Adaptation of K-Means Algorithm for Image Segmentation

Ali Salem Bin Samma and Rosalina Abdul Salam

Abstract—Image segmentation based on an adaptive K-means clustering algorithm is presented. The proposed method tries to develop K-means algorithm to obtain high performance and efficiency. This method proposes initialization step in K-means algorithm. In addition, it solves a model selection number by determining the number of clusters using datasets from images by frame size and the absolute value between the means, and additional steps for convergence step in K-means algorithm are added. Moreover, in order to evaluate the performance of the proposed method, the results of the proposed method, standard K-means and recently modified K-means are compared. The experimental results showed that the proposed method provides better output.

Keywords—Initialization, Modify K-Means, Segmentation, Standard K-Means.

I. INTRODUCTION

IMAGE techniques can be grouped under a general framework of image engineering (IE), which consists of three layers: image processing (low layer), image analysis (middle layer) and image understanding (high layer), as shown in Fig.1 [3].

Image segmentation is the first step and also one of the most critical tasks of image analysis. It is used either to distinguish objects from their background or to partition an image onto the related regions. Although it is one of the primary steps in object recognition, it is also considered to be one of the most popular problems in computer vision. There are different techniques that would help solve the image segmentation problem. Byoung [2] 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, neural networks [2]. The clustering approaches can be categorized into two general groups: partitional and hierarchical clustering algorithms (for details, please refer to [1]). Partitional clustering algorithms such as K-means and EM clustering are widely used in many applications such as data mining [3], compression, [4] image segmentation [4], [5] and machine learning [6]. 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, [7]. This paper is organized as follows: Section 2 reviews the related studies and briefly describes the family of K-means clustering algorithms. The method to

Ali Salem Bin Samma is with School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia (e-mail: Binsamma@cs.usm.my).

Rosalina Abdul Salam is with School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia. (phone:+ 604-6532486; fax: +604-6573335; e-mail: rosalina@cs.usm.my).
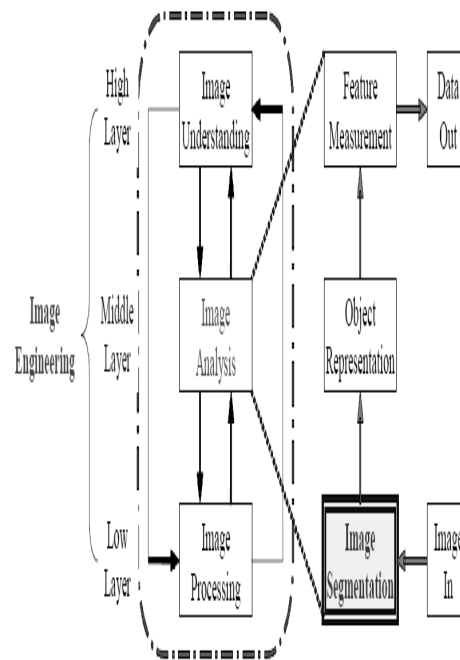

Fig. 1. General framework image engineering (IE).

modify K-means algorithm is presented in section 3. In Section 4, the experimental results and evaluation of proposed method in the area of image segmentation are provided. Finally, the conclusion and future work are presented.

II. RELATED WORKS

The K-means algorithm and its development algorithms are in the family of center base clustering algorithms. This family consists of several methods: expectation maximization, fuzzy K-means and harmonic K-means. A brief review of these algorithms is given in the following sub-sections.

A. General clustering algorithm

According to D. Malyszko [8], the general steps of the center base clustering are:

1) Initialize step with centers C.
2) For each data point xi, compute its minimum distance with each center cj.

$$C_j = \frac{\sum_{i=1}^{N} m(c_j|x_i)w(x_i)x_i}{\sum_{i=1}^{N} m(c_j|x_i)w(x_i)} \qquad (1)$$

3) For each center cj, recomputed the new center from all data points xi belong to this cluster.
4) Repeat steps 2 and 3 until convergence.

