VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2019 ISSN(e): 2309-3978;ISSN(p): 2411-6246
Volume 7, Number 1, January-December, 2019

pp. 16-24

EXTRACTING TRUE NUMBER OF CLUSTERS FOR SEGMENTING IMAGE THROUGH ADAPTIVE FINITE GAUSSIAN MIXTURE MODEL

M MASROOR AHMED[*1], SAJID NAEEM[2], SYED MUHAMMAD REHMAN HABIB[*2]
1 Department of Computer Science, Capital University of Science & Technology (CUST), Islamabad
2 School of Systems and Technology (SST), Department of Computer Science, University of Management & Technology (UMT), Lahore
masroorahmed@gmail.com

ABSTRACT: *Knowing exact number of clusters in a digital image significantly facilitates in precisely clustering an image. This paper proposes a new technique for extracting exact number of clusters from grey scale images. It analyzes the contents of the input image and adaptively reserves one distinct cluster for one distinct grey value. The total count of the grey values found in an image determines the exact number of clusters. Based on the contents of image, this number of clusters keeps on changing from image to image. After obtaining this number, it is given as an input to Gaussian Mixture Model (GMM) which clusters the image.GMM works with finite number of clusters and forms mixture of various spectral densities contained in that image. The proposed method facilitates GMM to adapt itself according to the changing number of clusters. Therefore, the proposed model along with the inclusion of GMM, is named as Adaptive Finite Gaussian Mixture Model (AFGMM). The clustering performance of AFGMM is evaluated through Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR). Both of these performance measuring methods confirmed that exact number of clusters is essentially important for reliably analyzing an image.*
Keywords: gaussian mixture model (GMM); clustering; segmentation; expectation maximization (EM) algorithm

1. **Introduction.** Formally, segmentation can be defined as a process for dividing an image into a number of constituent classes. This division is carried out on the basis of some predefined similarity criterion [1-3]. Generally, the criterion aims to minimize 'within class' variance and maximize 'between class' variance. Usually, preliminary image analysis is done either manually with the help of a human expert, as in the case of medical images, or automatically by employing computer based approaches.

The human experts, when analyzing an image, possess an innate capability to produce realistic results. However, the idea seems to be impractical as it is subjective, time consuming and prone to inconsistent data interpretation even in repeated attempts [4-5]. On the contrary, machine-based approaches offer time efficient processing of complex and large data but these approaches are not completely independent. In certain situations, for reaching to a justifiable decision, they may require an external opinion which is likely to differ from person to person.

In spite of these limitations, machine-based approaches seem to be the only choice and a number of methods have been proposed which improved overall clustering accuracy. For example, Genetic algorithm, which is known for optimizing search and for addressing large scale problems, is employed for finding exact number of clusters. Heuristic classification method attempts to optimize predefined index to automatically detect the number of constituent classes. Edge based method [6] emphasizes on defining precise edges around the objects in a given image. On the basis of these objects, number of clusters is established. In fuzzy clustering algorithm [7], association of a specific data element with a specific class is decided on the basis of its degree of membership. In probability based approaches, the observed data element is investigated in the light of its mean and variance which decides future relationship of observed data element with a specific class. Artificial Neural Network, which is a form of supervised learning, is also used for determining the number of clusters in fuzzy clustering [8].K-Means classification [9], which is a form of hard classification, forces the data to adjust itself according to the given number of classes. It relies on mean of the given data and prior information about the total number of classes.

Though these methods improved the situation, they still remained unsuccessful in providing a generalized rule for extracting the exact number of classes from an image [10] [11].
This paper addresses the issue of determining the exact number of clusters. The obtained number of clusters is used as an input for the subsequent segmentation process. It effectively controls the variations which are normally caused by guessing the number of clusters. The following section describes the proposed model.

