

# A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques For Large Data sets

Mrs Monika Jain<sup>1</sup>, Dr. S.K.Singh<sup>2</sup>

1Research scholar, Department of computer science, Mewar university, Rajasthan, India.

monika\_smech@yahoo.co.in

2Professor and Head of Department of Information Technology, HRIT Engineering college, Ghaziabad, India.

singhsks123@gmail.com

## ABSTRACT

*Content-based image retrieval (CBIR) is a new but widely adopted method for finding images from vast and unannotated image databases. As the network and development of multimedia technologies are becoming more popular, users are not satisfied with the traditional information retrieval techniques. So nowadays the content based image retrieval (CBIR) are becoming a source of exact and fast retrieval. In recent years, a variety of techniques have been developed to improve the performance of CBIR. Data clustering is an unsupervised method for extraction hidden pattern from huge data sets. With large data sets, there is possibility of high dimensionality. Having both accuracy and efficiency for high dimensional data sets with enormous number of samples is a challenging arena. In this paper the clustering techniques are discussed and analysed. Also, we propose a method HDK that uses more than one clustering technique to improve the performance of CBIR. This method makes use of hierarchical and divide and conquer K-Means clustering technique with equivalency and compatible relation concepts to improve the performance of the K-Means for using in high dimensional datasets. It also introduced the feature like color, texture and shape for accurate and effective retrieval system.*

## General Terms

*Content Based Image Retrieval , divide and conquer k-means, hierarchical*

## KEYWORDS

*Content Based Image Retrieval, Color, Texture, shape*

## 1. INTRODUCTION

Content-Based Image Retrieval (CBIR) is defined as a process that searches and retrieves images from a large database on the basis of automatically-derived features such as color, texture and shape. The techniques, tools and algorithms that are used in CBIR, originate from many fields

such as statistics, pattern recognition, signal processing, and computer vision. It is a field of research that is attracting professionals from different industries like crime prevention, medicine, architecture, fashion and publishing. The volume of digital images produced in these areas has increased dramatically over the past 10 decades and the World Wide Web plays a vital role in this upsurge. Several companies are maintaining large image databases, where the requirement is to have a technique that can search and retrieve images in a manner that is both time efficient and accurate (Xiaoling, 2009). In order to meet these requirements, all the solutions, in general, perform the retrieval process in two steps. The first step is the 'feature extraction' step, which identifies unique signatures, termed as feature vector, for every image based on its pixel values. The feature vector has the characteristics that describe the contents of an image. Visual features such as color, texture and shape are more commonly used in this step. The classification step matches the features extracted from a query image with the features of the database images and groups images according to their similarity. Out of the two steps, the extraction of features is considered most critical because the particular features made available for discrimination directly influence the efficacy of the classification task (Choras, 2007).

The CBIR focuses on Image 'features' to enable the query and have been the recent focus of studies of image databases. The features further can be classified as low-level and high-level features. The focus is to build a universal CBIR system using low level features. Users can query example images based on these features. By similarity comparison the target image from the image repository is retrieved. Meanwhile, the next important phase today is focused on clustering techniques. Clustering algorithms can offer superior organization of multidimensional data for effective retrieval. Clustering algorithms allow a nearest neighbour search to be efficiently performed.

The retrieval of content based image involves the following systems

### **A. Color-based retrieval**

Out of the many feature extraction techniques, color is considered as the most dominant and distinguishing visual feature. Generally, it adopt histograms to describe it. A color histogram describes the global color distribution in an image and is more frequently used technique for content-based image retrieval (Wang and Qin, 2009) because of its efficiency and effectiveness. Color histograms method has the advantages of speediness, low demand of memory space and not sensitive with the image's change of the size and rotation, it wins extensive attention consequently.

### **B. The retrieval based on texture feature**

The identification of specific textures in an image is achieved primarily by modeling texture as a two-dimensional gray level variation. Textures are characterized by differences in brightness with high frequencies in the image spectrum. They are useful in distinguishing between areas of images with similar color (such as sky and sea, or water, grass). A variety of methods has been used for measuring texture similarity; the best-established depend on comparing values of what are well-known as second-order statistics estimated from query and stored images. Essentially, these estimate the relative brightness of picked pairs of pixels from each image. From these it is

possible to measure the image texture such as contrast, coarseness, directionality and regularity [3] or periodicity, directionality and randomness [4].

### **C. The retrieval based on shape feature**

Shape information are extracted using histogram of edge detection. Techniques for shape feature extraction are elementary descriptor, Fourier descriptor, template matching, Quantized descriptors, Canny edge detection [5] etc. Shape features are less developed than their color and texture counterparts because of the inherent complexity of representing shapes. In particular, image regions occupied by an object have to be found in order to describe its shape, and a number of known segmentation techniques combine the detection of low-level color and texture features with region-growing or split-and-merge processes. But generally it is hardly possible to precisely segment an image into meaningful regions using low-level features due to the variety of possible projections of a 3D object into 2D shapes, the complexity of each individual object shape, the presence of shadows, occlusions, non-uniform illumination, varying surface reflectivity, and so on.[6]

### **D. The retrieval based on clustering techniques**

Clustering techniques can be classified into supervised (including semi-supervised) and unsupervised schemes. The former consists of hierarchical approaches that demand human interaction to generate splitting criteria for clustering. In unsupervised classification, called clustering or exploratory data analysis, no labeled data are available [9],[10]. The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of “natural,” hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution [11], [12]. This paper critically reviews and summarizes different clustering techniques.

#### *D.1. Relevance feedback:*

A relevance feedback approach allows a user to interact with the retrieval algorithm by providing the information of which images user thinks are relevant to the query [13][14][16]. Keyword based image retrieval is performed by matching keyword according to user input and the images in the database. Some images may not have appropriate keywords to describe them and therefore the image search will become complex. One of the solution in order to overcome this problem is “relevance feedback” technique [17] that utilize user feedback and hence reduces possible errors and redundancy [18][19]. This technique uses a Bayesian classifier [20][13] which deals with positive and negative feedback. Content based clustering methods cannot adopt to user changes, addition of new topics due to its static nature. To improve the performance of information retrieval log-based clustering approaches are brought into the application.

