Semi Supervised Image Segmentation Using Optimal Color Seed Selection

L. Sankari and C. Chandrasekar

***Abstract*—Image Segmentation refers to the process of partitioning the input image into several disjoint regions with similar characteristics such as intensity, color, and texture, shape etc. Semi supervised image segmentation is clustering the pixels of an image with some prior information or constraints. The Existing semi supervised method takes EM algorithm with mouse clicks as prior information. The drawback of EM algorithm is that it is prone to local maxima problem. Because of this reason the segmentation results will not be proper for certain kind of images. In this paper a new approach of optimal Semi Supervised Image Segmentation using Genetic algorithm is discussed. The optimal seeds are obtained and passed to EM algorithm. The Optimal seeds are nothing but color centers. The nearest colors are grouped together. The color classes are given in prior and the image is clustered using EM Clustering. In this paper Genetic algorithm is applied for finding optimal color classes so that the colors in the image are clustered sharply. Natural image data set from BSD images are taken and tested. The results of the proposed method are compared with Standard EM algorithm. The results show that the segmentation accuracy is improved.**

***Index Terms*—Image segmentation, EM clustering, semi supervised image segmentation**

I. Introduction

The image segmentation is an idea of grouping or classifying similar colors of an image and put them into the same group [1] [2]. That means it clusters (groups) the colors into several groups based on the closeness of color intensities inside an image. The objective of the image segmentation is to extract the image based dominant colors, texture, shape, spatial information etc.. The applications of image segmentation are diversely in many fields such as image compression, image retrieval, object detection, image enhancement, and medical image processing. Several traditional approaches have been already introduced for image segmentation [3]. The most popular method for image segmentation is EM clustering which is a model based clustering. Many researchers used Gaussian Mixture Model with its variant Expectation Maximization [4][5]. There are various methods existing for image segmentation in data mining areas like cluster based and classification based. But if the segmentation algorithms use only clustering technique, they may not produce proper clusters (groups) as requiredotherwise multiple possible groupings (over segmentation) may occur. Similarly if the image segmentation algorithms are based on only classification technique, sometimes insufficient labeled data may lead to poor classifier which will not give proper segmentation result (under segmentation/over segmentation). Therefore the idea of semi supervised [6] image segmentation (using both labeled and unlabeled data) is introduced in this paper.

Manuscript received June 22, 2012; revised October 29, 2012.

L. Sankari is with the Department of computer Science, Sri Ramakrishna College of Arts and Science For Women, Coimbatore, Tamilnadu India (e-mail: sankarivnm@gmail.com).

Dr. C. Chandrasekar is now with the Department of Computer Science, Periyar University, Salem, Tamilnadu, India (e-mail: ccsekar@gmail.com).

II. Semi Supervised Image Segmentation

A. Basic Concept

Semi supervised clustering [7] means grouping of objects such that the objects in a group will be similar to one another and different from (or unrelated to) the objects in other groups with related to certain constraints or prior information[8]. The following figure represents the same.

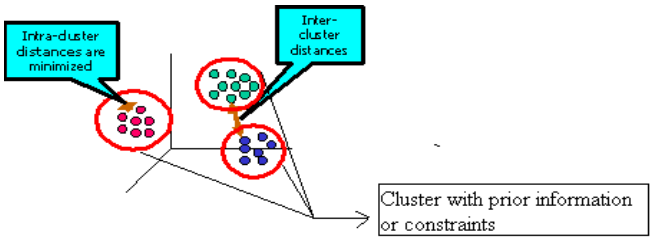

Fig. 1. Semi supervised clustering model.

A set of unlabeled objects, each described by a set of attributes and a small amount of domain knowledge is given as input. The image is read and the pixels are stored in a matrix. The prior information is nothing but the number of color groups for segmentation. With this information the EM clustering is done. Normally, for EM algorithm the cluster centroids are assigned randomly. This method was used in paper [9]. Since the cluster centroids were chosen randomly, for certain images the regions are not segmented properly. So in the proposed method the initial color centers are obtained using histogram and then optimal color centers are chosen. The following section discuss about the related previous works and proposed method.

