

Infrared Image Segmentation using Enhanced Fuzzy C-Means Clustering for Automatic Detection Systems

SITANSHU GUPTA

Department of Electronics and Communication Engineering,
National Institute of Technology, Allahabad
Allahabad -211004, India
Email: sitanshu.08@gmail.com

ASIM MUKHERJEE

Department of Electronics and Communication Engineering,
National Institute of Technology, Allahabad
Allahabad -211004, India
Email: asimmkj@yahoo.co.in, asimmkj@mnnit.ac.in

Abstract-This paper proposes Enhanced Fuzzy C-means technique (EFCM) based infrared image segmentation and its broad application in Automatic detection systems. The EFCM based image segmentation is able to approximate the exact number of clusters present in the image. EFCM based segmentation is applied on various infrared images that can be used for automatic detection systems and compared with widely used clustering techniques such as K-means and EM. Clustering performance has been compared in terms of well-proven and widely accepted validation indices, Global Silhouette Index and Separation Index. The segments or clusters obtained from above mentioned clustering methods have been assessed visually. Automatic Detection Systems based on EFCM can help in reducing complexities present in conventional systems.

Keywords: EFCM, K-means, EM, GS, SI.

I. INTRODUCTION

Clusters are the collection of similar objects such that each group or cluster is homogeneous; that is, objects belonging to the same group are similar to each other. Each group or cluster should be different from other clusters; that is, objects belonging to one cluster should be different from the objects of other clusters. Clustering technique [1] can be hard or fuzzy. A hard clustering algorithm assigns each data point to a particular cluster during its operation and in its output. A fuzzy clustering method allocates a degree of membership to each data point depending on its association with several other clusters.

Automatic Detection Systems, on which proposed technique can be used, are those that can control any process, weather forecast or detect diseases with the help of infrared images. These systems are discussed in Section-4 of this paper.

This paper is organized in four sections. An overview of the existing clustering algorithms is given in Section-2. The proposed technique is explained in Section-3. In Section-4, we present the simulation results of all the clustering techniques, including the proposed EFCM technique. Moreover, a comparison between the cluster quality of all clustering techniques is made. Finally, we present our conclusion in the last section.

II. AN OVERVIEW OF SOME CLUSTERING BASED SEGMENTATION TECHNIQUES

Some of the widely used clustering techniques are K-means Clustering [2], FCM Clustering [3], Gath-Geva Clustering [4] and EM Clustering [5].

K-means clustering algorithm was proposed by Forgy and MacQueen in 1967. It is one of the simplest unsupervised clustering algorithms, which has been widely used in various applications. It groups data into a fixed number of clusters where the centroid of one cluster is placed as far away as possible from another. Each data point in a cluster is associated to the nearest centroid. The Expectation Maximization (EM) clustering algorithm is based on the statistical model that uses the finite Gaussian mixtures model. This clustering algorithm utilizes the K-means initialization in which a fixed number of clusters are obtained through K-means clustering. Then the cluster parameters (weights, means and co-variances) are re-computed until a desired convergence value is achieved. In infinite mixture model, all attributes are assumed to be independent random variables. The Probabilistic Clustering technique [6] objects at fitting a fixed k-component Gaussian mixture model (GMM) to the data set. It assesses the probability of the components of a GMM such that whole data set can be categorized into components with high probability. This clustering algorithm employs the basic Expectation-Maximization Algorithm. Fuzzy C-Means (FCM) clustering was developed by Dunn in 1973 and later improved by J. C. Bezdek in 1981. FCM is frequently used in pattern recognition. In this technique, data points are bound to each cluster by means of membership functions, which give degree of association with each cluster. A data point is assigned to that cluster with which it is highly associated i.e. degree or membership is highest.

III. ENHANCED FUZZY C-MEANS CLUSTERING ALGORITHM (EFCM)

The FCM algorithm [3] could not identify the optimum number of clusters required, which is a major drawback. To overcome this drawback, a modified and improvised algorithm EFCM is presented in this paper. This algorithm has the capability to find the optimum number of clusters.

A. The Algorithm

The EFCM algorithm has four important steps:

- 1) Generation of subsets.
- 2) Estimation of optimum number of clusters.

