Efficient Modeling of Visual Art Color Image Clustering

Y. Poornima
Research Scholar,
Vinayaka Mission University Salem, India

P.S. Hiremath, Ph. D
Professor, Dept. of Computer Science
Gulbarga University, Gulbarga, India

ABSTRACT

Although there has been massive research work being conducted in the area of content-based image retrieval (CBIR) system using various sophisticated techniques, very little work has been witnessed for visual art images. From the literatures, it has also been witnessed that clustering algorithm has played a big role in justifying the outcome of various CBIR system. The objective of the proposed system is to introduce a new clustering technique which is implemented over a large set of visual art images. The proposed algorithm is implemented and its performance is measured with respect to two performance parameter namely,. recall and precision. The accomplished outcome of the study is also compared with two conventional clustering techniques that are frequently seen on literatures to understand where the proposed system stands. The accomplished results were seen to outperform conventional clustering technique.

Keywords

Content based image retrieval system, visual art image, K-Means algorithm, Clustering Techniques, Block Truncation Coding.

1. INTRODUCTION

Content-based image retrieval(CBIR) has been proposed as a viable alternative to text-based image retrieval. The objective of the retrieval of images are largely on the basis of automatically extracted visual features such as intensity, color, texture and shape, the strength of content-based image retrieval mainly stems from its ability to search for an image depending on metrics for comparing image or structure properties that can match human judgments of similarity. [1] Response times in the interaction between computer systems and human users are of great importance to user satisfaction. At present, this fact is not widely explicitly addressed in CBIR: many authors discuss mechanisms for reducing search time, but few quote actual times. The goal should be a suitable trade-off between response time and the quality of the results. Most current CBIRSs represent images as points in a multidimensional feature space. Image similarity is defined as the Euclidean or Mahalanobis distance between points in this vector space. The issue in focus of the proposed system is based on pruning problems from large datasets concerning visual art images retrieval system. The Optimization Problem refers to find the transformation that minimizes the dissimilarity between the transformed pattern and the other pattern. The Computation Problem relates for computing the dissimilarity between the two patterns of visual art images. Sometimes the time complexities to solve such issues are rather high, so that it makes sense to devise approximation algorithms:

- **Shape Approximation and Simplification:** construct a shape of fewer elements (points, segments, triangles, etc.), that is still similar to the original.
- **Shape Retrieval:** search for all shapes in a typically large database of shapes that are similar to a query shape.
- **Shape Alignment and Registration:** transform one shape so that it best matches another shape (optimization problem), in whole or in part.
- **Approximate Optimization Problem:** To find a transformation that gives dissimilarity between the two patterns that is within a constant multiplicative factor from the minimum dissimilarity. These problems play an important role in the following categories of applications.
- **Shape Recognition and Classification:** determine whether a given shape matches a model sufficiently close (decision problem), or which of k class representatives is most similar (k computation problems).

As discussed by Antani S et al. [1], a cluster is a collection of data points that are similar to one another within the same cluster and dissimilar to data points in other clusters. Clustering is a method of unsupervised classification, where data points are grouped into clusters based on their similarity. The goal of a clustering algorithm is to maximize the intra-cluster similarity and minimize the inter-cluster similarity.

In this paper, the objective is to propose a new clustering algorithm, which would come under the category of partitional clustering algorithms. The notion of ‘contribution of a data point’ is used for partitional clustering. The resultant algorithm requires only three passes and the time complexity of each pass is same as that of a single iteration of the k-means clustering algorithm. While the k-means algorithm optimizes only on the intra-cluster similarity, the proposed algorithm also optimizes on the inter-cluster similarity. The organization of the remaining part of the paper is as follows. In Section 2, an overview of related works is given. Section 3 highlights the proposed model. The implementation and results are discussed in Section 4. The comparative performance analysis of the proposed method with existing method is also given. Finally, in Section 5, the conclusions are presented.

2. RELATED WORK

The various clustering techniques that have been introduced in the literature for the purpose of pruning feature set in CBIR are discussed.

