Histogram based Evolutionary Dynamic Image Segmentation

Amiya Halder
Computer Science & Engineering Department
St. Thomas' College of Engineering & Technology
Kolkata, India
amiya_halder@indiatimes.com

Arindam Kar and Soumajit Pramanik
Computer Science & Engineering Department
St. Thomas' College of Engineering & Technology
Kolkata, India
{arindam04.kar, soumajit.pramanik}@gmail.com

***Abstract*—This paper describes an evolutionary approach for unsupervised gray-scale image segmentation that segments an image into its constituent parts automatically. The aim of this algorithm is to produce precise segmentation of images using intensity information along with neighborhood relationships. The proposed technique automatically determines the number of clusters. The proposed algorithm is evaluated on well known natural images and its performance is compared to that of DCPSO, MAMA based segmentation techniques etc. The proposed algorithm is very simple in implementation. Experimental results shown that the algorithm generates good quality segmented image.**

Keywords- Clustering; Segmentation; Thresholding; Histogram Analysis; Genetic Algorithm.

I. INTRODUCTION

Segmentation refers to the process of partitioning a digital image into multiple segments or regions. The goal of segmentation is to simplify the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images [1]. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Image segmentation is a very important field in image analysis, objects recognition, image coding and medical imaging. Segmentation is very challenging because of the multiplicity of objects in an image and the large variation between them. Image segmentation is the process of division of the image into regions with similar attributes. In many object based image segmentation applications, the number of cluster is known a priori, but our proposed scheme automatically determines the number of cluster produced during the segmentation of images. The proposed technique should be able to provide good results whereas K-means algorithm which may get stuck at values which are not optimal [16]. Some of the several unsupervised clustering algorithms developed include K-means [7,8], fuzzy K-means, ISODATA [12], self-organizing feature map (SOM) [10], Particle Swarm Optimization (PSO) [9], learning vector quantizers (LVQ) [11], GA based Clustering [5] etc.

This paper presents automatic image segmentation of gray scale images using Histogram Analysis and Genetic Algorithm based clustering. One natural view of segmentation is that we are attempting to determine which components of a data set naturally "belong together". Clustering is a process whereby a data set is replaced by clusters, which are collections of data points that "belong together". Thus, it is natural to think of image segmentation as image clustering i.e. the representation of an image in terms of clusters of pixels that "belong together". The specific criterion to be used depends on the application. Pixels may belong together because of the same color or similarity measure. The result of this algorithm produced a better result to compare with other techniques such as DCPSO [6] MAMA [15,17]. Various segmentation techniques have been developed for image segmentation [2,13,14,15,17].

The rest of this paper is organized as follows: - Section II, the concepts of clustering is provided. Section III describes the threshold method. Section IV gives the concepts of Histogram analysis and section V gives the concepts of Genetic Algorithm and section VI describes the Histogram based Genetic Algorithm analysis and section VII describes the proposed algorithm and section VIII describes the experimental results and section IX concludes the paper.

II. CLUSTERING

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. By clustering, one can identify dense and sparse regions and therefore, discover overall distribution patterns and interesting correlations among data attributes.

Clustering may be found under different names in different contexts, such as unsupervised learning (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences) and partition (in graph theory) [3]. By definition, "cluster analysis is the art of finding groups in data", or "clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data into subsets (clusters), so that the data in each subset (ideally) share some common trait-often proximity according to some defined distance measure" [4]. Clustering is a challenging field of research as it can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, cluster analysis serves as a pre-processing step

978-1-4673-1850-1/12/$31.00 © 2012 IEEE

for other algorithms, such as classification, which would then operate on detected clusters.

Clustering is a useful unsupervised data mining technique which partitions the input space into K regions depending on some similarity/dissimilarity metric where the value of K may or may not be known a priori. The main objective of any clustering technique is to produce a K × n partition matrix U(X) of the given data set X, consisting of n patterns,

X ={x1, x2. . . xn} [13].

III. THRESHOLDING

Thresholding refers to the selection of a range such that if a pixel is within the threshold distance from a known centroid then the pixel is said to belong to that centroid's cluster. For any pixel x, the membership to a cluster centroid, C_i is defined as

$$C_i=\{x:x \in f(x,y) \text{ and } |I(C_i) - I(x)| <= T\},$$

where C_i= ith cluster centroid,

x = pixel under consideration,

$I(C_i)$ = Intensity value of ith centroid,

I(x) = Intensity value of the pixel x,

T = Threshold value,

f(x,y) = Input image.

