An Efficient Content Based Image Retrieval System for Color and Shape Using Optimized K-Means Algorithm

A.Komali, R.Veera Babu, D.Sunil Kumar, K.Ganapathi Babu

Abstract
This paper deals with the Content Based Image Retrieval CBIR system which is the challenging research platform in the digital image processing. The important theme of using CBIR is to extract visual content of an image automatically, like color, texture, or shape. The simple process to retrieve an image from the image set, we use image search tools like Google images, Yahoo, etc. The main goal of view is based on the efficient search on information set. In the point of searching text, we can search flexibly by using keywords, but if we use images, we search using some features of images, and these features are the keywords. Color and shape image retrieval (CSIR) describes a possible solution for designing and implementing a project which can handle the informational gap between a color and shape of an image. This similar color and shape of an image is retrieved by comparing number of images in datasets. This CSIR can be developed by using K-Means algorithm for getting retrieval results of similar image efficiently. By using K-Means algorithm, more number of iterations occurred. In order to reduce the number of iterations we use codebook algorithm. This CSIR can be used in several applications such as photo sharing sites, forensic lab, etc. CLARANS is the normal method which is used to reduce the bugs occurred in the existing algorithms.

Index Terms
K-Means Algorithm, code book algorithm, CLARANS.

1. INTRODUCTION:

Content-based image retrieval information systems use information extracted from the content of query image. Content-based image retrieval system retrieves an image from a database using visual information such as color, texture, or shape of an image. The CBIR systems have a great importance in the criminal investigation. In the most retrieval systems, the user queries by presenting an example image that has the intended feature. The features of this sample image are taken into consideration. Although this discussed approach has advantages in effective query processing, it is inferior in expressive power and the user cannot represent all intended features in his query.

Before going to the information technology platform, a large amount of data had to be maintained, processed and stored. It was also textual and visual information. The efficiency of searching in information set is a very important point of view. In many cases if we want to search efficiently some data have to be recalled. The human is able to recall visual information more easily using for example the shape of an object, or arrangement of colors and objects. In case of texts we can search accurately by using keywords, but if we use images, in this case we search using some features of images, and these features are the keywords.

While using the images, the large amount of data and the management of those cause the problem. The processing space is enormous. For this our purpose is to develop a content based image retrieval system, which can retrieve using sketches in frequently used databases. The user has a drawing area where he can draw those sketches efficiently. These sketches act as the base of the retrieval method. In some cases we can recall our minds with the help of figures or drawing. This drawing consists of color, shape texture.

Color and shape image retrieval (CSIR) describes a possible solution for designing and implementing a project which can handle the informational gap between a color and shape of an image. In the following paragraph some application possibilities are analyzed.

The possible application area of sketch based information retrieval is the searching of analog circuit graphs from a big database [3]. The Sketch-based image retrieval (SBIR) was introduced in QBIC and Visual SEEK [4] systems. In these systems the user draws color sketches and blobs on the drawing area. The images were divided into grids, and the color and texture features were determined in these grids. The applications of grids were also used in other algorithms, for example in the edge histogram descriptor (EHD) method [2]. This method displays the variations occurred while changing the number.

2. PREVIOUS WORK

Content Based Image Retrieval (CBIR) is the process to search relevant images based on user input automatically. The input could be parameters, sketches or example images. A typical CBIR process first extracts the image features and stores them efficiently. Then it compares with images from the database and returns the results. In case of images, we search using some features of images, and these features are the keywords. Feature extraction and

Manuscript received April 5, 2015
Manuscript revised April 20, 2015

similarity measure are very dependent on the features used. In each feature, there would be more than one representation. Among these representations, histogram is the most commonly used technique to describe features.

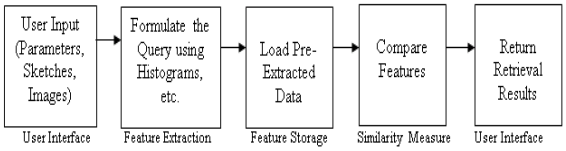

Fig.1 Flow chart of a typical CBIR process

Fig.1 describes the flow of a typical CBIR process although content based methods are efficient, they cannot always match user's expectation. Relevance Feedback (RF) techniques are used to adjust the query by user's feedback. RF is an interactive process to improve the retrieval accuracy by a few iterations. RF algorithms are dependent on feature representations, in this chapter, RF process and its histogram weighting method will be introduced.

Fig.2 The retrieval has to be robust in contrast of illumination and difference of point of view.