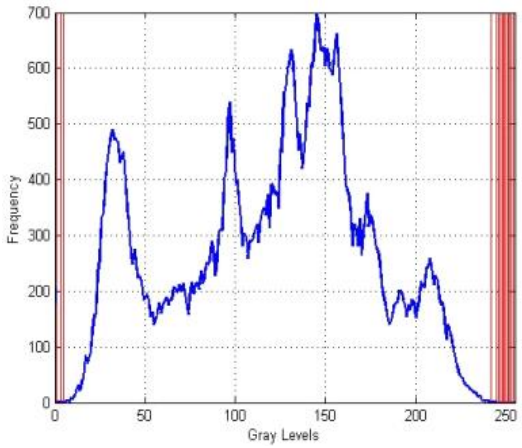
2 Proposed Model. The proposed model consists of two main parts. One deals with the automatic extraction of exact number of clusters from a given image, whereas, the other deals with the modeling of data followed by it classification into constituent classes. Both of these parts are discussed in the following sections.
Extracting the exact number of clusters. The proposed method automatically detects the exact number of clusters in a given image. It achieves this by considering all grey values in the given image and holds that all of them are fundamentally essential for drawing exact information from an image. Pixels with same intensity value are counted and are grouped together in distinct sets. One distinct cluster is allotted to one distinct set of grey value. The overall number of constituent classes is decided by counting the number of sets formed by grouping the observed grey values. The pseudocode given below describes the approach for reliably establishing the number of clusters[12]:

```
forgLevel=0:1:255
        imghist(gLevel+1)=0;
    for i=1:1:col
        for j=1:1:row
          if (x(j,i)== gLevel)
                                        imghist(gLevel+1)= imghist(gLevel+1)+1;
                                end
                        end
                end
end
```

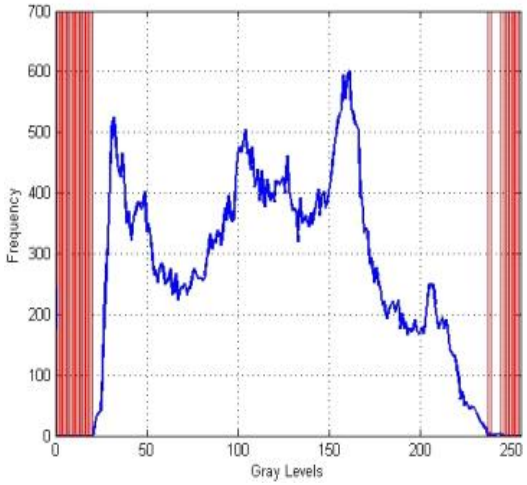
Generally a gray scale image contains gray values ranging from 0-255. It is not necessary that every single image contains all gray values. Some of the gray values may be missing. Knowledge about exact number of missing and available grey values significantly contribute in deciding about the exact number of clusters.This knowledge is obtained by using the following pseudocode [13].

```
y=find(imghist==0);
clusterNumb= size(imghist,2)-size(y,2)
fori=1:size(y,2)
     h = vline(y(i),'r');
    plot(h)
end
```

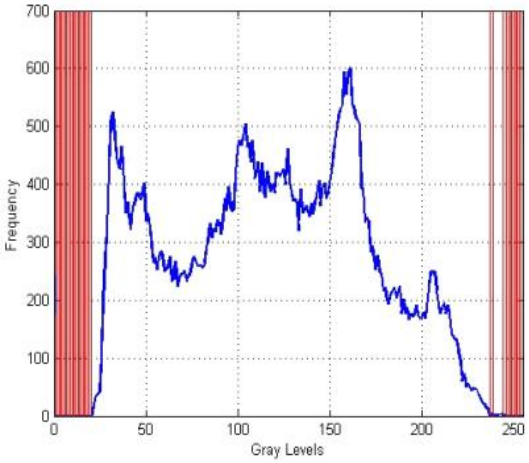
The above mentioned pseudocode plots intensity profiles of the given image as shown in Fig. 1. These profiles produce ‘0’ values against those gray values which are missing. The missing gray values are indicated by putting vertical red line. By the end of the process, the total number of these red lines is counted and is subtracted from the total number of gray values which are not ‘0’. This mutual difference establishes the exact number into which the given image should be clustered. From the intensity profiles shown in Fig.1/a and Fig.1/b, this is clear that majority of the images are missing some of the gray values except the ‘cameraman’ image. Therefore, the ‘cameraman’ image is required to be clustered by considering 255 clusters. The immediate advantage of this approach is that, on one hand, it introduces desired level of accuracy and, on the other hand, it effectively controls the issues of over-clustering and under-clustering.