1) Generation of subsets:

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m \leq \infty \quad (1)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i th of d -dimensional measured data, c_j is the d -dimension center of the cluster, C is no. of cluster, and $\|\cdot\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} by:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

and the cluster centers c_j are given by:

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m x_i}{\sum_{i=1}^N u_{ij}^m} \quad (3)$$

This iteration will stop when $\max_{ij} \{ |u_{ij}^{k+1} - u_{ij}^k| \} < \epsilon$, where ϵ is a termination criterion between 0 and 1, whereas k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m .

The algorithm comprises of four steps:

1. Initialize the $U = [u_{ij}]$ matrix. It represents $U^{(0)}$.
2. At k -step: calculate the centers vectors $Ce^{(k)} = [c_j]$ with $U^{(k)}$ using equation 3 given above, assuming $C=2$ initially.
3. Update $U^{(k)}$, $U^{(k+1)}$ using equation 2 mentioned above.
4. If $\|U^{(k+1)} - U^{(k)}\| < \epsilon$ then STOP; otherwise return to step 2.

In this manner, the algorithm iteratively converges on the most optimum cluster centers.

2) Estimation of Optimum number of Clusters:

Determine the optimum number of clusters, C using validity measure Global Silhouette Index (GS) [7,8] defined in Section 3.2. GS value is calculated for various number of clusters ($n=1,2,\dots,20$). Value of n for which GS is maximum

is assigned to C . Form optimum number of clusters C using the Steps 2 to 13. These are the required clusters for an infrared image.

B. Cluster Quality Measure

1) Global Silhouette Index

GS is calculated as shown below:

$$GS = \frac{1}{c} \sum_{m=1}^c S_m \quad (4)$$

S_m is calculated as:

$$S_m = \frac{1}{N_m} \sum_{i=1}^{N_m} s(i), \quad (5)$$

where N_m is number of data points in m^{th} cluster and $s(i)$ is :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (6)$$

where $a(i)$ is the average distance between the i^{th} sample and all of the samples included in X_m (m^{th} cluster); ‘max’ is the operator, and $b(i)$ is the minimum average distance between the i^{th} sample and all of the samples clustered in X_k ($k = 1, \dots, c; k \neq m$). From this formula it follows that $-1 \leq s(i) \leq 1$. For a good cluster GS value should be near to 1. When $s(i)$ is close to 1, one may infer that the i^{th} sample has been “well-clustered”, i.e. it was assigned to an appropriate cluster. When $s(i)$ is close to zero, it suggests that the i^{th} sample could also be assigned to the nearest neighboring cluster.

2) Separation Index (SI) [9]

It uses a minimum-distance separation for partition validity. A lower value of SI indicates a better partition. For good quality clusters SI should have the least possible value.

$$SI = \frac{\sum_{m=1}^c \sum_{j=1}^n (\mu_{jm})^2 \|\bar{\mathbf{x}}^j - \bar{\mathbf{z}}_m\|^2}{n \cdot \min_{k,m} \|\bar{\mathbf{z}}_k - \bar{\mathbf{z}}_m\|^2} \quad (7)$$

where,

$\bar{\mathbf{z}}_m$ is the m^{th} cluster center,

μ_{jm} is the membership of data point j in cluster m .

$$\mu_{jm} \in \{0,1\}; 1 \leq j \leq n; 1 \leq m \leq c$$

$$\sum_{m=1}^c \mu_{jm} = 1$$

IV. RESULTS

The widely used clustering algorithms discussed above are tested on various synthetic and natural infrared images. Simulations are done using MATLAB on a PC with 2GB RAM, using a 2.00 GHz processor. Prior to the experiments, no pre-processing was done on these images. RGB features are used in clustering. The effectiveness of clusters is compared in terms of GS and SI values of cluster formed.

Example 1: In this example we deal with an infrared image of a boiler. Fig.1 shows why four is chosen as an optimum number of clusters. Table 1 shows the performance of different clustering techniques. Table 2 shows the clusters formed by different clustering techniques. From Table 2 we can easily visualize that the yellow part (1st cluster) which shows the hot water part is segmented properly using EFCM.

Example 2: Here we are concerned with an infrared image showing breast cancer. Yellow part shows the cancer prone area in the breast. Fig. 2 shows how the optimum number of cluster is four. Table 3 shows the performance of different clustering techniques used in this paper. Table 4 shows the clusters formed by different clustering methods. From this table we can see that cancer prone part (3rd cluster) is removed properly using EFCM. Validity of number of clusters can be assessed visually.

