J.Thi-Qar Sci. Vol.3 (1) July/2011

ISSN 1991- 8690 الترقيم الدولي ١٩٩١ - ٨٦٩٠

Website: http://jsci.utq.edu.iq Email: utjsci@utq.edu.iq

Applying New Method for Computing Initial Centers of *k*-Means Clustering with Color Image Segmentation

Abbas H. Hassin Alasadi* Moslem Mohsinn Khudhair**

Dept. of Comp Sci., College of Sci., University of Basra, Basra, Iraq

Abstract

As a classic clustering method, the traditional k-Means algorithm has been widely used in image processing and computer vision, pattern recognition and machine learning. It is known that the performance of the k-means clustering algorithm depends highly on initial cluster centers. Generally initial cluster centers are selected randomly, so the algorithm could not lead to the unique result. In this paper, we present a method to compute initial centers for k-means clustering. Our method based on an efficient technique for estimating the modes of a distribution. We apply the new method in segmentation phase of color images. The experimental results appeared quite satisfactory.

Keywords: clustering; k-Means algorithm; Image segmentation; Color spaces.

1.Introduction

Image segmentation is the first step of the most critical tasks of image analysis, as shown in Figure (1). It is used either to distinguish objects from their background or to partition an image onto the related regions. There are different techniques that would help solve the image segmentation problem.Jeon *el al***. [1]** in his review of the previous related studies, categorized these techniques into the following: thresholding approaches, contour based approaches, region based approaches, clustering based approaches and other optimization based approaches using a Bayesian framework and neural networks. The clustering approaches can be categorized into two general groups: partitional and hierarchical clustering algorithms.

*Email: abbashh2002@yahoo.com **Email: Mos1970@yahoo.com

Partitional clustering algorithms such as *k*-means and Exception Maximize (EM) clustering are widely used in many applications such as data mining, compression, image segmentation, and machine learning.

Therefore, the advantage of clustering algorithms is that the classification is simple and easy to implement. Similarly, the drawbacks are of how to determine the number of clusters and decrease the numbers of iteration.**[2]**

The goal of image segmentation is partitioning of the image into homogeneous and connected regions without using additional knowledge on objects in the image. Homogeneity of regions in color image segmentation involves colors and sometimes textures. In the segmented image, the regions have, in contrast to single pixels, many interesting features, like shape, texture, and so forth. A human being recognizes objects in the environment using the visual system and segmenting color images.**[3]**

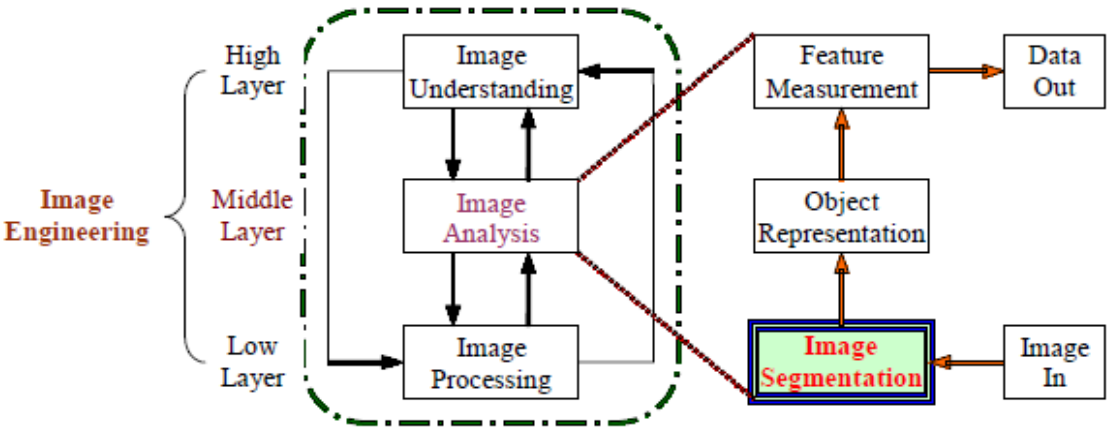

Figure (1): Image Engineering and Image Segmentation

2.Clustering analysis

Clustering analysis is one of the major data analysis methods widely used in many practical applications of emerging areas. Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. [4,5]

There are two main branches of clustering: (1) hierarchical and (2) partitional [6]. In this paper, we concentrate on partitional clustering. Particularly, a popular partitional clustering method called K-means clustering. The problem of clustering is to partition a data set consisting of *n* points embedded in m-dimensional space into *K* distinct set of clusters such that the data points within a cluster are more similar among them than to data points in other clusters. There are a number of

proximity indices that have been used as similarity measures [7]. Unfortunately, K-means algorithm is extremely sensitive to the initial choice of cluster centers, and a poor choice of centers may lead to a local optimum that is quite inferior to the global optimum. [8,9].