Murthy et al. [2] have presented an image retrieval system that takes an image as the input query and retrieves semantically-relevant images from an image database based on automatically-derived image features. The unique aspect of the system is the utilization of hierarchical and k-means clustering techniques. This procedure consists of two stages. Firstly, filtering most of the images in the hierarchical clustering and then, secondly, applying the clustered images to K-Means, so that better favored image results are obtained.

Tonge [3] have illustrated an approach for Content-based image retrieval system using k-means clustering. Clustering is very efficient and powerful technology to handle large data sets. It assists faster image-retrieval and also allows the search for most relevant images in large image database. 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. By using k-means user can select the closer group of image so that they gate fast result.

Analoui and Beheshti [4] are concerned with an open problem in the area of Content Based Image Retrieval (CBIR) and present an original method for noisy image data sets by applying an artificial immune system model. In this regard, appropriate feature extraction methods, in addition to a beneficial similarity criterion, contribute to retrieving images from a noisy data set precisely. The results show some improvement and resistance in the noise tolerance of content based image retrieval in a database of various images.

Samathal and Mohanraj [5] have proposed a framework of unsupervised clustering of images based on the color feature of images. Test has been performed on the feature database of color moments and Block Truncation Coding (BTC). The K-means clustering algorithm is applied over the extracted dataset. Results are quiet acceptable and show that performance of BTC algorithm is better than color moments.

Murthy et al. [6] have presented an image retrieval system that takes an image as the input query and groups all the images in the database based on their similarity. This helps the user to keep a faster track of his required images, after which he can opt to retrieve images from the group that would be closest to his visual interpretation. The unique aspect of the system is the utilization of hierarchical and k-means clustering techniques. The proposed procedure consists of two stages.

Balan and Devi [7] have carried out the study and analysis of image mining, image retrieval, image clustering of textile images. The retrieval method is designed based on relevance feedback, color layout, scalable color and edge histogram. The algorithm for image clustering is based on k-means algorithm and the developed software prototype allows one to retrieve the images of the textile based on categories such as shirts, t-shirts, pants and sarees and color descriptors.

Quynh et al. [8] have presented a technique to improve the retrieval process by image regions matching. They carried out an experiment on an image database containing 8000 images. The experimental results show that their proposed technique is more effective than the other retrieval techniques such as color histogram based and color based clustering based techniques.

Malakar and Mukherjee [9] have proposed a new multi feature image clustering technique which will help us to classify the large volume data with high accuracy level. Firstly, they extract color moments feature from an image, and then they consider histogram analysis and make a summation of each color bin. They have used canny edge detection technique. Lastly they combine all features in a matrix and perform clustering algorithm to cluster data.

Komali and Babu [10] have dealt with the important theme of using CBIR which 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, they use image search tools like Google images, Yahoo, etc. The main goal is based on the efficient search on information set. In the point of searching text, they can search flexibly by using keywords. However, if images are used, then search is done using some features of images, and these features are the keywords.

Jadhav and Ahmad [11] have demonstrated an effective approach to retrieve the images by using the low level and high-level features is proposed. Initially, the low-level features such as color, texture, shape, and homogeneity are extracted from the database images and from the query images. After that, the images which are relevant to the given query image are retrieved from the database based on these low-level features and high level features. By means of the query, keyword, which is a high-level feature, is generated and then by using this query keyword the images, which are relevant to the query keyword, are extracted.

Ravindran and Shakila [12] have proposed a strategy for automatically extracting visual patterns from a histology image collection. The foundation of the method is a Bag-of-Features (BOF) representation that builds a codebook which gathers the building blocks that explain the visual content of the image collection. A state-of-the-art feature selection process is applied to find a set of discriminative codeword's. The codeword's are related to high-level concepts individually, using conditional probabilities, and collectively using bolstering.