In the proposed algorithm the threshold value is taken to be 5. Now, a 5 × 5 image with pixel values as shown in given below. Assume the three known centroids are 26, 80 and 134 respectively. We denote their cluster membership as I, II and III respectively.

26	30	134	138	82
26	30	130	136	83
25	26	129	135	80
24	26	129	135	80
23	29	132	133	81

The corresponding membership pattern is as shown in below.

I	I	III	III	II
I	I	III	III	II
I	I	III	III	II
I	I	III	III	II
I	I	III	III	II

The value of the threshold is selected as 5 for natural images based on experimental results.

IV. HISTOGRAM ANALYSIS

In traditional k-means, the new centroid value is computed as the mean of the pixel values of all the pixels that belong to a particular cluster.

By histogram analysis, the mode of the distribution of the pixel values of a cluster is calculated instead of the mean. The new centroid is taken as the pixel value that is repeated the highest number of times in the cluster.

The basis of histogram analysis stems from the fact that the mode is a more robust representative of the cluster than the mean and so a single very unrepresentative pixel in a cluster will not affect the mode value, which will affect the mean value significantly. Since the mode value must actually be the value of the pixel occurring the maximum number of times in the cluster, the histogram analysis does not create new unrealistic pixel values when there is wide variation in pixel values. Thus new intensity is never generated. Suppose, for a cluster of 1000 pixels, the following distribution shown in Fig.1.

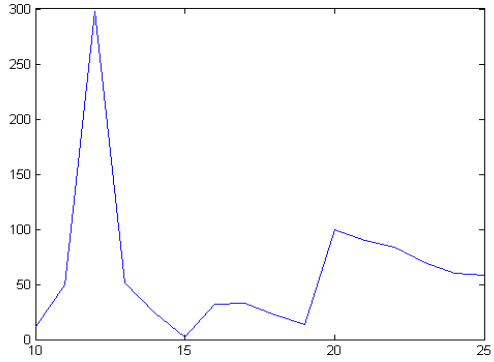

Figure. 1-Histogram (intensity vs. frequency)

Collect all the pixels of the each cluster which is defined by threshold operation. Find the pixel x that the occurrence is maximum and corresponding intensity I(x) is selected for the centroid's of that cluster.

From Fig.1, the mode value is: 12 with frequency of occurrence 298 i.e. 298 which is the maximum. The mean value is: 17.25 with frequency of occurrence 33. For this reason, mode is a better choice than mean.

V. GENETIC ALGORITHM

Genetic Algorithm (GA) is a population-based stochastic search procedure to find exact or approximate solutions to optimization and search problems. Modeled on the mechanisms of evolution and natural genetics, genetic algorithms provide an alternative to traditional optimization techniques by using directed random searches to locate optimal solutions in multimodal landscapes [18,19]. Each chromosome in the population is a potential solution to the problem. Genetic Algorithm creates a sequence of populations for each successive generation by using a selection mechanism and uses operators such as crossover and mutation as principal search mechanisms - the aim of the algorithm being to optimize a given objective or fitness function. An encoding mechanism maps each potential solution to a chromosome. An objective function or fitness function is used to evaluate the ability of each chromosome to provide a satisfactory solution to the problem. The selection procedure, modeled on nature's survival-of-the-fittest mechanism, ensure that the fitter chromosomes have a greater number of offspring in the subsequent generations. For crossover, two chromosomes are randomly chosen from the population. Assuming the length of the chromosome to be *l*, this process randomly chooses a point between 1 and *l*-1

and swaps the content of the two chromosomes beyond the crossover point to obtain the offspring. A crossover between a pair of chromosomes is affected only if they satisfy the crossover probability.

Mutation is the second operator, after crossover, which is used for randomizing the search. Mutation involves altering the content of the chromosomes at a randomly selected position in the chromosome, after determining if the chromosome satisfies the mutation probability.

In order to terminate the execution of GA we specify a stopping criterion. Specifying the number of iterations of the generational cycle is one common technique of achieving this end.