3. Problem Definition

There is no commonly accepted or standard method to determine either the number of clusters or the initial starting point values. The resulting set of clusters, both their number and their centroids, depends on the specified choice of initial starting point values. Two simple approaches to cluster initialization are either to select the initial values randomly or to choose the first *k* samples of the data points.[5]

To solve this problem, proposed an algorithm based on selecting *J* of subsamples, cluster them independently producing *J* estimates of the true cluster locations, and apply the k-means on this result to get the refined initial starting points. While this method can improve the final results but the final results depend on the quality of the selected subsamples, number of samples and the size of sample [5].

4. Related Work

Several attempts were made by researchers to improve the effectiveness and efficiency of the *k*-means algorithm. There are many researchers suggest initialized method of cancroids of k-means algorithm.

Stephen el al **[10]** proposed a method for choosing *K* instances randomly from database as seeds. The drawback of this technique is the computational complexity; several iterations of the *k*-means algorithm are needed after each instance is assigned, which in a large database is extremely burdensome.

Douglas el al **[11]** proposed a method to select a good initial solution by partitioning dataset into blocks and applying *k*-means to each block. But the time complexity is slightly more. Though the above algorithms can help finding good initial centers for some extent, they are quite complex and some use the *k*-means algorithm as part of their algorithms, which still need to use the random method for cluster center initialization.

5. Proposed Method

The standard *k*-means is a prototype-based, simple partitional clustering technique which attempts to find a user-specified *k* number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. The algorithm consist of two phases: the first phase is to define *k* centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to nearest centroid. The traditional *k*-means algorithm works as follows:

1. Select *K* points as initial centroids.
2. **repeat**
3. Form *K* clusters by assigning each point to its closest centroid.
4. Recomputed the centroid of each cluster.
5. **until** Centroids do not change.

We proposed a method for implementing the *k*-means algorithm. It can produce better clustering results on most cases. Our method scans the dataset block by block; produces *k* representative objects for each block, the method keeps the results from each block, and applies the *k*-means algorithm to the collected results

from all blocks to get the initial starting points. In other words, the proposed method compressed the data into smaller dataset by producing *k* means from each block, if the dataset contains *j* blocks, then the compressed data will contains *k*j* objects. Our method scans the original dataset two times and produces better clusters.

The main idea of the proposed method is to compress the dataset into finite number of representative. Each representative is the mean value of some data points form a small cluster. In this method we compress the dataset of size ***N*** into smaller data set of size ***k*m***; where ***k*** is the required number of partition for each block, ***m*** is the number of blocks. This process has done at the first phase. In the second phase we apply the *k*-means on the compressed dataset, to get the ***k*** representative points that will be the initial starting points for the *k*-means on the full dataset. The idea of compression of dataset comes from the BIRCH algorithm.**[12]**

Figure (3) exhibits the pseudo code of proposed method.

```
1.  Set the size of the block
2.  i=0
3.     while not end of file
4.        read the Block_i
5.        k-means (Block_i,k)
6.        (write/append) the means of output file
7.        i=i+1
8.     end while
9.  k-means (compressed dataset, k)
10. k-means (dataset, final means,k)
```

Figure (3): Pseudo code of proposed method.

Note that, in Figure (3), the method determines the size of each block and the user should determine the required number of partitions in each block. The value of *k* may be changed when applying the *k*-means on the compressed and the full dataset. From experimental results it will be better when we use small value for *k* at the first phase of our method, at the second phase we changed the value of *k* to the required number of clusters as a final results. So the value of *k* in steps 9, 10 may be different from the value of *k* in step 5. In line 10, the *k*-means start with the *k* points generated from the compressed dataset in line 9 and applied on the full dataset.