Raghatate and Janwe [13] have proposed technique they use clustering methods to cluster the images based on their low level visual features. Here the number of comparisons is reduced. Then the similarity measure between the clustered images and the query image are finding out and get the retrieved results. It works faster than the previous approaches. Thus using hierarchical and K-Means techniques together not only facilitates the user not to overlook the image he may require but also to obtain accurate refined Image results.

3. PROPOSED SYSTEM

The main purpose of the proposed system is to design a feature set pruning for content-based image retrieval by considering visual art images. A simple procedure of querying a content-based image retrieval system is to offer any visual art images. The system then recovers all visual art images in the database that are equivalent in content with the queried visual art image. In this proposed system, the prime objective of study the design of a framework for feature clustering to achieve content-based image retrieval. A massive collection of visual art images is partitioned into a number of visual art image clusters. Given a queried visual art image, the proposed framework will attempt to recover all visual art images from the cluster that is nearest in content to the queried visual art images. The schematic diagram of the proposed system is exhibited in Fig. 1. The contribution-based clustering algorithm is applied to visual art image retrieval and compares its performance with that of the conventional clustering mechanism.

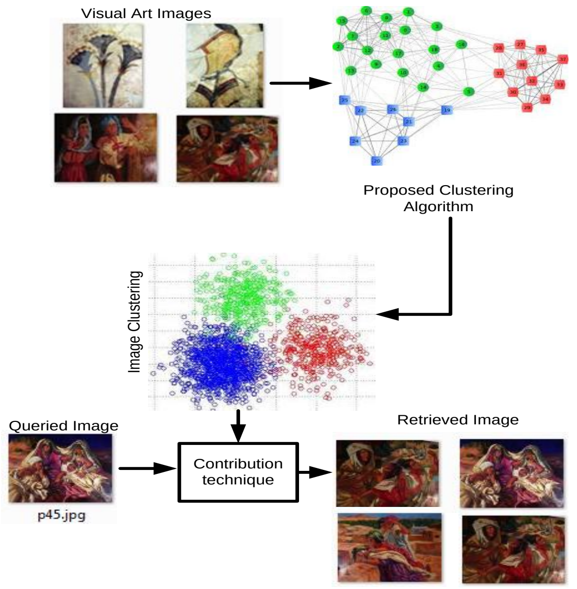

Fig 1: Proposed CBIR System

The proposed system considers each visual art images in the database to be represented by a visual content descriptor consisting of a set of visual features. The system then chooses to use a similarity / dissimilarity score to recover visual art images whose features are very nearest to that of the queried visual art images. A common distance / dissimilarity metric of Euclidean distance is considered in this study. In order to represent the visual content of a visual art image, the system uses a RGB color histogram where the color coordinates of the RGB color space are consistently quantized into a number of bins.

The proposed algorithm for CBIR is based on partitional clustering that aims at partitioning a group of data points into disjoint clusters optimizing a specific criterion [14]. When the score of clustering data points is large, a brute force method that considers all possible combinations would be computationally expensive. Instead, heuristic methods can be deployed to investigate the optimal partitioning issues. Therefore, the criterion function deployed for partitional clustering is the sum of squared error function represented as follows

$$\gamma = \sum_{i=1}^{N_C} \sum_{D_p \in CL} (D_p - Cent_i)^2 \qquad (1)$$

Where, Nc is the quantity of defined clusters, CLi is the ith cluster, Dp is a data point and Centi is the centroid of the ith cluster. The formulation of equation (1) is based on the frequently used squared-error based algorithm [14]. The contribution of a data point is illustrated to be a data point belonging to a cluster as the influence that it has on the quality of the cluster. This factor is then utilized to attain a best set of ‘Nc’ cluster from the pre-defined set of data points in the proposed formulation. The idea of the term ‘contribution’ is basically derived from the work done in [15] where the clustering is done using data-mining concept along with game theory. The formation of cluster is mapped to association formation in cooperative games and has utilized the resultant concept of shapely value to investigate the best quantity of clusters for a predefined set of data points. While this work uses the concept of contribution to find the best possible number of cluster, the proposed system deploys the motivated concept in a quite different manner for best possible partitioning of the data points into a predetermined number of clusters. Given a cluster CLi with Np points and centroid centi, the mean homogenous-cluster scattering can be represented as follows

$$\Psi(CL_i) = \frac{1}{N_p} \sum_{D_p \in CL_i} (D_p - Cent_i)^2 \qquad (2)$$

Therefore the contribution of a data point Dp ∈ CLi is estimated as

Contribution (Dp, CNi)=ψ(CLi-{DP})- ψ(CLi) (3)