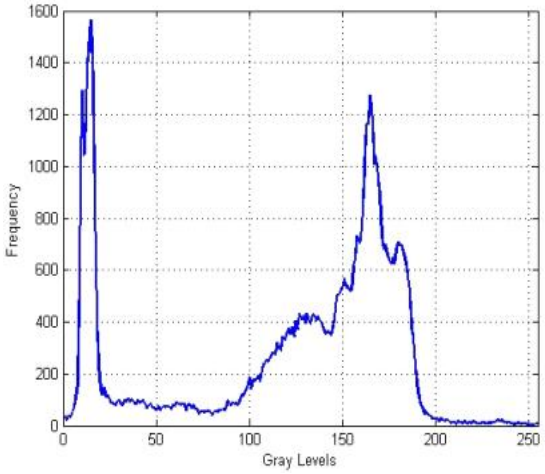
Lena



Barbara

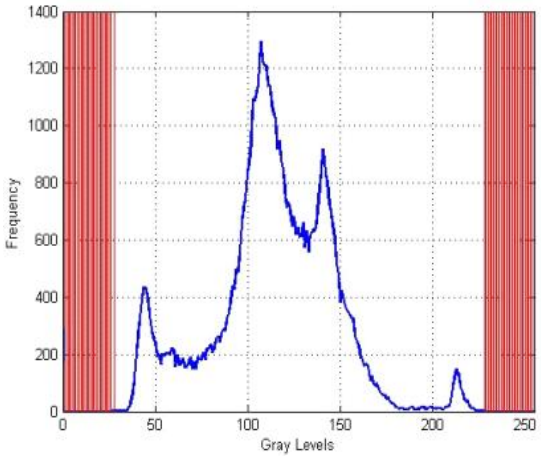


Baboon

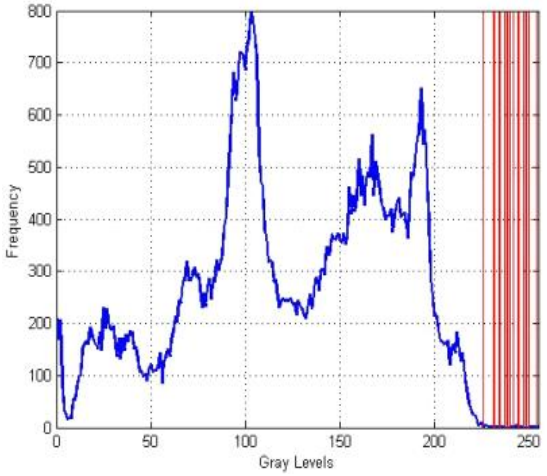


Cameraman

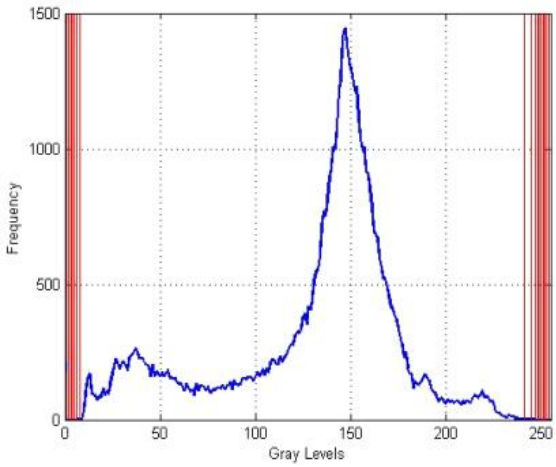
Fig. 1/a: Representation of Observed and Missing Grey Values