The system was designed for databases containing relatively simple images, but even in such cases large differences can occur among images in file size or resolution. In addition, some images may be noisier, the extent and direction of illumination may vary (see Fig. 2), and so the feature vectors cannot be effectively compared. In order to avoid it, a multistep preprocessing mechanism precedes the generation of descriptors.

A. ALGORITHM USED: k-mean

Disadvantage of k-mean algorithm-

- More number of iterations
- Does not work well with non globular clusters.
- Non-globular- in contrast to globular cluster do not have well defined centers. Non- globular cluster can have a chain like shape.
- Fixed no of clusters can make it difficult to predict what the K should be.
- Unable to handle noisy data and outliers.
- Used for small database

The k-means method uses centroid to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data. In order to overcome these problems we use medoids to represent the cluster rather than centroid. A medoid is the most centrally located data object in a cluster. Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. Or in other words, medoids move in each iteration. This process is continued until no any medoid move. As a result, k clusters are found representing a set of n data objects. An algorithm for this method is given below.

Input: 'k', the number of clusters to be partitioned; 'n', the number of objects.

Output: A set of 'k' clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Steps:

1. Arbitrarily choose 'k' objects as the initial medoids;
2. Repeat,
   - Assign each remaining object to the cluster with the nearest medoid;
   - Randomly select a non-medoid object;
   - Compute the total cost of swapping old medoid object with newly selected nonmedoid object.
   - If the total cost of swapping is less than zero, then perform that swap operation to form the new set of kmedoids.
3. Until no change.

3. NORMAL METHOD:

Partitioning Methods in Large Databases: From *k*-Medoids to CLARANS [11]

An advanced algorithm PAM (Partitioning Around Mediods) is used in K-mediods partitioning algorithm which works efficiently for small data sets, but does not scale well for large data sets. In case of larger data sets, a sampling-based method, called CLARA (Clustering LARge Applications), can be used. The idea behind CLARA is as follows: a small portion of the actual data is chosen as a representative of the data, there is no need of taking the whole set of data into consideration, Medoids are then chosen here by using PAM. If the sample is chosen as random manner, it should closely represent the original data set. The representative objects (medoids) chosen will likely be similar to those that would have been chosen from the whole data set. CLARA draws multiple samples of the data set, applies PAM on each sample, and returns its best clustering as the output. As expected, CLARA can deal with larger data sets than PAM. The time complexity of each iteration now becomes

$$O(ks^2+k(n-k)),$$

where s is the size of the sample, k is the number of clusters, and n is the total number of objects.

The effectiveness of CLARA depends on the sample size. Notice that PAM searches for the best k medoids among a given data set, whereas CLARA searches for the best k medoids among the selected sample of the data set. Consider; if an object oi is one of the best k medoids but is selected during sampling, CLARA will never find the best clustering. This is, therefore, a trade-off for efficiency.

CLARANS (Clustering Large Applications based upon Randomized Search) is the K-mediod algorithm which was proposed, which combines the sampling technique with PAM. At each stage in the search CLARA has a fixed sample, in each step of the search CLARANS draws a sample with some randomness. Conceptually, the clustering process can be viewed as a search through a graph, where each node is a potential solution (a set of k medoids).

The current node is then replaced by the neighbor with the largest descent in costs. Because CLARA works on a sample of the entire data set, it examines fewer neighbors and restricts the search to sub graphs that are smaller than the original graph. The number of neighbors to be randomly sampled is restricted by a user specified parameter. If a better neighbor is found (i.e., having a lower error), CLARANS moves to the neighbor's node and the process starts again; otherwise, the current clustering produces a local minimum. If a local minimum is found, CLARANS starts with new randomly selected nodes in search for a new local minimum. Once a user-specified number of local minima has been found, then it is said to be a local minimum, that is having the lowest cost. CLARANS has been experimentally shown to be more effective than both PAM and CLARA. CLARANS also enables the detection of outliers. However, the computational complexity of CLARANS is about O(n2), where n is the number of objects. Furthermore, its clustering quality is dependent on the sampling method used.

4. PROPOSED WORK:

The CSIR can be developed by using K-Means algorithm for getting retrieval results of similar image efficiently. By using K-Means algorithm, more number of iterations are occurred. In order to reduce the number of iterations we use codebook algorithm. The codebook algorithms are like Linde Buzo and Gray (LBG), Kekre's Fast codebook Generation (KFCG) Algorithms.

