Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Image Retrieval: K-Means and Contribution Based Clustering

Mr.S.P. Bhonge*
Asst. Prof. Dept. Of E & TC,
P. R. Pote (Patil) College of Engg. & Mgt.
Amravati, India

Mr.A.K. Sapkal, Mr.S.S. Jain, Mr. P.K. Gakare
Asst. Prof. Dept. Of E & TC,
DattaMeghe Institute of Engg.
Technology & Reaserch Wardha, India

Mr. Arun Katara
Asst. Prof. Dept. of EE ,
DattaMeghe Institute of Engg.
Technology & Reaserch Wardha, India

Abstract—Contribution based clustering algorithm used to retrieve image from large scale image database by measuring common visual features of the images. The resemblance between images is measured by measuring similarity within extracted features. Statistical feature extraction technique is used for the texture feature extraction. Clustering is an unendorsed classification method which place the similar object in a cluster and dissimilar objects are place in different clusters. Contribution based clustering algorithm is partition based clustering algorithm which gives the better results than that of k-mean clustering algorithm. In this paper, contribution based clustering approach for minimum distance findings among images are used. The results of experimental study of proposed algorithm are shown with analysis of resultant image features. The images are retrieved based on selection of images with maximum similarity features.

Index Terms— Contribution Based Clustering, GLCM, K-means clustering

I. INTRODUCTION

A substantial amount of image data has been produced in diversified areas due to modernisation in engineering and science practices. It becomes difficult and imperative problem in searching images from varying collection of image features [1]. Clustering is the process of grouping the similar type of objects or data points in one group while the objects in different groups are less similar. There are number of algorithms available for clustering which differs in there process of finding the clusters. Well-liked methods of clustering comprise groups with low distances among the cluster members. Clustering as such is not an automatic task, but it is an iterative process of knowledge detection or interactive multi-objective optimization that involves examination and failure [1]. Clustering can therefore be formulated as a multi-objective optimization problem. The major problems occurs in clustering are

- Distance function to be used for calculating the similarity among the data points.
- Selection of optimal threshold for clustering
- Expected number of clusters
- As the data sets are constantly becoming larger, it prevents easy analysis and validation of results.

There are two main challenges in captivating the concept of inferring common visual themes to creating a scalable and effective algorithm. The first challenge involved image processing required and seconds the need of evolving the mechanism for retrieval of images based on their similarity matches [4].

The transformations of raw pixel data to a small set of image regions were provided to image retrieval by applying segmentation. Regions are coherent in colors and texture. These region properties were used for image retrieval [6]. The descriptor and detector were developed for faster computations and comparisons. It was found that the correspondence between two images with respective repeatability, distinctiveness and robustness was helpful. Here corners, blob and T-junction of images were considered or selected as point of interest, then feature vector was created having representation of neighbourhood of every interest point. Lastly minimum distances were found by measuring Euclidian distance and depending on minimum distance matching between different images were carried out [2]. W. Zhou *et al.* provide canonical image selection by selecting subset of photos, which represents most important and distinctive visual word of photo collection by using latent visual context learning. In canonical image selection, images were selected in greedy fashions and used visual word of images and Affinity propagation [8] clustering for similarity findings. In this paper image retrieval methods using k-means and contribution based clustering for Image retrieval is covered. This is followed by experimental results and discussions worth.

II. Feature Extraction

To ensure the usefulness of k-means and contribution based algorithm for image retrieval in real sense, experiments were conducted using MatLab 7.10 environment on the images collected directly through Google Image. It was concentrated

on the 751 medium size image databases with seven different query images. In these four images from collection of database images were retrieved based on their Texture and Color features.

A. Feature Generation and Representation

The texture features were measured using Gray-Level Co-occurrence Matrix (GLCM), It considered the spatial relationship of pixels. The number of occurrence of pixel pairs with certain values and specified spatial relationship occurred in an image provides characteristics of texture values by creating GLCM [9].

