RESEARCH ARTICLE OPEN ACCESS

K-MED: An Algorithm of Uniform Clustering for Efficient Content-Based Image Retrieval

Rashmi Chauhan [1], Shashank Kumar Som [2], Ankush Mittal [3]
Department of Computer Science and Engineering [1] & [3]
Graphic Era University, Dehradun
Department of Computer Science and Engineering [2]
College of Engineering Roorkee
U.K - India

ABSTRACT

Currently, search and retrieval of the relevant images efficiently from the huge amount of data is a challenging task for the researchers. Clustering is much useful in image segmentation to segregate a digital image into discrete regions that can be used to perform content-based image retrieval. In this paper, a new clustering algorithm named as K-MED clustering is proposed and implemented for performing uniform image clustering. It is compared with k-means clustering algorithm in some aspects and observed that the time complexity is reduced using this algorithm. The experimental analysis of K-MED clustering shows that some cases for the problem of local minima that arise in k-means clustering seem to be resolve.

Keywords :- Content-based Image Retrieval; Image Segmentation; Clustering; K-MED; Image Analysis.

I. INTRODUCTION

The content-based image retrieval (CBIR) is the challenging research area in the field of image processing. Mostly the existing image search systems such as Google Images and Yahoo! Image search are based on textual annotation of images [1]. There is a requirement of an appropriate technique to search and retrieve the images from the large collection of the data. Indexing and retrieval of image data can be done statically using manual text annotation [2]. The text is submitted as query to search the images because of most of the images being annotated with several tags. It is not easy to represent the images with a small number of keywords. Also, annotating the images manually is very biased, indefinite and imperfect. CBIR is a technique to resolve these issues. CBIR is the process of browsing, searching and redirection of images from a huge image databases according to their visual contents [3]. Generally, low-level image features are used in CBIR systems such as color, texture, shape, edge, to perform indexing and retrieval of the images due to easy and automatic computation of low-level features [4]. Currently, several CBIR systems exist for retrieving the image. Some of these systems are discussed here. QBIC (Query By Image Content) developed by IBM [5-6]. VIR Image Engine developed by Virage Inc., in which images are retrieved on the basis of primitive attributes such as colour, texture and structure. VisualSEEK and WebSEEK developed by the Department of Electrical Engineering, Columbia University. In these systems, colour and spatial location matching as well as texture matching are supported.

NeTra developed by the Department of Electrical and Computer Engineering, University of California which supports colour, shape, spatial layout and texture matching as well as image segmentation. MARS (Multimedia Analysis and Retrieval System) developed by University of Illinois which supports colour, spatial layout, texture and shape matching [6]. Singh et al. [7] also described various techniques for CBIR such as text-based image retrieval and content-based image retrieval. Semantic image retrieval is also discussed on the basis of the content and required the semantic interpretation of a particular image. A CBIR model based on Wavelet Transform is studied by Giveki et al. [8] and Gonde et al. [9]. Verma and Balasubramanian [10] also proposed a new technique of image retrieval for texture, face and medical images in which the symmetric local binary pattern is extracted from the actual image to obtain the local information.

Texture-based image retrieval is performed by Yadav et al. [11]. They proposed a model of image retrieval using DC coefficients for compressive sensing in medical domain and used compressive sampling to perform the retrieval. Mittal and Cheong [12] focused various issues in extraction of syntactic features and semantic-level features for image retrieval. They designed a framework for synthesizing semantic-level indices to perform content-based image retrieval. In that framework, the synthesis of its large set of

elemental features is done to construct a high-level index. A few principal features are used for characterization and mapping of images or videos. Identification of the appropriate features is done through the medium of Bayesian Network. The experiments performed show that the framework has no need of an expert knowledge base and it enables a stronger coupling between the feature extraction and meaningful high-level indices. The ideas of CBIR and image extraction are used in a hybrid manner [13]. A different approach of image clustering is presented so that the speed of the image retrieval system can be enhanced.

