Content Based Image Retrieval Using Hybrid Technique

Gaganjot Kaur Dhillon, Ishpreet Singh Virk
Department of Computer Science, BBSBEC, Fategarh Sahib, Punjab
sandhu.gaganjot@gmail.com, Ishpreet.virk@bbsbec.ac.in

Abstract : Content base image retrieval has become an active research area today. We present a Hybrid approach for content base image retrieval using the K-mean and Hierarchical clustering methods for better and efficient access of images. We design and implement hybrid clustering methods in JAVA. The proposed system demonstrated a promising and better retrieval method on a database containing random images according to some categories like clouds, landscape, fish and trees etc. Images are in JPEG, PNG and GIF format. Images are storing in backend database and fetching according to the attributes of images. To extract the features of content base image retrieval using Mean for location measures and Euclidean distance(ED) for distance measures. The performance of system has been evaluated using precision and recall methods.

Keywords –
Related work, Methodology, Proposed method, Feature extraction, Database, Experimental results, Summary and conclusion and Discussion and Future work.

1. INTRODUCTION

By the search we analyzes Content based means that you search the content of the image instead of the metadata such as keywords, tags or descriptors associated with the image. The term 'content' refers to colors, shape, texture or any other information which can be derived from the image itself. The increased need of content based image retrieval technique can be found in a number of different domains such as Data Mining, Education, Medical Imaging, Crime Prevention, Weather forecasting, Remote Sensing and Management of Earth Resources. [1] The earliest use of the term content-based image retrieval in the literature seems to have originated in 1992, when T. Kato used it to describe his experiments into automatic retrieval of images from a database by color and shape feature. Since then, the term has been used to describe the process of retrieving desired images from a large collection of images based on syntactical image features. The technique and algorithms that are used for CBIR originate from fields such as statistics, pattern recognition, signal processing, and computer vision.

2. RELATED WORK

There is a growing demand of image retrieval systems by text, content, shape and texture and so on. And descriptive statistics for feature extraction has been providing a major contribution in all the above mentioned areas since long time. But the quest of betterment never ends. In later research, we have studied various methods for content based image retrieval [1], [3], [5], [6], [7], [8], [11], [13]. Also, many approaches have been discussed by researchers, based on various features, in the literature for image mining [2], [6], [9], [10]. We have observed many of discussed methods and approaches are effective and stable across different images for image retrieval. We have observed various studies provide us particular category oriented content based image retrieval and applications [4], [10], [12]. From the literature survey, various gaps has been found and it is observed that many researchers have proposed different techniques retrieve images based on text, shape, pattern and content basis. It has been found that text based image retrieval has limitations of power languages and content based image retrieval is mostly in demand. It has been analyzed that we are using various techniques for image retrieval, but hybrid clustering which is combination of hierarchical and k-means clustering is more efficient for content based image retrieval. So, the work is being done with hybrid clustering, which is best suited for retrieval images based on content. We use Euclidean distance and precision for feature extraction and performance measures.

3. METHODOLOGY

Image retrieval system uses the Hybrid algorithm that is the combination of the K-Mean algorithm and the hierarchical algorithm. Given a desired query image to the system and using Hierarchical clustering technique perform filtration of image based on RGB color content. These three different segments of query image are matched with the database images according to RGB value and we get better favored results using K-mean clustering technique. The details on implementation of this hybrid clustering technique are presented below.

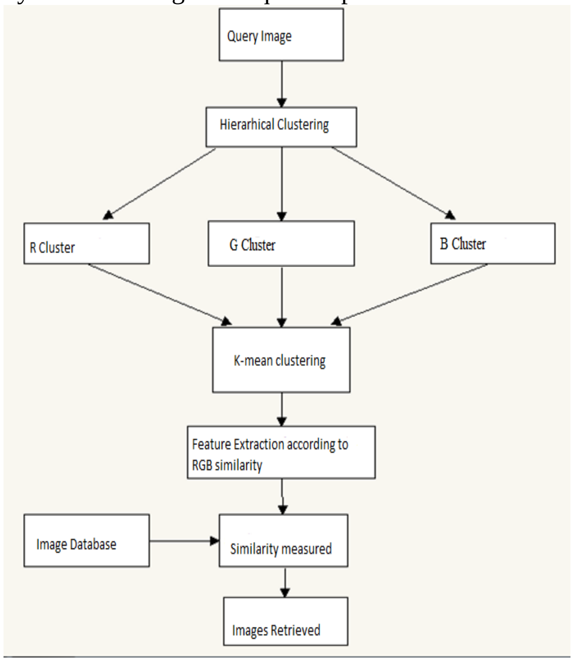


Figure-1: Block diagram for proposed image retrieval system

4. PROPOSED METHOD

The following is the proposed method/algorithm for hybrid technique:

Step 1: Indexing the database

In first step we perform indexing of the database by browsing it from the system from which you want to search a desired image.