#### *D.2. Log –Based Clustering:*

Images can be clustered based on the retrieval system logs maintained by an information retrieval process [21]. The session keys are created and accessed for retrieval. Through this the session clusters are created. Each session cluster generates log –based document and similarity of image couple is retrieved. Log –based vector is created for each session vector based on the log-based documents[22]. Now, the session cluster is replaced with this vector. The unaccessed documents creates its own vector.

A hybrid matrix is generated with at least one individual document vector and one log-based clustered vector. At last the hybrid matrix is clustered. This technique is difficult to perform in the case of multidimensional images. To overcome this hierarchical clustering is adopted.

### D.3 Hierarchical Clustering

Hierarchical clustering (HC) algorithms organize data into a hierarchical structure according to the proximity matrix. The results of HC are usually depicted by a binary tree or dendrogram as shown in Fig 1 where A, B, C, D, E, F, G are objects or clusters. It represents the nested grouping of patterns and similarity levels at which groupings change. The root node of the dendrogram represents the whole data set and each leaf node is regarded as a data object. The intermediate nodes, thus, describe the extent that the objects are proximal to each other; and the height of the dendrogram usually expresses the distance between each pair of objects or clusters, or an object and a cluster. The ultimate clustering results can be obtained by cutting the dendrogram at

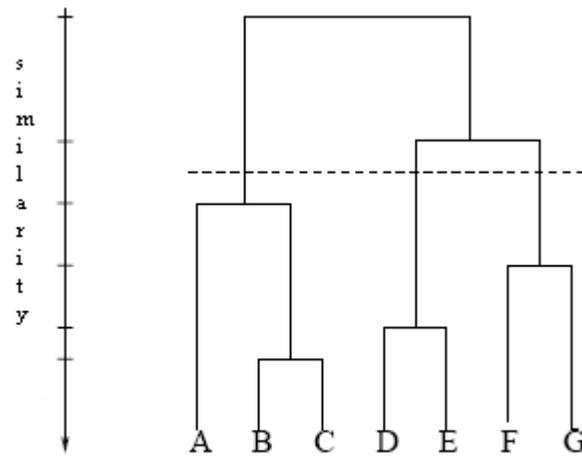


Fig 1. The dendrogram obtained using HC algorithm.

different levels Ref [64]. This representation provides very informative descriptions and visualization for the potential data clustering structures, especially when real hierarchical relations exist in the data, like the data from evolutionary research on different species of organisms. HC algorithms are mainly classified as agglomerative methods and divisive methods. Agglomerative clustering starts with clusters and each of them includes exactly one object. A series of merge operations are then followed out that finally lead all objects to the same group. Divisive clustering proceeds in an opposite way. In the beginning, the entire data set belongs to a cluster and a procedure successively divides it until all clusters are singleton clusters. For a cluster with objects, there are  $2^{N-1}-1$  possible two-subset divisions, which is very expensive in computation [30]. Therefore, divisive clustering is not commonly used in practice. In recent years, with the requirement for handling large-scale data sets in data mining and other fields, many new HC techniques have appeared and greatly improved the clustering performance. Typical examples include CURE [65], ROCK [66], Chameleon [67], and BIRCH [68].

#### *D.4. Retrieval Dictionary Based Clustering*

A rough classification retrieval system is formed. This is formed by calculating the distance between two learned patterns and these learned patterns are classified into different clusters followed by a retrieval stage. The main drawback addressed in this system is the determination of the distance.

To overcome this problem a retrieval system is developed by retrieval dictionary based clustering [23]. This method has a retrieval dictionary generation unit that classifies learned patterns into plural clusters and creates a retrieval dictionary using the clusters. Here, the image is retrieved based on the distance between two spheres with different radii. Each radius is a similarity measure between central cluster and an input image. An image which is similar to the query image will be retrieved using retrieval dictionary.

#### *D.5. Ncut Algorithm*

Ncut method attempts to organize nodes into groups so that the within the group similarity is high, and/or between the groups similarity is low. This method is empirically shown to be relatively robust in image segmentation [24]. This method can be recursively applied to get more than two clusters. In this method each time the subgraph with maximum number of nodes is partitioned (random selection for tie breaking). The process terminates when the bound on the number of clusters is reached or the Ncut value exceeds some threshold  $T$ . The recursive Ncut partition is essentially a hierarchical divisive clustering process that produces a tree[25]. Nonetheless, the tree organization here may misleading a user because there is no guarantee of any correspondence between the tree and the semantic structure of images. Furthermore, organizing image clusters into a tree structure will significantly complicate the user interface.

#### *D.6. K Means clustering*

This nonhierairchal method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters. The K-means algorithm is very simple and can be easily implemented in solving many practical problems. It can work very well for compact and hyperspherical clusters. The time complexity of K-means is  $O(NKd)$ . Since  $K$  and  $d$  are usually much less than  $N$ , K-means can be used to cluster large data sets. Parallel techniques for K-means are developed that can largely accelerate the algorithm [70], [71], [72]. Incremental clustering techniques for example (Bradley et al., 1998) do not require the storage of the entire data set, and can handle it in a one-pattern-at-a-time way. If the pattern displays enough closeness to a cluster according to some predefined criteria, it is assigned to the cluster. Otherwise, a new cluster is created to represent the object.

#### *D.7 Graph theory based clustering*

The concepts and properties of graph theory [73] make it very convenient to describe clustering problems by means of graphs. Nodes of a weighted graph correspond to data points in the pattern space and edges reflect the proximities between each pair of data points. A graph-based clustering method is particularly well suited for dealing with data that is used in the construction of minimum spanning tree MST. It can be used for detecting clusters of any size and shape without specifying the actual number of clusters. Well known algorithms in clustering are Zhan's

Minimum Spanning Tree based clustering [74], [75], and clustering editing method [78] [79], HCS algorithm [80], etc. Current research is focused on clustering using divide and conquers approach [81]. Usually this clustering methodology is used to detect irregular clustering boundaries in clustering results. Zhan [74] proposes to construct an MST and delete the inconsistent edges, i.e. the edges weight values are significantly larger than average weight of the nearby edges in the tree. The inconsistency measure [82] is applied to each edge to detect and remove the inconsistency edges, which results as a set of disjoint subtrees, each subtree will represent a separate cluster

#### D.8. Divide and Conquer K-Means

When the size of a data set is too large, it is possible to divide the data into different subsets and to use the selected cluster algorithm separately to these subsets. This approach is known as divide and conquer [26, 27]. The divide and conquer algorithm first divides the entire data set into a subset based on some criteria. The selected subset is again clustered with a clustering algorithm K-Means. The advantage is to accelerate search and to reduce complexity which depends on number of samples. Methods based on *subspace clustering* may help to ease the problem of clustering high-dimensional data, but they are not adapted at obtaining a large number of clusters [28][29]. A possible solution to this issue, is to cluster hierarchically (obtain a small number of clusters and then cluster again each of the clusters obtained). The proposed enhanced clustering method HDK which uses the combination of unsupervised clustering methods is one of the method that can largely accelerate the CBIR system. In next section problem definition and framework describes. In sections 3,4,5 and 6 evolution, recent work, open areas and then conclusion are given respectively.

