ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

Performance Analysis of Improved K-Means & K-Means in Cluster Generation

K.Shanthi[1], Dr.Sivabalakrishnan.M[2]

Assistant Professor, Department of ECE, MNM Jain Engineering College, Chennai, India[1]

Associate Professor, School of Computing Science and Engineering, VIT University, Chennai, India[2]

ABSTRACT: K-means is the well known and most familiar algorithm among the other partition based clustering algorithms. It typically shows spectacular results even in significantly massive information sets of Image segmentation supported associate adaptive K-means clustering algorithm is conferred. The proposed method tries to develop K-means algorithm to get high performance and potency. This technique proposes data formatting step in K-means algorithmic rule. additionally, it solves a model choice variety by deciding the quantity of clusters victimization datasets from image by frame size and also the definite quantity between the means that, and extra steps for convergence step in K-means algorithm are supplementary. Moreover, so as to judge the performance of the proposed technique, the results of the proposed technique, customary K-means and recently changed K-means are compared. The experimental results showed that the proposed technique provides higher output.

KEYWORDS: INITIALIZATION, IM K-MEANS, SEGMENTATION, STANDARD K-MEANS.

I. INTRODUCTION

A general framework for image techniques are club under image engineering, image processing (low layer), image analysis (middle layer) and image understanding (high layer) are the three layers of image processing. In image processing, first stage is Image segmentation and also image analysis has serious task in that process. Next stage is differentiating object from their background or else matching with their pattern. These are all only a basic step for object recognition, it is also considered to be serious trendy issues in computer vision.

Image segmentation has various ways to solve; some of the methods can be reviewed. Byoung [1] on his survey based on Bayesian framework, neural networking and categorized techniques as follows, thresholding approaches, contour based approaches, region based approaches, clustering based approaches and other optimization based approaches. Partitional and hierarchical clustering algorithms are two general group of clustering approaches. (for details, please refer to [2]). data mining [3], compression, [4] image segmentation [4], [5] and machine learning [6] these are the applications of K-means and EM clustering of type partitional clustering method. Therefore, clustering algorithms is that the classification is simple and easy to implement are the advantages. Likely, the determination of the number of clusters and reduce the numbers of iteration are the drawbacks, [7].

This paper is organized as follows: The related survey are reviewed and in brief express the family of K-means clustering algorithms are explained in section 2. In section 3 IM K-means algorithm is presented. In Section 4, the comparative results K-means and IM K-means algorithms are analysis with help of image segmentation. Finally, in section 5, the conclusion and future work are presented

II. RELATED WORK

The family of axis pedestal clustering algorithm is k-means and its improvement algorithms. This family consists of several methods: expectation maximization, fuzzy K-means and harmonic K-means are members of the family. Template matching[8], Mean-shift[9], Particle filter[10] and K-means tracker[11] are delegates of object tracking algorithms. The conventional method is template matching and often used for a towering accuracy for spots of mechanical and electrical side. Even though efficiency is increased but not shows sufficient performance in digital camera application.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

On other way, in digital camera following methods have high performance. The methods are Mean-shift, Particle filter and K-means tracker. However, An Object changes appearance and scale are not sufficient in Mean-Shift. For particle filter have high computation cost, it is difficult to apply in digital camera. We selected K-means tracker as the object tracking algorithm because the algorithm is applicable for wide variation of scale and appearance change with reasonable computation cost.