A. LBG Algorithm [6], [7]:

- In this algorithm centroid is computed as the first code vector for the training set.
- In Fig. 1 two vectors v1 & v2 are generated by adding constant error to the code vector.
- Euclidean distances of all the training vectors are computed with vectors v1 and v2 thus two clusters are formed based on nearest of v1 or v2.
- This procedure is repeated for every cluster.
- . This results in inefficient clustering.

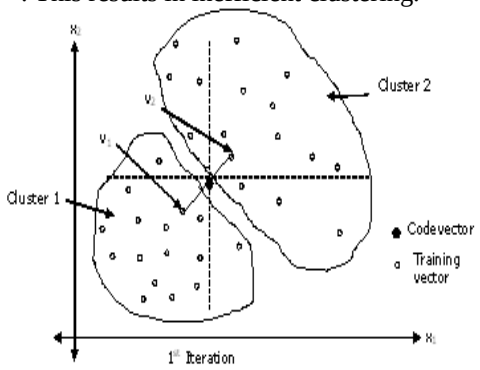

Fig.1. LBG for 2 dimensional case

B. Kekre's Fast Code book Generation algorithm (KFCG) [8], [9]:

This algorithm reduces code book generation time.

- Initially we have one cluster with the entire training vectors and the code vector C1 which is centroid.
- In the first iteration of the algorithm, the clusters are formed by comparing first member of training vector with first member of code vector C1.
- The vector Xi is grouped into cluster 1 if xi1< c11 otherwise vector Xi is grouped into cluster 2.

- In second iteration, the cluster 1 is split into two by comparing second member xi2 of vector Xi belonging to cluster 1 with that of the member c12 of the code vector C1.
- Cluster 2 is split into two by comparing the member xi2 of vector Xi belonging to cluster 2 with that of the member c22 of the code vector C2.
- This procedure is repeated untill the codebook size is reached to the size specified by the user.

It is observed that this algorithm gives minimum error and requires least time to generate codebook as compared to other algorithms.

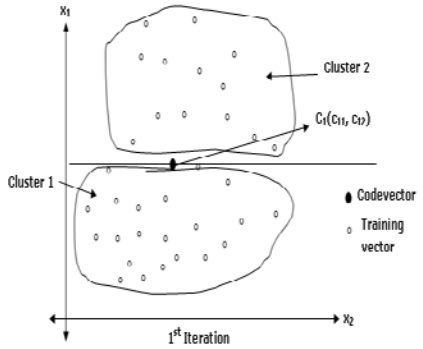

Fig.2a.

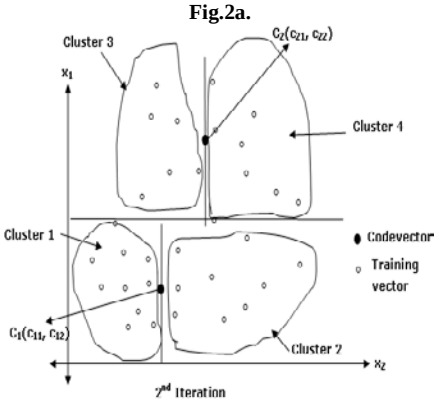

Fig.2b.

Fig. 2 KFCG algorithm for 2 dimensional cases

C. K-Means Algorithms [10]:

Select k random vectors from the training set and call it as code vectors. Find the squared Euclidean distance of all the training vectors with the selected k vectors and k clusters are formed. A training vector Xj is put in ith cluster if the squared Euclidean distance of the Xj with ith code vector is minimum. In case the squared Euclidean distance of Xj with code vectors happens to be minimum for more than one code vector then Xj is put in any one of them. Compute centroid for each cluster. Centroids of each of cluster form set of new code vectors as an input to K-Means algorithm for the next iterations. Compute MSE for each of k clusters. Compute net MSE. Repeat the above process till the net MSE converges. This algorithm takes very long time to converge and to obtain minimum net MSE if we start from random k vectors selection. Instead of this random selection we are giving codebook generated from LBG and KFCG algorithms.

Following are the steps for proposed method

1. Obtain codebook containing k code vectors using LBG or KFCG or any other codebook generation algorithm.
2. Give the above codebook as an input to K-Means algorithm (i.e. Instead of this random selection we are giving codebook generated from LBG and KFCG algorithms).
3.Find the squared Euclidean distance of all the training vectors with the k code vectors and k clusters are formed. A training vector Xj is put in ith cluster if the squared Euclidean distance of the Xj with ith code vector is minimum. In case the squared Euclidean distance of Xj with code vectors happens to be minimum for more than one code vector then Xj is put in any one of them.
4. Compute centroid for each cluster.
5. Compute MSE for each of k clusters and net MSE.
6. Repeat the steps 3 to 5 till the two successive net MSE values are same.