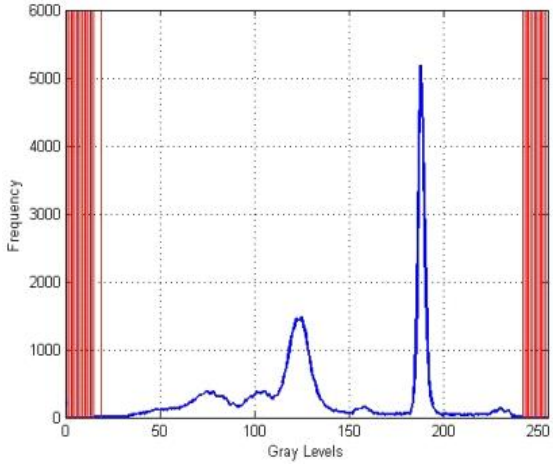
Einstein



Peppers



Boat



House

Fig. 1/b: Representation of Observed and Missing Grey Values

Adaptive Finite Gaussian Mixture Model (AFGMM). This is the second stage of the proposed method. It makes use of conventional Gaussian Mixture Model (GMM) for clustering the given image. In the present case, the GMM is both adaptive and finite. It is finite because it works with limited number of clusters and it is adaptive becuase the number of clusters which is one of the fundamental input is neither fixed nor randomized. This input is dynamically computed by looking into the composition of image. With the inclusion of these features, the proposed model is known as Adaptive Finite Gaussian Mixture Model (AFGMM). The model follows Gaussian distribution of data and works with an assumption that the underlying data is independent and is identically distributed. On top of it, the model is known for capturing higher level of details. Statistically, the model is simple and is characterized by three important elements, i.e. mean, variance and the mixing weights.

Mean (μ_k) is the average value of the given data, variance ($\sum_k$)is the spread of a data element from its mean position and mixing weight ($mixWt_k$) is the prior knowledge with which a particular data element is associated with a particular cluster. Collectively, all these three components are known as parameters of the model.

Therefore, with this information, AFGMM can formally be described by supposing that $y_1, y_2, \ldots, y_n$are data elements with their initial parameters represented by$mixWt_k^{(0)}$, $\mu_k^{(0)}$, $\sum_k^{(0)}$. The probability density function of the model is defined as [14]:

$$f(y) = \sum_{k=1}^{K} mixWt_k f_k(y; \mu_k, \sum_k) \quad (1)$$

Equation (1) represents the model as linear combination of mixtures of Guassians. This mixture is also known as probability density function (PDF). This is a constrained model which works under the influence of constraint shown in (2).

$$\sum_{k=1}^{K} mixWt_k = 1 \quad (2)$$

Initially, the parameters of the model are randomized, and in latter iterations they are regularly updated. For updating them,expectation maximization (EM) algorithm is used. According to this algorithm, at time‘t’, the model computes the expected (probabilistic) value by using the following (3):

$$p_k^t(x) = \frac{mixWt_k f_k(y_x, \mu_k^t, \Sigma_k^t)}{\sum_{l=1}^{K} mixWt_l f_l(y_l, \mu_l^t, \Sigma_l^t)} \quad (3)$$

This probabilistic value is obtained by dividing the PDF of the observed data value with the sum of PDFs of all remaining clusters. In subsequent iteration, i.e. at ‘t+1’, the model parameters of AFGMM are updated by using the following update (4-6):

$$mixWt_k^{(t+1)} = \frac{\sum_{x=1}^{N} p_k^{(t)}(x)}{N} \quad (4)$$

$$\mu_k^{(t+1)} = \frac{\sum_{x=1}^{N} p_k^{(t)}(x) y_x}{\sum_{x=1}^{N} p_k^{(t)}(x)} \quad (5)$$

$$\sum_k^{(t+1)} = \frac{\sum_{x=1}^{N} p_k^{(t)}(x) \left(y_x - \mu_k^{(t)}\right)' \left(y_x - \mu_x^{(t)}\right)}{\sum_{x=1}^{N} p_x^{(t)}(x)} \quad (6)$$