VI. HISTOGRAM BASED GA CLUSTERING

In this proposed scheme, we consider a gray level image of size m×n. The basic steps of the GA-clustering algorithm for clustering image data are as follows:

A. Population initialization

After thresholding and histogram analysis we have a set of centroids. Let, that set is C={C_1,C_2,……..C_k} and all centroids belonging to that set are marked as ungrouped. Suppose μ_1 is the minimum of all the ungrouped centroid belonging to C i.e.

$$\mu_1 = \min(C_i),\ \ where\ i = 1,2,..k \text{ -------}(1)$$

Create a group for μ_1 and put all the ungrouped centroids which have a difference of value □ or less with μ_1, into that group (S_1):

$$S_1 = \{\ C_i \mid\ C_i \in C\ \ and\ \ (C_i - \mu_1) \le \theta)\ \} \text{ --}(2)$$

Mark elements of S_1 as grouped. Next time, again find the minimum μ_2 of all ungrouped centroids in C and repeat this process until all centroids are grouped in this manner. Each centroid of set S_i is considered as chromosome of the population S_i. Now apply GA on centroids in each group S_i to find out the optimal centroid representing that group in final clustering,

B. Fitness value calculation

For each group S_i, segment the whole image using only the centroids belonging to S_i, using any standard method (such as, K-means). The fitness value for each chromosome (centroid) is the no. of pixels in its cluster.

Let, S_i = {C_1,C_2..C_n} and pixels of the cluster having centroid C_1={x_1,x_2..x_m}. Then the fitness value of C_1 is m:

$$f(C_1) = m \text{ ----------}(3)$$

In this way, the fitness function assigns a fitness value to each centroid belonging to S_i.

C. Selection

This fitness level is used to associate a probability of selection with each individual chromosome. We apply Roulette Wheel selection, a proportional selection algorithm where the no of copies of a chromosome that go into the mating pool for subsequent operations is proportional to its fitness.

If f(C_1) is the fitness of individual C_i in the population, its probability of being selected is,

$$p_i = \frac{f(C_i)}{\sum_{j=1}^{n} f(C_j)}, \text{ ----------}(4)$$

Where n is the number of individuals in the population.

D. Crossover

For crossover at first the selected chromosomes are represented as 8bit binary nos. Then we use single-point crossover with fixed crossover probability μ_c. Select a random no, rnd_1 such that $0 < rnd_2 < 8$ and do crossover with bits starting from rnd_2th bit position between centroids C_p and Cp+1 of Group S_i ,where p=1,3,5....[(no of members of Group S_i)- 1]

E. Mutation

Each chromosome undergoes mutation with a fixed probability μ_m. Select a random bit position of C_i and invert that with probability μ_m.

Here, the random bit position is 2 and μ_m =0.05 .Then the 2nd bit position of C_2 will be inverted with probability 0.05.

F. Terminal Criterion

We execute the processes of fitness computation, selection, crossover, and mutation for a predetermined number of iterations. In every generational cycle, the fittest chromosome till the last generation is preserved - elitism. Thus on termination, this chromosome gives us the best solution encountered during the search. That is, after this process we get the best centroid for each group S_i.

VII. PROPOSED ALGORITHM

Step 1: Take a grayscale image f(x,y) as input.

Step 2: **Thresholding:** For each and every pixel of the image, compare them to find similarity (a maximum deviation of (+/-5)
- 2.1: If true, put them in the same group.
- 2.2: Else, form a different group.

Step 3: Histogram Analysis: For each and every group, find the mode of all the pixel values belonging to that group. This is the new centroid of the group.

Step 4: Applying GA: Form groups of centroids found in step 3 and from each and every group find the fittest centroid using GA as discussed in section VI.

Step 5: Replacing Pixel Values: Cluster the image using optimal centroids proposed by G.A. Replace the image's pixel values with the centroids of the clusters to which they belong.