B. Previous Related Works

Recently there are many papers focusing the importance of semi supervised image segmentation. Among them a few papers are analyzed. According to paper [10], the semi-supervised C-Means algorithm is introduced to solve three problems in the domains like choosing and validating the correct number of clusters, insuring that algorithmic labels correspond to meaningful physical labels tendency to recommend solutions that equalize cluster populations. The algorithm used MRI brain image for segmentation.

DOI: 10.7763/IJET.2012.V4.496

In this [11] paper, the author proposed how the popular k-means Clustering algorithm can be modified to make use of the available information with some artificial constraints. This method was implemented for six datasets and it has showed good improvement in clustering accuracy. This method was also applied to the real world problem of automatically detecting road lanes from GPS data and observed dramatic increases in performance.

In paper [12] a novel semi-supervised Fuzzy C-means algorithm is proposed. A set called as seed set which contains a small amount of labeled data is used. First, an initial partition in the seed set is done, then use the center of each partition as the cluster center and optimize the objective function of FCM using EM algorithm. Experiments results show that the defect of fuzzy c-means is avoided that is sensitive to the initial centers partly and give much better partition accuracy.

In Paper [13], Semi-supervised clustering was used with a small amount of labeled data to aid and bias the clustering of unlabeled data. Here labeled data is used to generate initial seed clusters along with the constraints generated from labeled data to guide the clustering process. It introduces two semi-supervised variants of KMeans clustering that can be viewed as instances of the EM algorithm, where labeled data provides prior information about the conditional distributions of hidden category labels. Experimental results demonstrate the advantages of these methods over standard random seeding and COP-KMeans, a previously developed semi-supervised clustering algorithm.

This paper [14] focused on semi-supervised clustering, where the goal is to cluster a set of data-points given a set of similar/dissimilar examples. Along with instance-level equivalence (similar pairs belong to the same cluster) and in-equivalence constraints (dissimilar pairs belong to different clusters) feature space level constraints (how similar are two regions in feature space) are also used for getting final clustering. This task is accomplished by learning distance metrics (i.e., how similar are two regions in the feature space?) over the feature space which that are guided by the instance-level. A bag of words models, which are nothing but code words (or visual-words) are used as building blocks. The proposed technique showed that non-parametric distance metrics over code words from these equivalence (and optionally, in-equivalence) constraints, which are then able to propagate back to compute a dissimilarity measure between any two points in the feature space. Thus this work is more advanced than previous works. First, unlike past efforts on global distance metric learning which try to transform the entire feature space so that similar pairs are close. This transformation is non-parametric and thus allows arbitrary non-linear deformations of the feature space. Second, while most Mahalanobis metrics are learnt using Semi-Definite Programming (SDP), this paper discuss about a Linear Program (LP) and in practice, is extremely fast. Finally, Corel image datasets were used (MSRC, Corel) where ground-truth segmentation is available. Over all, this idea gave improved clustering accuracy.

III. New Approach For Image Segmentation

In order to eliminate local maxima problem the idea of proposed method is developed. For optimization of color cluster centers the GA(Genetic Algorithm) is applied.

Fig. 2. Overall process of the proposed method.

Introduction to Genetic Algorithm

Genetic algorithm [15] is mainly useful for optimization and it is an evolutionary approach. Genetic algorithms (GA) were formally introduced in the United States in the year 1970s by John Holland at University of Michigan. In particular, genetic algo`rithms work very well on mixed (continuous *and* discrete), combinatorial problems. They are less susceptible to get 'stuck' at local optima than gradient search methods. But they tend to be computationally expensive.

To use a genetic algorithm, we must represent a solution of the problem as a *genome* (or *chromosome*). The genetic algorithm [16,17] then creates a population of solutions and applies genetic operators such as mutation and crossover to evolve the solutions in order to find the best one(s).

A typical genetic algorithm requires:

1) The genetic representation of the solution domain,
2) A fitness function to evaluate the solution domain.

A standard representation of the solution is as an array of bits.