The detailed survey is performed related to clustering techniques and their application in image search and retrieval. Mishra et al. [14] have been described various techniques of clustering and also discussed the steps to be followed in performing CBIR. Clustering is the technique of grouping a set of objects such that the objects in a particular group i.e. a cluster, are related to each other more than to those in other clusters [15]. Clustering is useful in many applications such as data mining and statistical data analysis. It is used in various areas, including machine learning, pattern recognition, image analysis, information retrieval, image retrieval, information indexing and bioinformatics [16]. The literature survey is performed related to image retrieval systems, image processing and image annotation, clustering and clustering based image retrieval. Several issues have been identified by the researchers. So, there is a requirement of a new effective technique for retrieving relevant images. Thus we are motivated to propose a new technique and algorithm to perform clustering and image retrieval based on the algorithm.

II. METHOD OF CLUSTERING

All K-MED clustering method is proposed in the present study. This clustering is a technique of categorizing the objects into K sets. The categorization is to be performed by minimizing the sum of Euclidean distances between the points and their corresponding centroid. Euclidean distance is the length of the connecting line segment between two points. Segmentation of image is represented in Fig. 1. In Cartesian coordinates, if $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ are two points in Euclidean n-space, the Euclidean distance (d) from p to q or q to p is given by the Pythagorean formula (equation 1).

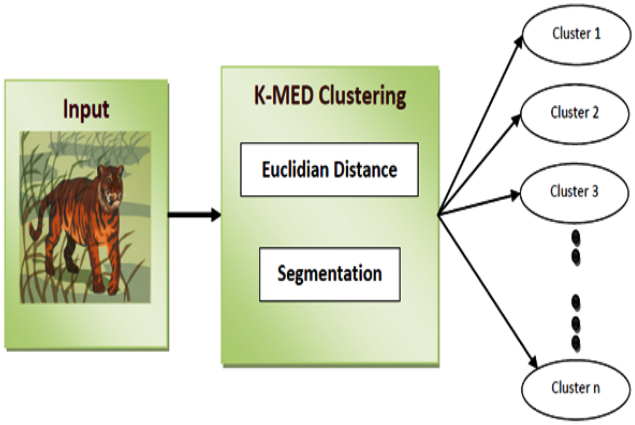

Fig. 1 Segmentation of an image into 'n' clusters

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

III. K-MED CLUSTERING

K-MED clustering is a methodology of grouping together the input items into pre-specified number of groups (K) on the basis of the similarity and commonality between the properties of the input items. In accordance with the numerical value of K supplied along with the input items, the centroids of the groups are chosen in such a way that the centroids remain uniformly distributed instead of initial non-uniform distribution as done in most of the other methods of clustering. Clustering of data points of an image with cluster centroids is shown in Fig. 2. After choosing the initial clusters wisely, K-MED adjusts the centroids on the basis of the Euclidean distances with each and every input item, closer is the item to the centroid more is its belongingness to the cluster and finally the item is assigned to that cluster with respect to which it has the minimum Euclidean.

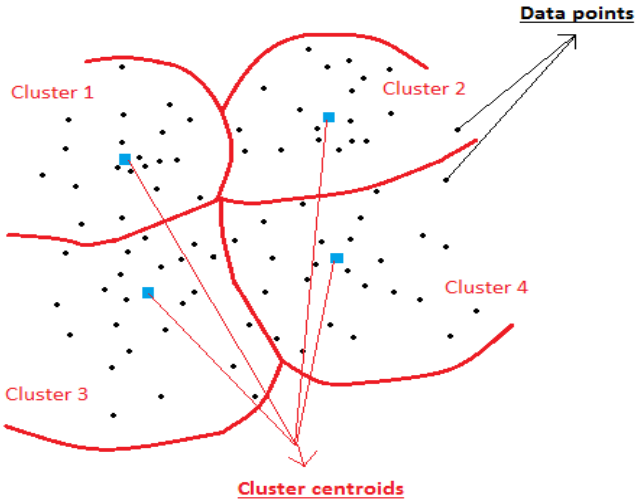

Fig. 2 Clustering of several data points of an Image with cluster centroids

A. Centroid selection

A centroid is a word that is most often used in mathematics and physics to represent "the center of mass" of any object in space, though here, the centroids are not exactly "the center of mass" type centroids instead in K-MED clustering the centroids are the carefully calculated central values of the input items. The centroids for each clusters of an image are shown in Fig. 2. The process of choosing the initial centroids for the pre-specified number of clusters can be done in the following ways:

1) ***Dynamically choosing the initial centroids:*** This method is the best choice when the size of the input data is dynamic which means that it can shrink or expand its boundary limits. In this method, the initial centroids of the pre-specified number of clusters are chosen from the first few of the input items. For example, if the supplied value of K is 3, then it means that the input items must be clustered into 3 groups and the initial centroids of these groups will be the first 3 items of the input items.