After the updation of model parameters, EM aims to maximize log likelihood function for estimating the GMM’s parameters. It achieves this with the help of following (7):

$$label\ (y_c) = argmax_{1 \le k \le K} p_k(x) = argmax_{1 \le k \le K} mixWt_k f_k(y_x;\ \mu_k, \Sigma_k) \quad (7)$$

This covers basic explanation of the proposed idea. Additionally, it may be noted that modern imaging devices produce high resolution images [15]. These images carry higher level of details as a result of which they are larger in size. Consequently, these images demand large memory space and larger processing time. Though the quality of data and the level of information these images are holding forces us to accommodate these requirements, but practically this is unrealistic to have unlimited memory space and time. To deal with this situation, generally the acquired data is compressed. Compression phenomenon primarily aims to maintain visual quality of the data and drops some of the data elements which are believed to have insignificant impact over the entire data. By dropping certain part of the data, the overall data size gets reduced. This reduction creates more space for storing new incoming data, makes transmission of the data easy and reduces the processing time.

Performance of AFGMM. The performance of the proposed model was evaluated by looking into the values of mean square error (MSE) and peak signal to noise ratio (PSNR). A lower possible value of MSE and highest possible value of PSNR is desired. Both of these performance measuring values were computed by using Equations (8) and (9) respectively. The trend of the results indicated that with increased number of clusters the accuracy gradually increases [21].

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |I(i,j) - segImage(i,j)|^2 \quad (8)$$

$$PSNR = 10 * log_{10}\left(\frac{255}{MSE}\right) \quad (9)$$

Where M and N determines the number of rows and columns, I represents the input image, (i, j) represent the location of a specific pixel, and $segImage$ represents the segmented image.

3 Results. Table 1 describes clustering results obtained by AFGMM. Standard images were used to investigate the performance of proposed AFGMM. From these results, it can be extracted that ideally the images of Lena, Barbara, Baboon, Cameraman and Einstein should be clustered by having clusters equal to 240, 222, 202, 255 and 200, respectively. This cluster number was dynamically produced on the basis of gray values found in the image. A higher degree of accuracy was found when clustering procedure was carried out on the basis of this dynamically obtained number of clusters, whereas, it influences MSE and PSNR when some different cluster number is used.