Algorithm 1. Standard Genetic Algorithm

```
Step 1.Choose the initial population set
Step 2. Evaluate the fitness of each individual in
that population set
Step 3.Repeat until some criteria is met
    begin
      Calculate fitness function
      Perform crossover and mutation
     Evaluate the individual fitness function
            for new child
     Replace the unfit population with new
          one
    end
```

The objective function to be optimized:

$$F(c)=\sum_{i=1}^{K}\min_{j=1}^{N}(Region_i, center_j) \quad (1)$$

where,

K = Number of clusters (regions)

N = Total No. of pixels of same color

a)Initialization:

The population is generated randomly covering the entire solution space

b) Selection:

Individual solutions are to be selected through a fitness function.

$$Fitness= 1 \Big/ \sum_{i=1}^{N}(d) \quad (2)$$

where,

N = No. of pixels of same color.
d = Distance between each pixel of same color

A. Reproduction

The next step is to generate a second generation population of solution by crossover (also called recombination), and mutation. These steps are repeated in the next generation population of chromosomes that is different from the initial generation. Generally the average fitness will be increased by this procedure for the population

B. Termination

These procedures are repeated until certain condition is reached. In the exiting paper semi supervised algorithm for image segmentation is discussed with image quantization, prior information, EM clustering, Region merging operation.

In the proposed method the manual labeling is given in a different form. That is, from the histogram the most dominant color value is chosen and optimized those center points by genetic algorithm. The same output is given to EM clustering. The following algorithm explains these steps.

Algorithm 2: Proposed Algorithm

Step 1: Read the image
Step 2: Color quantization
Step 3: Manually assign number of clusters.
Step 4: Draw histogram and find out certain number of peaks using some threshold value.
Step 5: Assume these values as center points
Step 6: Find out optimal cluster centers Using Genetic Algorithm.
Step 7: These optimal color cluster centers Given as input for EM algorithm.
Step 8: Apply EM Algorithm for segmentation.

The above steps are explained below

1) Reading the given image

The image is read in Matlab. Bergley's image data set is taken and tested for various images.

2) Quantization

The RGB image cannot be processed as such. It is converted to L*a*b color space so that the colors are reduced. The L*a*b* color space is a perceptually uniform color space in which L* represents brightness and a* and b* represents chromatic information. It is obtained from RGB color space.

3) Initial label and histogram

The number of color clusters is given manually. For example the given image is to be separated (segmented) into 4 colors means assign 4.

Draw histogram of the given image and select some peeks at regular intervals (distances).

4) Initial centers

From Histogram all peeks cannot be taken as cluster centers. Since the histogram peaks are displayed with very minimum deviations. By assuming some threshold value peeks are selected. The most dominant color peeks are taken for color centers. Among these optimal centers are chosen by GA.

5) Optimal colors using GA

Fitness function: MATLAB rastriginsfcn
Selection : Random
Cross over: Scattered
Mutation : Guasian mutation

6) EM clustering

The Standard EM Algorithm:

Initialize the parameters to $\mu=0$, $i = 0$.
While (convergence condition not satisfied)
Begin

E-step:

Compute membership probabilities using the current μ *(current)* values.

M-step:

Compute new parameters μ *(new)* using the membership probabilities from the E-step.

Calculate the log likelihood using new parameter value and check the relative increase since the last iteration is below some threshold. If so, halt and return the current parameter. If not, continue to iterate.

End

The log likelihood is,

$$L(\theta, X) = \sum_{i=1}^{N} \log \sum_{i=1}^{K} p(c_j) P(x_i \mid c_j, \theta_j) \qquad (3)$$

K = No. of clusters (regions)
N = Total No. of pixels of same color

IV. Experimental Results And Discussion

The results are shown in Figure 4. The proposed results are compared with standard EM algorithm. It is clear from the result that the proposed method gives better visual appearance than the existing one, even though it takes more time (represented in Fig. 3.).

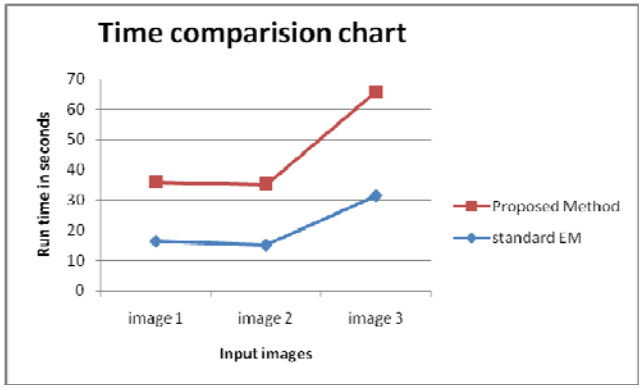

Fig. 3. Performance chart.

