Image segmentation using Fuzzy C-Mean and K Mean clustering technique

[1]Nikita Patil, [2]Ramesh Karandikar

[1]Almuri Ratnamala Institute of Technology and Engineering, Asangoan
[2] K.J. Somaiya College of Engineering, Vidyavihar

Abstract- Segmentation of an image entails the division or separation of the image into regions of similar attribute. The most basic attribute for segmentation of an image is its luminance amplitude for a monochrome image and color components for a color image. Clustering is one of the methods used for segmentation. This paper is review of K mean and Fuzzy c mean clustering.

Keywords— K-Means clustering, Optimal Fuzzy C- Means Clustering, Segmentation

I. INTRODUCTION

The image segmentation is a key process of the image analysis and the image comprehension. Because of the influence of the complicated background, the object characteristics diversity and the noise, the image segmentation is the difficult and hot research issues on the image processing. The process of partitioning a digital image into multiple regions (sets of pixels) is called image segmentation. Actually, partitions are different objects in image which have the same texture or color. The result of image segmentation is a set of regions that collectively cover the entire image, or a set of contours extracted from the image. All of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics. Some of practical applications of image segmentation are: image processing, computer vision, face recognition, medical imaging, digital libraries, image and video retrieval, etc [8].

Image segmentation methods fall into five categories: Pixel based segmentation [7], Region based segmentation [6], Edge based segmentation, Edge and region Hybrid segmentation and Clustering based segmentation [1]. The K-Means clustering technique is a well-known approach that has been applied to solve low-level image segmentation tasks. This clustering algorithm is convergent and its aim is to optimize the partitioning decisions based on a user-defined initial set of clusters that is updated after each iteration. This procedure is computationally efficient and can be applied to multidimensional data but in general the results are meaningful only if homogenous non-textured color regions define the image data. The applications of the clustering algorithms to the segmentation of complex color textured images are restricted by two problems. The first problem is generated by the starting condition (the initialization of the initial cluster centers), while the second is generated by the fact that no spatial (regional) cohesion is applied during the space partitioning process. Fuzzy clustering techniques have been effectively used in image processing, pattern recognition and fuzzy modeling .

II. COLOR IMAGE SEGMENTATION USING K-MEANS CLUSTERING

The K-Means clustering technique is a well-known approach that has been applied to solve low-level image segmentation tasks. This clustering algorithm is convergent and its aim is to optimize the partitioning decisions based on a user-defined initial set of clusters. The k-means clustering was proposed by Bo Zhao, ZhongxiangZhu, Enrong Mao and Zhenghe Song [12]. The selection of initial cluster centers is very important since this prevents the clustering algorithm to converge to local minima, hence producing erroneous decisions. The most common initialization procedure selects the initial cluster centers randomly from input data.

This procedure is far from optimal because does not eliminate the problem of converging to local minima and in addition the segmentation results will be different any time the algorithm is applied. In this paper, a different approach to select the cluster centers by extracting the dominant colors from the color histograms. The developed procedure is generic and proved to be very efficient when applied to a large number of images. The second limitation associated with the K-Means (and in general clustering algorithms) is generated by the fact that during the space partitioning process the algorithm does not take into consideration the local connections between the data points (color components of each pixel) and its neighbors. This fact will restrict the application of clustering algorithms to complex color-textured images since the segmented output will be over-segmented. In this regard to sample the local color smoothness, the image is filtered with an adaptive diffusion scheme while the local texture

complexity is sampled by filtering the input image with a gradient operator. Thus, during the space partitioning process, the developed algorithm attempts to optimize the fitting of the diffusion-gradient distributions in a local neighborhood around the pixels under analysis with the diffusion (color) gradient distributions for each cluster. This process is iteratively applied until convergence is reached. This is applied to the developed spatial clustering algorithm on a large selection of images with different level of texture complexity and on test data that has been artificially corrupted with noise.

A. The Algorithm

K-means is one of the most popular clustering algorithms. It is simple and fairly fast. K-means is initialized from some random or approximate solution. Each iteration assigns each point to its nearest cluster and then points belonging to the same cluster are averaged to get new cluster centroids. Each iteration successively improves cluster centroids until they become stable. Formally, the problem of clustering is defined as finding a partition of D into k subsets such that

$$\sum_{i=1}^{n} \zeta(ti;Cj) \qquad (1)$$

is minimized, where Cj is the nearest cluster centroid of ti.