5. Conclusion and Discussions:

The Color Shape for Image Retrieval is Mainly Based on the Content Based Image Retrieval System. And it First Compares the Shape (Sketch) of an Image and then it Compares colors of the given Image with shape matched Images in Dataset. The drawn image without modification cannot be compared with color image, or its edge representation. So, it needs to be preprocessed before Comparison. In this the K-Means Algorithm takes Random Images and compare with query image. It Leads to Number of Iterations more to satisfy the User.

K-means algorithm is an optimization algorithm. It reaches optimal value if there is only one minima. The time taken for the optimal solution depends upon the initial starting point. If there is no prior knowledge of the optimal point one has to start by randomly choosing the initial values. Hence it takes extremely large time for convergence as the initial value is invariably too far off from optimal solution. In this paper we are proposing K-means algorithm for optimization of codebook which already exist so that the convergence time is reduced considerably.

In this project before Comparison of given image with Dataset Images, In the Dataset, Image Features are extracted and based on the features similar images are formed as clusters by using LBG Vector Quantized algorithm or Kekre's Fast Codebook Algorithm. These formed clusters are taken as initial clusters by k-means algorithm to compare the images based on extracted

features. So, by this we can reduce the number of iterations in K-means algorithm to get the result and user can satisfy in less number of iterations. Kekre's Fast Codebook Generation Algorithm is better than LBG Vector Quantized algorithm. Means, from the results it is obvious that KFCG codebook takes lesser number of iteration in most cases as compared to LBG codebook. This indicates that KFCG codebook is closer to the optimum.

References:

[1] B. Sz´ant´o, P. Pozsegovics, Z. V´amossy, Sz. Sergy´an "Sketch4Match – Content- based Image Retrieval System Using Sketches" O´ buda University/Institute of Software Technology, Budapest, Hungary.

[2] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur+ and Marc Alexa "An evolution of descriptors for largescale image retrieval from sketched features lines" TU Berlin + Telecom ParisTech - CNRS LTCI

[3] Gy¨orgy Gyorok "Embedded Hybrid Controller with Programmable Analog Circuit" IEEE 14th International Conference on Intelligence Systems.

[4] John R. Smith and Shih-Fu Chang, "VisualSEEk a fully automated content-based image query" ACM Multimedia 96, Boston, MA, November 20, 1996

[5] Dr. H.B. Kekre, Ms. Tanuja K. Sarode "Vector Quantized Codebook Optimization using K-Means" Dr.H.B. Kekre et al / International Journal on Computer Science and Engineering Vol.1(3), 2009, 283-290

[6] Y. Linde, A. Buzo, and R. M. Gray.: 'An algorithm for vector quantizer design," IEEE Trans. Commun.', vol. COM-28, no. 1, pp. 84-95, 1980.

[7] A. Gersho, R.M. Gray.: 'Vector Quantization and Signal Compressio', Kluwer Academic Publishers, Boston, MA, 1991.

[8] J. Z. C. Lai, Y.C. Liaw, W. Lo, Artifact reduction of JPEG coded images using mean-removed classified vector quantization, Signal Process. vol. 82, No.10, pp. 1375–1388, 2002.

[9] H. B. Kekre, Tanuja K. Sarode, "Fast Codebook Generation Algorithm for Color Images using Vector Quantization," International Journal of Computer cience and Information Technology (IJCSIT), Vol. 1, No. 1, pp: 7-12, Jan 2009.

[10] J.B.MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley symposium on Mathematical Statistics and Probability", Berkely, University of California Press, vol 1, pp281-297, 1967.

[11] Clustering Large Applications, "Data Mining: Concepts and Techniques", by Jiawei Han.

A.Komali, pursuing her (M.Tech in Computer Science Engineering) at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. Her research interest in Image Processing.

R.Veera Babu, Asst.Prof , Department of CSE, Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. His research interest includes Networks and Image Processing.

D.Sunil Kumar, pursuing his (M.Tech in Computer Science Engineering) at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. His research interest in Image Processing.

K.Ganapathi Babu, pursuing his (M.Tech in Computer Science Engineering) at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. His research interest in Image Processing.