III. PARTITION BASED CLUSTERING METHODS

In one step Partition based clustering methods can create the clusters. From that one cluster different clusters are created internally with help of partition based methods. From that different clusters one set of cluster is allotted for output remaining clusters are allotted for input. For example consider N object dataset Given as input to partition based [12] clustering algorithm then it construct the data of k partition. From this process objective function is optimized. From this analysis any proposed solution has high performance with aid of these clustering methods. This live of quality may be average distance between clusters or another metric. One common determine of such kind is that the squired error metric, that measures the squired distance from every position to the centroid for the associated cluster. Partition based mostly clustering algorithms attempt to regionally improve a precise criterion. the bulk of them may be thought-about as greedy algorithms, i.e., algorithms that at every step select the most effective resolution and should not result in best ends up in the tip. The most effective resolution at every step is that the placement of a precise object within the cluster that the representative purpose is nearest to the thing. This family of bunch algorithms includes the primary ones that appeared within the data processing Community. The foremost usually used are K means [JD88,KR90][13], PAM (Partitioning Around Medoids) [KR90], CLARA (Clustering LARge Applications) [KR90] and CLARANS (Clustering LARge ApplicatioNS) [NH94]. All of them are applicable to data sets with numerical attributes.

3.1. STANDARD K-MEANS CLUSTERING ALGORITHM

K-means could be a fashionable algorithm for clustering; it partitions information set into k sets. The membership for every datum belongs to its nearest center, betting on the minimum distance. This membership determines as, [14]: There are many ways to boost the quality K-means algorithm associated with many aspects. standard K-Means formula consists of 4 steps: format, classification, machine and convergence condition.

Basically, the format step has received the foremost attention compared to the further steps. Stephen [15], indicates that the most basic references to formatting the K means that algorithm was by Forgy in 1965 UN agency selected points willy-nilly. as an example, it may be purpose of close to a cluster centre or far point. MacQueen, [16], introduced what's such as an internet learning strategy to see a collection of cluster seeds. Tou and Gonzales[17], urged the easy Cluster Seeking (SCS) technique. Linde et al. [18], proposed a Binary splitting (BS) technique that was supposed to be used within the style of Vector Quantizes codebooks. Kaufman and Rousseeuw [19], urged choosing the primary seed because the most centrally situated instance. babu and Murty [20], printed a technique of close to best seed choice victimization genetic programming. However, the matter with genetic algorithms is that the results vary considerably with the selection of population size, and crossover and mutation possibilities [21].

Huang and Harris [22], the Direct Search Binary splitting (DSBS) technique was proposed. This technique is comparable to the Binary splitting method higher than except that the dividing step is increased through the utilization of Principle component Analysis (PCA). Katsavounidis et al. [23], proposed, what has been termed by some because the KKZ method. This method starts by selecting a degree x, ideally one on the 'edge' of the information, because the initial seed. the purpose that is furthest from x is chosen because the second seed. Daoud and Roberts [24], to divide the total input domain into 2 disjoint volumes. In every topological space, it's assumed that the points are arbitrary distributed which the seeds are placed on a daily grid. Thiesson et al. [25], urged taking the mean of the complete dataset and arbitrary heavy it K times to provide the K seeds. Bradley and Fayyad [19], conferred a way that begins by arbitrary breaking the information into ten, or so, subsets. Then it performs a K means that clustering on every of the ten subsets, all beginning at constant set of initial seeds, that are measure chosen victimization Forgy's method. Likas et al. [26], gift a worldwide K means that method that aims to step by step increase the quantity of seeds till K is found.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

Khan and Ahmad [27], delineate a Cluster Centre Initialization Method (CCIA) employing a Density-based Multi Scale data Condensation (DBMSDC) that was introduced.

DBMSDC involves estimating the density of the information at a degree and then sorting the points in step with their density. From the purpose. Then it moves on to following purpose that has not been cropped from the list and also the method is perennial till a desired variety of points stay. The authors select their seeds by examining every of the m attributes singly to extract an inventory of K0 > K doable seed locations. Next the DBMSDC formula is invoked and points that square measure approximate square measure unified till there are measure solely K points remaining.

The K-means formula works as follows:

a. Choose initial centres of the K clusters. Repeat steps b through c till the cluster membership stabilizes.
b. Generate a replacement partition by distribution every information to its nearest cluster centres.
c. Calculate new cluster centres because the centroids of the clusters.