Normalized probability density $P_\delta(i, j)$ of the co-occurrence matrices can be defined as follows.

$$P_\delta(i,j) = \frac{\#\{[(p,q),(p+r,q+r)] \in G | f(p,q)=i, f(p+r,q+r)=j]\}}{\#G} \quad (1)$$

Where, p, q = 0, 1…M-1 are co-ordinates of the pixel, i, j = 0, 1…L-1 are the gray levels, G is set of pixel pairs with certain relationship in the image. The number of elements in G *is* obtained as *#G*. r is the distance between two pixels i and j. P_δ *(i, j)* is the probability density that the first pixel has intensity value i and the second j, which separated by distance δ*= (rp, rq).* [9]

Energy measures textural uniformity i.e. pixel pairs repetitions. Energy is ranging 0 to 1 being 1 for a constant image. It returns the sum of squared elements in the GLCM. Energy is given by

$$Energy = \sum_{i,j} P_{(i,j)}^2 \quad (2)$$

Contrast is the difference in luminance and color that makes an object distinguishable. It measures the local variations in the Gray-Level Co-occurrence Matrix. Contrast is 0 for a constant image and it is given by

$$\text{Contrast} = \sum_{i,j} |i-j|^2 P_{(i,j)} \quad (3)$$

A correlation function is the correlation between random variable at two different points in space or time, usually as a function of the spatial or temporal distance between the points.

$$\text{Correlation} = \frac{\sum_{i,j} (i-\mu i)(j-\mu j) P_{(i-j)}}{\sigma_i \sigma_j} \quad (4)$$

Where μ_i, μ_j, σ_i, σ_j are the means and standard deviations of P_i and P_j respectively. *Pi* is the sum of each row in co-occurrence matrix and P_j is the sum of each column in the co-occurrence matrix.

Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. It has Range from 0 to 1 and homogeneity is 1 for a diagonal GLCM. Homogeneity is given by

$$Homogeneity = \sum_{i,j} \frac{P_{(i,j)}}{1+|i-j|} \quad (5)$$

Color features contain values of R, G and B. For better results rather taking color feature matching test for complete image, divided it into eight subregions. So that color feature contain in 8 × 3 matrix, measured values of R, G, B for 8 subregions as shown in Fig. 1.

Fig. 1 Color Feature Extraction from Small Regions of Image

B. Effecting Clustering

K-means is commonly used simplest algorithm which employs the square error criterion. In this algorithm the number of partitions is initially defined. The cluster centers are randomly initialized for predefined number of clusters. Each data point is then assigned to one of the nearest cluster. The cluster centers are then re-estimated and new centroid is calculated. This process is repeated until the convergence has been reached or until no significant change occurs in cluster center [5]. K-means is easy to implement and its time complexity is less. But the output of k-means algorithm depends on selection of initial number of clusters as shown in Fig. 2. If the initial number of clusters is not properly chosen then the output of algorithm may converge to false cluster locations and completely different clustering result [3] [10].The following figure shows how the output of k-means clustering is depends on the selection of initial number of clusters.

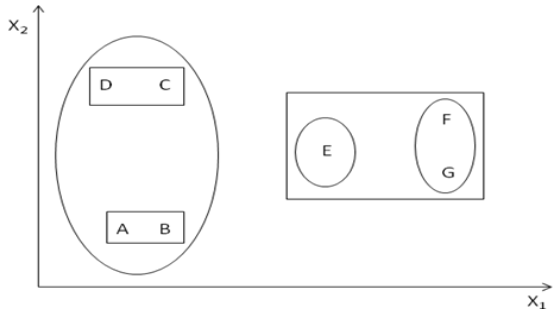


Fig. 2 k-means Algorithm is Sensitive to the Initial Partition

If the algorithm select A, E and F as initial means then the resulted clusters will be {{A, B, C, D}, {E}, {F, G}} represented by ellipse. The square error criteria value is much larger for this partition than for the partition {{A, B}, {C, D}, {E, F, G}} represent by the rectangles, if we choose A, C and E as initial means.
Fig. 3 shows K-means clustering flowchart. Where, k is the number of clusters and x is the number of centroid.

