

Content Based Image Retrieval using K-Means Clustering Techniques

Shaheen Fatima

Associate Professor of Electronics, Government College (Autonomous), Kalaburagi, Karnataka India.

Abstract :

In this paper we present an image retrieval system that takes an image as the input query and retrieves images based on image content. Content Based Image Retrieval is an approach for retrieving semantically-relevant images from an image database based on automatically-derived image features. The unique aspect of the system is the utilization of k-means clustering techniques. Image retrieval systems, which compare the query image exhaustively with each individual image in the database, We present a clustering based indexing technique, where the images in the database are grouped into clusters of images, with similar color content using a clustering algorithm. At search time, the query image is not compared with all the images in the database, but only with a small subset. Experiments show that this clustering-based approach offers a superior response time with high retrieval accuracy.

Keywords: CBIR, Clustering, K-Means Clustering.

Introduction

Image retrieval is the process of browsing, searching and retrieving images from a large database of digital images. The collection of images in the web are growing larger and becoming more diverse .Retrieving images from such large collections is a challenging problem. One of the main problems they highlighted was the difficulty of locating a desired image in a large and varied collection. While it is perfectly possible to identify a desired image from a small collection simply by browsing, more effective techniques are needed with collections containing thousands of items. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be meta tags, color distribution in images, region/shape attributes, etc. Unfortunately, image retrieval systems have not kept pace with the collections they are searching. The shortcomings of these systems are due both to the image representations they use and to their methods of accessing those representations to find images. The problems of image retrieval are becoming widely recognized, and the search for solutions an increasingly active area for research and development.

In recent years, with large scale storing of images the need to have an efficient method of image searching and retrieval has increased. It can simplify many tasks in many application areas such as biomedicine, forensics, artificial intelligence, military, education, web image searching. Most of the image retrieval systems present today are text-based, in which images are manually annotated by text-based keywords and when we query by a keyword, instead of looking into the contents of the image, this system matches the query to the keywords present in the database [2].

This technique has its some disadvantages:

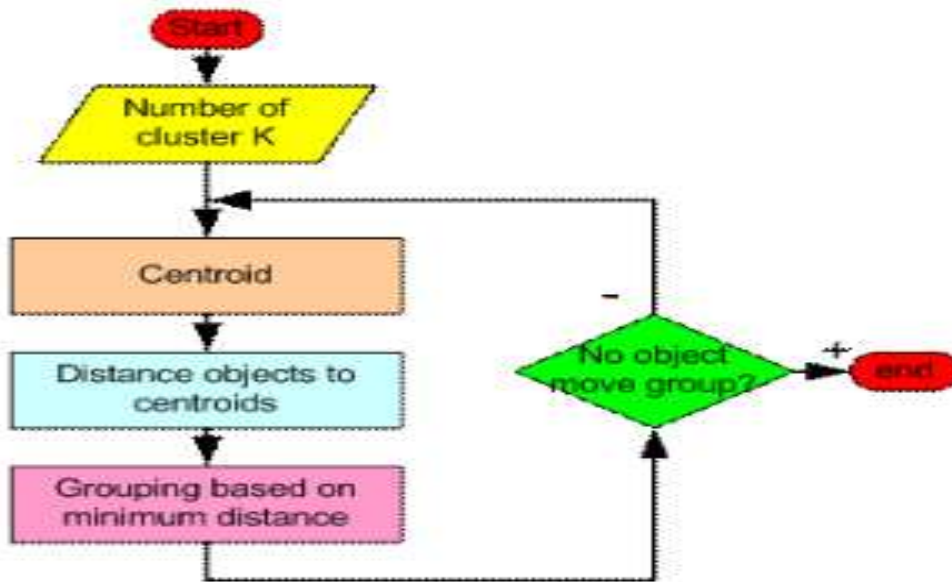
- a) Firstly, considering the huge collection of images present, it's not feasible to manually annotate them
- b) Secondly, the rich features present in an image cannot be described by keywords completely.