The quality of a clustering model is measured by the sum of squared distances from each point to the cluster where it was assigned [6]. This quantity is proportional to the average quantization error, also known as distortion.

The quality of a solution is measured as:

$$q(C)=1/n \sum_{i=1}^{n} \zeta(ti; Cj) \qquad (2)$$

This can be computed from R as,

$$q(R, W) = \sum_{j=1}^{k} Wj \sum_{i=1}^{d} Rij \qquad (3)$$

In general, spatial partitioning methods are implemented using iterative frameworks that either attempt to minimize the variation within the clusters or attempt to identify the optimal partitions based on a set of Gaussian Mixture Models. In this paper focus the implementation of the K-Means algorithm, although the methodology detailed in this paper can be applied to other clustering schemes such as fuzzy clustering [3] or competitive agglomerative clustering [2]. The K-Means is a nonhierarchical clustering technique that follows a simple procedure to classify a given data set through a certain number of K clusters that are known a priori. The K-Means algorithm updates the space partition of the input data iteratively, where the elements of the data are exchanged between clusters based on a predefined metric (typically the Euclidian distance between the cluster centers and the vector under analysis) in order to satisfy the criteria of minimizing the variation within each cluster and maximizing the variation between the resulting K clusters. This clustering algorithm, in its standard formulation consists mainly of four steps that are briefly described below:

B. Steps of the classical K-Means clustering algorithm

1. Initialization – generate the starting condition by defining the number of clusters and randomly select the initial cluster centers.
2. Generate a new partition by assigning each data point to the nearest cluster center.
3. Recalculate the centers for clusters receiving new data points and for clusters losing data points.
4. Repeat the steps 2 and 3 until a distance convergence criterion is met.

The K-means clustering is a partitioning method for grouping objects so that the within-group variance is minimized. By minimizing dissimilarity of each subset locally, the algorithm will globally yield an optimal dissimilarity of all subsets [3].

The algorithm, as applied to image threshold, is given by the following steps:

1).Initialize the (K) class centers. For simplicity, un equal-distance method is used to define the initial class centers:

$$Center_I^0 = GL\ min + [(i-i/2)\ (GL\ max - GL\ min)/k)] \quad i = 1, 2\ldots k \qquad (4)$$

Where $Center_I^0$ is the initial class center for the i th class, GL max and GL min are the maximum and minimum of the gray value GL in the sample space, respectively.

2) Assign each point to its closest class center. The criterion to assign a point to a class is based on the Euclidean distance in the feature (GL) space using:

Distance i, j = abs (GL j – Center i)

$$i = 1,2,\ldots, K;\ j = 1,2,\ldots, N . \qquad (5)$$

Where Distance i, j is the distance from the j th point to the i th class, and N is the total number of points in the sample space.

3). Calculate the (K) new class centers from the mean of the points that are assigned to it. The new class centers are alculated by

$$Center_i^m = 1/Ni \sum_{i=1}^{Ni} GLj \quad j = 1, 2, \ldots..K. \qquad (6)$$

Where Ni is the total number of points that are assigned to the i th class in step 2.

4) Repeat step 2 if any class centers change, otherwise end the circulation.

5) The threshold value is defined as the average of the K th class center and the (K-1) th class center:

Threshold=1/2(center k+ center k-1). (7)

III. FUZZY CLUSTERING

Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed. Some examples of values that can be used as similarity measure include distance, connectivity, and intensity.

A. The Unsupervised Optimal Fuzzy C-Means Clustering Simple and well-known Fuzzy clustering algorithms, which is also widely used. In fact, there are two main shortcomings in the FCM algorithm. First, the numbers of resulting clusters need to be specified in advance, which in practice, such as in the unsupervised classification. Secondly, the FCM is very limited according to the restriction to spherical clusters. There are several derivatives of the FCM: the PCM algorithm, the UFP-ONC algorithm [3], and the KNN-IFCMA algorithm. All of them adopted a two step approach to obtaina good clustering result. The FCM algorithm is a core part of these algorithms. Especially, the results of PCM do depend on initialization which is obtained by using the FCM algorithm in advance. An unsupervised optimal fuzzy clustering (UOFC) algorithm [13], which can also be regarded as an improvement of the FCM, is proposed here to overcome these disadvantages [12].