V. Conclusion

In this paper a new optimal semi supervised algorithm for color seed selection is proposed and tested with Bergley's natural images. This segmentation algorithm may be useful for finding objects in satellite images, medical image analysis. According to the nature of genetic algorithm, it takes more time but gives better results. The results are better in visual appearance than the standard EM algorithm. In future some other fitness function may be selected or some constraints may be given for selecting colors.

References

[1] L. Grady, "Random Walks for Image segmentations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783. November 2006.
[2] W. Skarbek and A. Koschan, *Colour Image Segmentation - A survey*, October 1994.

[3] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and Semi-supervised Clustering: a Brief Survey," August 15, 2005.
[4] Dr. K. Revathy and V. S. Roshini, "Applying EM algorithm for segmentation of textured images," *Proceedings of the World Congress on Engineering*, 2007, pp. 702-707.
[5] C. Fraley and A. E. Raftery, Technical Report No. 329 Department of Statistics University of Washington Box 354322 Seattle, WA 98195-4322 USA 'How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis".
[6] R. Wilkins, School of Engineering and Technology National University "Semi-Supervised Clustering".
[7] S. Basu, "Semi-supervised Clustering: Probabilistic models, algorithms and experiments," PhD Thesis, Department of Science, the University of Texas at Austin, 2005.
[8] K. L. Wagstaff, "Intelligent clustering with instance-level constraints," Ph.D. thesis, Cornell University, 2002.
[9] A. Y. Qian and W. Si, School of Computer Science, Zhejiang University, Hangzhou,310027, China, "Semi-supervised Color Image Segmentation Method," *IEEE International Conference*, Sept 11-14 2005.
[10] A. M. Bensaid and L. O. Hall, Department of Compute Science and Engineering University of South Florida Tampa, *Partially Supervised Clustering for Image Segmentation*, Pattern Recognition, vol. 29, no. 5, pp. 859-871, September 1994.
[11] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means Clustering with Background Knowledge," *Proceedings of the Eighteenth Nternational Conference on Machine Learning*, 2001, pp. 577-584.
[12] K. Li, Z. Cao, L. Cao, and R. Zhao, Coll. Of Electron.and Inf. Eng., Hebei Univ., Baoding, China, "A novel semi-supervised fuzzy c-means clustering method," *IEEE Explorer*, June 17-19, 2009.
[13] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised Clustering by Seeding," in *proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, Sydney, Australia, July 2002, pp. 19-26.
[14] D. Batra, R. Sukthankar, and T. Chen, *Semi-Supervised Clustering via Learnt Codeword stances*, 2008
[15] Matthew. Introduction to Genetic Algorithm. [Online]. Available: lancet.mit.edu/~mbwall/ presentations
[16] F. Pernkopf, *Member, IEEE* and D. Bouchaffra, *Senior Member, IEEE*, "Genetic-Based EM Algorithm for Learning Gaussian Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27 August 2005.
[17] S. Saha and S. Bandyopadhyay, "MRI brain image segmentation by fuzzy symmetry based genetic clustering technique," *IEEE Congress on Evolutionary Computation*, pp. 4417-4424, 2007

Mrs. L. Sankari is currently working as an Assistant Professor in the Department of Computer Science, Sri Ramakrishna College of Arts and Science for women, Coimbatore- 641 044, Tamilnadu, India. She is about 16 years of teaching experience. She has published four national and four international research papers. Her research interest area includes image processing, Data mining, Pattern classification and optimization techniques.

Dr. C. Chandrasekar has received his Ph.D. degree from Periyar University, Salem. He has been working as Associate Professor at Dept. of Computer Science, Periyar University, Salem – 636 011, TamilNadu, India. His research interest includes Wireless networking, Mobile computing, Computer Communication and Networks. He was a Research guide at various Universities in India. He has been published more than 60 research papers at various National/ International Journals.