## 2. PROBLEM DEFINITION

Clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories). There is no universally agreed upon definition [30]. Most researchers describe a cluster by considering the internal homogeneity and the external separation [31], [32], [33], i.e., patterns in the same cluster should be similar to each other, while patterns in different clusters should not. Both the similarity and the dissimilarity should be examinable in a clear and meaningful way. Here, we give some simple mathematical descriptions of several types of clustering, based on the descriptions in [32]. Given a set of input patterns  $\mathbf{X}=\{\mathbf{x}_1,\dots,\mathbf{x}_j,\dots,\mathbf{x}_N\}$ , where  $\mathbf{x}_j=(f_{j1},f_{j2},\dots,f_{jd})^T \in \mathbb{R}^d$  and each measure  $f_{ji}$  is said to be a feature (attribute, dimension, or variable).

• (Hard) partitional clustering attempts to seek a K-partition of  $\mathbf{X}, C=\{C_1,\dots,C_K\}(K \leq N)$ , such that

- 1)  $C_i \neq \phi, i=1,\dots,K;$
- 2)  $\bigcup_{i=1}^K C_i = \mathbf{X};$
- 3)  $C_i \cap C_j = \phi, i,j=1,\dots,K$  and  $i \neq j.$

## 2.1 Image Retrieval

Image Retrieval using HDK algorithm from the image collections involved with the following steps.

1. Pre-processing is based on RGB color Components using Hierarchical clustering Method.
2. Apply divide and conquer k means.

Some hypothesis has been considered:

- H1 - Proposed method would be able to group samples of same no of clusters and find similarity among them.
- H2 - Proposed method is faster and accurate than single step clustering due to use of divide and conquers technique.
- H3 – Proposed method would allow Euclidean distance to be used in high dimensional data.

In this study we assume that the space is orthogonal and dimensions for all objects are the same and finally we use ordinal data type because of the application. We group images based on number of clusters and clusters are retrieved based on color which is one of the most widely used features for image similarity retrieval. The advantage is nearly linear trend means stability in execute of time . The proposed algorithm is scalable because it follows K-Means complexity that is linear with n number of samples, T number of iteration, d number of dimensions and k number of clusters:  $O(T.K.N.D)$ . Both HC and K-Means use pair's similarity measurement that is very time consuming. By dividing space into subspaces, run time would be reduced because pair's similarity measurement doesn't need to apply for whole samples.

## 2.2 Framework

In this study, our framework includes three main steps as shown in fig.2.

1. Preprocessing is done by applying hierarchical clustering algorithm based on color feature.
2. After preprocessing find the no of clusters for subspace sampling
3. And in final step we separately cluster each subspace sampled using k means clustering technique.

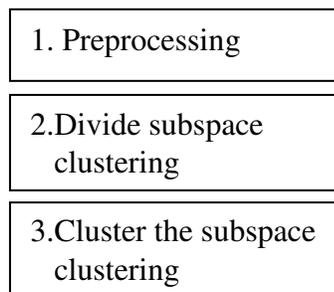


Fig.2 Framework for clustering high dimensional data set.

### 3.EVOLUTION

In the early 1990s, as a result of advances in the Internet and new digital image sensor technologies, the volume of digital images produced by scientific, educational, medical, industrial, and other applications available to users increased dramatically. Most early work on CBIR was based on computing color, shape, and texture based features and using them to define a similarity between the images. A 2008 survey on CBIR highlights the different approaches used for CBIR through time (Datta et al., 2008). Recent approaches for CBIR use key point based features. For example, SIFT (Lowe, 2004) descriptors can be used to represent the images. However, once the size of the image database increases (10 million), and assuming 10 ms to compute the matching score between an image pair, a linear search would take approximately 30 h to answer one query. This clearly is unacceptable. A novel approach for image retrieval is to convert the problem into a text retrieval problem. The key points from all the images are first clustered into a large number of clusters (which is usually much less than the number of key points themselves). These are called visual words. An image is then represented by a histogram of visual words, i.e., the number of keypoints from the image that are in each word or each cluster. The difficulties faced by text based retrieval became more and more severe. The efficient management of the rapidly expanding visual information became an urgent problem. Greg Pass Ramin Zabih [34] in 1996, described histogram refinement for comparing images. Histogram refinement splits the pixels in a given bucket into several classes, based upon some local property. Within a given bucket, only pixels in the same class are compared. Here describe a split histogram called a color coherence vector (CCV), which partitions each histogram bucket based on spatial coherence. After that in 1997, Chad Carson, Serge Belongie, Hay it Greenspan, and Jitendra Malik [35], they present a new image representation which provides a transformation from the raw pixel data to a small set of localized coherent regions in color and texture space. This so-called “blobworld” representation is based on segmentation using the Expectation Maximization (EM) algorithm on combined color and texture features. Minakshi Banerjee, Malay K.Kundu [36] in 2003 discussed the common problem in content based image retrieval (CBIR) is selection of features. Image characterization with lesser number of features involving lower computational cost is always desirable. Color is an important attribute for image matching and retrieval. Recently there has been an increased interest in color research, e.g. classification and segmentation. Swain and Ballard [37] have proposed a color matching method in their paper on 'color indexing'. Their method, known as *histogram intersection*, is based on matching of color histograms and the core idea in their technique is to compute:

$$H(I,M)=\frac{\sum_{j=1}^n \min(I_j, M_j)}{\sum_{j=1}^n M_j}$$

where  $H(I,M)$  is the match value and I and M are image (query image) and model (an image in the database) histograms, respectively, each containing n bins. The match value is computed for every model histogram and the value is closer to unity if the model image is similar to the query image.