Example 3: In this example we consider segmentation of an infrared image taken at Tromso satellite station showing advancing of bloom. Yellow and red part shows warmer water and advancement of bloom. Here five is chosen as the optimum number of clusters. From Table 6 we can see that the red region (1st cluster) is completely separated from the image.

A. Comparison of performance

The results of infrared image segmentation indicate that EFCM is successful in segmenting the image properly with optimum number of clusters. The Global Silhouette Index value of EFCM is the best in most of the cases. Results show that EFCM could retrieve most of the clusters properly. The cluster centers in case of Fuzzy C-means are widely separated. It is able to retrieve all the basic clusters, giving less redundant clusters. The disadvantage with Fuzzy C-Means is that it cannot find out the optimum number of clusters. K-means result is also good but the most important disadvantage of K-means is that it generally does not give proper clusters in the first run. This is due to random selection of cluster centers which results in consumption of more time while dealing in large images. EM results are not better than that of EFCM and K-means.

B. Practical Application

The main aim of the paper is to propose a technique which can be conveniently used in automatic detection systems. Here, the phrase "Detection Systems" is concerned with systems like detection of hot water level in boilers, detection

of geographical features on any planet or even on unreachable areas of earth, weather forecasting systems and medical systems. Systems related to the examples mentioned above are described below:

- 1) Detection of water level in a boiler: Infrared images work on the principle of temperature sensing. They map a temperature range to various colors. This detection system takes the infrared image of the boiler and uses EFCM technique to segment the image properly for optimum number of clusters. For each cluster, average R, G, B value is calculated and corresponding temperature of that color is obtained from predefined scale. Then it compares the predefined temperature of water (i.e. 100°C) with the temperature of various clusters and finds out the required cluster. Next step is to check the height of water level in boiler which is proportional to the height of color in the required cluster (1st cluster). Now water supply can be controlled on the basis of height of color. Real time implementation of this system can reduce the complexity of conventional boiler control system designs.
- 2) Detection of breast cancer: Medical thermal imaging (MTI) can indicate life threatening diseases like breast cancer in their developing states, earlier than anatomical examinations. This type of detection system takes the infrared image, processes it for EFCM, calculates the average R, G, B values and obtains corresponding temperatures as described earlier. Subsequently, it checks for a predefined temperature region of the affected area to obtain the required cluster. From the cluster, affected area can be easily visualized. In Example 2 yellow color (3rd cluster) shows the cancer prone area which is faithfully separated.
- 3) Detection of Geographical Features: Example 3 shows the infrared image taken at Tromso Satellite station. This image shows the bloom. Advancement of red color can give the information about direction of movement of bloom and its intensity. Usually satellite sends the whole image (very large data) to earth station. The large amount of data to be transmitted is very likely to be corrupted by noise. So it is better to send only required information which is depicted in red color (1st cluster) in this case. If the satellite is equipped with EFCM processor then it will take the image, cluster it properly as shown in Table 6, and send only red color image to the earth station which is very small volume of data as compared to the amount of data required to represent the entire image. Continuous tracking of red color can give information about the advancement and intensity of bloom. This system can help in reducing the quantity of data transferred from the satellite to earth station because we can send only the required part of the image to earth.

TABLE 1
PERFORMANCE OF DIFFERENT CLUSTERING METHODS

Original image	Clustering Methods	GS	SI
Optimum number of clusters=4			
	K-Means	.7800	.0221
	EM	.5786	.0554
	EFCM	.7914	.0210