2) ***Randomly choosing the initial centroids:*** As the name suggests, this method chooses the initial centroids randomly. The initial centroids are the randomly chosen items from the available number of input items and the number of randomly chosen items needless to say depends upon the supplied value of K.

3) ***Initial centroids selection from upper and lower bounds:*** In this method, the extreme values of the input items are chosen as the initial centroids and this ensures that there is maximum distance between the two centroids and minimum overlapping while calculating the belongingness of any item to its respective centroid.

K-MED clustering does not use any of the existing methods of initial centroids selection as described previously in the paper for initial partitioning. It uses its own method named as Uniformly Distributed Partitioning for the process of choosing the initial centroids. This method is clustered the input items into their respectively desired groups. Unlike other methods of clustering, K-MED uniformly partitions the items into the clusters straight from the beginning of the algorithm. When the items to be clustered are in large amount enough to assign them their respective clusters, it works excellently to follow its varying property of uniformly distributed clusters.

B. Requirements of K-MED Clustering

There are some certain pre-requisites to process the data successfully and obtain the best results:

(i) Each Cluster contains at least one item which means that Blank clusters are not allowed.

(ii) Number of clusters in which the input items are to be clustered must always be less than or equal to the number of input items. If the number of clusters (the value of K) is greater than the number of input items, K-MED do not work properly and display ERROR.

So, the inputs of the algorithm must be chosen carefully to avoid unnecessary distortions in the output clusters.

C. Algorithm

The function K-MED(X, K) partitions the points in 'X' matrix of N×P data points and K clusters on the basis of Euclidean Distance. It uniformly distributes the cluster centroids throughout the space instead of choosing them as random points initially. The rows and columns of 'X' matrix represent the points and variables respectively.

When 'X' matrix is a row oriented vector, K-MED deals with it as N×1 data matrix. K-MED returns a resultant 'N×(P+1)' matrix Y containing the cluster indices of each point in the $(P+1)^{th}$ column.

Inputs: A matrix 'X' of the order N×P which represents N points with P dimensional properties of each point provide the basis of the clustering. K is number of clusters in which these N points are to be classified.

Output: Matrix of the order (N×(P+1)), where $(P+1)^{th}$ column indicates the cluster assigned to that image.

Step1. Initializing centroids: Calculate the Euclidean distance of each point from one single reference point. K-MED uses origin (0, 0) as the initial reference point and perform sorting of the input data on the basis of the corresponding Euclidean distance.

Select the initial centroid(s) of the supplied Input data on the basis of the number of clusters required. Position of the

centroids is decided on the basis of K and it is given by N/K i.e. every (N/K) th location is the centroids' location.

Step2. Initialize all the other sample points with any sentinel value (in K-MED as -1) and centroid points with their own cluster value.

Step3. For each cluster, K-MED finds out all the points belonging to that cluster, then sort them according to the Euclidean distance from their centroid and the centroids are repositioned by following step 4.

Step4. With the inclusion of the new points in a particular cluster C_i, the centroid is recalculated and the centroid will either shift left or right from its current position in the cluster depending upon the points included as given below:

for each cluster C_i do
for each point 'x_j' in C_i do
if x_j > ctrd(C_i)
then
ctrd(C_i) shifts right according to x_j
otherwise
ctrd(C_i) shifts left according to x_j
end for
end for

where, C_i is the i^{th} cluster and i={1,2,3….K}
x_j is the j^{th} point in i^{th} cluster and j={1,2,3….N}

Step5. *for each data point q_k do*
for each cluster C_i do
Recalculate and reassign the Euclidean distance ED (q_k, ctrd(C_i))
end for
end for

where, q_k is the k^{th} input data point and k={1,2,3….N}

Step6. Stop.