In 1995, Ref[38] Mohan S. Kankanhalli, T Babu M. Mehtre And Jian Kang Wut proposed new method in which color clustering can be used to find out clusters and to assign a representative

color to each of these clusters. For each pixel, compute the color distance to the different clusters. Assign the pixel to the cluster for which color distance is minimum. Thus, every pixel is assigned to one of the clusters. Unfortunately, for a given set of color images, there is not enough information about the number and population of clusters.

Ref[28] feature selection methods select only the most relevant of the dimensions from a dataset to reveal groups of objects that are similar on only a subset of their attributes. While quite successful on many datasets, feature selection algorithms have difficulty when clusters are found in different subspaces. It is this type of data that motivated the evolution to subspace clustering algorithms. These algorithms take the concepts of feature selection one step further by selecting relevant subspaces for each cluster separately. All clustering algorithms require that the user set some parameters and make some assumptions about the clusters to be discovered. For detecting data subsets combined with a set of dimensions suitable for local reduction, subspace clustering is used. Clustering in general is used to find natural object groupings in data. The objects contained in such groups are good candidates for generating minimal bounding regions in index structures. Subspace clustering goes beyond fullspace clustering: It was proposed first in [agarwal 1998]. A subspace cluster is a grouping of objects that are only similar in a subset of their dimensions and thus minimum bounding regions that are induced by subspace clusters are very compact. The CLIQUE algorithm [93] was one of the first subspace clustering algorithms. The key idea of the CLIQUE algorithm is that if there are dense units in  $k$  dimensions, there are dense units in all  $(k-1)$  dimensional projections, various subspace clustering algorithms have been published. Some of them extend CLIQUE in different directions[cheng1999,nagesh1999,woo2002].

Subspace clustering algorithms allow users to break the assumption that all of the clusters in a dataset are found in the same set of dimensions.

There are many potential applications with high dimensional data where subspace clustering approaches could help to uncover patterns missed by current clustering approaches. Applications in bioinformatics and text mining are particularly relevant and present unique challenges to subspace clustering.

Segmentation is very important to Image Retrieval, shape feature depend on good segmentation. After extracting features next step in content based image retrieval system is efficient multi-dimensional techniques. But there are two main difficulties occur in exploring multi-dimensional indexing techniques for image retrieval system and they are:

- 1) High dimensionality (The dimensionality of the feature vectors is normally of the order of  $10^2$ ).
- 2) Non-Euclidean similarity measure (Since Euclidean measure may not effectively simulate human perception of a certain visual content, various other similarity measures, such as Histogram Intersection, Cosine, Correlation, etc. need to be supported).

To overcome these problems, we need to perform dimensional reduction and then use appropriate multi-dimensional indexing techniques which are capable of supporting Non-Euclidean similarity measures. The dimension of the feature vectors in image retrieval system is very high, the embedded dimension is much lower [28] before we utilize any indexing technique, it is beneficial to first perform dimension reduction. There are two approaches for dimensional reduction Karhunen-Loeve Transform (KLT) and column-wise clustering. After identifying the embedded

dimension of the feature vectors, we need to select appropriate multi-dimensional indexing algorithms to index the reduced but still high dimensional feature vectors. There are many popular multi-dimensional indexing techniques present; they are Bucketing algorithm-d tree, priority k-d tree [39], quad-tree, K-D-B tree, hb-tree, R-tree and its variants  $R^+$ -tree and  $R^*$ -tree [83, 84, 85, 86, 87]. In addition to the above approaches, clustering and neural nets, widely used in pattern recognition, are also promising indexing techniques [88, 89]. With highdimensional data, indexing schemes such as k-d trees do not work well, to the point where examining every data point (i.e. using no special data structure for acceleration) can be much faster than algorithms intended for low-dimensional acceleration. The most popular and the simplest partitioning algorithm is K-means. K-means has a rich and diverse history as it was independently discovered in different scientific fields by Steinhaus (1956), Lloyd (proposed in 1957, published in 1982), Ball and Hall (1965), and MacQueen (1967). Even though K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity. In 1988, Jain and Dubes gives simple steps of k-means algorithm. Since partitioning algorithms are preferred in pattern recognition due to the nature of available data. Although the topic has been studied for the past 15 years or so, there has been only limited success. The most efficient current method for high-dimensional k-means clustering is Elkan's algorithm [97]. His method does not use any indexing structure, but keeps a number of distance bounds that allow it to avoid unnecessary distance computations. Greg Hamerly proposed another modified and simplified of Elkan's k-means algorithm. As such, he used efficiently updated distance bounds and the triangle inequality to avoid point-center distance calculations.

Because most clustering applications tend to apply to a massive size of datasets, we are interested in finding a method to partition a data set by K-means algorithm in such a way that we can temporarily ignore some resulting partitions in order to work on a few chosen specific partitions without sacrificing clustering quality. A motivation is to save some significant working space and we believe, by means of doing so, we can find a way to add a new portion of a data set into the existing partitions without having to rerun the K-means algorithm from the start. Some partially-related work has already been studied in [97]. Example algorithms include BIRCH (Zhang et al., 1996), divide-and-conquer (Steinbach et al., 2000), coresets K-means (Har-peled and Mazumdar, 2004), and coarsening methods (Karypis and Kumar, 1995). Bradley et al. (1998) presented a fast scalable and singlepass version of K-means that does not require all the data to be fit in the memory at the same time. X-means (Pelleg and Moore, 2000) automatically finds K by optimizing a criterion such as Akaike Information Criterion (AIC) or Bayesian information Criterion (BIC). In K-medoid (Kaufman and Rousseeuw, 2005), clusters are represented using the median of the data instead of the mean. Kernel K-means (Scholkopf et al., 1998) was proposed to detect arbitrary shaped clusters, with an appropriate choice of the kernel similarity function. In 2009, Madjid Khalilian, Farsad Zamani Boroujeni, proposed a method k-means divide and conquer in which objects are clustered base on their size, in fact, they are using subspaces for clustering Ref[40].