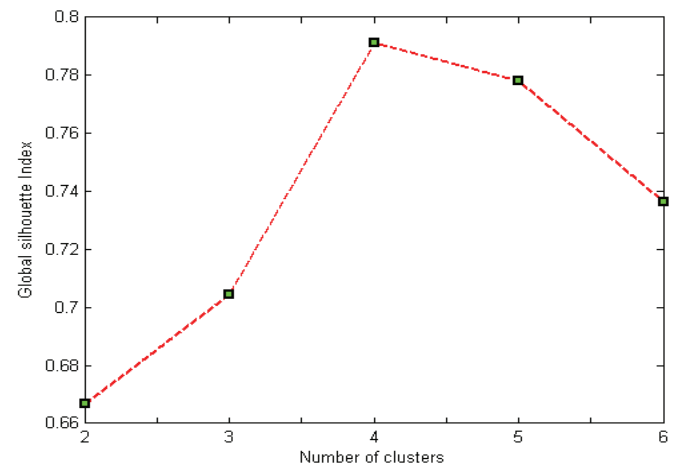


Fig. 1: Variation of GS with the number of clusters

TABLE 2
CLUSTERS OF INFRARED BOILER IMAGE FORMED BY DIFFERENT CLUSTERING TECHNIQUES

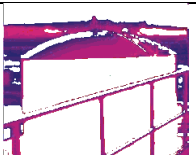
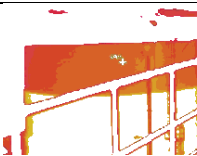
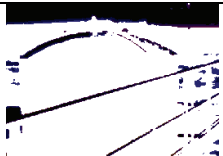
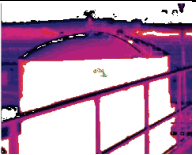
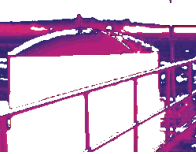
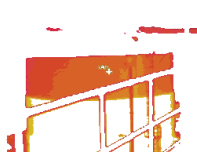
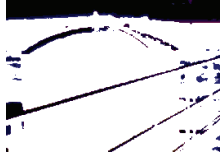
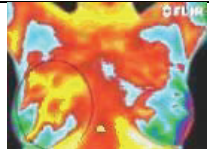
Clustering Method	1 st Cluster	2 nd Cluster	3 rd Cluster	4 th Cluster
K-means				
EM				
EFCM				

TABLE 3
PERFORMANCE OF DIFFERENT CLUSTERING METHODS

Original image	Clustering Methods	GS	SI
Optimum number of clusters=4			
	K-Means	.7097	.0347
	EM	.7094	.1122
	EFCM	.7128	.0317

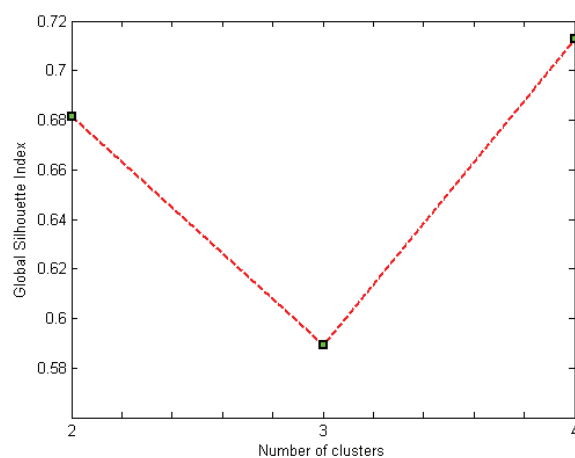


Fig. 2: Variation of GS with the number of clusters

TABLE 4
CLUSTERS OF INFRARED BREAST CANCER IMAGE FORMED BY DIFFERENT CLUSTERING TECHNIQUE

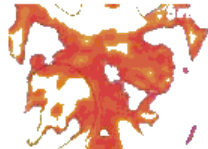
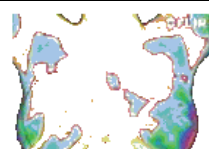
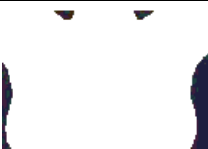
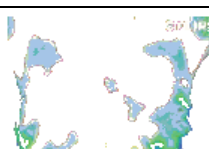
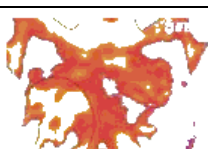
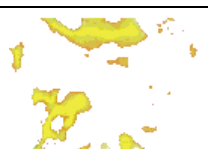
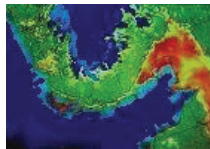
Clustering Method	1 st Cluster	2 nd Cluster	3 rd Cluster	4 th Cluster
K-means				
EM				
EFCM				

TABLE 5
PERFORMANCE OF DIFFERENT CLUSTERING METHODS

Original image	Clustering Methods	GS	SI
Optimum number of clusters=5			
	K-Means	.6480	.0256
	EM	.4031	.1913
	EFCM	.6498	.0263