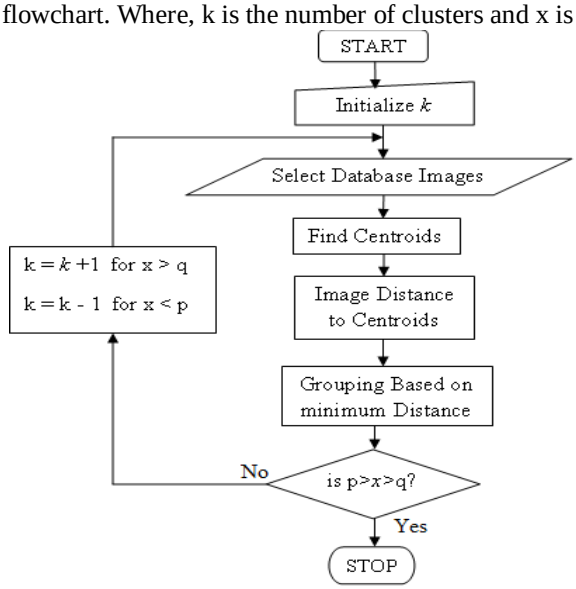


Fig. 3 Flowchart for K-means Clustering

The main idea is to define k centroid for k clusters, one for each cluster. The better choice is to place them as much as possible far away from each other. Here we initially made two centroids.
In contribution based clustering algorithm, inter-clustering distance is minimized and intra-cluster is maximized by finding positive and negative contribution points [7]. The When query was fired then based on query and cluster features, query finds the group of similar images having minimum image distance.

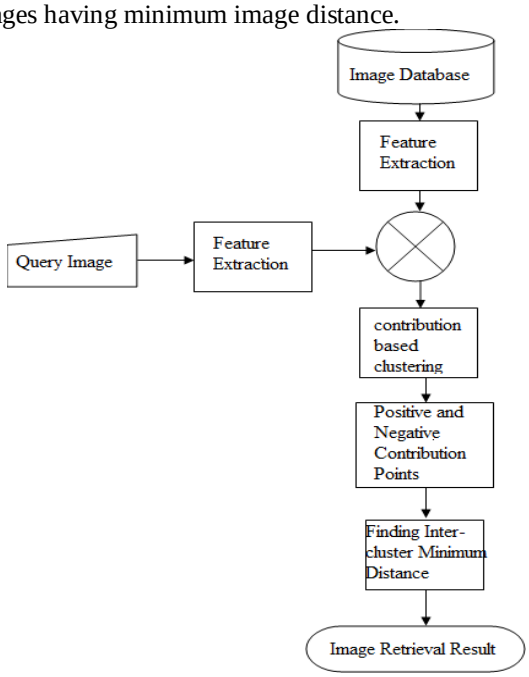


Fig. 4 Contribution Based Clustering Algorithm Flowchart

The retrieval results are returned based on minimum distance between the images inside cluster with query image.

III. RESULTS AND DISCUSSIONS

In feature extraction, the color features were measured by dividing original images into 16 sub regions and color feature contains R, G and B components. Due to which each sub region having 1×3 values of color feature, so 16 subregions are containing 16×3 values. Total 48 values for entire image are measured.