#### **4. RECENT WORK**

In a small database ,a simple sequential scan is usually employed for k nearest – neighbor (K-NN) search. But for large data set , efficient indexing algorithms are imperative. High dimensional data is increasingly in many common fields. As the number of dimensions increase, many clustering techniques begin to suffer from the curse of dimensionality, degrading the

quality of the results. In high dimensions, data becomes very sparse and distance measures become increasingly meaningless. There is a general categorization for high dimensional data set clustering: 1-Dimension reduction, 2-Parsimonious models, 3-Subspace clustering [44]. Feature selection and feature extraction are most popular techniques in dimension reduction. It is clear that in both methods we will have losing information which naturally affects accuracy. Ref. [45] reviewed the literature on parsimonious models and Gaussian models from the most complex to simplest which yields a method similar to the K Means approach. When we have low dimensional spaces these methods aren't able to work well. There are two main approaches for subspaces methods: in first class centers are considered on a same unknown subspace and in second each class is located on specific subspace [46]. The idea of subtopics or subgroups is appropriate for document clustering and text mining [47]. Tensor factorization as a powerful technique has been used in [48]. Inconsistency has been shown in those public data sets because of outlier. Ref. [49] proposed a framework which integrates subspace selection and clustering. Equivalency between kernel K-Means clustering and iterative subspace selection has been shown. Ref. [50] proposed a general clustering improver scheme which is included two main steps. In first step it uses intrinsic properties of the data set for dimension reduction after that several iteration of a clustering algorithms are applied, each with different parameter. Base on BIC criterion the best result will be selected. There are some weaknesses for this method e.g. since BIC fits a model to specific data distribution it cannot be used to compare models of different data sets. Ref. [51] presented a semi-supervised clustering method base on spherical K-Means via feature projection which is tailored for handling sparse high dimensional data. They first formulated constraint-guided feature projection then applied the constraint spherical K-Means algorithm to cluster data with reduced dimension. Ref.[52] proposed two methods of combining objective function clustering and graph theory clustering.– Method 1: incorporates multiple criteria into an objective function according to their importance, and solves this problem with constrained nonlinear optimization programming.– Method 2: consists of two sequential procedures:

- (a) A traditional objective function Clustering for generating the initial result
- (b) An auto associative additive system based on graph theory clustering for modifying the initial result.

Ref. [53] Proposed sequential combination methods for data clustering

– In improving clustering performance they proposed the use of more than one clustering method.

– They investigated the use of sequential combination clustering as opposed to simultaneous combination and found that sequential combination is less complex and there are improvements without the overhead cost of simultaneous clustering. In clustering points lying in high dimensional spaces, formulating a desirable measure of “similarity” is more problematic. Recent research shows that for high dimensional spaces computing the distance by looking at all the dimensions is often useless, as the farthest neighbor of a point is expected to be almost as close as its nearest neighbor [54]. To compute clusters in different lower dimensional subspaces, recent work has focused on projective clustering, defined as follows: given a set  $P$  of points in  $R^d$  and an integer  $K$ , partition into subsets that best classify into lower dimensional subspaces according to some objective function. Instead of projecting all the points in the same subspace, this allows each cluster to have a different subspace associated with it [55]. Proposed model will be more effective and achieves significant performance improvement over traditional method for most

clustering. It will be able to cluster samples of same no of clusters and level and be more efficient and accurate than a single one pass clustering. There are some delimitations for this method. First space should be orthogonal it means there is no correlation among attributes of an object. Second base on application that is used all attributes in an object have the same kind of data types. Without losing generality it is possible to extend proposed method to other kinds of data types. Samples are considered in a same number of dimensions. The proposed method HDK select subspaces and perform clustering based on these subspaces.

Fred and Jain [56] use a hierarchical clustering with a single-link method is then applied to the co-occurrence matrix. The final number of clusters is taken as the one that corresponds to the longest lifetime on the dendrogram of the hierarchical algorithm. The method of Fred and Jain is well adapted when the number of clusters is approximately known a priori. If the number of clusters is sequentially changed from 2 to the number of samples, the co-association matrix will change towards a near diagonal matrix with small values out of diagonal. So, the more clusters used to build the coassociation matrix the more clusters result from the combination. The same kind of experiments but with another combination scheme is made by Topchy *et al.* in [58]. Lange and Buhmann [59] optimize a probabilistic model of the co-association matrix. The *EM* algorithm optimizes model parameters and needs  $O(I^2)$  operations for each iteration, where  $I$  is the number of data samples. It makes difficult to apply this approach to a high volume of data. Many of the presented methods need to know a priori about data to combine clusterings or to set manually some parameters for a combination scheme. This provided us with the motivation to pose the problem of combination in a form which will not depend on any parameter and a priori knowledge. Our formulation of the problem is based on a co-association matrix. It allows us to process a huge volume of data as well as clusterings.

#### **4. OPEN AREAS**

Similarity search is an important topic of research, which finds many applications and not surprisingly there is a vast literature about the subject. It is already been discussed that the existing techniques may be used to improve the quality of image retrieval and the understanding of user intentions. An approach that combines two different approaches to image retrieval, together with active use of context information and interaction has been proposed. An important aspect of the research work outlined here is to design and evaluate a system that, in addition to combine the use of TBIR with CBIR [60]. Use of the hybrid feature including color, texture and shape as feature vector of the regions to match images can give better results. Results on a database of 1000 general-purposed images demonstrate the efficiency and effectiveness of the image representation for region based image retrieval [61]. The technique called self taught multiple-instance learning (STMIL) [62] that deals with learning from a limited number of ambiguously labeled examples is also a effective area to work with for efficient results. The problem of bridging the semantic gap between high level query which is normally in terms of an example image and low level features of an image such as colour, texture, shape and object forced to apply techniques to reduce the semantic gap. 3D informatics[90] incorporates additional research areas including content-based retrieval, image understanding, indexing, data mining, and data management. Existing techniques for image retrieval based on divide and conquer k-means and subspace clustering is a novel approach for retrieving the images based on their contents. HDK algorithm can be used in various areas to improve the performance of the system and to achieve better results in different applications.