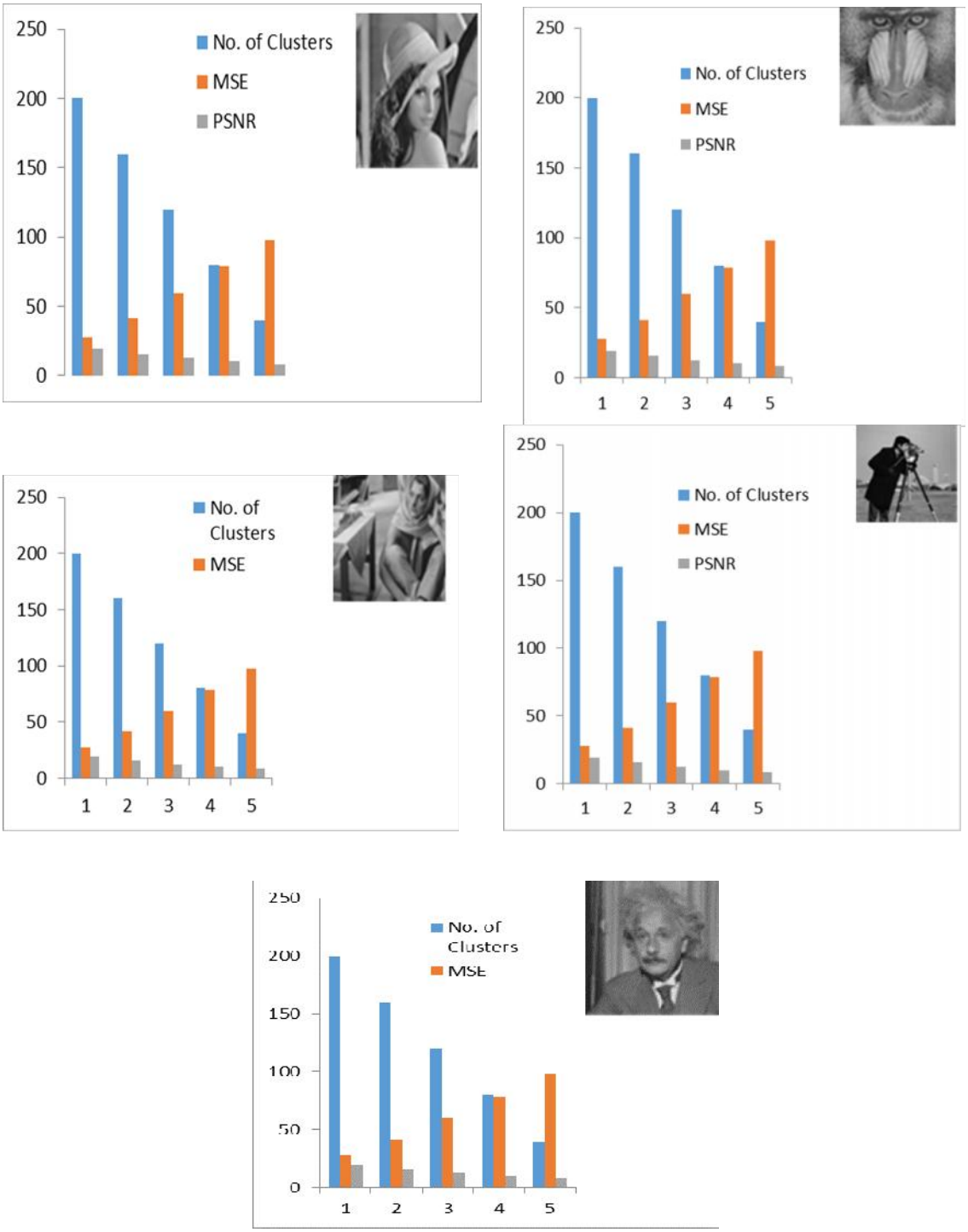


Fig 2: *Showing Images to be Clustered, Desired Number of Clusters, MSE and PSNR of Clustered Image*

4 Conclusion. Precise image clustering is a challenging issue [18-20]. A higher level of precision can be maintained by knowing the exact number of clusters in the data under investigation. This number can be obtained either through domain expert or by using computer. Manual extraction of this number is found to be unrealistic mainly because of huge volume and/or variations in the data under investigation. Therefore, an alternate approach, i.e. computer based investigation is frequently relied upon. A wide range of research was conducted to optimize the

process for extracting the exact number of clusters. This research resulted in introducing some state-of-the-art methods to achieve the desired objectives but the question of having a generalized rule for extracting the exact number of clusters remains unanswered. The present study attempts to address this question and proposes a method **AFGMM**. The distinguishing feature of the proposed method is that it dynamically adapts itself according to the contents of the data and decides about the exact number of clusters. The extracted number of clusters is used as an input for GMM to model the given data. The parameters of GMM were updated by using EM algorithm. The initial results obtained by the proposed model were appreciative.

REFERENCES

[1] Ahmed, M. M., & Mohamad, D. B. (2008). Segmentation of brain MR images for tumor extraction by combining kmeans clustering and perona-malik anisotropic diffusion model. *International Journal of Image Processing*, *2*(1), 27-34.

[2] Razak, Z., Zulkiflee, K., Noor, N. M., Salleh, R., & Yaacob, M. (2009). Off-line handwritten Jawi character segmentation using histogram normalization and sliding window approach for hardware implementation. *Malaysian Journal of Computer Science*, *22*(1), 34-43.