VIII. VALIDITY INDEX

The cluster validity measure used in the paper is the one proposed by Turi [2]. It aims at minimizing the validity index given by the function,

$$V = y \times \frac{intra}{inter} \quad ----------(5)$$

The term intra is the average of all the distances between each pixel x and its cluster centroid z_i which is defined as

$$\text{intra} = \frac{1}{N} \sum_{i=1}^{k} \sum_{x=C_i} \| x - z_i \|^2 \quad ----------(6)$$

Where $||x - z_i||$ means the Euclidean distance, and N is the total number of pixels, C_i is the cluster number, z_i is the centroids of cluster C_i, k is the total number of clusters. Intra cluster dependency is the sum of square of Euclidean distance of every element from the centroids of the cluster to which it belongs.

On the other hand, inter is the inter cluster dependency which gives the idea about the extent to which each clusters are related. The higher this value the better clustering is obtained. It is represented as

$$inter = \min(\| z_i - z_j \|^2), \; where \; i = 1,2,..k-1$$

$$j = i+1, i+2,..K \quad ----------(7)$$

Where z_i and z_j are the centroids. Intra cluster dependency is the minimum of the square of Euclidean distances of each centroids from the other.

Lastly, y is given as

$$y = c \times N(2,1) + 1$$

Where c is a constant (selected value is 25), N(2,1) is a Gaussian distribution function with mean 2 and standard deviation 1.This validity measure serves the dual purpose of

- Minimizing the intra-cluster spread, and
- Maximizing the inter-cluster distance.

Moreover it overcomes the tendency to select a smaller number of clusters (2 or 3) as optimal, which is an inherent limitation of other validity measures such as the Davies-Bouldin index or Dunne's index.

IX. EXPERIMENTAL RESULTS

The algorithm developed has been simulated using MATLAB. The input images are considered to be pgm images. The precision is assumed to be 8 i.e. the no of bits per pixel is 8. All the images files that we have tested are natural images. All the results have been reported in Fig.2, Fig.3 and Fig.4. These results have been compared to those of MAMA [15] and DCPSO [6]. The optimal range for the number of clusters for the images of Lena, mandrill and peppers has also been copied from [2] which are based on visual analysis by a group of ten people. Number of cluster obtained by this proposed method always gives range between optimal ranges [15].

In this paper we compare the validity index with other techniques. Our algorithm gives the better results.

X. CONCLUSIONS

This paper presented a new approach for unsupervised segmentation for gray-scale image that can successfully segment the images. In this paper, that the user does not need to predict the number of clusters, required to partition the dataset, in advance. Comparison of the experimental results with that of other unsupervised clustering methods, show that the technique gives satisfactory results when applied on well known natural images. Moreover results of its use on images from other fields (MRI, Satellite Images) demonstrate its wide applicability.

REFERENCES

[1] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, Pearson Education, 2002.

[2] R. H. Turi, "Clustering-Based Color Image Segmentation", PhD Thesis, Monash University, Australia, 2001.

[3] S.theodoridis and K.koutroubas, "Pattern Recognition", Academic Press, 1999.

[4] http:// en. wikipedia. org/ wiki/ cluster_analysis, Wikipedia-Cluster Analysis.

[5] Hwei-Jen Lin, Fu-Wen Yang and Yang-Ta Kao, "An Efficient GAbased Clustering Technique", in Tamkang Journal of Science and Engineering Vol-8 No-2, 2005.

[6] Mahamed G. H. Omran, Andries P Engelbrecht and Ayed Salman, "Dynamic Clustering using Particle Swarm Optimization with Application in Unsupervised Image Classification", PWASET Volume 9, 2005.

[7] E Forgy, "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification", Biometrics, Vol. 21, 1965.

[8] JA Hartigan, Clustering Algorithms, John Wiley & Sons, New York, 1975.

[9] DW van der Merwe, AP Engelbrecht, "Data Clustering using Particle Swarm Optimization".

[10] T Kohonen, "Self-Organizing Maps", Springer Series in Information Sciences, Vol 30, Springer-Verlag, 1995.

[11] LV Fausett, "Fundamentals of Neural Networks", Prentice Hall, 1994.

[12] G Ball, D Hall, "A Clustering Technique for Summarizing Multivariate Data", Behavioral Science, Vol. 12, 1967.

[13] Indrajit Saha, Ujjwal Maulik and Sanghamitra Bandyopadhyay," An Improved Multi-objective Technique for Fuzzy Clustering with Application to IRS Image Segmentation", EvoWorkshops 2009, LNCS 5484, pp. 426–431, 2009.