3.2. ALGORITHM FOR GETTING INITIAL CENTROIDS

Now let's review the quality k-means formula.
Input: the quantity of categories and also the population U that.
Output: k categories that satisfy the smallest amount square error.

The process of the formula is contains n objects.
(1) Choose k objects arbitrary from the population U as .the initial centroids.
(2) Repeat (3) and (4) till no object changes the category t belongs to.
(3) Calculate the distances between every object & and every one centroids, and if one object has the shortest distance from one centroids with regkd to the opposite centroids then it's constant name because the centroid; all of those objects that have constant name belong to constant category.
(4) Average all the vectors of objects happiness to constant category and type the new centroids. the quality k-means formula altimeters between distribution the data-points to their nearest centroid (the E-step) and moving every centroid to the mean of its allotted data-points (the M-step).

Because the quality k-means method gets simply unfree during a native minimum and completely different initial centroids result in different results, if we discover sure initial centroids that are according to the distribution of knowledge, then a stronger cluster may be obtained. The aim of k-means method is to partition objects into many categories and to form the distances between objects within the same category nearer than the distances between objects in several categories. therefore if certain centroids within which every centroid represents a cluster of comparable objects may be obtained, we'll determine the centroids according to the distribution of knowledge. Let U be a data-point set. The initial centroids may be gotten by the subsequent steps. first of all calculate the distances between every data-point and every one of the opposite data-points in U. second determine the 2 data-points between that the space is that the shorkst and type a data-point set AI that contains these two data-points, then we tend to delete them from the population U. Thirdly. calculate distances between every data-point in AI and every data-point in U, determine the data-point that's nearest to the data-point set AI (i.e. of all distances, thy distance between this data-point and sure data-point in A1 is shortest), delete it from U and add it to AI. Repeat the third step until the quantity of data-point in A1 reaches sure threshold. Then we tend to attend step 2 and type another data-point set until we tend to get k data-point sets. Finally the initial centroids may be gotten by averaging all the vectors in every data-point set.

3.2.1. THE ALGORITHM OF GETTING THE INITIAL CENTROIDS

1.The distance is employed during this paper. the space between one vector X=(xl,x2,. . .,xn) and also the alternative vector Y=(yl,y2, ...,y n) is delineate as follows.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

The distance between a data-point X and a data-point d(X, V)=min(d(X,Y), Y EV) Suppose there are measure of n data-points within the population U and that we need to partition U into k categories. Set m=l. Then the formula is delineate as follows. calculate distances between every information-point and every one of the opposite data-points in U; realize the 2 data points between that the space is that the shortest and type a data-point set Am (l<m<k) that contains these 2 data-points; delete these two data-points from U.
(2) realize the data-point in U that's nearest to the data-point set Am, add it to Am and delete it from U.
(3) Repeat step (2) until the quantity of data-points in Am reacbes :$a \times n/k(0 < a \leq 1)$
(4) If m<k, then m=m+l; realize another try of data-points between that the space is that the shortest in U and type another data-point set Am and delete them from U then attend step (2)
5) For each Am (I<m<k) sum the vectors of data-points and divide the sum by the number of data-points in *Am,* then each data-point set outputs a vector and we select these vectors as the initial centroids.
(6) The method of the quality k-means formula from step 2.
The value of a is completely different with relation to different information. If the worth of a is just too tiny, all the centroids is also obtained within the same region that contains several similar data-points; however if the worth of a is just too massive, the centroids might stray removed from the region that contains several similar data-points. in step with the results of our experiment, better cluster will typically be obtained if the worth of a is ready to be 0.75.