K-MED algorithm continues to converge with every step and finally terminates when all the points are assigned to their clusters. Each cluster must contain at least one point and the same point being the centroid of the cluster in such case. K-MED uniformly distributes the centroids throughout the available domain space.

IV. RESULTS AND ANALYSIS OF ALGORITHM

In present study, K-MED algorithm is proposed for image clustering which is a method of grouping the input items on the basis of their closeness to the chosen cluster centroids. A particular item is assigned to the cluster to which it is closest to. This closeness is usually measured in terms of squared distances more often called as the Euclidean distance as it depicts an appropriate method to judge the closeness of any item with the cluster centroid in the multidimensional coordinate system. Important features of K-MED algorithm are described as:

A. Efficiency

K-MED depicts better efficiency for smaller number of sample items and also in case of high value of K. The test samples of 18 two-dimensional points are considered for analyzing K-MED algorithm.

SAMPLES = [1.0 1.0; 1.5 2.0; 3.0 4.0; 5.0 7.0; 3.5 5.0; 4.5 5.0; 3.5 4.5; 1.5 3.5; 3.5 4.5; 1.5 1.5; 2.5 2.5; 3.0 3.0; 5.5 3.5; 2.0 3.5; 1.1 1.1; 1.2 1.1; 1.3 1.4; 1.5 1.4];

The performance of K-MED and k-means algorithm in MATLAB is shown in Fig. 3. When clustering performs on the "SAMPLES", the function K-MED takes less time than k-means clustering. Moreover, the difference is even greater when the number of clusters is increased for the "SAMPLES". Time taken by these two algorithms with variation of the number of clusters is shown in Table I for taking SAMPLES as the input items.

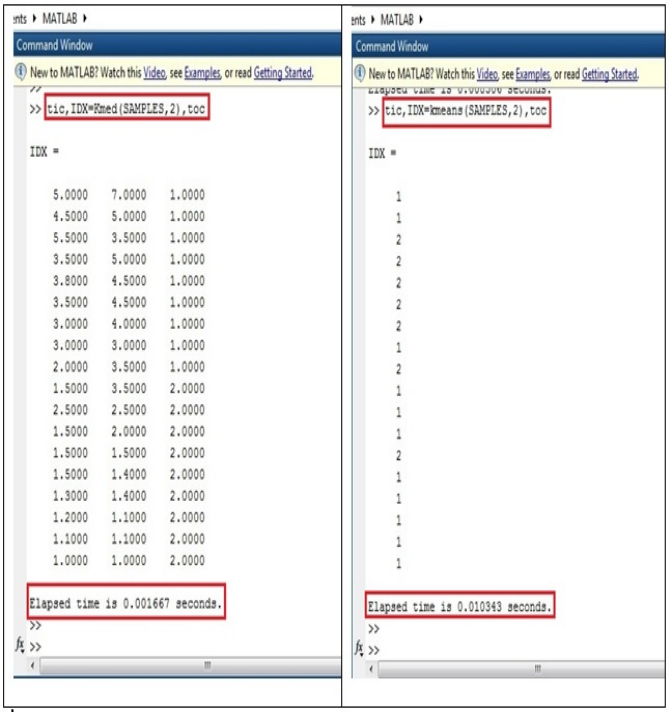

Fig. 3 The performance of K-MED and k-means Algorithm in MATLAB

It is evident that the time taken by K-MED is considerably less as compared to k-means for clustering smaller number of sample items and in case of high value of desired number of clusters. Efficiency of K-MED can further be enhanced by using better searching and sorting algorithm than that which are currently used in it.

B. Local Minima

Another positive feature of K-MED algorithm is that it does not allow the occurrence of Local Minima while clustering the input items. Local minima is the minimum value within the set of points but it need not necessary be the minimum value among all the points of the complete set. In other words, a local minima need not necessary be the global minima too. Hence, a local minima is also called as the relative minima because it is the minimum value with respect to the subset of the complete global set. Finding out the global minima is an iterative process in which the process has to run for exponential runtime for all the possible combinations within the set. Local maxima and minima are not same as the global maxima and minima respectively.