B. Standard K-means clustering algorithm

K-means is a popular algorithm for clustering; it partitions data set into k sets. The membership for each data point belongs to its nearest center, depending on the minimum distance. This membership determines as, [8]:

$$AKM(x,c) = \sum_{i=1}^{N} min_{j \in (1...k)} ||x_i - c_j||^2 \qquad (2)$$

There are several methods to improve the standard K-means algorithm related to several aspects. Standard K-means algorithm consists of four steps: initialization, classification, computational and convergence condition. Basically, the initialization step has received the most attention compared to the other steps. Stephen [9], indicates that the earliest references to initializing the K means algorithm was by Forgy in 1965 who chose points randomly. For example, it can be a point near a cluster centre or outlying point. MacQueen, [10], introduced what is akin to an online learning strategy to determine a set of cluster seeds. Tou and Gonzales [11], suggested the Simple Cluster Seeking (SCS) method. Linde et al. [12], proposed a Binary Splitting (BS) method which was intended for use in the design of Vector Quantizes codebooks. Kaufman and Rousseeuw [13], suggested selecting the first seed as the most centrally located instance. Babu and Murty [17], published a method of near optimal seed selection using genetic programming. However, the problem with genetic algorithms is that the results vary significantly with the choice of population size, and crossover and mutation probabilities [5]. Huang and Harris [15], proposed the Direct Search Binary Splitting (DSBS) method. This method is similar to the Binary Splitting algorithm above except that the splitting step is enhanced through the use of Principle Component Analysis (PCA). Katsavounidis et al. [16], proposed, what has been termed by some as the KKZ algorithm. This algorithm starts by choosing a point x, preferably one on the 'edge' of the data, as the first seed. The point which is furthest from x is chosen as the second seed. Daoud and Roberts [18], proposed to divide the whole input domain into two disjoint volumes. In each subspace, it is assumed that the points are randomly distributed and that the seeds will be placed on a regular grid. Thiesson et al. [19], suggested taking the mean of the entire dataset and randomly perturbing it K times to produce the K seeds. Bradley and Fayyad [13], presented a technique that begins by randomly breaking the data into 10, or so, subsets. Then it performs a K means clustering on each of the 10 subsets, all starting at the same set of initial seeds, which are chosen using Forgy's method. Likas et al. [20], present a global K means algorithm which aims to gradually increase the number of seeds until K is found. Khan and Ahmad [21], described a Cluster Centre Initialization Method (CCIA) using a Density-based Multi Scale Data Condensation (DBMSDC) which was introduced. DBMSDC involves estimating the density of the data at a point, and then sorting the points according to their density. From the top of the sorted list was choosing a point and pruning all points within a radius inversely proportional to the density of that point. Then it moves on to the next point which has not been pruned from the list and the process is repeated until a desired number of points remain. The authors choose their seeds by examining each of the m attributes individually to extract a list of K0 ¿ K possible seed locations. Next the DBMSDC algorithm is invoked and points which are close together are merged until there are only K points remaining.

C. Expectation maximization

Expectation maximization algorithm uses a linear combination of Gaussian distribution as centers [22]. Its minimization is:

$$EM(x,c) = \sum_{i=1}^{N} \log(\sum_{j=1}^{k} p(x_i|c_j)p(c_j)) \qquad (3)$$

This algorithm has a constant weight that gives all data point to its nearest center.

D. Fuzzy K-means

Fuzzy K-means algorithm is also called fuzzy c-means. It is adaptation of the K-means algorithm and use soft membership function. This algorithm determines a data point belongs to any centers depends on its membership as [8]:

$$EM(x,c) = \sum_{i=1}^{N} \sum_{j=1}^{k} U_ij^r ||x_i - c_j||^2 \qquad (4)$$

This algorithm has a soft membership and constant weight that gives all data point to the closed center.

E. Harmonic K-means algorithm

The harmonic K-means algorithm is a method which is similar to the standard K-means. It uses the harmonic mean of the distance from each data point to all centers as [23]:

$$HKM(x,c) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{||x_i - c_j||^2}} \qquad (5)$$

This algorithm has a soft membership and weight function to points that are far away from every center.