IV. IMPROVED K MEANS

Original K-means formula select k points as initial cluster centers, completely different points might get dissimilar solutions. so as to diminish the sensitivity of initial purpose alternative, we tend to use a mediod [11], that is that the most centrally situated object during a cluster, to get higher initial centers. The demand of random sampling is of course bias the sample to just about represent the first dataset, that's to mention, samples drawn from dataset can't cause distortion and might mirror original data's distribution. Scrutiny two solutions generated by cluster sample drawn from the first dataset and itself victimization K-means severally, the placement of cluster centroids of those two are nearly similar. So, the sample-based technique is applicable to refine initial conditions. so as to minimize the influence of sample on selecting initial beginning points, following procedures are utilized. First, drawing multiple sub-samples (say J) from original dataset (the size of every sub-sample isn't over the potential of the memory, and also the total for the scale of J sub-samples is as shut as doable to the scale of original dataset) . Second, use K-means for every sub-sample and manufacturing a bunch of mediods severally. Finally, scrutiny J solutions and selecting one cluster having lowest worth of square-error perform because the refined initial points. To avoid dividing one massive cluster into 2 or a lot of ones for adopting square-error criterion, we tend to assume the quantity of bunch is K' (K > K, K' depends on the balance of bunch quality and time). In general, larger K' will expand looking out space of resolution area, and cut back things that there don't seem to be any initial worth close to some extremum. later on, re-clustering the dataset through K-means with the chosen initial conditions would manufacture K' mediods, then merging K' clusters (which are nearest clusters) till the quantity of clusters reduced to k.

V. PROPOSED APPROACH

Original K-means algorithm select k points as primary cluster centers, completely different points might get different solutions. So as to diminish the sensitivity of initial purpose alternative, we tend to use a mediod [I1], that is that the most centrally situated object during a cluster, to get higher initial centers. The demand of random sampling is of course bias the sample to just about represent the first dataset, that's to mention, samples drawn from dataset can't cause distortion and might mirror original data's distribution so as to reduce the influence of sample on selecting initial beginning points, following procedures are utilized. First, drawing multiple sub samples (say J) from original dataset (the size of every sub-sample isn't over the potential of the memory, and also the total for the scale of J sub-samples is as shut as doable to the scale of original dataset) . Second, use K means for every sub-sample and manufacturing a bunch of mediods severally. Finally, scrutiny J solutions and selecting one cluster having lowest worth of square-error perform because the refined initial points.

To avoid dividing one massive cluster into 2 or a lot of ones for adopting square-error criterion, we tend to assume the quantity of cluster is K' (K > K, K' depends on the balance of cluster quality and time). In general, larger K' will expand looking out space of resolution area, and cut back things that there don't seem to be any initial worth close to

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

some extremum. later on, re-clustering the dataset through K-means with the chosen initial conditions would manufacture K' mediods, then merging K' clusters (which square measure nearest clusters) till the quantity of clusters reduced to k.