The equation (3) implies that if the contribution of a data point is negative, it has an ill effect on its cluster. On the other hand, a positive value of contribution indicates that the removal of the data point from the cluster would corrupt its perceptual quality. In the proposed framework, the clustering algorithm considers data-points with negative value of contribution quite dissimilar from those with positive value of contribution factor. In the situation of a negative value of contribution data point, the data point is reallocated to a cluster, where its contribution is the highest, possibly with positive value. On the other hand, for a positive value of contribution point, a multi-objective optimization criterion is considered, where optimization is done on both the homogenous and heterogeneous cluster scattering factor measures.

4. IMPLEMENTATION AND RESULTS

The proposed system is executed on 32 bit processor with 1.84 GHz speed and programming platform is considered in Matlab. It is also known that 64-bit architectures perform 40% faster computation than 32 bit architectures. Although the computer industry is encountering transition from 32-bit architectures to 64-bit architecture, but still majority of the users (academicians, researchers, commercial user, corporate users) still use 32-bit processor and therefore experimentation is on 32 bit processor. In the proposed system, an algorithm is primarily built considering two parameters for optimization.

1) Parameter δ for homogenous scattering factor of data points defined by,

$$\delta = \frac{1}{N_p} \sum_{D_p \in CL_i} (D_p - Cent_i)^2$$

2) Parameter φ for heterogeneous scattering factor of data points defined by

$$\varphi = \frac{1}{N_p} \sum_{i=1}^{N_C} (Cent_i - \overline{Cent_i})^2$$

The algorithm for the proposed contribution based feature pruning CBIR system is given below:

Algorithm: Contribution based feature pruning CBIR system

Input: Visual art image

Output: Set of best recognized retrieved visual art images similar to query image.

Start

1. Input colored visual art query images

2. Compute 3 dimensional histogram

3. Estimate data-points using the criterion

$$\gamma = \sum_{i=1}^{N_C} \sum_{D_p \in CL} (D_p - Cent_i)^2$$

```
4. Arbitrarily choose $N_c$ centroids ($cent_1, cent_2, \ldots, cent_{Nc}$).
5. For each point $D_p$
6.   For $1 \le l \le D_p$, distance ($D_p$, $centi_l$) is minimum
7.   Add $D_p$ to cluster $CL_i$ and update centroid $cent_l$.
8. End
9. For each cluster $CL_i$
10.  For each data-point $D_P \in CL_i$.
11.     If Contribution ($D_p$, $CL_i$)<0
12.        Shift $D_p$ to Cluster $CL_p$|Contribution($D_p$, $CL_p$) is Highest value
13.        Revise new value of centroid $Cent_p$ and store it.
14.     End If
15.  End For
16. End For
17. For each cluster $CL_i$
18.  For each data-point $D_P \in CL_i$.
19.     If Contribution ($D_p$, $CL_i$)<0
20.        Shift $D_p$ to a cluster $C_p$ so that
              $\frac{\delta - \delta_{new}}{\delta} + \frac{\varphi_{new} - \varphi}{\varphi_{new}}$ is maximum
21.           Revise new value of centroid $Cent_p$ and store it.
22.           End If
23.     End For
24. End For
End
```

It can be seen that δ_{new} and φ_{new} are the updated values of δ and φ after the data point D_p is shifted to the cluster CL_p.