B. The Optimal Fuzzy Clustering

Let us consider a collection of n patterns constituting vectors in the p-dimensional space of real numbers, namely x1,x2,……xnЄRp, forming the input data set X.then the new modified generalized objective function proposed based on [9] is given as follows:

$$J(U,V;X) = \sum_{i=1}^{C}\sum_{k=1}^{n}(\mu ik)m \{ (1-g) * (\| xk-Vi \| l)l + g *(\sum_{j=1}^{r} \{(xk-Vi)\bullet sij \}2)\}, l\geq 1 \quad (8)$$

Where, c is the number of the clusters, and other notations are described as follows.

Vi, (i=1, 2… c) is the prototype of the ith cluster. If one pattern xk belongs to the ith cluster, that means the distance between xk and Vi is smallest. The exponent parameter m is used to control the influence of intermediate membership values on the objective function.

And $1 < m < \infty$.

U = { μik } is the fuzzy membership matrix. Where, μik denotes the grade of membership of the k th pattern in the i th cluster. And it should satisfy the following two conditions:

- $\sum_{K=1}^{n} \mu ik = 1$ for all k.
- $n > \sum_{K=1} \mu ik > 0$ for all i.

g Є [0,1] is a weighted value whose role is to keep the balance between two basic components in the above equation. depending on the data set, a change of g could affect the resulting shapes of the obtained clusters.

The symbol (•) means the inner product.

It can be seen that Eq.(8) there are two components:

The first one characteristic the distance between the prototype Vi and the kth pattern xk, which presents the dissimilarity between Vi and xi. Unlike the simple Euclidean distance measure used in the FCM algorithm, UOFC algorithm adopted a more general l-norm distance measure. Apparently, if l=2, the distance measure is the well-known Euclidean distance. The advantage is that it needs not to be restricted in to the spherical clusters. Actually the UOFC algorithm can be applied in to arbitrary-shaped clusters.

The second term represents a linear variation which goes through the prototype Vi and is spanned by the collection of r linearly independent vectors si1, si2,…….,sir. These r vectors are the eigenvectors of the generalized within cluster scatter matrix Ei, corresponding to its first r largest eigen values which give the cohesiveness of the cluster.

$$Ei = \sum_{K=1}^{N} (\mu ik) m (xk-Vi) (xk-Vi)^T \quad (9)$$

These r eigenvectors, seen as the r principle eigenvectors determining the whole cluster approximately, give the most important directions, along which most of the patterns xk, (k =1, 2,….n) in the i-th cluster scatter. By introduction this special term, the principle directions of the cluster are emphasized .

As a result, the speed of searching the prototype of the cluster is improved. Especially for a large number of input patterns, the value of r can be increased to significantly elevate the convergence speed of this clustering algorithm. Differentiating the objective function J with respect to each Vi and μik ,we can obtain Eqs.(10),(11) used for updating the membership degrees and the prototypes in an iterative procedure until the difference between the new membership matrix and the

old one in the previous iteration step is less than a given tolerance bound.

$$Vi = \sum_{k=1}^{N} (\mu_{ik}^{t-1}) \, m \, xk \Big/ \sum_{k=1}^{n} (\mu_{ik}^{t-1})^m \qquad (10)$$

$$\mu_{ik}^{t} = 1 \Big/ \{ \sum_{j=1}^{r} \delta ik^t / \delta^t_{jk} \}^{(1/(m-1))} \qquad (11)$$

Where t is the iterative step number, and

$$\delta_{ik}^{t} = (1-g)* (\| xk-Vit \| l)l + g*(\sum_{J=1}^{r} \{(xk - Vit) \bullet s^t_{ij}\}^2 \qquad (12)$$

Obiviously, these two values are the necessary conditions for J to have a local minimum. However, minimization of J with Eq. (8) forms a class of constrained nonlinear optimization problems whose solution is unknown. This problem arises by placing the l-norm in the objective function.

For the present research, only l=2 norm is considered for testing the UOFC algorithm while other norm measures will be studied in future work. It can be seen that the two main advantages of the UOFC algorithm are

(1) Its additional linear in the generalized objective function;

(2)The l-norm distance measure used.

By minimizing the objective function J, we can quickly group great number of input patterns along with their r largest scattering direction. The computational precision can then be improved while the time and memory requirement can be greatly decreased [1]. In addition, it can process not only the hyper sphere-shaped and hyperllipsoidal-shaped clusters, but also any arbitrary-shaped clusters due to the l-norm introduced in distance measure.

