siti Naraini Sulaiman, Nor Ashidi Mat Isa, "Denoising based Clutering Algorithms for Segmentation of Low level of Salt and Pepper Noise Corrupted Images", *IEEE Transactions on Consumer Electronics*, Vol. 56, No.4, November 2010.

[13] J. Harikiran, et.al, "A New Method of Gridding for Spot Dectection in Microarray Images", *Computer Engineering and Intelligent Systems*, Vol 5, No 3, pp. 25-33.

[14] Nor Ashidi Mat Isa, Samy A. Salamah, Umi Kalthum Ngah., "Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation", *IEEE Transaction on Consumer Electronics*, 12/2009; DOI: 10.1109/TCE.2009.5373781.

[15] B. Saichandana, Dr. K. Srinivas, Dr. R. Kiran Kumar, “De-noising based clustering Algorithm for Classification of Remote Sensing Image”, *Journal of Computing*, Volume 4, Issue 11, November 2012.

[16] Zhengjian DING, Juanjuan JIA, DIA LI , “Fast Clustering Segmentation Method Combining Hill Climbing for Color Image”, *Journal of Information and Computational Sciences*, Vol 8, pp. 2949-2957.

[17] B. Saichandana, K. Srinivas, R. Kiran Kumar, “Clustering Algorithm combined with Hill Climbing for Classification of Remote Sensing Images”, *International Journal of Electrical and Computer Engineering*, vol 4, No. 6, pp 923-930.

BIOGRAPHIES OF AUTHORS

J. Harikiran received B.Tech and M.Tech degree from JNTU Hyderabad and Andhra University in the year 2005 and 2008 respectively. He is currently working as Assistant professor in the Department of IT, GIT, Gitam University. His research interest include Image Segmentation, Microarray Image Procesing etc. Currently he is persuing phd from JNTU Kakinada.

Dr. P.V. Lakshmi received M.Tech and PhD degrees from Andhra University. Her research interest include Cryptography, Algorithms in Bioinformatics etc.. Currently She is working as Professor and Head, Department of Information Technology, GIT, GITAM University.

Dr. R. Kiran Kumar received MCA, M.Tech and Phd degrees from Andhra University, JNTU kakinada and Acharya Nagarjuna University. His research interest include image processing and bioinformatics. Currently he is working as assistant Professor, Department of Computer science, krishna University, Machilipatnam.