Our test data consisted of 804 visual art images belonging to multiple groups obtained from retrieval evaluation campaign in Yale University Art Gallery [16]. Each group contained varying number of visual art images. All the visual art images contained various features description mentioning the salient foreground objects. The images were clustered using proposed technique with the initial centroids chosen at arbitrary. The cluster whose centroid was nearest in distance to the given test visual image was determined and the images belonging to the cluster were retrieved. The results were then compared with images retrieved using the conventional clustering algorithm with the same set of initial centroids.

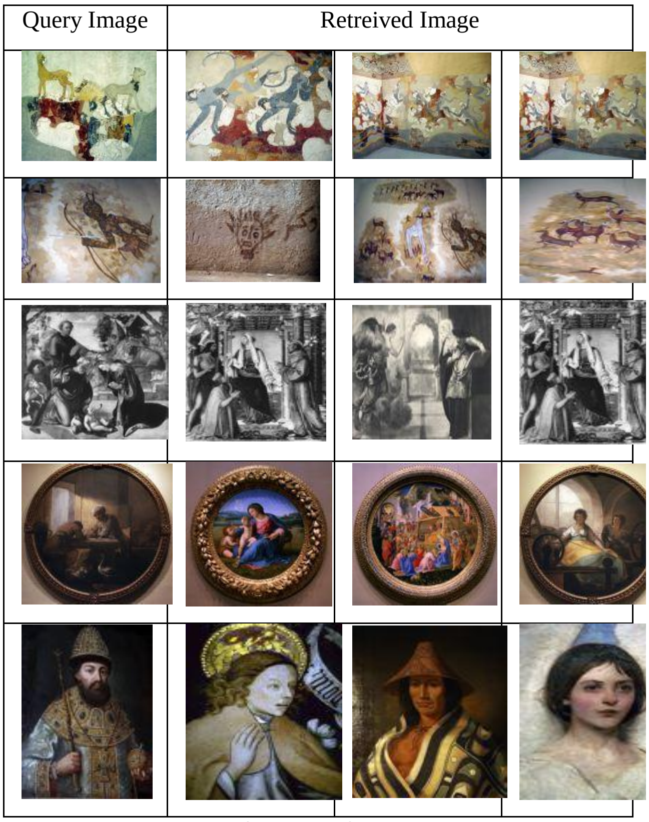

Fig 2: Result of Retrieval for proposed system

The evaluation is performed for 22 groups of visual art images The sample results are shown in Fig.2. The following performance measures were used to evaluate the performance of the algorithm,

- Precision=(Total number of retrieved relevant images)/(Total number of retrieved images)
- Recall=(Total number of retrieved relevant images)/(Total number of relevant images)

The results accomplished after complete testing 804 visual art images for the proposed contribution based technique are as tabulated in Table 1 as follows:

Table 1. Results for Contribution based Feature Set Pruning

#	Group	Recall (%)	Precision (%)
1	image group-1	98	97.2
2	image group-2	97	95.3
3	image group-3	99	97.7
4	image group-4	97	97.6
5	image group-5	85	86.1
6	image group-6	89	90.1
7	image group-7	87	84.7
8	image group-8	95	97.5
9	image group-9	94	93.2
10	image group-10	96	96.4
11	image group-11	94	93.7

12	image group-12	87	86.3
13	image group-13	96	95.7
14	image group-14	93	98.5
15	image group-15	97	97.6
16	image group-16	92	98.7
17	image group-17	83	82.8
18	image group-18	85	90.7
19	image group-19	89	92.8
20	image group-20	99	98.7
21	image group-21	94	99.6
22	image group-22	97	98.3
Average		92.86	94.05

For the purpose of comparative performance analysis, the proposed system is compared with certain prior research work that has claimed optimal output of pruning of CBIR system. We consider our database and experiment it using all the considered prior approaches in CBIR. Following are the approaches considered for performance evaluation.