[14] Mofakharul Islam, John Yearwood and Peter Vamplew, "Unsupervised Color Textured Image Segmentation Using Cluster Ensembles and MRF Model" Advances in Computer and Information Sciences and Engineering, 323–328, 2008.

[15] Sreya Banerjee, Amiya Halder and Ayan Banerjee, "An Efficient Automatic Hierarchical Image Segmentation Algorithm based on Modal Analysis and Mutational Agglomeration", ICCCT 2010, pp. 216-219, Allahabad,India.

[16] S. Z. Selim, M. A. Ismail, " K-means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", IEEE Trans. Pattern Anal. Mach.Intell. 6, (1984), 81-87.

[17] Amiya Halder, Soumajit Pramanik, Swastik Pal, Nilabha Chatterjee and Arindam Kar, "Modal and Mutational Agglomeration based Automatic Colour Image Segmentation ",ICMV 2010, December 28 - 30, 2010, Hong Kong, China.

[18] 6. M. Srinivas, Lalit M. Patnaik, "Genetic Algorithms: A Survey".

[19] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.

(a) (b) (c) (d)

Figure.2: Experimental results of some natural images, (a)-Original image, using (b) MAMA, (c) DCPSO and (d) Our proposed method.

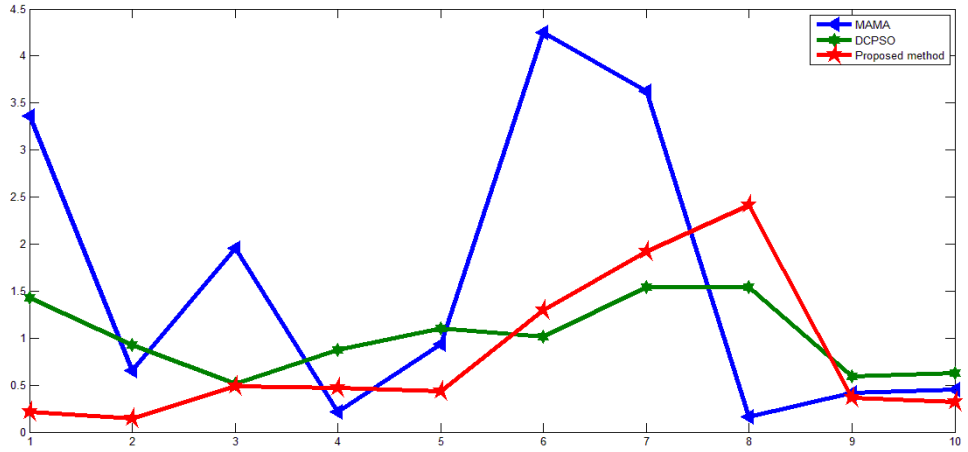

Figure..3: Validity index curve for MAMA, DCPSO and Proposed Method.

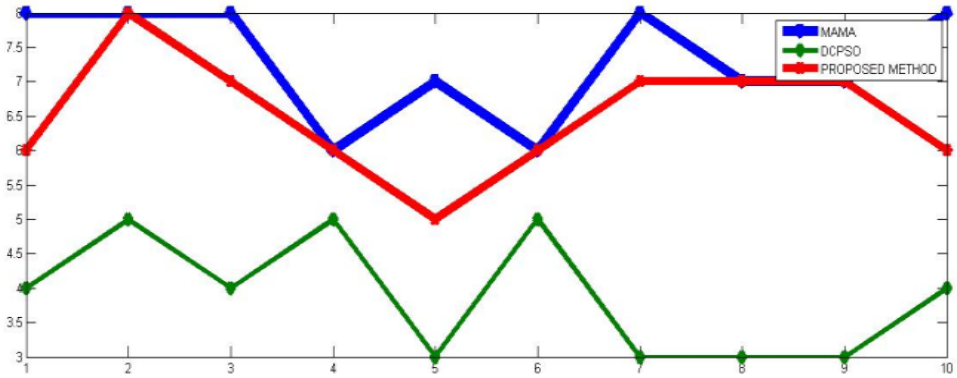

Figure.4: Number of clusters values for MAMA, DCPSO and Proposed Method.