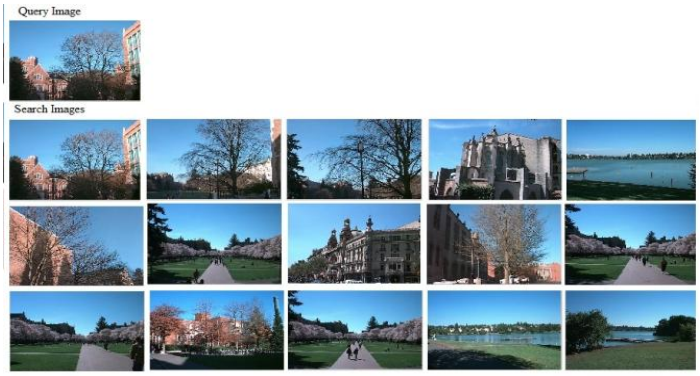
The Gray Level Co-occurrence Matrix (GLCM) was computed in four directions for 0^0, 45^0, 90^0 and 135^0. Based on the GLCM four statistical parameters energy, contrast, correlation and homogeneity were computed in four directions at four points, so total 64 values of texture features are returned for each image.

After completing feature extraction and storage of database images, query image was fired and same features of query image were measured. Fig. 4 shows the image retrieval results for k-means and contribution based clustering. The energy, contrast, correlation and homogeneity were having total 64 texture values, but single color (RGB) feature was containing total 48 values for each image. Comparing to texture features, color feature were highly matched among query image and retrieval images as shown in Fig.4 (e). The effectiveness of k-means and contribution based clustering measured by calculating precision and recall value.

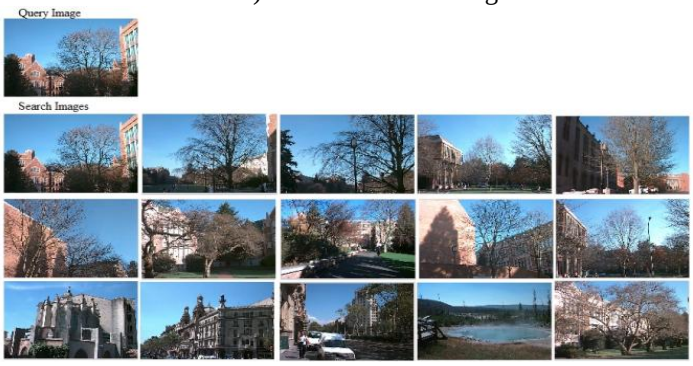
$$\text{Precision} = \frac{Total\ number\ of\ relevant\ retrieved\ images}{Total\ number\ of\ retrieved\ images}$$

$$\text{Recall} = \frac{Total\ number\ of\ relevant\ retrieved\ images}{Total\ number\ of\ relevant\ images}$$

In our work precision and recall value are measure for different number of clusters with their maximum and average values as shown in Fig. 5.