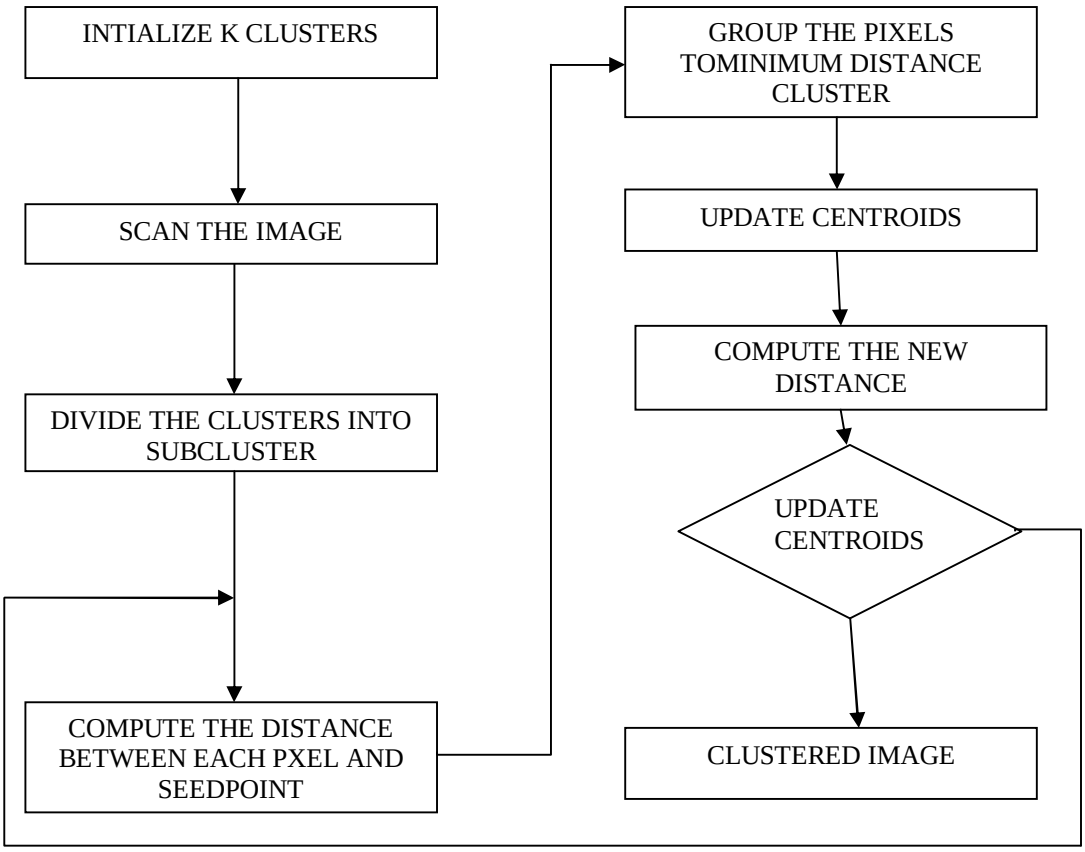


Fig 1:proposed block diagram

VI. EXPERIMENTAL RESULTS

In This part presents the results of the comparison conducted for reviewing the proposal method. The goal of this comparison is to show the performance of the proposal method implemented in MATLAB .The IM is just the extension of K-means to provide the many clusters to be generated by the K means algorithm. It also provides the initial set of means to K-means. Therefore it has been decided to make a comparative analysis of the clustering quality of IMK-means with conventional K-means. The main difference between the two algorithms is that in case of IM-K-means it is not necessary to provide the many cluster to be generated in earlier and for K-means, users have to provide the number of clusters to be generated.

To test the algorithms thoroughly, separate programs were developed for IMK-means and conventional K-means.

Fig:2, Input image for compare K-means and IM K-means

10.15662/ijareeie.2014.0309049

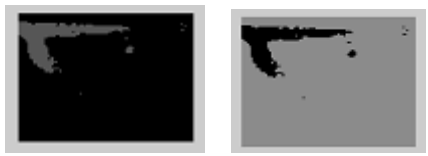
ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

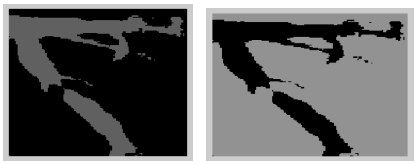
For compare K-means and IM K-mean the input image has been shown in the fig 2. It can be compare with different cluster numbers in these algorithms. The results are shown in following diagram.



(a) **(b)**

Fig: 3, output of K-means at cluster value 4 in (a) and 5 in (b)

In fig 3 shows the way the cluster number increase the input image modifies according to the number of cluster. If cluster number increase image detection rate also increased in K-means algorithm but false detection is high in this algorithm.



(a) **(b)**

Fig: 4, Output of IM K-means at cluster value 4 in (a) and 5 in (b)

To reduce the false positive of k-means, proposed method is developed that is IM K-means. It can be show in the fig 4. In this method also follow the property of K-means but false positive is much more reduced that is get from the output of IM K-means.