TABLE I

Time taken by k-means and K-MED algorithm with variation of number of clusters

No. of Clusters (K)	Time taken by k-means (seconds)	Time taken by K-MED (seconds)
2	0.010343	0.001667
3	0.341743	0.001790
4	0.017194	0.000423
5	0.147903	0.012697
6	0.146749	0.012575
7	0.127256	0.012937
8	0.018904	0.002083
9	0.156455	0.012596
10	0.148527	0.012782
11	0.125412	0.012896
12	0.125670	0.012787
13	0.144476	0.013104
14	0.146039	0.012614

The k-means and K-MED algorithm are applied on a famous example of Color-based segmentation for obtaining the blue nuclei from an image of tissue stained with hemotoxylin and eosin (H&E). In this example, the k-means clustering yields wrong output image because of occurrence of local minima as shown in Fig. 4. In the case of K-MED, there is no such problem for the occurrence of local minima as shown in Fig. 5, though it takes slightly more time in case of Color-Based segmentation. K-MED is also able to perform all the mundane clustering tasks just as similar to the k-means algorithm in usual conditions of clustering.

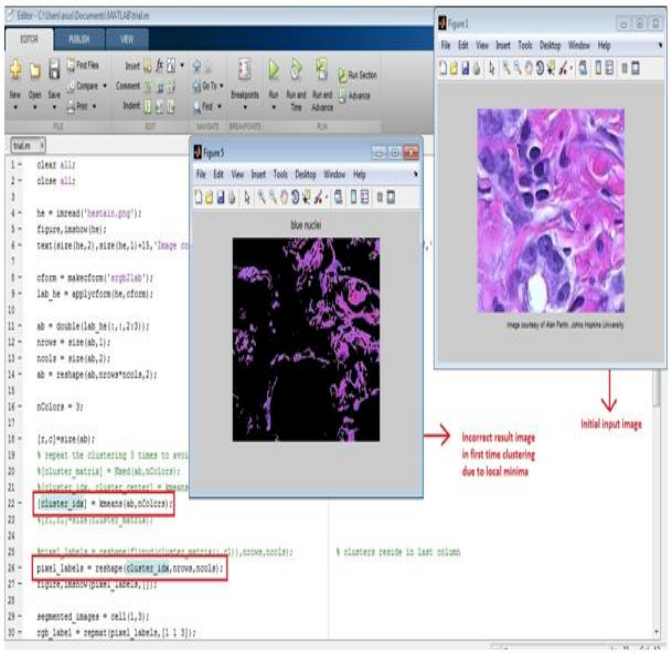

Fig. 4 The clustering of color image using k-means algorithm

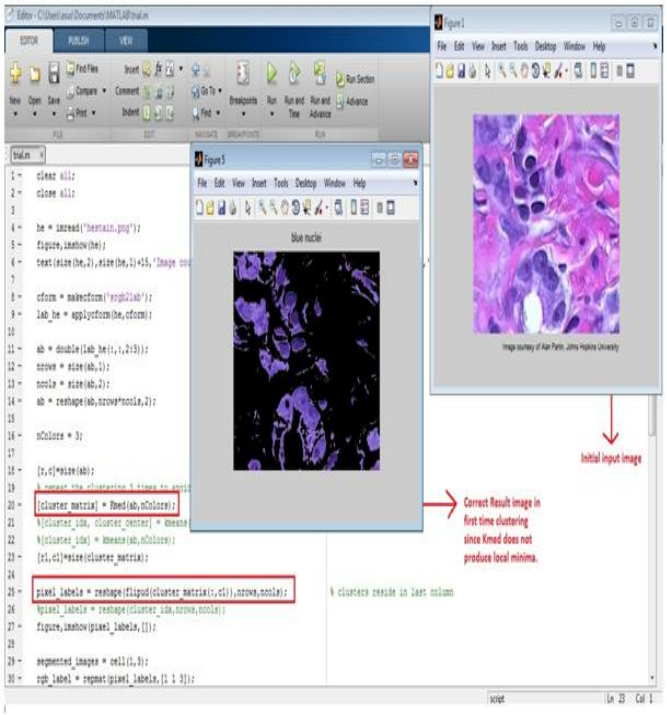

Fig. 5 The clustering of color image using K-MED algorithm

V. CONCLUSIONS

Nowadays, Image Retrieval is an important topic of research for searching the images. The existing systems provide the searched results but always the information is not completely relevant. Sometimes the efficiency of the searching system is not excellent. From the perspective of improving the search efficiency and to increase the relevancy of retrieved results, an algorithm of image

clustering is proposed named as K-MED. In this study, K-MED algorithm is presented, implemented and analyzed to perform uniform image clustering. The clustering of this algorithm represents the better efficiency for smaller number of sample items and also for high value of clusters. K-MED clustering is also compared with the k-means clustering and observed that the proposed clustering resolved the problem of local minima. Therefore, K-MED clustering provides the required results of image clustering.

REFERENCES

[1] S. Asha, S. Bhuvana, and R Radhakrishnan, "A survey on content based image retrieval based on feature extraction," *International Journal of Novel Research in Engineering & Pharmaceutical Sciences*, vol. 1, pp. 29-34, 2014.

[2] J. P. Eakins, "Towards intelligent image retrieval," *Pattern Recognition*, vol. 35, pp. 3-14, 2002.

[3] N. Shrivastava, and V. Tyagi, "Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching," *Information Sciences*, vol. 259, pp. 212-224, 2014.