BTC Based Feature Set Pruning: Silakaris et al. [17] have discussed block truncation based feature extraction in VBIR process which is basically a framework of unsupervised clustering of images based on the color feature of image. Test has been performed on the feature database of color moments and BTC (Block Truncation Coding). The K-means clustering algorithm is applied over the extracted dataset. Results are quite acceptable and showing that performance of BTC algorithm is better than color moments. The results accomplished using this approach for the proposed visual art based CBIR technique is as shown in Table 2.

Table 2. Result for BTC based Feature Set Pruning

#.	Groups	Recall (%)	Precision(%)
1	image group-1	98	97.2
2	image group-2	97	95.3
3	image group-3	99	97.7
4	image group-4	97	97.6
5	image group-5	85	84.6
6	image group-6	89	87.7
7	image group-7	87	84.7
8	image group-8	95	94.8
9	image group-9	94	93.2
10	image group-10	96	96.4
11	image group-11	94	93.7
12	image group-12	87	86.3
13	image group-13	96	95.7
14	image group-14	93	98.5
15	image group-15	97	96.8
16	image group-16	92	91.6
17	image group-17	83	82.8
18	image group-18	85	84.9
19	image group-19	89	87.4
20	image group-20	99	98.7
21	image group-21	94	99.6
22	image group-22	97	96.9
Average		92.86	92.82

Conventional K-Means Algorithm: Prasad et al. [18] has discussed k-means clustering algorithm for the purpose of accomplishing CBIR. It is a system that is developed for retrieving images similar to a query image from a large set of distinct images. It follows an image segmentation based approach to extract the different features present in an image. These features are stored in vectors called feature vectors and compared to the feature vectors of query image and thus, the image database is sorted in decreasing order of similarity

Table 3. Results for K-Means based Feature Set Pruning

#	Groups	Recall (%)	Precision (%)
1	image group-1	98	97.2
2	image group-2	97	95.3
3	image group-3	99	78.6
4	image group-4	97	97.6
5	image group-5	85	83.4
6	image group-6	89	87.7
7	image group-7	87	84.7
8	image group-8	95	94.8
9	image group-9	94	90.4
10	image group-10	94	96.4
11	image group-11	94	93.7
12	image group-12	80	80.7
13	image group-13	97	95.7
14	image group-14	93	98.5
15	image group-15	97	96.8
16	image group-16	87	91.6
17	image group-17	83	80.4
18	image group-18	84	84.9
19	image group-19	89	81.7
20	image group-20	99	98.7
21	image group-21	94	99.6
22	image group-22	84	96.9
Average		91.63	91.15

The Fig.3 shows the comparative performance analysis for the three existing techniques with proposed technique. Silakari et al. [17] have performed feature extraction using color moments and then BTC is applied along with k-means clustering. Prasad et al. [18] has used discrete wavelet transformation for feature extraction followed by k-means clustering. The k-means algorithm has been used to cluster the feature vectors into several classes with every class corresponding to one region in the segmented image. Although, this technique has improved the search results considerably and also consistent with the human perception of an image, yet recall and precision rate are found to be very low compared to other works. The comparative performance analysis is as exhibited in Figure 3.

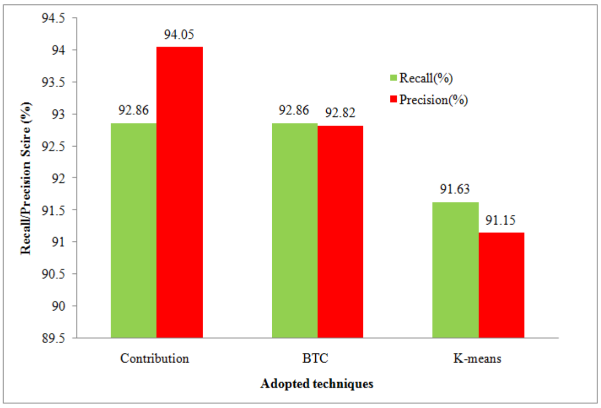

Fig 3: Comparative analysis of proposed system with other existing methods

5. CONCLUSION

The proposed system comprises a novel partitional clustering algorithm based on the idea of ‘contribution of a data point.’ Unlike conventional clustering algorithm like k-means algorithm, the proposed technique produces highly refined retrieval results for both the homogenous-cluster and heterogeneous-cluster similarity measures and need smaller amount of trial with each trials having the lesser computational complexity as that of the k-means and BTC algorithm. The clustering technique is applied to content-based image retrieval. Experimental results exhibit that the algorithm enhances the CBIR performance significantly.