C. The Algorithm

Fuzzy partition is carried out through an iterative optimization:

(1) Choose initial cluster centroids (seeds) V_i

(2) Compute the degree of membership of all feature vectors in all the clusters:

$$u\_\{ij\} = \frac{(1/d\text{^}2(Xj, Vi))\text{^} (1/(q-1))}{\text{Sum } (k=1, K) \, (1/d\text{^}2(Xj, Vk)) \text{^} (1/(q-1))}$$

(3) Compute new centroids Vi_new:

$$V_{i}\_new = \frac{\text{Sum } (j=1, N) \, (u\_\{ij\}) \text{^}q \, Xj}{\text{Sum } (j=1, N) \, (u\_\{ij\}) \text{^}q}$$

(4) When the movement of centroids (relative changes) is less than a predetermined threshold MOVETHRS, stop the iteration. Otherwise go to step 2. The algorithm will also terminate when a maximum number of iterations is reached.

(5) Finally, a data point Xj is assigned to cluster i if the fuzzy membership u_{ij} >= u_{kj} for all k clusters.

IV. CONCLUSION

In this paper, study of two clustering techniques was performed. The K-Means clustering and Optimal Fuzzy C-Means clustering techniques were chosen for evaluation. Using these two techniques, the performance for different images can be evaluate by using the parameters like error rate and resolution. A higher performance is achieved by OFCM compared with K-Means clustering method. In future this Optimal Fuzzy C-Means clustering technique is to be applied for obtaining better performance in the applications like face recognition and video retrieval.

In different ways, the performances of a traditional OFCM could be improved. Perhaps, the performances of most of the algorithm primarily depend on the chosen cluster centers. By properly fixing the cluster centers through domain knowledge, or some other mean, the convergence time along with the accuracy could be improved. So here the point is to find a better mechanism to fix the cluster centers by prior knowledge.

REFERENCES

[1] J.Bezdek, A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithm, IEEE Trans.Pattern. Anal. Mach. Intel, 2:1-8, 1980.

[2] GAO Xin-bo.Fuzzy cluster analysis and its applications [M].Xi'an, China: House ofXidian University, 2004. (in Chinese)

[3] Ahmed M N, Yamany S M, Mohamed N, etal. A modified fuzzy Cmeans algorithm for bias field estimation and segmentation of MRIdata [J].IEEE Trans on MedicalImaging, 2002, 21(3):193-199.

[4] DINGZhen, HU Zhong-shan, YANG Jing-yu, TANG Zhen-min, WU Yong-ge.Afuzzy-clustering-based approach to image 1997, 34(7):536-541. (in Chinese)

[5] F.Meyer, Color image segmentation, In IEEE International Conference on Image Processing and its Applications, pages 303–306, May 1995.Maastricht, The Netherlands.

[6] A.Moghaddamzadeh and N. Bourbakis, A Fuzzy Region Growing Approach for Segmentation of Color Images,Pattern Recognition, 30(6):867-881, 1997.

[7] G.A. Ruz, P.A.Estevez, and C.A. Perez, A Neurofuzzy Color Image Segmentation Method for Wood Surface Defect Detection, Forest Prod. J. 55 (4), 52-58, 2005.

[8] A.A. Younes, I. Truck, and H. Akdaj, Color Image Profiling Using Fuzzy Sets, Turk J Elec Engin, vol.13, no.3, 2005.

[9] Refael C.Gonzalez and Richard E.Woods, Digital Image processing (second edition) Pearson Education Asia Limited and Publishing House of Electronics Industyt, 2007.

[10] Pham DL and Prince JL, An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities, Pattern Recognit Lett, 1999, 20(1):57—68.

[11] B. Zhang,"Generalized K-harmonic Means–boosting in Unsupervised Learning, Technical report (HPL-2000–137), Hewlett-Packard Labs, 2000.

[12] ZhongxiangZhu ,Bo Zhao ,Enrong Mao and Zhenghe Song, Image segmentation based on Ant colony optimization and K-means clustering, Proceedings of the IEEE international conference on Automation and ogistics, August 18-21,2007,jinan,china.

[13] Gath,I. and Geva A. B. (1989), Unsupervised Optimal fuzzy clustering,IEEE Transaction on Pattern Analysis Machine Intelligence, 11(7): 773-781.