F. Early stop K-means

Early stop K-means algorithm is the first one to handle a convergence step in the standard K-means algorithm. It consists of associating the square error values to a convergence condition. It gets action when there are two consecutive iterations and the square error of the last iteration exceeds that of the preceding iteration. It finds a solution at least as good as that of the standard K-means with a number of iterations smaller than or equal to that of standard K-means algorithm [24].

G. Modified K-means

Modified K-means algorithm is a new algorithm for K-means based on the optimization formulation and a novel iterative method. The steps of this algorithm represented as [25]:

1) Dividing data set (D) into K parts:

$$D = \bigcup_{k=1}^{k} S_k, S_k1 \bigcap S_k2 = \varnothing, k1 \neq k2 \qquad (6)$$

2) Let be initial clustering centers calculate by:

$$X_{(k)}^{(0)} = \sum_{d^{(j)} \in S_k} d^{(j)} / |S_k|, k = 1, ..., K \qquad (7)$$

3) Decide membership of the patterns in each one of the K clusters according to the minimum distance from cluster center.

4) Calculate new centers using the iterative formula below:

5) Repeat step 3 and 4 till there in no change in cluster center.

III. PROPOSAL METHOD

With the development of approaches for image segmentation, Zhang [7], indicated to the numbers of studies for image segmentation algorithms. The proposed method tries to modify the K-means algorithm in two aspects: firstly to improve initialization step and secondly to develop convergence step.

A. Initialization step

Since the K Means algorithm results are dependent on the choice of the initial cluster centers, the modification method is used to estimate the number of cluster and their values in k means algorithm. This method depends on the data and works well to find the best number of cluster and their centroids values. It starts by reading the data as 2D matrix, and then calculates the mean of the first frame size F1=300x300, F2=150x150, F3=100x100, F4=50x50, F5=30x30, F6=10x10 or F7=5x5. Then, it keeps the value of means in an array called means array even at the end of the data matrix. After that it sorts the values in the means array in an ascending manner. In cases where the values are similar, they are removed to avoid an overlap. In other words, only one value is kept. It will then calculate the number of elements in the means array: this number is the number of clusters and their values are the centroids values as indicated in the steps below:

1) Read the data set as a matrix.
2) Calculate the means of each frame depending on the frame size and putting them in the means array.
3) Sort the means array in an ascending way.
4) Comparing between the current element and the next element in the means array. If they are equal, then keep the current element and remove the next, otherwise, keep both.
5) Repeat step 4 until the end of the means array.
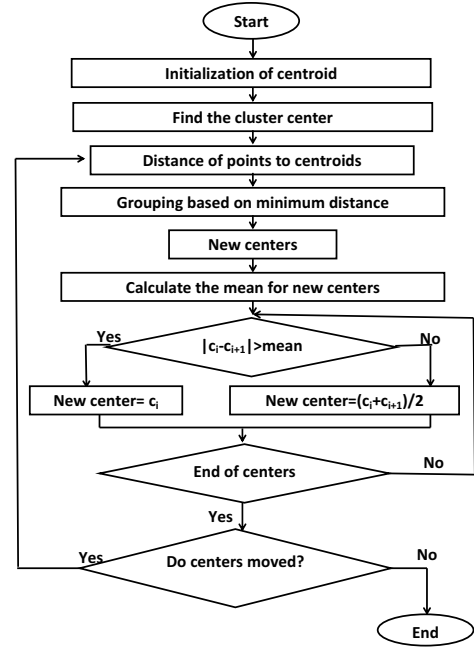6) Count how many elements remain in the means array. These are equal to the number of clusters and their values.

Fig. 2. Steps of Adaptation K-means.

B. Convergence step

Adaptation of the K-means clustering algorithm (AKM) for the convergence step is proposed for fast and reliable target detection in dataset. This proposal method involves less computational complexity and can detect targets in dataset. Therefore, the additional steps to recalculate the new centers in K-means algorithm are added. Those steps are as mentioned below:

1) After re-computing the cluster centers with the new assignment of elements, then calculate the mean for new centers.
2) Calculate the absolute value between the current center and the next center , if the absolute is greater than the mean in the previous step, then keep this center as the new center; otherwise, the new center equals the average of the current center and the next center.
3) Repeat the previous steps until the end of the centers number.