[4] M. E. ElAlami, "A new matching strategy for content based image retrieval system," *Applied Soft Computing*, vol. 14, pp. 407-418, 2014.

[5] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovie, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *Computer (IEEE)*, pp. 23-32, 1995.

[6] Y. Chen, J. Z. Wang, and R. Krovetz, "An unsupervised learning approach to content-based image retrieval," *Seventh International Symposium on Signal Processing and its applications (IEEE)*, Vol. 1, pp. 197-200, 2003.

[7] A. Singh, P. Sohoni, and M. Kumar, "A review of different content based image retrieval techniques,' *International Journal of Engineering Research and General Science*, vol. 2, pp. 266-275, 2014.

[8] D. Giveki, A. Soltanshahi, F. Shiri, and H. Tarrah, "A new content based image retrieval model based on wavelet transform," *Journal of Computer and Communications*, vol. 3, pp. 66-73, 2015.

[9] A. B. Gonde, R. P. Maheshwari, and R. Balasubramanian, "Complex wavelet transformation with vocabulary tree for Content Based Image Retrieval," *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2010, New York, USA, pp. 359-366.

[10] M. Verma, and R. Balasubramanian, "Center symmetric local binary co-occurrence pattern for texture, face and bio-medical image retrieval," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 224-236, 2015.

[11] K. Yadav, A. Srivastava, A. Mittal, and M. A. Ansari, "Texture-based medical image retrieval in compressed domain using compressive sensing,' *International Journal of Bioinformatics Research and Applications*, vol. 10, pp. 129-144, 2014.

[12] A. Mittal, and L. Cheong, "Framework for synthesizing semantic-level indices" *Multimedia Tools and Applications*, vol. 20, pp. 135-158, 2003.

[13] A. Kannan, V. Mohan, and N. Anbazhagan, "Image clustering and retrieval using image mining techniques," *IEEE International Conference on Computational Intelligence and Computing Research*, 2010.

[14] P. Mishra, Sonam, and S. Vijayalakshmi, "Content based image retrieval using clustering technique: a survey," *International Journal of Research in Computer Engineering and Electronics*, vol. 3, 2014.

[15] S. Kalyani, and K. S. Swarup, "Particle swarm optimization based K-means clustering approach for security assessment in power systems," *Expert Systems with Applications*, vol. 38, pp. 10839-10846, 2011.

[16] D. Carlos, G. Pedronette, J. Almeida, and R. S. Torres, "A scalable re-ranking method for content-based image retrieval," *Information Sciences*, vol. 265, pp. 91-104, 2014.