6. k-means Clustering

Clustering is the process of partitioning a set of objects (pattern vectors) into subsets of similar objects called clusters. Pixel clustering in three-dimensional color space on the basis of color similarity is one of the popular approaches in the field of color image segmentation. Colors, dominated in the image, create dense clusters in the color space in a natural way. Many different clustering techniques, proposed in the pattern recognition literature can be applied to color image segmentation[113]. One of the most popular and fastest clustering techniques is the *k*-means technique.

The *k*-means technique was proposed in the 1960s [14]. The first step of this technique requires determining a number of clusters k and choosing initial cluster centers C_i :

Where

$$C_i = [R_i, G_i, B_i], i = 1,2,\cdots,k \ldots (1)$$

During the clustering process, each pixel x is allocated to cluster K_j with the closest cluster center using a predefined metric the Euclidean metric. For pixel x, the condition of membership to the cluster K_j during the nth iteration can be formulated as follows:

$x \in K_j(n) \Leftrightarrow \forall i = 1,2,..., j-1, j+1,..., k...$ wh

$\|x - C_j(n)\| < \|x - C_i(n)\|$ ...(2)

ere C_j is the center of the cluster K_j

The main idea of *k*-means is to change the positions of cluster centers as long as the sum of distances between all the points of clusters and their centers will be minimal. For cluster K_j , the minimization index J can be defined as follows:

$$J_j = \sum_{x \in K_j(n)} \left\| x - C_j(n+1) \right\|^2 ...(3)$$

After each allocation of the pixels, new positions of cluster centers are computed as arithmetical means. From Equation 3, we can calculate the arithmetical means of color components of the pixels belonging to the center of the cluster K_j formed after $n+1$ iterations as:

$$C_{j_R}(n+1) = \frac{1}{N_j(n)} \sum_{x \in K_j(n)} x_R$$

$$C_{j_G}(n+1) = \frac{1}{N_j(n)} \sum_{x \in K_j(n)} x_G$$

$$C_{j_B}(n+1) = \frac{1}{N_j(n)} \sum_{x \in K_j(n)} x_B$$

... (4)

where $N_j(n)$ is the number of pixels in cluster K_j after *n* iterations.

In the next step, checking the difference between new and old positions of the centers. If the difference is larger than a threshold ***T***, for example (***T***=$\mathbf{10^{-6}}$), then

start the next iteration, and calculate the distances from the pixels to the new centers, pixels membership, and so forth. If the difference is smaller than threshold ***T***, then stop the clustering process. During the last step of the *k*-means processes, the color of each pixel is turned to the color of its cluster center. The number of colors in the segmented image is reduced to *k* colors.

7. Experimental Results

The results of segmentation by *k*-means depend on the position of the initial cluster centers. In the case of semi automated version of *k*-means, the input data can be defined by the human operator. In the case of the automated version of *k*-means, the initial centers can be chosen randomly from all the colors of the image. There are also other possibilities for the choice of centers, including *k* colors from the first pixels in the image and *k* gray levels from the gray line uniformly partitioned into *k* segments.

Table (1) shows the experimental results of proposed methods which implemented on four samples of images. The results show that each segmented image has its own number of clusters and own number of iteration, which depends on the density of the color and its gradation.

We have evaluated our method on several different standard images, as shown in Table (1). We have compared our results with that of *k*-means algorithm in terms of the total execution time and quality of clusters. Our experimental results are reported on PC 2.0GMHz CPU, 2.0GB RAM, 512 kB Cache.

In Figure (2), we compared the CPU time of the proposed method with the standard *k*-means methods. The execution time of proposed method was much less than the average execution time of *k*-means when used random initialization. Our proposed method provides higher accuracy than the other methods and takes moreover equal time.

Table (1): Experimental Results of proposed method.

Original Image	# Clusters	# Iterations	*Segmented Image*
	5	8	
	12	35	
	12	23	
	8	15	

In Figure 3, demonstrate that the proposed method provides better cluster accuracy than the existing methods. It shows the proposed method performs much better than the random initialization algorithm. The experimental images show the effectiveness of our approach. This may be due to the initial cluster centers generated by proposed method are quite closed to the optimum solution.

8. Conclusion

k-means algorithm is a popular clustering algorithm applied widely, but do not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids. Experimental results show that selecting centroids by our method can lead to a better clustering. Moreover, the computational complexity of the standard algorithm is objectionably high owing to the need to reassigning the data points a number of times, during every iteration of the loop.