The whole adaptation steps on the standard K-means algorithm are indicated as shown in Fig.2, and the entire steps of the proposal method are:

1) Initialization step, like the steps in section A of the proposal method (steps from 1 until 6), determines the number of cluster for this proposal method.
2) Calculate the membership for each data point xi, belonging to cj, depends on the minimum Euclidean distance as:

$$AKM(X, C) = \sum_{i=1}^{N} min_{j \in (1..k)} ||X_i - C_j||^2 \qquad (8)$$

3) Calculate the new center for each cluster.

4) Run the steps in section B of the proposal method (from steps 1 until 3).

5) Repeat steps 2, 3 and 4 until no centers changed.

Measure the quality of the proposal method (AKM), standard K-means and modify the K-means (MKM) is presented. Each algorithm minimizes a different objective and experiments use the fish photograph, baby image and Lens image as the datasets. By using several images of fish and different initializations method are presented to illustrate the convergence properties of the different algorithm and to show the improvement of the K-means algorithm.

IV. EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted for assessing the proposal method with initialization step and adaptation of convergence condition for the K-means algorithm proposed in this paper. The results are obtained from different images by using the Java program to estimate the number of clusters in K-means algorithm to show that it is dependent on the data. There are different ways to show the number of clusters that are dependent on the frame size and the absolutes values between the means. The processing time that indicates the best way to accelerate the process has also been obtained. Also, the results of the comparison between the proposal methods, standard K-means and to modify the K-means in two different ways: in terms of number of iteration and quality of segmentation are reported.

A. Results of proposal method for initialization step

The target of the result of the experiment is to illustrate the initialization steps of the different methods like random method for the K-means algorithm. Since the K-means algorithm results are dependent on the choice of the initial cluster centers, the modification method is used to estimate the number of cluster and their values in K-means algorithm. The different sizes of the frames and absolute values were used to see the number of clusters and to show the time of process. First frame sizes of F1=300x300, F2=150x150, F3=100x100, F4=50x50, F5=30x30, F6=10x10 or F7=5x5 were used, and were tested with different absolute values V1=100, V2=75, V3=50, V4=25 or V5=10. The results are shown in Table 1 and Fig.3.

TABLE I
THE RESULT BETWEEN THE ABSOLUTE VALUE AND THE
FRAME SIZE TO SEE THE NUMBER OF CLUSTERS

Value	F1	F2	F3	F4	F5	F6	F7
V1	1	1	2	2	2	2	2
V2	1	2	2	3	3	3	3
V3	2	3	4	4	4	5	5
V4	5	7	9	9	9	10	10
V5	9	16	24	25	24	25	25

B. Results of proposal method for the convergence step

In order to compare the results of the proposal method (Adaptation K-means algorithm AKM), standard K-means
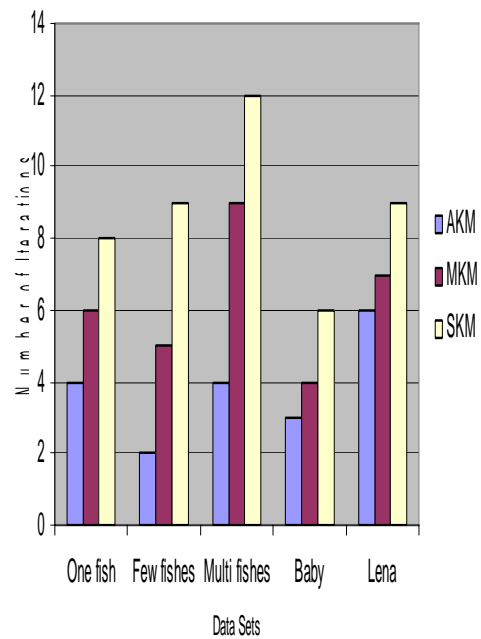

Fig. 3.   Number of Iteration by Random Initialization.

algorithm (SKM) and the modified K-means (MKM), the fish photographs with different numbers of fish in the images, Lena image and baby image were used to show the results related to the number of iterations and to obtain the quality of the solution in the task of image segmentation. The objectives of this experiment are to show the performance of the three algorithms (AKM, SKM and MKM) with the proposal method for initialization steps and with random initialization, and also to determine how many iterations of each algorithm are required to obtain good grouping region of image data. The comparison between the three algorithms with different data sets and with two initialization methods is represented. The use of random initialization on the result to compare between AKM, MKM and SKM is shown in Fig.3. While the proposal method for the initialization step is represented in Fig.4.