## 5. CONCLUSIONS

The purpose of this survey is to provide an overview of the functionality of content based image retrieval systems. Combining advantages of HC and divide and conquer K-Means strategy can help us in both efficiency and quality. HC algorithm can construct structured clusters. Although HC yields high quality clusters but its complexity is quadratic and is not suitable for huge datasets and high dimension data. In contrast K-Means is linear with size of data set and dimension and can be used for big datasets that yields low quality. Divide and conquer K-Means can be used for high dimensional data set . In this paper we present a method HDK to use both advantages of HC and Divide and conquer K-Means by introducing equivalency and compatible relation concepts. Using two steps clustering in high dimensional data sets with considering no of clusters based on color feature helps us to improve accuracy and efficiency of original K-Means clustering. For this purpose we should consider orthogonal space. HDK algorithm has been used extensively in various areas to improve the performance of the system and to achieve better results in different applications.

## REFERENCES:

- [1] Rui Xu, Donald Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions On Neural Networks, Vol. 16, No. 3, May 2005.
- [2] Ritendra Datta, Dhiraj Joshi, Jia Li, And James Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, Vol. 40, No. 2, Article 5, Publication date: April 2008.
- [3] Tamura et al. "Texture Features Corresponding to Visual Perception"- IEEE Trans on system, Man and cyber 8-460-472-1978.
- [4] Sanjoy Kumar Saha et al. "CBIR Using Perception Based Texture And Colour measures "CSE Department; CST Department Jadavpur Univ., India; B.E. College, Unit ISI, Kolkata, India - -2003.
- [5] Canny, J., "A computational approach to edge detection", IEEE Trans on Pattern Analysis and Machine Intelligence, 8:679-698, 1986.
- [6] S.Nandagopalan, Dr. B.S. Adiga, and N. Deepak "A Universal Model for Content-Based Image Retrieval" World Academy of Science, Engineering and Technology 46 2008.
- [7] Mohammed Eisa and Ibrahim Elhenawy and A. E. Elalfi and Hans Burkhardt, "Image Retrieval based on Invariant Features and Histogram Refinement", ICGST International Journal on Graphics, Vision and Image Processing, March 2006, pp. 7-11.
- [8] V. Castelli and L. D. Bergman (Eds.), "Image Databases: Search and Retrieval of digital Imagery", Wiley: New York, 2002.
- [9] J B. Everitt, S. Landau, and M. Leese, "Cluster Analysis", London:Arnold, 2001.
- [10] A. Jain and R. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [11] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks—Part I and II," IEEE Trans. Neural Netw., vol. 13, no. 3, pp.645–677, May 2002.
- [12] V. Cherkassky and F. Mulier, "Learning From Data: Concepts, Theory, and Methods", New York: Wiley, 1998.
- [13] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papatomas and P.Yianilos, "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments", IEEE Trans. Image Processing, vol. 9, no. 1, pp. 20-37, 2000.
- [14] Y. Rui, T. S. Huang, M. Ortega and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval", IEEE Trans. Circuits and Video Technology, vol. 8, no. 5, pp. 644-655, 1998.
- [15] S. Thilagamani and N. Shanthi , "A Survey on Image Segmentation Through Clustering", International Journal of Research and Reviews in Information Sciences , Vol. 1, No. 1, March 2011 .

- [16] Rui Y, Thomas, S. Huang. "Content-based image retrieval with relevance feedback in MARS", In proceedings of IEEE international conference on image processing, pp. 815-818, 1997.
- [17] Hoi, C.-H. and Lyu, M. R. 2004b. "A novel log based relevance feedback technique in content based image retrieval". In Proc. ACM Multimedia,2004.
- [18] Zhou XS, Huang TS., Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst*; 8: 536-544, 2003.
- [19] P.Chundi, U.Dayal, , M.Sayal,, M.Hsu, "A *Document Clustering Method and System*:" google patents,US20077181678, 2007.
- [20] D. Melas and S. Wilson, "Double Markov Random Fields and Bayesian Image Segmentation," *IEEE Trans. Signal Processing*,vol. 50, no. 2, pp. 357-365, Feb. 2002.
- [21] Huiyu Zhou, Abdul H. Sadka, Mohammad R. Swash, Jawid Azizi and Abubakar S. Umar., "Content Based Image Retrieval and clustering: A Brief Survey" school of Engineering and Design, Brunel University, Uxbridge, UB8 3PH, UK
- [22] Hoi, C.-H. and Lyu, M. R. 2004a. "Group-based relevance feedbacks with support vector machine ensembles" In Proc. IEEE ICPR,2004.
- [23] Li Wenchao Zhou Yong Xia Shixiong China Univ. of Min. & Technol., Xuzhou, "A Novel Clustering Algorithm Based on Hierarchical and K-means Clustering" Print ISBN: 978-7-81124-055-9, On page(s): 605, 2009.
- [24] Jianbo Shi Malik, J.Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA , "Normalized cuts and image segmentation", ISSN: 0162-8828, On page(s): 888 – 905,2000.
- [25] Emma Regentova, Dongsheng Yao, and Shahram Latifi, "Image segmentation using Ncut in the wavelet domain", *International journal of image and graphics*,world scientific Publishing company,vol.6(4),pp.569-582,2006.
- [26] Guha, Meyerson, A. Mishra, N. Motwani, and O. C. ."Clustering data streams: Theory and practice ." *IEEE Transactions on Knowledge and Data Engineering*, vol. 15,pp. 515-528, 2003.
- [27] A. Jain , M. Murty , and p. Flynn " Data clustering: A review.," *ACM Computing Surveys*, vol. 31, pp. 264-323,1999.
- [28] L. Parsons, E. Haque and H. Liu., "Subspace clustering for high dimensional data: a review." *SIGKDD Explor. Newslett.* 6 (1), pp. 90–105. 2004.
- [29] C. Bouveyron, S. Girard and C. Schmid, "High-dimensional data clustering. *Computational Statistics & Data Analysis*", vol. 52, pp.502–519. 2007.
- [30] B. Everitt, S. Landau, and M. Leese, "Cluster Analysis" . London:Arnold, 2001.,
- [31] Classification, 2nd ed. London, U.K.: Chapman & Hall, 1999.
- [32] P. Hansen and B. Jaumard, "Cluster analysis and mathematical programming", *Math. Program.*, vol. 79, pp. 191–215, 1997.
- [33] A. Jain and R. Dubes, "Algorithms for Clustering Data". Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [34] Greg Pass, Ramin Zabih, "Histogram refinement for content based image retrieval" WACV '96.
- [35] Chad Carson, Serge Belongie, Hay it Greenspan, and Jitendra Malik, "Region Based Image Querying," this work is supported by an NSF digital library grant (IRI 94-11334) 1997 IEEE.
- [36] Minakshi Banerjee 1, Malay K. Kundu "Edge based features for content based image retrieval" 0031-3203/2003.
- [37] M. J. Swain and D. H. Ballard, "Color indexing" , *Int. d. Comput. Vis.* 7(1), 11-32 (1991).
- [38] Mohan S. Kankanhalli,T Babu M. Mehtre and Jian Kang Wut, "Cluster-Based Color Matching", *Pattern Recognition*. Vol. 29, No. 4, pp. 701 708, 1996.
- [39] Madjid Khalilian, Norwati Mustapha, MD Nasir Suliman, MD Ali Mamat," A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets",*Proceedings of the IMECS 2010*.
- [40] Madjid Khalilian, Farsad Zamani Boroujeni, Norwati Mustapha, Md. Nasir Sulaiman, "K-Means Divide and Conquer Clustering", *International Conference on Computer and Automation Engineering*.
- [41] H. T. Shen, X. Zhou, and A. Zhou. "An adaptive and dynamic dimensionality reduction method for high-dimensional indexing" *VLDB J.*, 16(2):219{234,2007.
- [42] Y. Tao, K. Yi, C. Sheng, and P. Kalnis., "Quality and efficiency in high dimensional nearest neighbor search",.In *SIGMOD*, pages 563{576, 2009.

