International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 4, Sep 2013, 17-26

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

HYBRID *K*-MEANS CLUSTERING FOR COLOR IMAGE SEGMENTATION

ABBAS H. HASSIN ALASADI[1], HINDROSTOM MOHAMMED[2], EBTESAM N. ALSHEMMARY[3] & MOSLEM MOHSINN KHUDHAIR[4]

[1,4]Department of Computer Science, Science College, Basra University, Basrah, Iraq

[2]Department of Computer Science, Mathematical and Computer Sciences College, University of Kufa, Najaf, Iraq

[3]Information Technology Research and Development Center, University of Kufa, Najaf, Iraq

ABSTRACT

Colour image segmentation is an important problem in computer vision and image processing. Variant application such as image processing, computer vision, pattern recognition and machine learning widely used classical clustering method which is considered traditional k-means algorithm. K-means algorithm is famous clustering algorithm; it divided data into *k* clusters.

The initial centroids are random selected, so the algorithm could not lead to the unique result. In this paper, we proposed a new algorithm for colour image segmentation using hybrid k-means clustering method which combine between two methods, geometric and block method.

Hybrid method is to compute initial centers for k-means clustering. Geometric method depends on equal areas of distribution. Block method segments the image into uniform areas. The proposed method can overcome the drawbacks of both method (geometric and block).

Furthermore, we have presented a simple validity measure based on the *intra-cluster* and *inter-cluster* distance measures which allows determining the number of clusters. The proposed method looks for the first local maximum in the validity measure. The experimental results appeared quite satisfactory.

KEYWORDS: Clustering, *K*-Means Algorithm, Image Segmentation

INTRODUCTION

Segmentation is the process of partitioning an image into disjoint and homogeneous regions. The homogeneous regions, or the edges, are supposed to correspond to actual objects, or parts of them, within the images. Image segmentation is the first step 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 [1].

The process of image segmentation is defined as: “the search of homogenous regions in an image and later the classification of these regions”. It also means the partitioning of an image into meaningful regions based on homogeneity or heterogeneity criteria. Image segmentation techniques can be differentiated into the following basic concepts: Pixel-oriented, Contour-oriented, Region-oriented, Model-oriented, Color-oriented and Hybrid. Most gray level image segmentation techniques can be extended to color images, such as histogram thresholding, clustering, region growing, edge detection, fuzzy approaches and neural networks.

Gray level segmentation methods can be directly applied to each component of the color space, and then the results can be combined in some ways to obtain a final segmentation result (see Figure (1)).

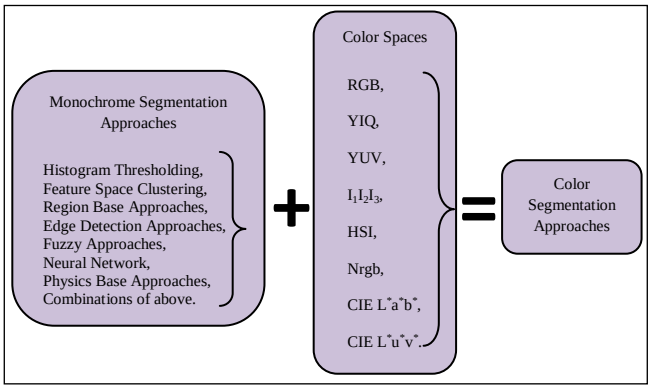

Figure 1: Commonly Used Color Image Segmentation Approaches

However, there are two critical issues for color image segmentation

- What segmentation method should be utilized?
- What color space should be adopted [2].

Color segmentation of image is a crucial operation in image analysis and in many computer vision, image interpretation, and pattern recognition system, with applications in scientific and industrial field(s) such as medicine, Remote Sensing, Microscopy, content based image and video retrieval, document analysis, industrial automation and quality control. The performance of color segmentation may significantly affect the quality of an image understanding system. The most common features used in image segmentation include texture, shape, grey level intensity, and color [2,3 and 4].

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.

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 [5] 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 and, Michael [6] 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.

Haralick and Shapiro [7] suggested that there is no full theory of clustering, and, therefore, no full theory of image segmentation. They have established the following qualitative guideline for good image segmentation:

- Segmented regions should be unified and homogeneous with respect to some characteristic such as gray level or texture.

- Region interiors should be simple and without many small holes.
- Adjacent segmented regions should have significantly different values with respect to the characteristic on which they are considered unified.
- Boundaries of each segment should be simple, not ragged, and must be spatially accurate.

This means that image segmentation techniques are generally ad hoc and differ on how they emphasize one or more of the desired properties. In the end, each method tries to balance one property against another. Therefore, the final implementation of each image segmentation algorithm depends very much on the end of the application. However, these differences usually center on the choices of parameters or methods of how to adapt certain parameters to the image.

