A Survey on: Image Retrieval Based on Featuresand Clustering Techniques

N . Silpa
CSE Department, KMMITS,
Tirupathi, India.

K . Santhi
CSE Department, S V College of Engineering,
Tirupathi, India.

***Abstract* – Imagescan hide valuable information. Theneed for image retrieval is high in view of the fastgrowing amounts of image data. Image mining dealswith the extraction of image patterns from a largecollection of images in database. Clearly, imagemining is different from low-level computer visionand image processing techniques because the focusof image mining is in extraction of patterns fromlarge collection of images according to user queries,whereas the focus of computer vision and imageprocessing techniques is in understanding and / orextracting specific features from a single image.In image mining, the goal is thediscovery of image patterns that are significant in agiven collection of images as per user queries.In this paper the clustering techniquesare discussed and analyzed. Also, we propose a method HDK that uses more than one clustering techniqueto improve the performance of image retrieval.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.**

Keywords–Image retrieval;color;texture;shape;association rule;MST;divide and conquer k-means;hierarchical; HDK;

I. INTRODUCTION

The image mining was introduced to extract implicitknowledge, image and data relationship. Imagemining is an extension of data mining. In text basedimage retrieval system only find out the images thoseare just concerned with the accurate text that isdescribed by human or relevant query, insteadwithout looking into the content of related images.Images have many more duplications and user is notaware about it.WWW having largest global imagerepository. So remove this drawback with the help of Image retrieval. The users are always not satisfied with the given technologies they used in present time they alwayslook forward for further enhancement. The CBIR focuses on image features. The Features are furtherclassified as low-level and high –level features. User simply put the query regarding that features such ascolor, shape, region etc. and retrieve the required images. After that we have to focuses on clustering tocombine the related images into one cluster and other images into another cluster for fast retrieval.

II. OVERALL PROCESS OF IMAGE MINING

Image mining overall process can be divided into the following parts:-

A. Data preprocessing

A lot of dirty and noisy data exist in large image databases, for instance, images that are extremely unclear. Those data often cause chaos in miningprocess and give birth to worse mining results, so it isNecessary to preprocess data, clean up the noisy,dirty data to highlight the features of that image.

B. Extracting multi-dimensional feature vectors

Using image processing technologies such as imageSegmentation, picking up the edge to extract task relatedfeature vectors, form multi-dimensionalfeature vectors.

C. Mining on vectors and acquire high-levelknowledge

Various methods such as object recognition, imageindexing and retrieval, image classification andclustering, neural network are used on feature vectorsfor mining and acquiring hidden and valuable high-levelknowledge, then evaluate and explain that exactquery related knowledge.

III. LITERATURE REVIEW

Various researches have been carried in Image mining .In this section we present a surveyon different image retrieval using features and clustering techniques.

A. Color-based retrieval

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

B. Texture –based retrieval

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, orwater, 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 measures the image texture such as contrast, coarseness, directionality and regularity [1] or periodicity, directionality and randomness [2].

C. Shape-basedretrieval

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 [3] 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.[4]

D. Clustering-based retrieval

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 [5] [6]. 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 [7], [8]. This paper reviews and summarizes different clustering techniques.

- *Log –Based Clustering*

  Images can be clustered based on the retrieval system logs maintained by an information retrieval process [9]. 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[10]. 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.

- *Association rule*

  Association rule deals with the extraction of image pattern from a large database of images. This method help us for Prediction We will discuss this with an example if sky contains black clouds so there are 64% chances it will rain.

  The method is as follows:-

  - It segmented the images into blobs (regiondescriptor) where blob is equal to an object.
  - Compare blob with all other blobs with an id.This works as a pre-processing algorithm
  - After That Create Auxiliary images with identified objects.
  - Apply data mining techniques to produce object association rule.

  Basically this technique used for selecting images for a particular field (eg. Weather, medical images).

- *Minimum spanning tree –based clustering*

  The minimum spanning tree clustering algorithm is proficient of detecting clusters with irregular boundaries. The author presented a minimum spanning tree depending on the clustering technique using weighted Euclidean distance for edges, which is vital constituent in constructing the graph from image. The technique constructs 'k' clusters with segments. This approach is very much capable of protecting detail in low variability image regions while not considering detailin high variability regions which is the mainadvantage of this approach. This approach hashandled the problems of undesired clusteringstructure and redundant huge number of clusters.Effective research in the field of image retrieval andmining has turned out to be a significant researcharea because of significant applications in digitalimage databases.

- *Hierarchical clustering*

  Hierarchical clustering is an agglomerative (top down) clustering method. As its name suggests, the idea of this method is to build a hierarchy of clusters, showing relations between the individual members and merging clusters of data based on similarity. In the first step of clustering, the algorithm will look for the two most similar data points and merge them to create a new "pseudo-datapoint", which represents the average of the two merged datapoints. Each iterative step takes the next two closest datapoints (or pseduo-datapoints) and merges them. This process is generally continued until there is one large cluster containing all the original datapoints. Hierarchical clustering results in a "tree", showing the relationship of all of the original points.