This paper presents an enhanced *k*-means algorithm which combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. The previous improvements of the *k*-means algorithm compromise on either accuracy or efficiency. A limitation of the proposed method is that the value of *k*, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points. Evolving some statistical methods to compute the value of *k*, depending on the data distribution, is suggested for future research. Methods for refining the computation of initial centroids is worth investigating.

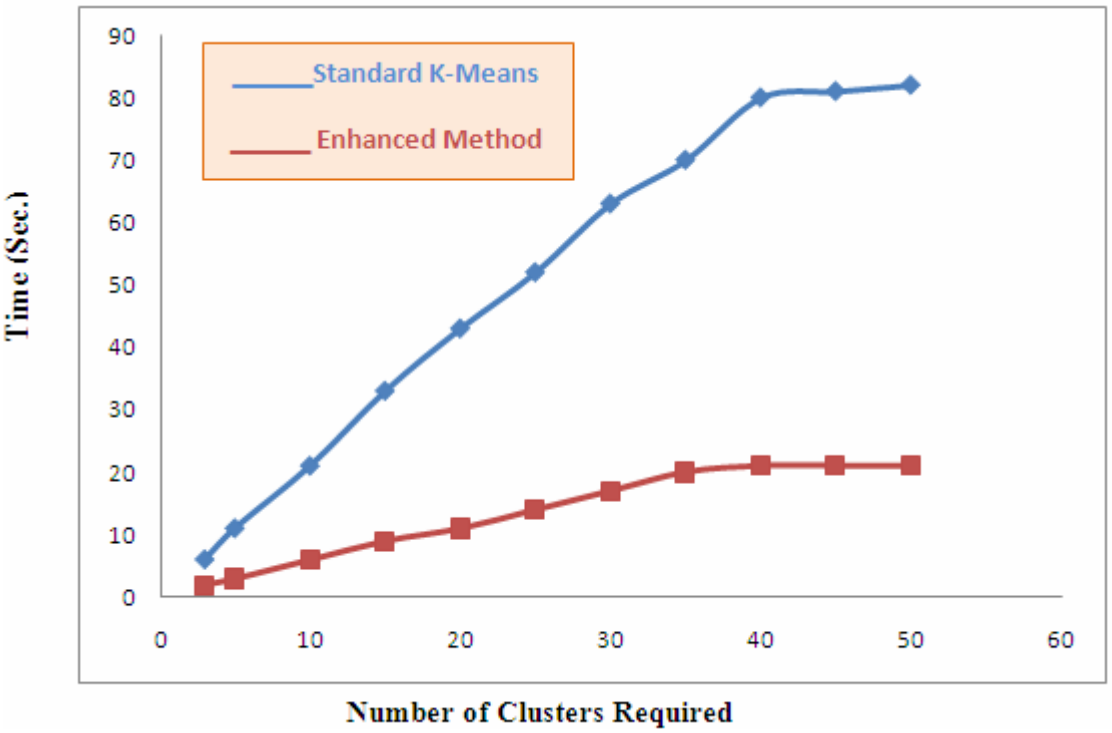

Figure (2): Execution Time (Lena Image).