[3] Benrabh, M., Bouroumi, A., & Hamdoun, A. (2005). A fuzzy validity-guided procedure for cluster detection. *Malaysian Journal of Computer Science*, *18*(1), 31-39.

[4] Jung, C., Kim, C., Chae, S. W., & Oh, S. (2010). Unsupervised segmentation of overlapped nuclei using Bayesian classification. *IEEE Transactions on Biomedical Engineering*, *57*(12), 2825-2832.

[5] Yi, W., Yao, M., & Jiang, Z. (2006, November). Fuzzy particle swarm optimization clustering and its application to image clustering. In *Pacific-Rim Conference on Multimedia* (pp. 459-467). Springer, Berlin, Heidelberg.

[6] Patil, R. V., & Jondhale, K. C. (2010, July). Edge based technique to estimate number of clusters in k-means color image segmentation. In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on* (Vol. 2, pp. 117-121). IEEE.

[7] Le Capitaine, H., & Frelicot, C. (2010, August). On selecting an optimal number of clusters for color image segmentation. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 3388-3391). IEEE.

[8] Erilli, N. A., Yolcu, U., Eğrioğlu, E., Aladağ, Ç. H., & Öner, Y. (2011). Determining the most proper number of cluster in fuzzy clustering by using artificial neural networks. *Expert Systems with Applications*, *38*(3), 2248-2252.

[9] Wang, L., Leckie, C., Ramamohanarao, K., & Bezdek, J. (2009). Automatically determining the number of clusters in unlabeled data sets. *IEEE Transactions on knowledge and Data Engineering*, *21*(3), 335-350.

[10] Rosenberger, C., & Chehdi, K. (2000). Unsupervised clustering method with optimal estimation of the number of clusters: Application to image segmentation. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on* (Vol. 1, pp. 656-659). IEEE.

[11] Chen, Y. H., Ho, Y. W., Wu, C. H., & Lai, C. C. (2009, May). Aerial image clustering using genetic algorithm. In *Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA'09. IEEE International Conference on* (pp. 42-45). IEEE.

[12] Ali, L., Hussain, A., Li, J., Shah, A., Sudhakr, U., Mahmud, M., ... & Rajak, M. (2014, December). Intelligent image processing techniques for cancer progression detection, recognition and prediction in the human liver. In *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*(pp. 25-31). IEEE.

[13] Beale, M. H., Hagan, M. T., & Demuth, H. B. (2012). Neural network toolbox™ user's guide. In *R2012a, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098,, www. mathworks. com*.

[14] Leung, S., Liang, G., Solna, K., & Zhao, H. (2009). Expectation-maximization algorithm with local adaptivity. *SIAM journal on imaging sciences*, *2*(3), 834-857.

[15] Tran, T. N., Wehrens, R., & Buydens, L. M. (2005). Clustering multispectral images: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *77*(1-2), 3-17.

[16] Rocha, A., & Room, C. D. T. MO444/MC886.

[17] Naidu, V. P. S., & Raol, J. R. (2008). Pixel-level image fusion using wavelets and principal component analysis. *Defence Science Journal*, *58*(3), 338.

[18] Srivisal, C., & Lursinsap, C. (2009, April). Predicting Number of Unsupervised Clusters by Supervised Function. In *2009 International Joint Conference on Computational Sciences and Optimization* (pp. 726-730). IEEE.

[19] Vinh, N. X., & Epps, J. (2009, June). A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. In *Bioinformatics and BioEngineering, 2009. BIBE'09. Ninth IEEE International Conference on* (pp. 84-91). IEEE.

[20] Langan, D. A., Modestino, J. W., & Zhang, J. (1998). Cluster validation for unsupervised stochastic model-based image segmentation. *IEEE Transactions on Image Processing*, *7*(2), 180-195.

[21] Ahmed, M. M., Zain, J. M., & Rana, M. T. A. (2012, November). Context Independent Expectation Maximization Algorithm for Segmentation of Brain MR Images. In *Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on* (pp. 436-441). IEEE.