V. CONCLUSION

The proposed scheme presents the relationship between the frame sizes with the absolute values to estimate the number of cluster dependent on the values of pixels as indicated by the results in Table 1. Thus, it can be concluded that with a large absolute value, there are small numbers of cluster. Similarly, with large sizes of frame there are small numbers of clusters. However, it should be noted that there are small differences between the results of the frame size in the same absolute value. The number of iteration affects most methods for its computational cost performance for reaching convergence. Therefore, adaptation K-means algorithm (AKM) is proposed to improve the standard K-means algorithm (SKM) using a modification and additional steps for convergence condition. Furthermore, the comparison between proposal method (AKM), recent method (MKM) and standard method (SKM) is
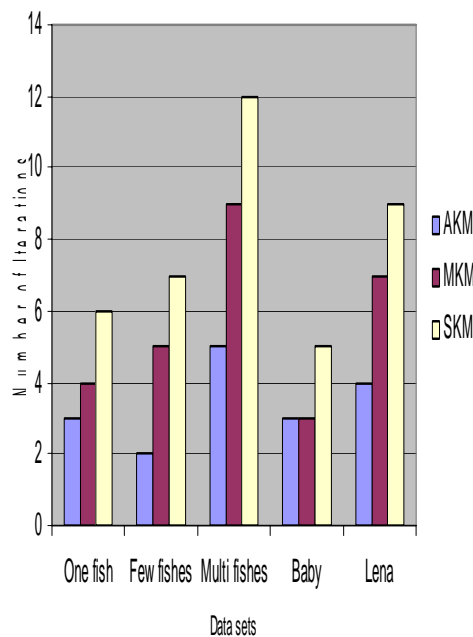
Fig. 4. Number of Iteration by Proposal Method Initialization.

presented. In summary, it can be concluded that the proposed scheme was successful in estimating the number of clusters, decreasing the number of iteration for K-means algorithm, therefore, increasing the speed of the execution process.

REFERENCES

[1] M. G. H. Omran, A. Salman and A. P. Engelbrecht "Dynamic clustering using particle swarm optimization with application in image segmentation", Pattern Anal Applic (2006) 8: 332-344.

[2] B. Jeon, Y. Yung and K. Hong "Image segmentation by unsupervised sparse clustering, " pattern recognition letters 27science direct,(2006) 1650-1664.

[3] Zhang, Y. J. (2002a). "Image engineering and related publications" International Journal of Image and Graphics,(2002a) 2(3), 441-452.

[4] G. B. Coleman, H. C. Andrews (1979) "Image segmentation by clustering", Proc IEEE 67:773-785.

[5] A. K. Jain, M. N. Murty, P. J. Flynn, 1999. "Data clustering: A review", ACM Comput. Surveys 31 (3), 264-323.

[6] C. Carpineto, G. Romano (1996) "A lattice conceptual clustering system and its application to browsing retrieval", Mach Learn 24(2):95-122

[7] Y. J. Zhang, (2006). "A study of image engineering", In M. Khosrow-Pour (Ed.), Encyclopedia of information science and technology (2nd ed.)

[8] D. Ma?yszko, S. T. Wierzcho? "Standard and Genetic K-means Clustering Techniques in Image Segmentation", (CISIM'07) 0-7695-2894-5/07 IEEE 2007.

[9] [9] S. J. Redmond, C. Heneghan, "A method for initialising the K-means clustering algorithm using kd-trees. Science direct", Pattern Recognition Letters 28 (2007) 965-973.

[10] J. B. MacQueen,1967 "Some methods for classification and analysis of multivariate observation", In: Le Cam, L.M., Neyman, J. (Eds.), University of California.