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. [3, 8]

There are two main branches of clustering: (1) hierarchical and (2) partitional [9]. 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 [10]. 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 [11, 12].

K-MEANS CLUSRTING ALGORITHM

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 [13]. 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 :

$$C_i = [R_i, G_i, B_i], i = 1,2,\cdots,k \tag{1}$$

During the clustering process, each pixel x is allocated to cluster K_j with the closest cluster center using a predefined metric (e.g., the Euclidean metric, the city-block metric, the Mahalanobis metric, etc.). For pixel x, the condition of membership to the cluster K_j during the n th iteration can be formulated as follows:

$$x \in K_j(n) \Leftrightarrow \forall i = 1,2,..., j-1, j+1,..., k...$$
$$\|x - C_j(n)\| < \|x - C_i(n)\| \quad (2)$$

Where 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:

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

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:

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

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

$$C_{j_B}(n+1) = \frac{1}{N_j(n)} \sum_{x \in K_j(n)} x_B \quad (4)$$

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

In the next step, check the difference between new and old positions of the centers. If the difference is larger than a threshold ***T***, 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.

The operation of *K*-means algorithm is illustrated in Figure (2) which shows: (a) two dimensional input data with three seed points selected as cluster centers and initial assignment of data points to clusters; (b) to (e) intermediate iterations updating cluster labels and their centers; (f) final clustering obtained by *K*-means algorithm at convergence [15].

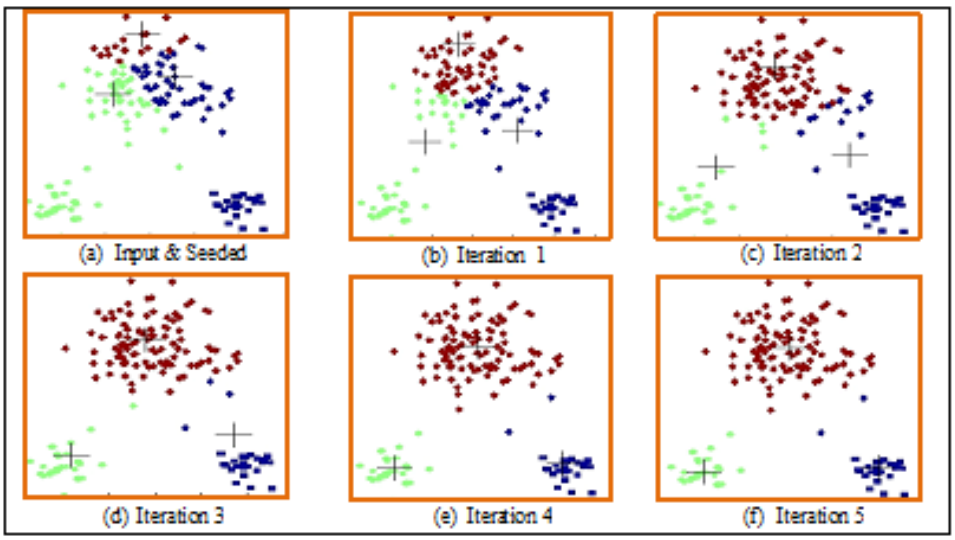

Figure 2: Shows an Illustration of *K*-Means Algorithm on a 2-Dimensional Dataset with Three Clusters [15]

PROPOSED METHOD

We proposed an efficient algorithm which consists of three methods to compute initial centers for *k*-means clustering. First one is called geometric method which depends on equal areas of distribution. The second is called block method which segments the image into uniform areas. The last method called hybrid which combined between first and second methods.

It 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 algorithm 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 algorithm scans the original dataset two times and produces better clusters.

The main idea of the proposed algorithm is to compress the dataset into finite number of representative. Each representative is the mean value of some data points form a small cluster. 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 (see, Figure 3). The idea of compression of dataset comes from the BIRCH algorithm [16]. Figure (4) exhibits the pseudo code of proposed method. Note that, in Figure (4), 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 full dataset.

From experimental results it will be better to use a small value for *k* at the first phase of our methods, 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.