- *Graph theory- based clustering*

  The concepts and properties of graph theory [11] 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 withdata 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 MinimumSpanning Tree based clustering [12], [13],and clustering editing method [14] [15], HCS algorithm [16], etc. Current research is focused on clustering using divide and conquers approach [17]. Usually this clustering methodology is used to detect irregular clustering boundaries in clustering results. Zhan [12] 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[18] is applied to each edge to detect and remove the inconsistence edges, which results as a set of disjoint sub trees, each sub tree will represent a separate cluster.

- *Divide and Conquer K-Means- based clustering*

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 [19, 20]. 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 [21][22].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 methods that can largely accelerate the CBIR system.

- *HDK Algorithm*

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 samplesof 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 distanceto 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 numberof clusters: 0(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.

This includes three main steps.They are

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

IV. FUTURE ENHANCEMENT

Image mining is an extension of data mining technique. Most of the image processing algorithms include image mining. Therefore, image mining is always an emerging field and it has attracted a lot of researchers to investigate its applications in recent years.

V. CONCLUSION

This paper presents a survey on various image mining techniques that was proposed earlier. The purpose of this survey is to provide an overview of the functionality of image retrieval. 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 forhigh dimensional data set. In this paper we present a method HDK to use both advantages of HCand 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 oncolor feature helps us to improve accuracy and efficiency of original K-Means clustering. For thispurpose we should consider orthogonal space. HDK algorithm has been used extensively invarious areas to improve the performance of the system and to achieve better results in differentapplications.

VI. REFERENCES

[1] Tamura et al. "Texture Features Corresponding to Visual Perception"- IEEE Trans on system, Manand cyber 8-460-472-1978.
[2] Sanjay Kumar Saha et al. "CBIR Using Perception Based Texture And Colour measures "CSEDepartment; CST Department Jadavpur Univ., India; B.E. College, Unit ISI, Kolkata, India -2003.
[3] Canny, J., "A computational approach to edge detection", IEEE Trans on Pattern Analysis andMachine Intelligence, 8:679-698, 1986.
[4] S.Nandagopalan, Dr. B.S. Adiga, and N. Deepak "A Universal Model for Content-Based ImageRetrieval" World Academy of Science, Engineering and Technology 46 2008.
[5] B. Everitt, S. Landau, and M. Leese, "Cluster Analysis", London:Arnold, 2001.
[6] A. Jain and R. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, NJ: Prentice-Hall, 1988.
[7] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks—Part I and II," IEEE Trans. Neural Netw., vol. 13, no. 3, 645–677, May 2002.
[8] V. Cherkassky and F. Mulier, "Learning From Data: Concepts, Theory, and Methods", New York:Wiley, 1998
[9] Huiyu Zhou, Abdul H. Sadka, Mohammad R. Swash, JawedAzizi and Abubakar S. Umar., "ContentBased Image Retrieval and clustering: A Brief Survey" school of Engineering and Design, BrunelUniversity, Uxbridge, UB8 3PH, UK
[10] Hoi, C.-H. andLyu, M. R. 2004a. "Group-based relevance feeedbacks with support vector machineensembles" In Proc. IEEE ICPR, 2004.
[11] F. Harary, *Graph Theory*. Reading, MA: Addison-Wesley, 1969
[12] C.T Zahn(1971): Graph-theoretical methods for detecting and describing clusters, IEEE trans onComputers c(20):68-86
[13] 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
[14] J.Gramm; J.Guo(2003): Graph modeled data clustering: fixed parameter algorithms for clique generation, In Lecture Noted on Computer Science(LNCS),pages 109-118,Springer
[15] R. R.Shamir; D.Tsur(2002): Cluster graph modification problems. In Lecturer notes in computerscience(LNCSI),pp379-390,springer
[16] E. Hartuy; R. Sharmir. A clustering algorithm based on graph connectivity. Informationprocessing,pp175-181
[17] Xiaochun Wang; D. Mitchell Wilkes (2009): "A Divide-and-Conquer Approach for MinimumSpanning tree-Based Clustering", Member,IEEE Transactions on knowledge and data Engineering,Vol 21 No7
[18] Oleksandr; Grygorash(2006); Yan Zhou: "Minimum Spanning tree Based Clustering", IEEE Toolswith artificial intelligence,pp 3-81
[19] 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.
[20] A. Jain, M. Murty, and p. Flynn " Data clustering: A review." ACM Computing Surveys, vol. 31,pp. 264-323,1999.
[21] L. Parsons, E. Haque and H. Liu., "Subspace clustering for high dimensional data: a review."*SIGKDD Explor. Newslett*. 6 (1), pp. 90–105. 2004.
[22] C. Bouveyron, S. Girard and C. Schmid, "High-dimensional data clustering. Computational Statistics& Data Analysis", vol. 52, pp.502–519. 2007.