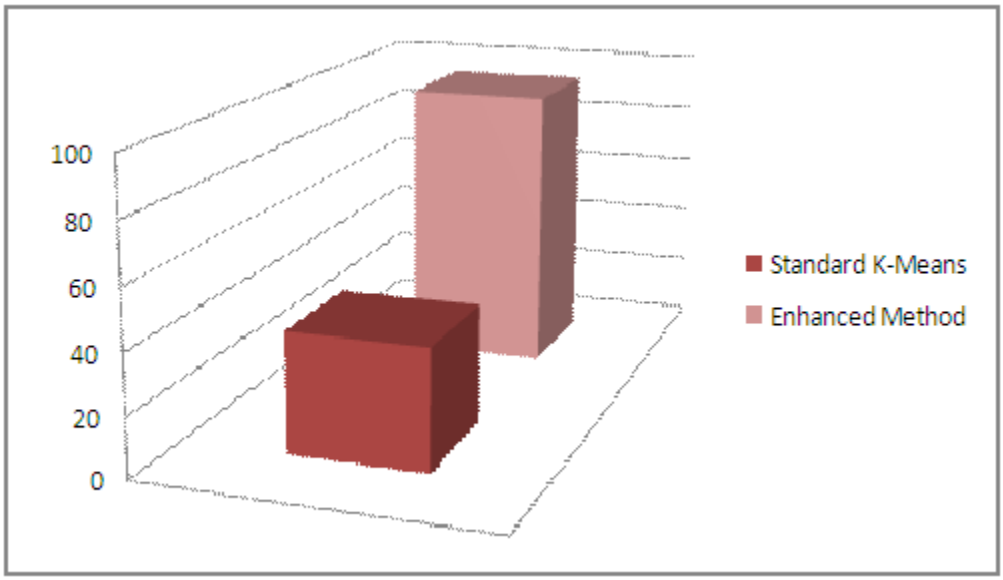

Figure (3): Efficiency and Accuracy of the Algorithms.

References

1. B. Jeon, Y. Yung and K. Hong ”Image segmentation by unsupervised sparse clustering, ” pattern recognition letters 27science direct,(2006) 1650-1664.
2. Ali Salem Bin Samma and Rosalina Abdul Salam. "Adaptation of *K*-Means Algorithm for Image Segmentation", International Journal of Signal Processing 5:4 2009, P 270-274.
3. Rastislav Lukac, Konstantions N. P. , "Color Image Processing: Methods and Applications", CRC Press is an imprint of Taylor & Francis Group, an Informa business, 2007.
4. Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in *K*-Means Clustering and Application in Colour Image Segmentation", School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia.
5. Fisher, D., 1987. Knowledge acquisition via incremental conceptual clustering. Mach. Learn. 2, 139–172.
6. Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Comput. Surveys 31 (3), 264–323.
7. Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press Inc.
8. Xiaoping Qing, Shijue Zheng, "A new method for initialising the *K*- means clustering algorithm", Second International Symposium on Knowledge Acquisition and Modeling, 2009, P 41-44.
9. Fahim A.M., Salem A. M., Torkey F. A., Ramadan M. A., Saake G., An Efficient K-Means with Good Initial Starting Points", Georgian Electronic Scientific Journal: Computer Science and Telecommunications 009, No.2(19), P47-57.

10. Stephen J. Redmond, Conor Heneghan, "A method for initialising the K-Means clustering algorithm using kd-trees", Department of Electronic Engineering, University College Dublin, Bel_eld, Dublin 4, Ireland.
11. Douglas Steinley, Michael J. Brusco, "Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques", Journal of Classification 24:99-121 (2007), DOI: 10.1007/s00357-007-0003-0.
12. Zhang T., Ramakrishnan R., Linvy M.: "BIRCH: An Efficient Data Clustering Method for Very Large Databases". Proc. ACM SIGMOD Int. Conf. on Management of Data, ACM Press, New York, 1996, pp.103- 114.
13. A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.
14. J. Mac Queen, Some methods for classification and analysis of multivariate observations, in Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics, and Probabilities, Berkeley and Los Angeles, CA, Vol. I, University of California, Berkeley, CA, USA, 1967, pp. 281–297.

تطبيق طريقة جديدة لحساب المراكز الأولية لخوارزمية (*k*-means) للعنقدة واستخدامها في تقسيم الصور الملونة

عباس حنون حسن الاسدي مسلم محسن خضير

قسم علوم الحاسبات – كلية العلوم – جامعة البصرة

الخلاصة

تعتبر خوارزميات (K-means) العنقدة من الطرق التقليدية واسعة الاستخدام في مجالات عديدة مثل معالجة الصور ، تمييز الأنماط والتنقيب عن البيانات ..الخ.

أن كفاءة وانجازيه هذه الخوارزميات تعتمد بشكل كبير على القيمة الأولية لاختيار نقاط التمركز الأولية في بداية عمل الخوارزمية. وكانت الطريقة التقليدية المتبعة لاختيار هذه النقاط تتم بصورة عشوائية.

في هذا البحث ، اقترحنا طريقة جديدة يتم من خلالها احتساب نقاط التمركز على أساس تجزئة الصورة إلى مساحات متساوية.

تم استخدام هذه الطريقة لتقسيم الصور الملونة وكانت النتائج مرضية من ناحية الزمن والكفاءة.

الكلمات المفتاحية :العنقدة.K-means; خوارزمية التعديل; تقطيع الصورة;فضاء الألوان