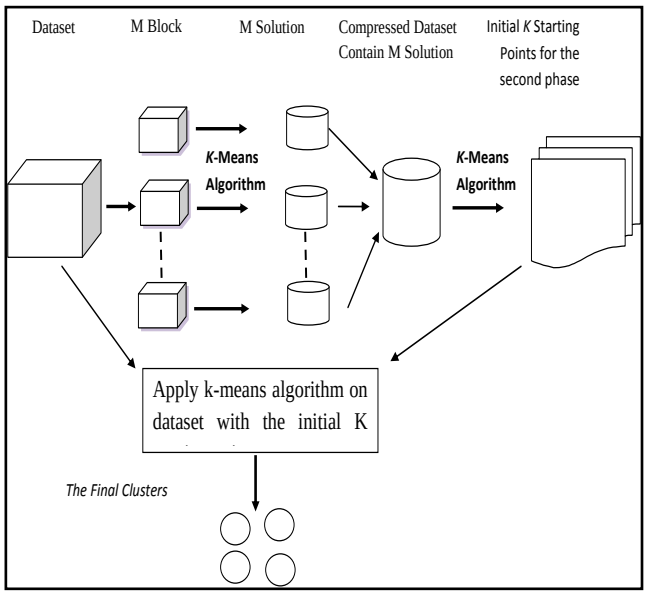

Figure 3: An Overview of the *k*-Means Block

```
1.  Set the size of the block
2.   i=0
3.     while not end of file
4.       read the Block,
5.       k-means (Block, 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 4: Pseudo Code of Proposed Method

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 algorithm 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 methods 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 Pentium Duel-Core CPU 2.0GHz, 2.0GB RAM, 512 KB Cache.

Table 1: Experimental Results of Proposed Method

Original Images	Segment Result	Smooth & Lap.	*K*
			5
			12
			15
			24

Figure (5) demonstrates that the proposed method provide better cluster accuracy than the standard *K*-means. It shows that the proposed method performs much better than the randomly initialized algorithm. This is due to the initial cluster centers generated by the proposed methods which are quite closed to the optimum solution.

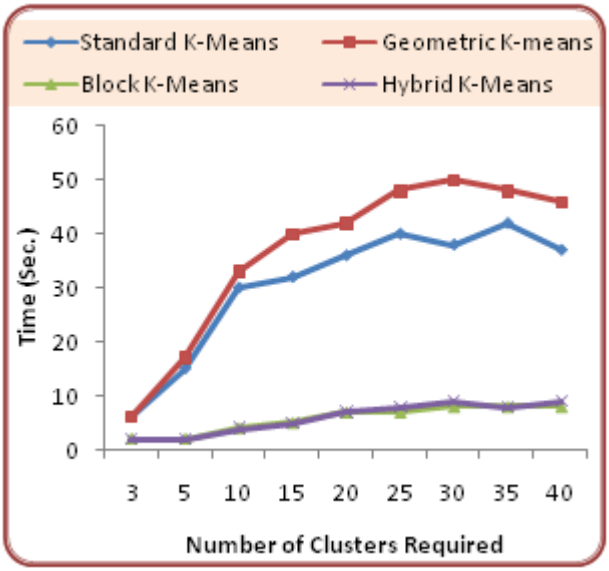

Figure 5: Execution Time (Lena Image-Size 64*64)

Figure (6) demonstrates that the proposed method (hybrid) provide better cluster accuracy than the standard *K*-means. It shows that the proposed method performs much better than the randomly initialized algorithm. This is due to the initial cluster centers generated by the proposed methods which are quite closed to the optimum solution. Figure (7) demonstrates that the proposed method (hybrid method) has high efficiency and is characterized as same as block method but the execution time is less compared with it.

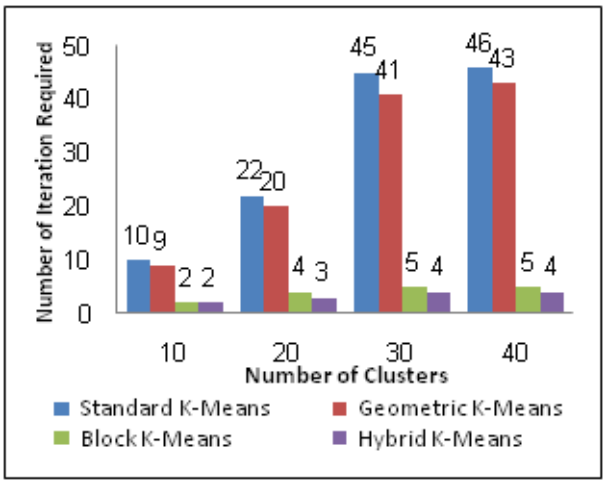

Figure 6: Number of Iterations of All Methods (Lena Image)

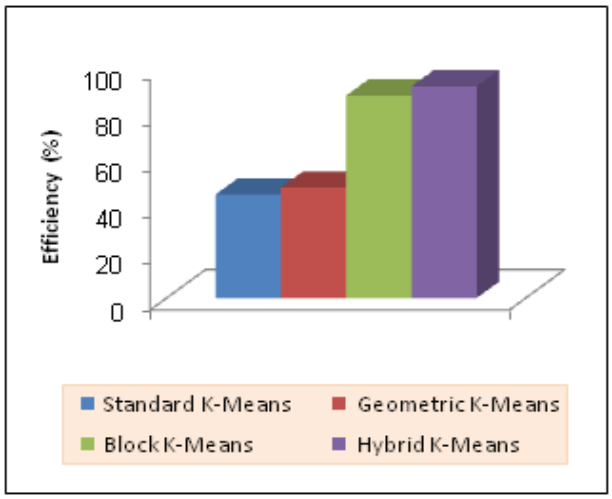

Figure 7: Efficiency of Four Methods (Lena Image)

CONCLUSIONS

K-means algorithm is a popular clustering algorithm which widely applied, but do not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids.

This paper presents a new algorithm for computing initial centers for *k*-means clustering. This algorithm ensures the entire process of clustering in *O(n2)* time without sacrificing the accuracy of clusters. The previous improvements of the *k*-means algorithm compromise on either accuracy or efficiency.

A limitation of the proposed algorithm 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. In both block and hybrid methods, it can be noted that perfect number of clusters range (*3-9*). Also the amount of noise increases with the increase of this range.

Therefore, it can be recommended using small value for *k* when we apply the *K*-means on the blocks. Another important matter is the size of block. It has been found that sizes (*8* x *8*) and (*16* x *16*) are more suitable in applications. Also the proposed methods perform well when the dataset contains large number of clusters.

All accuracy of result depends on quality of original images and their size. In other words, this matter has more effect in aspects of problem noise. Hence, it can achieve superior result with images that have equal size or more than (*256*

x *256*) and not subject to compression or resize processes. It can be notice that the performance of our proposed methods (block and hybrid) are better than standard *K*-means algorithm which we can see in Figure (5&6).