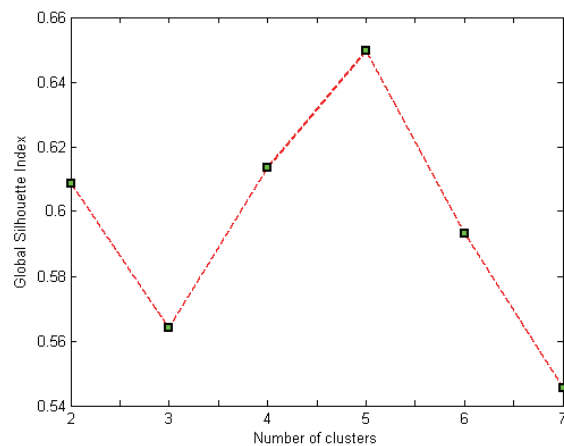
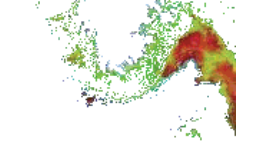
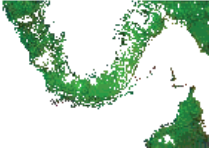
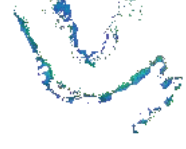


Fig.3: Variation of GS with the number of clusters

TABLE 6
CLUSTERS OF INFRARED GEOGRAPHICAL IMAGE FORMED BY DIFFERENT CLUSTERING TECHNIQUES

Clustering Method	1 st Cluster	2 nd Cluster	3 rd Cluster	4 th Cluster	5 th Cluster
K-means					
EM					
EFCM					

V. CONCLUSION

In this paper, we present EFCM based infrared image segmentation. We have applied some well-known clustering techniques, viz., K-Means, EM for the segmentation of colored images and compared their performance in terms of Global Silhouette Index and Separation Index that they achieve. We have obtained that EFCM was able to estimate the number of clusters accurately. Its performance on the basis of GS and SI is better than other techniques used for comparison. Hence this technique can be preferably used for the Automatic Detection Systems. Automatic Detection Systems described in this paper can make data transfer much faster and easier (in Example 3), can reduce the size and complexity of conventional control systems for boiler (in Example 1) and can even detect the very dangerous and rare diseases that cannot be easily diagnosed in early stages.

ACKNOWLEDGMENT

The authors would like to thank Mr. Sayed Atif Ali, undergraduate student at National Institute of Technology Allahabad, India, for helping to revise the paper thoroughly.

REFERENCE

- [1] A. K. Jain, M.N. Murthy, P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.
- [2] J.A. Hartigan and M.A. Wong, "A K-means clustering algorithm", *Appl. Stat.*, 28, 126-130, 1979.
- [3] M. Razaz, "A fuzzy C-means clustering placement algorithm", *Circuits and Systems*, 1993., ISCAS '93, 1993 *IEEE International Symposium*, 3-6 May 1993 Page(s): 2051 -2054 vol.3.
- [4] J. Abonyi, R. Babuska, F. Szeifert, "Modified Gath-Geva Clustering for Identification of Takagi-Sugeno Fuzzy Models", *IEEE Trans. on*

Systems, Man, and Cybernetics-partB: cybernetics, Vol.32, no5, October 2002.

- [5] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J.RoyStat.Soc. B*, Vo. 39, pp.1-38, 1977.
- [6] G. M. Tao, A .Tan, M.Hanmandlu, "The Model-Selection Problem", submitted to *IEEE Trans. on Fuzzy Systems*.
- [7] P.J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *J. Comp App. Math*, Vol. 20, 1987, pp. 53-65.
- [8] N. Bolshakova and F. Azuaje, "Clustering validation techniques for genome expression data", *Genomic signal processing*, Vol. 83, Issue 4, April 2003, pp. 825-833.
- [9] A.M Bensaid, L.O.Hall, J.C. Bezdek, L.P. Clarke, M.L. Silbiger, J.A.Arrington, and R.F. Murtagh. "Validity-guided (Re) Clustering with applications to image segmentation", *IEEE Transactions on Fuzzy Systems*, 4:112-123, 1996.