- [43] R. Weber, H.-J. Schek, and S. Blott., "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces.", In VLDB, pages 194-205, 1998.
- [44] C. Bouveyran, S. Girard, C. Schmid: Technical Report 1083M, LMC-IMAGE 2008.
- [45] C. Fraley and A. Raftery, "Model based clustering discriminate analysis and density estimation", journal of American statistics association, 97, 611-631, 2002.
- [46] H. Bock, on the interface between cluster analysis, principal component clustering and multidimensional scaling, 1987.
- [47] Jeffery L. Solka: Text data mining Theory and methods, vol 2, 94-112, 2008
- [48] H. Huang, C. Ding, D. Luo, T. Li, "simultaneous tensor subspace selection and clustering", KDD08, lasvegas USA, 2008.
- [49] J. Ye, Z. Zhao, M. Wu, "Discriminative K-Means for clustering", proceeding of the annual conference, 2007.
- [50] R. Varshavsky, D. Itorn and M. Linial, "Cluster algorithm optimizer: A framework for large datasets", ISBRA pp 85-96, springer 2007.
- [51] W. Tang, H. Xiong, S. Zhang, J. Wu, "Enhancing semisupervised clustering", KDD07 california USA, ACM 2007.
- [52] Yuntao Qian and Ching Y. Suen, "Clustering Combination Method", IEEE 2000.
- [53] Yuntao Qian and Ching Y. Suen, "Sequential combination methods for data clustering", Journal of computer science, 2002.
- [54] A. Raftery and N. Dean, "variable selection for model based clustering", journal of American statistical association, 101(473):168-178, 2006.
- [55] Pankaj K. Agarwal and Nabil H. Mustafa, "K-Means Projective clustering", PODS Paris France, ACM 2004.
- [56] A. L. Fred and A. K. Jain., "Combining multiple clusterings using evidence accumulation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6):835-850, 2005.
- [57] Ivan O. Kyrgyzov, Henri Maître and Marine Campedel, "A Method of Clustering Combination Applied to Satellite Image Analysis".
- [58] A. P. Topchy, A. K. Jain, and W. F. Punch., "A mixture model for clustering ensembles", In *SDM*, 2004.
- [59] T. Lange and J.M. Buhmann, "Combining partitions by probabilistic label aggregation", In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 147-156, New York, NY, USA, 2005. ACM Press.
- [60] Christian Hartvedt, "Using Context to Understand User Intentions in Image Retrieval" 2010.
- [61] ZhiYong Zeng, ZhiQiang Yao, ShiGang Liu, "An Efficient and effective Image Representation for Region-Based Image Retrieval" ICIS 2009, November.
- [62] Qifeng Qiao, Peter A. Beling, "Localized Content Based Image Retrieval with Self-taught Multiple Instance Learning" 2009 IEEE.
- [63] Nidhi Singhai, Prof. Shishir K. Shandilya, "A Survey On: Content Based Image Retrieval Systems", International Journal of Computer Applications (0975 - 887) Volume 4 - No.2, July 2010.
- [64] Steven M. Holland, "Cluster Analysis", Department of Geology, University of Georgia, Athens, GA 30602-2501
- [65] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 1998, pp. 73-84.
- [66] "ROCK: A robust clustering algorithm for categorical attributes," *Inf. Syst.*, vol. 25, no. 5, pp. 345-366, 2000.
- [67] G. Karypis, E. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *IEEE Computer*, vol. 32, no. 8, pp. 68-75, Aug. 1999.
- [68] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in *Proc. ACM SIGMOD Conf. Management of Data*, 1996, pp. 103-114.
- [69] Ball, G., Hall, D., 1965. ISODATA, a novel method of data analysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA.
- [70] E. Dahlhaus, "Parallel algorithms for hierarchical clustering and applications to split decomposition and parity graph recognition," *J. Algorithms*, vol. 36, no. 2, pp. 205-240, 2000.