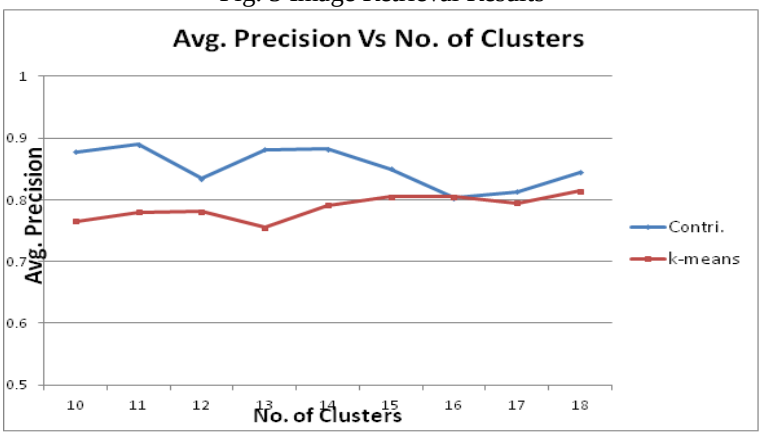


a) k-means clustering

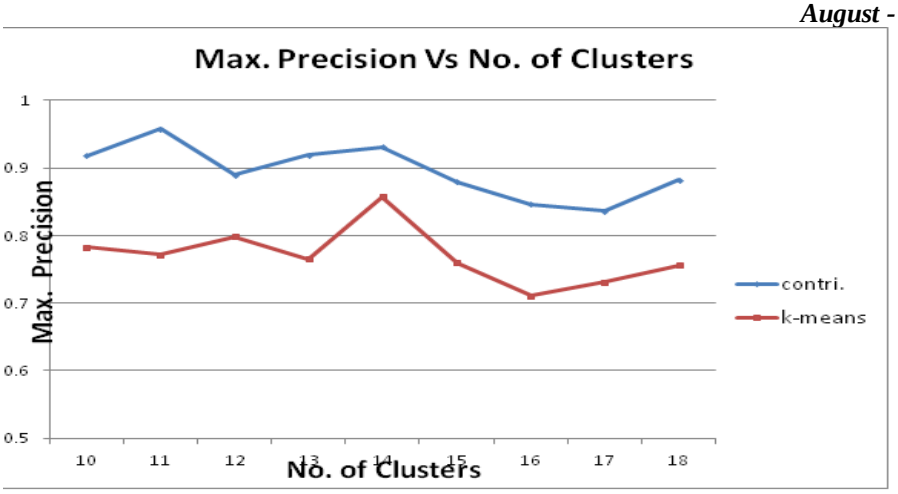


b) contribution based clustering

Fig. 5 Image Retrieval Results



a) Avg. Precision Vs no. of Clusters



b) Max. Precision Vs no. of Clusters

IV. CONCLUSION

The contribution and k-means clustering provide simple mechanism for image retrieval by taking in to account minimum distances among the images. After using Contribution based clustering, the relevant images were returned are more than k-means clustering and if irrelevant images present are returned at the bottom in image search results. The similarity measurement of images was based on the common visual feature between the images.

In future perspective, the numbers of clusters are selected dynamically also the convergence of algorithm is fast. In the real time operation where the human interaction reduces the speed of execution. This algorithm can be used for such fully automated real time operations.

REFERENCES

[1] S. Anitha, Akilandeswari. J and Sathiyabhama. B, "A survey on partition clustering algorithm", *International Journal of Enterprise Computing and Business System International Systems, vol. 1, pp. 1-13*, 2011.

[2] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026-1038, Aug. 2002.

[3] M. Ferecatu, "Image retrieval with active relevance feedback using both visual and keyword-based descriptors", Ph. D. Thesis, University of Versailles Saint-Quentin-En-Yvelines, France.

[4] Jain. M and Singh S. K., "A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques for Large Data sets", *International Journal of Managing Information Technology (IJMIT) Vol.3, No.4, pp. 23-39*, November 2011.

[5] Wang Juntao and Su Xiaolong, *An improved K-Means clustering algorithm* School of computer science and technology china university of mining and Technology, pp.44-46, 2011.

[6] H. Bay, T. Tuytelaars, and L.V. Gool, "Surf: Speeded Up Robust Features," *Proc. Ninth European Conf. Computer Vision, pp. 404-417, 2006.*

[7] Narasimhan, Ramraj, 'Contribution-Based Clustering Algorithm for Content-Based Image Retrieval', *5th International Conference on Industrial and Information Systems, ICIIS 2010, pp. 442-447*, 2010.

[8] W. Zhou, Y. Lu. H. Li and Q. Tian. "Canonical Image Selection by Visual Context Learning" *International Conference on Pattern Recognition 2010.*

[9] Dr. H. B. Kekre, S. D. Thepade, T. K. Sarode and V. Suryawanshi, 'Image Retrieval using Texture Features extracted from GLCM, LBG and KPE', *International Journal of Computer Theory and Engineering*, **2**(5), October, 2010.

[10] W. Triggs, "Detecting key points with stable position, orientation and scale under illumination changes," in *Proceedings of the European Conference on Computer Vision*, vol. 4, pp. 100–113, 2004.