These disadvantages of text-based image retrieval techniques call for another relatively new technique known as Content-Based Image Retrieval (CBIR). CBIR is a technology that in principle helps organize digital image archives according to their visual content. This system distinguishes the different regions present in an image based on their similarity in color, pattern, texture, shape, etc. and decides the similarity between two images by reckoning the closeness of these different regions. The CBIR approach is much closer to how we humans distinguish images. Thus, we overcome the difficulties present in text-based image retrieval because low-level image features can be automatically extracted from the images by using CBIR and to some extent they describe the image in more detail compared to the text-based approach [2]. Image classification or categorization has often been treated as a preprocessing step for speeding-up image retrieval in large databases and improving accuracy, or for performing automatic image annotation. Image clustering inherently depends on a similarity measure, image categorization has been performed by varied methods that neither require nor make use of similarity metrics. Image categorization is often followed by a step of similarity measurement, restricted to those images in a large database that belong to the same visual class as predicted for the query. In such cases, the retrieval process is intertwined, whereby categorization and similarity matching steps together form the retrieval process. Similar arguments hold for clustering as well, due to which, in many cases, it is also a fundamental “early” step in image retrieval [3] .

We have used the K-means clustering procedures which can be applied for image retrieval from large databases. Both these clustering algorithms have been frequently used in the pattern recognition literature. Clustering Algorithm is used to group similar images into clusters to increase the retrieval speed. K Means is an iterative refinement heuristic algorithm that works faster. A common method is to run the algorithm several times regain the best clustering found. Color is one of the most widely used features for image similarity retrieval, Color retrieval yields the best results, in that the computer results of color similarity are similar to those derived by a human visual system that is capable of differentiating between infinitely large numbers of colors. One of the main aspects of color feature extraction is the choice of a color space. A color space is a multidimensional space in which the different dimensions represent the different components of color [4] .Most color spaces are three dimensional. Example of a color space is RGB, which assigns to each pixel a three element vector giving the color intensities of the three primary colors, red, green and blue. The space spanned by the R, G, and B values completely describes visible colors, which are represented as vectors in the 3D RGB color space. As a result, the RGB color space provides a useful starting point for representing color features of images.

Proposed Method



Block Diagram for proposed Image Retrieval System

We have used clustering algorithms, the K-means clustering algorithms to group the images into clusters based on the color content. Both these clustering algorithms have been frequently used in the pattern recognition literature. Here we are going to filter most of the images in the clustering and then apply the clustered images to K-Means, so that we can get better favored image results. K-means cluster analysis is a set an algorithm that groups similar objects into groups called clusters. The end point of cluster analysis is a set of clusters, where each cluster is distinct from each other cluster and the object within each cluster are broadly similar to each other. Brief details on the implementation of this clustering algorithms is presented below.

A K-Means Combined Algorithm

- Step 1: Specify the number of clusters (k).
- Step 2: Set K (cluster size) choose the number of clusters.
- Step 3: Calculate centroids for K Clusters (randomly).
- Step 4: Partition the data in K clusters by comparing the data points with centroids.
- Step 5: Allocate objects to clusters.
- Step 6: Update the centroid based on similarity measure.
- Step 7: Compute cluster means.
- Step 8: Allocate each observation to the closest cluster center.
- Step 9: Repeat step 3 to 8 until the solution converges.

Flow chart for K mean algorithm:

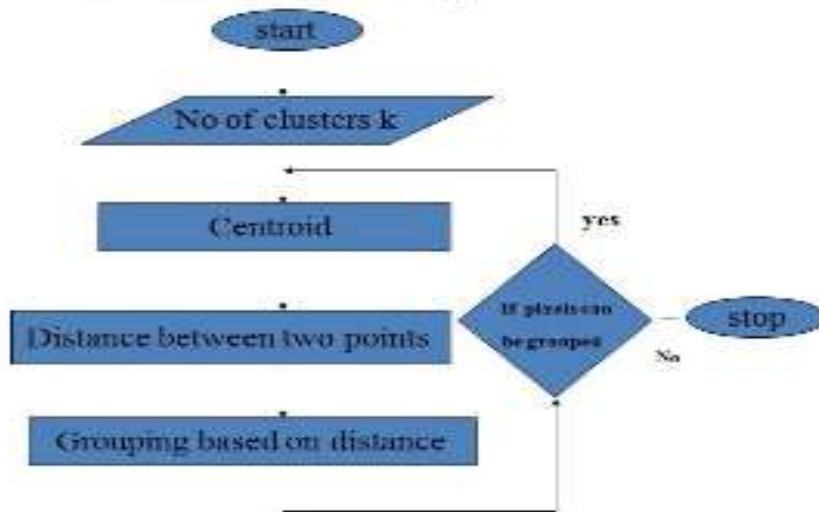


Image Similarity and Measures

Searching large databases of images is a challenging task especially for retrieval by content. Most search engines calculate the similarity between the query image and all the images in the database and rank the images by sorting their similarities.

The retrieval time is the sum of two times: T_{sim} and T_{sort} . T_{sim} is the time to calculate the similarity between the query and every image in the database, and T_{sort} is the time to rank all the images in the database according to their similarity to the query.

$$T_{total} = nT_{sim} + O(n \log n)$$

Where n is the number of images in the database, T_{sim} is the time to calculate the similarity between two images, and $O(n \log n)$ is the time to sort n elements.

When the images in the database are clustered, the time to calculate the similarity between the query and the images in the nearest clusters images.

Therefore the total search time is:

$$T_{cluster} = kT_{sim} + lT_{sim} + O(l \log l)$$

Here k is the number of clusters; l is the number of images in the clusters nearest to the query. Since $k \ll n$ and $l \ll n$, $T_{cluster} \ll T_{total}$ [1].

Retrieval accuracy with Clustering

Clustering is a mutually exclusive partitioning process of the feature space of feature vectors in a meaningful way for the application domain context. With the clusters, we may perform nearest neighbor search efficiently. The unique aspect of this system is the utilization of k -means clustering techniques. Here we are going to filter most of the images in the clustering and then apply the clustered images from the clustering to K-Means, so that we can get better favored image results. After clustering and selecting the cluster centers, the given query image is first compared with all the cluster centers. The clusters are ranked according to their

similarity with the query. Then the query image is compared directly with the images in these clusters. Thus, the number of comparisons is reduced considerably from comparing the query with all the images in the database. The number of similarity comparisons required depends on the sizes of the clusters and the number of clusters being examined [5]. A user instead of searching through a large database is concerned in only clustered image results. Now, we apply clustered images from the clustering to the k-means algorithm which takes the input parameter k , and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high. An object is assigned to the cluster to which it is the most similar one. This object assignment is based on the distance between the object and the center it's closest to. It then computes the new centroid and in this way each center finds the centroid of its own points. This process iterates until the criterion function converges. Thus, the retrieval will be very accurate with the K-Means clustering. It leads to the better performance than by using individual algorithmic methods.

Conclusion

In this paper we have presented an approach for Content Based Image Retrieval using K-Means clustering techniques where images are initially clustered into groups having similar color content and then the preferred group is clustered using K-Means. Clustering assists faster image retrieval and also allows the search for most relevant images in large image databases. K-Means is a clustering method based on the optimization of an overall measure of clustering quality is known for its efficiency in producing accurate results in image retrieval. Since each cluster obtained is a unique set of similar images, the user can select an image set of his choice and further refine the search by applying K-Means technique. Thus using K-Means techniques together not only facilitates the user not to overlook the image he may require but also to obtain accurate favored image results.

References

- [1.] Santhana Krishnamachari, Mohamed Abdel -Mottaleb, Hierarchical clustering algorithm for fast image retrieval, Part of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII, San Jose, California, January 1999. pp427-435.
- [2.] Subhankar Biswas ,A system for Content-Based Image Retrieval.
- [3.] 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.
- [4.] Daniela Stan & Ishwar K. Sethi, Image Retrieval Using a Hierarchy of Clusters.
- [5.] Mohamed Abdel-Mottaleb, Santhana Krishnamachari, Nicholas J. Mankovich, "Performance Evaluation of Clustering algorithms for scalable image retrieval", Appeared in IEEE Workshop on Empirical Evaluation of Computer Vision Algorithms, CVPR 1998.