- [71] C. Olson, "Parallel algorithms for hierarchical clustering," *Parallel Comput.*, vol. 21, pp. 1313–1325, 1995.
- [72] K. Stoffel and A. Belkoniene, "Parallel K-means clustering for large data sets," in *Proc. EuroPar'99 Parallel Processing*, 1999, pp.1451–1454.
- [73] F. Harary, *Graph Theory*. Reading, MA: Addison-Wesley, 1969
- [74] C.T Zahn(1971) : Graph-theoretical methods for detecting and describing clusters , IEEE trans on Computers c(20):68-86
- [75] G. Karypis; E.Han(1999) : "A hierarchical clustering Algorithm using dynamic modeling, IEEE Trans on Computers:", Special issue on Data analysis and Mining, 32(8):68-75
- [78] J.Gramm; J.Guo(2003) : Graph modeled data clustering: fixed parameter algorithms for clique generation, In Lecture Notes on Computer Science(LNCS),pages 109-118,Springer
- [79] R. R.Shamir; D.Tsur(2002): Cluster graph modification problems. In Lecturer notes in computer science(LNCSI),pp379-390,springer
- [80] E. Hartuy ; R. Sharmir. A clustering algorithm based on graph connectivity. Information processing,pp175-181
- [81] Xiaochun Wang; D. Mitchell Wilkes (2009) : "A Divide-and-Conquer Approach for Minimum Spanning tree-Based Clustering", Member,IEEE Transactions on knowledge and data engineering, Vol 21 No7
- [82] Oleksandr; Grygorash(2006); Yan Zhou: "Minimum Spanning tree Based Clustering", IEEE Tools with artificial intelligence,pp 3-81
- [83]. A. Guttman, R-tree: "A dynamic index structure for spatial searching", in *Proc. ACM SIGMOD, 1984*.
- [84]. T. Sellis, N. Roussopoulos, and C. Faloutsos, "The RC-tree: A dynamic index for multi-dimensional objects", in *Proc. 12th VLDB, 1987*.
- [85]. D. Greene, "An implementation and performance analysis of spatial data access", in *Proc. ACM SIGMOD, 1989*.
- [86]. N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, The R<sup>+</sup>-tree: "An efficient and robust access method for points and rectangles", in *Proc. ACM SIGMOD, 1990*.
- [87]. Y. Rui, K. Chakrabarti, S. Mehrotra, Y. Zhao, and T. S. Huang, "Dynamic clustering for optimal retrieval in high dimensional multimedia databases", in *TR-MARS-10-97, 1997*.
- [88] R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Chap. 6, Wiley, New York, 1973.
- [89]. H. J. Zhang and D. Zhong, "A scheme for visual feature based image retrieval", in Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995.
- [90] 3D Medical Informatics, Information Science in Multiple Dimensions chapter12 pp 333-347, Terry S. Yoo
- [91] Jayant Mishra, Anubhav Sharma and Kapil Chaturvedi, "An Unsupervised Cluster-based Image Retrieval Algorithm using Relevance Feedback", International Journal of Managing Information Technology (IJMIT) Vol.3, No.2, May 2011
- [92] Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan, Prabhakar, 1998. "Automatic subspace clustering of high dimensional data for data mining applications." In: Proc. ACM SIGMOD, pp. 94–105.
- [93] Guojun Gan., "Subspace clustering for high dimensional categorical data", May 2003. Talk Given at SOSGSSD.
- [94] Harsha S. Nagesh., "High performance subspace clustering for massive data sets." Master's thesis, North-western Univeristy, 2145 Sheridan Road, Evanston IL 60208, June 1999.
- [95] K. G. Woo and J. H. Lee. FINDIT: "a Fast and Intelligent Subspace Clustering Algorithm using Dimension Voting". PhD thesis, Korea Advanced Institute of Science and Technology, Taejon, Korea, 2002.
- [96] Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang., "Entropy-based subspace clustering for mining numerical data", In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 84{93. ACM Press, 1999.

- [97] Charles Elkan., "Using the triangle inequality to accelerate k-means", In Tom Fawcett and Nina Mishra, editors, ICML, pages 147–153. AAAI Press, 2003.
- [98] Greg Hamerly, "Making k-means even faster", SIAM.
- [99] Lloyd, S., 1982., "Least squares quantization in PCM. IEEE Trans. Inform. Theory 28, 129–137. Originally as an unpublished Bell laboratories Technical Note (1957).
- [100] MacQueen, J., 1967, "Some methods for classification and analysis of multivariate observations", In: Fifth Berkeley Symposium on Mathematics. Statistics and Probability. University of California Press, pp. 281–297.
- [101] Steinhaus, H., 1956. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci. IV (C1.III), 801–804.
- [102] Zhang, Tian, Ramakrishnan, Raghu, Livny, Miron. 1996. "BIRCH: An efficient data clustering method for very large databases.", In: Proc.1996 ACM SIGMOD Internat. Conf. on Management of data, vol. 25, pp. 103–114.
- [103] Steinbach, Michael, Tan, Pang-Ning, Kumar, Vipin, Klooster, Steve, Potter, Christopher, 2003., "Discovery of climate indices using clustering", In: Proc Ninth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining.
- [104] Bradley, P.S., Fayyad, U., Reina, C., 1998., "Scaling clustering algorithms to large databases", In: Proc. 4th KDD.
- [105] Har-peled, Sarel, Mazumdar, Soham, 2004., "Coresets for k-means and k-median clustering and their applications", In: Proc. 36th Annu. ACM Sympos. Theory Comput., pp. 291–300.
- [106] Karypis, George, Kumar, Vipin, 1995. , "A fast and high quality multilevel scheme for partitioning irregular graphs.", In: Proc. Internat. Conf. on Parallel Processing, pp.113–122.
- [107] Kaufman, Leonard, Rousseeuw, Peter J., 2005., Finding groups in data: An introduction to cluster analysis. Wiley series in Probability and Statistics.
- [108] Pelleg, Dan, Moore, Andrew, 2000., "X-means: Extending k-means with efficient estimation of the number of clusters.", In: Proc. Seventeenth Internat. Conf. on Machine Learning. pp. 727–734.
- [109] Scholkopf, Bernhard, Smola, Alexander, Muller, Klaus-Robert, 1998., "Nonlinear component analysis as a kernel eigenvalue problem", Neural Computer. 10 (5),1299–1319.
- [110] Choras, R.S. (2007), "Image feature extraction techniques and their applications for CBIR and biometrics systems", International Journal of Biology and Biomedical Engineering, Vol.1, Issue 1, Pp. 6-16.
- [111] Xiaoling, W. (2009), "A Novel Circular Ring Histogram for Content-Based Image Retrieval", First International Workshop on Education Technology and Computer Science, Vol. 2, Pp.785-788.
- [112] Wang, S. and Qin, H. (2009), "A Study of Order-Based Block Color Feature Image Retrieval Compared with Cumulative Color Histogram Method", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD '09, IEEE Xplore, Vol. 1, Pp. 81 – 84.