VII. CONCLUSION

The most acceptable and extensively used cluster method is K-Means. Even though acceptable it has lot of drawbacks, like algorithm choose randomly k object from the image this leads no consistent result for different execution of same input. Here comparative result shows that cluster number can increase lead to better result in the output image. Comparative results show that our algorithm can produce best clustering by choosing centroids. The limitations of K-means algorithm are computing resource, time and huge dataset. These are all overcome by the proposed method. The improved K-means algorithm presented in this paper is a solution to handle large scale data, which can select initial clustering center purposefully, cut back the sensitivity to isolated point, avoid dissevering massive cluster, and overcome deflexion of data in some degree that caused by the disparity in data partitioning as a result of adoption of multi-sampling.

REFERENCES

1. B. Jeon, Y. Yung and K. Hong “Image segmentation by unsupervised sparse clustering, ” pattern recognition letters 27science direct,(2006) 1650-1664
2. 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.
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. YS. Chen, YP. Hung, CS. Fuh, “Fast block matching algorithm based on the winner-update strategy”, *IEEE Transactions on Image Processing,2001*:pp.1212-1222.
9. B. Georgescu, I. Shimshoni, P. Meer, “Mean shift based clustering in high dimensions: a texture classification example”, *Ninth IEEE International conference on Computer Vision, 2003*:pp.456-463.
10. MS. Arulampalam, S. Maskell, N. Gordon, t. Clapp, “A tutorial on particle filteres for online nonlinear/non-Gaussian Bayesian tracking”,*IEEE Transactions on signal processing, 2002*:pp.174-188.
11. C.Hua, H.Wu, Q.Chen, T.Wada, “Object Tracking with Target and Background Sample”, *IEICE Transactions on Information and Systems,2007 E90-D(4)*:pp.766-774.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

12. Data Clustering and Its Applications, Raza Ali, Us-man Ghani, Aasim Saeed.
13. MacQueen, J. Some methods for classification and analysis of multivariate observations. Proc: 5th Berkeley Symp. Math. Statist, Prob, 1:218-297, 1967.
14. D. Ma³yszko, S. T. Wierzchoñ "Standard and Genetic K-means Clustering Techniques in Image Segmentation", (CISIM'07) 0-7695-2894-5/07 IEEE 2007
15. S. J. Redmond, C. Heneghan, "A method for initializing the K-means clustering algorithm using kd-trees. Science direct", Pattern Recognition Letters 28 (2007) 965–973
16. J. B. MacQueen,1967 "Some methods for classification and analysis of multivariate observation", In: Le Cam, L.M., Neyman, J. (Eds.),University of California.
17. J. Tou, R. Gonzales, 1974. "Pattern Recognition Principles",Addison-Wesley, Reading, MA.
18. Y. Linde, A. Buzo, R. M. Gray, 1980 "An algorithm for vector quantizer design", IEEE Trans. Commun. 28, 84–95.
19. 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
20. 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.
21. A. K. Jain, M. N. Murty, P. J. Flynn, 1999. "Data clustering: A review",ACM Comput. Surveys 31 (3), 264–323.
22. C. Huang, R. Harris 1993 "A comparison of several codebook generation approaches", IEEE Trans. Image Process. 2 (1), 108–112.
23. 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.
24. M. B. A. Daoud, S. A. Roberts, 1996. "New methods for the initialization of clusters", Pattern Recognition Lett. 17 (5), 451–45.
25. B. Thiesson, B. Meck, C. Chickering, D. Heckerman, D., 1997."Learning mixtures of bayesian networks", Microsoft Technical Report TR-97-30, Redmond, WA.
26. A. Likas, N. Vlassis, J. J. Verbeek, 2003. "The global K-means clustering algorithm", Pattern Recognition 36, 451–461.
27. S. S. Khan, A. Ahmad, 2004 "Cluster center initialization algorithm for k means clustering", Pattern Recognition Lett. 25 (11), 1293–1302.