[11] J. Tou, R. Gonzales, 1974. "Pattern Recognition Principles", Addison-Wesley, Reading, MA.

[12] Y. Linde, A. Buzo, R. M. Gray, 1980 "An algorithm for vector quantizer design", IEEE Trans. Commun. 28, 84-95.

[13] [13] P. S. Bradley, U. M. Fayyad, 1998 "Refining initial points for K-means clustering", In: Proc. 15th Internat. Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 91-99. Available from: http://citeseer.ist.psu.edu/bradley98refining.html

[14] C. Kleinn , F. Vilc?ko "A new empirical approach for estimation in k-tree sampling", Science direct. Forest Ecology and Management 237 (2006) 522-533.

[15] C. Huang, R. Harris 1993 "A comparison of several codebook generation approaches", IEEE Trans. Image Process. 2 (1), 108-112.

[16] I. Katsavounidis, C. C. J. Kuo, Z. Zhen, 1994 "A new initialization technique for generalized lloyd iteration", Signal Process. Lett. IEEE 1 (10), 144-146.

[17] G. P. Babu, M. N. Murty, 1993 "A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm", Pattern Recognition Lett. 14 (10), 763-769.

[18] M. B. A. Daoud, S. A. Roberts, 1996. "New methods for the initialization of clusters", Pattern Recognition Lett. 17 (5), 451-45.

[19] B. Thiesson, B. Meck, C. Chickering, D. Heckerman, D., 1997. "Learning mixtures of bayesian networks", Microsoft Technical Report TR-97-30, Redmond, WA.

[20] A. Likas, N. Vlassis, J. J. Verbeek, 2003. "The global K-means clustering algorithm", Pattern Recognition 36, 451-461.

[21] S. S. Khan, A. Ahmad, 2004 "Cluster center initialization algorithm for k means clustering", Pattern Recognition Lett. 25 (11), 1293-1302.

[22] C. M. Bishop, " Neural networks for pattern recognition ", Clarendon Press, Oxford, 1995.

[23] B. Zhang, "Generalized k-harmonic means - Boosting in unsupervised learning", Technical Report HLP-2000-137", Hewlett-Packard Labs, 2000.

[24] J. Prez, R. Pazos , L. Cruz, G. Reyes, R. Basave, and H. Fraire "Improving the fficiency and Efficacy of the K-means Clustering Algorithm Through a New Convergence Condition", Gervasi and M. Gavrilova (Eds.): ICCSA 2007, LNCS 4707, Part III, pp. 674-682. Springer-Verlag Berlin Heidelberg 2007.

[25] W. Li "Modified K-means clustering algorithm", 978-0-7695-3119-9/08 , 2008 IEEE, DOI 10.1109/CISP.2008.349

Ali Salem Bin Samma received his Bachelors degree in Computer Engineering from Amman University, Jordan, in 1997, the M.Sc. degree in Computer Science from Universiti Sains Malaysia, USM, Penang, Malaysia in 2006 and is currently a Ph.D. candidate in the School of Computer Sciences, Universiti Sains Malaysia, USM, Penang, Malaysia. His research interest focuses on the field of Artificial Intelligence and Image Processing.

Rosalina Abdul Salam is an associate professor at the School of Computer Sciences, Universiti Sains Malaysia and a member of Artificial Intelligence Research Group. She received her Bachelors degree in Computer Science in 1992 from Leeds Metropolitan University, United Kingdom. She was a system analyst in Intel Penang, from 1992 to 1995. She returned to the United Kingdom to further her studies. She received her Masters degree in Software Engineering from Sheffield University, United Kingdom in 1997. She completed her PhD in 2001 from Hull University in the area of computer vision. She has published more than 70 papers in journals and conferences. She is a member of the International Computational Intelligence Society and World Informatics Society. Recently she joined the editorial board of the International Journal of Computational Intelligence and the International Journal of Signal Processing. Presently, she is continuing her teaching, graduate supervisions and her research. Her current research area is in the area of artificial intelligence, image processing and bioinformatics applications. The most recent project that she is working is on underwater images and cellular images.