Really, many researchers have agreed that evaluating the number of clusters is a difficult problem in the meantime. This is one of the important drawbacks of K-means clustering. Nevertheless, it have been investigated the clustering validity depending on intra and inter clustering through experiment execution. There are many possible potential avenues for further research that can be arisen from this work. Investigating the work of *K*-means algorithm with a different type of similarity measurement or extension of the combined Euclidean distance and other similarity measure. The extension of color clustering to a Euclidean distance and vector angle hybrid as well as optimization for various color spaces can be a good topic for research. Methods for refining the computation of initial centroids are worth investigating.

REFERENCES

1. Ali Salem Bin Samma and Rosalina Abdul Salam. ***"Adaptation of K-Means Algorithm for Image Segmentation"***, International Journal of Signal Processing, Vol. 5, No. 4, p.p 270-274, 2009.
2. Anil Z. Chitade, Dr. S.K. Katiyar, ***"Colour Based Image Segmentation Using K-Means Clustering"***, International Journal of Engineering Science and Technology, Vol. 2, No. 10, p.p 5319-5325, 2010.
3. Siddheswar Ray and Rose H. Turi, ***"Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation",*** School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia.
4. Anderberg, M.R., ***"Cluster Analysis for Applications"***, Academic Press Inc., 1973.
5. Stephen J. Redmond, Conor Heneghan, ***"A Method for Initializing the K-Means Clustering Algorithm Using KD-Trees"***, Department of Electronic Engineering, University College Dublin, Bel_eld, Dublin 4, Ireland.
6. Douglas Steinley, Michael J. Brusco, ***"Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques"***, Journal of Classification DOI: 10.1007/s00357-007-0003-0, 24:99-121 (2007).
7. Robert M. Haralick and Linda G. Shapiro, ***"Computer and Robot Vision"***, Vol. 1, Addison-Wesley, MA, 1992.
8. Fisher, D., ***"Knowledge Acquisition Via Incremental Conceptual Clustering"***, Mach. Learn. 2, 139–172, 1987.
9. Jain, A.K., Murty, M.N., Flynn, P.J., ***"Data clustering: A review"***, ACM Comput. Surveys, Vol. 31, No. 3, p.p 264–323, 1999.
10. Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, Milu Acharya, ***"A Hybridized K-Means Clustering Approach for High Dimensional Dataset"***, International Journal of Engineering, Science and Technology, Vol. 2, No. 2, p.p 59-66, 2010.
11. Xiaoping Qing, Shijue Zheng, ***"A New Method for Initializing the K- Means Clustering Algorithm",*** Second International Symposium on Knowledge Acquisition and Modeling, , p.p 41-44, 2009.
12. 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, Vol. 19, No. 2, p.p 47-57.
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, pp. 281–297, University of California, Berkeley, CA, USA, 1967.

15. Anil K. Jain, ***" Data Clustering: 50 Years Beyond K-Means"***, This paper is based on the King-Sun Fu Prize lecture delivered at the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, December 8, 2008.

16. 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, pp.103- 114, New York, 1996.