6. REFERENCES

[1] Antani, S., Long, L.R., Thomas, G.R2004. Content-Based Image Retrieval for Large Visual art Image Archives, MEDINFO

[2] Murthy, V. S. V. S. 2010.Content based image retrieval using Hierarchical and K-means clustering techniques, International Journal of Engineering Science and Technology 2.3,pp. 209-212

[3] Tonge, Vanita G. 2011.Content based image retrieval by K-Means clustering algorithm." Int J Eng Sci Tech,Vol. 2, pp. 209-212

[4] Analoui, M., Beheshti, M.2011.Content-based Image Retrieval Using Artificial Immune System (AIS) Clustering Algorithms, In Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1

[5] Samathal, S., Mohanraj, N. 2010.BTC with K means classifier using color image clustering”. Journal of Computer Application, 5

[6] Murthy, V. S. V. S., Vamsidhar, E., Rao, P. S., Raju, G. S. V.2010. Application of hierarchical and K-means techniques in Content based image retrieval, International Journal of Engineering Science and Technology,Vol. 2(5), pp.749-755

[7] Balan, S., Devi, T. 2012.Design and Development of an Algorithm for Image Clustering In Textile Image Retrieval Using Color Descriptors, International Journal of Computer Science, Engineering and Applications (IJCSEA), Vol. 2(3)

[8] Huu, Q. N., Thu, H. N. T., Quoc, T. N.2012.An efficient content based image retrieval method for retrieving images, International Journal Of Innovative Computing Information And Control,Vol. 8(4), pp. 2823-2836

[9] Malakar, A., Mukherjee, J.2013.Image Clustering using Color Moments, Histogram, Edge and K-means Clustering, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064

[10] Komali, A., Babu, R. V.2013.An Efficient Content Based Image Retrieval System for Color and Shape Using Optimized K-Means Algorithm, International Journal of Application or Innovation in Engineering & Management (IJAIEM),Vol.2, Issue.8

[11] Jadhav, S. H., Ahmed, S. A.2012.A Content Based Image Retrieval System using homogeneity Feature extraction from Recency-based Retrieved Image Library, IOSR Journal of Computer Engineering (IOSRJCE), Vol. 7, Issue. 6, pp. 13-24

[12] Ravindran, U., Shakila, T.2013. Content Based Image Retrieval For Histology Image Collection Using Visual Pattern Mining, International Journal of Scientific & Engineering Research, Vol. 4, Issue 4, 2013

[13] Raghatate, K. S., Janwe, J.2013.Content Based Image Retrieval With Relevance Feedback Using Clustering, International Journal of Recent Advances in Engineering & Technology (IJRAET) ISSN, Vol. 1, Issue -2

[14] R. Xu and D. Wunsch.2005.Survey of clustering algorithms, IEEE Transactions on Neural Networks, Vol.16, Issue 3, pp. 645– 678

[15] V.K. Garg.2009.Pragmatic data mining: Novel paradigms for tackling key challenges, Project Report, Computer Science & Automation (CSA), Indian Institute of Science

[16] http://guides.library.yale.edu/content.php?pid=47735&sid=354520

[17] Silakari, S., Motwani, M., Maheshwari, M.2009. Color Image Clustering using Block Truncation Algorithm, IJCSI International Journal of Computer Science Issues, Vol. 4, No. 2, 2009

[18] Prasad, B. G., K. K. Biswas, and S. K. Gupta.2004Region-based image retrieval using integrated color, shape, and location index, Computer vision and image understanding, Vol.94(1),pp. 193-233.