Step 2: Given a query image to system

In second step we input the desired query image by browsing it from the system which we want to search. The reason behind inputting the query image is to find the images from the database which are most similar to the query image.

Step 3: Perform Hierarchical clustering

After providing the query image to the system, the search process is performed. This is the process in which RGB value of query image is calculated and compared its values with each RGB values of the database images. Basically the similarity of query image to the database images is calculated based on color content. Then the resultant images are grouped based on the similarity levels this provides random image groups.

Step 4: Apply K-mean clustering for better results

After apply hierarchical clustering we get the random images from the database based on similarity of RGB color to the query image. Then we can apply K-mean clustering on the resultant images because K-mean produces tighter clusters than the hierarchical clustering, so that we can get better results. By this step the images that are most similar to query image are found. [2].

Step 5: Resultant images similar to query

This step displays all the resultant images find by the k-mean clustering technique.

5. FEATURE EXTRACTION

In the developed system, images are analyzed based on the color feature. We currently use RGB color feature. We use of descriptive statistics parameters for feature extraction. Example of Statistical feature extraction techniques include mean and standard deviation computations we are using mean in this system for measure of location [2].

5.1 Location Measures

Location statistics describe where the data is located. The most common functions include measures of central tendency like the mean, median, and mode.

- Mean: For calculating the mean of element of vector x.

$$Mean(x) = SUMi \quad x(i)/N \qquad (1)$$

If x is a matrix, compute the mean of each column and return them into a row vector.

5.2 Distance metrics

Distances metrics are now an important problem in information retrieval. The performance of algorithms for data classification often depends heavily on the availability of a good metric. In CBIR, the space of features is a vector space, but it is not obvious how to introduce a norm because of the incommensurability of the components. Similarity between descriptors is usually computed with either the Euclidean or the cosine angle distance.

However, an image can usually be perceived with different meanings and therefore, the similarity between the same pair of images may change when the concept being queried changes. Understanding the relationship among different distance measures is helpful in choosing a proper one for a particular application. We are using Euclidean distance for measure distance in this system [2].

- Euclidean Distance: If u = (x_1,y_1) and v = (x_2,y_2) are two points, then the Euclidean distance between u and v is given by

$$EU(u,v) = \left(\sqrt{(x1 - x2)^2 + (y1 - y2)^2}\right) . \qquad (2)$$

Instead of two dimensions, if the points have n-dimensions, such as $a = (x_1, x_2, \ldots\ldots x_n)$ and $b = (y_1, y_2, \ldots\ldots y_n)$.

then above equation can be generalized by defining the Euclidean distance between a and b as:

$$EU(a,b) = \sqrt{(x1 - y1)2 + (x2 - y2)2 + \cdots \ldots + (xn - yn)2}. \qquad (3)$$

$$EU(a,b) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \qquad (4)$$

6. DATABASE

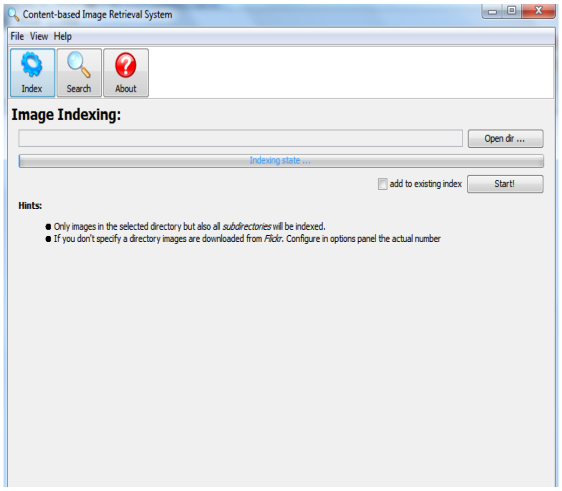
Currently, we have 100 images from 5 different categories in our database, collecting mainly from the web. The number of images for each category (like clouds, landscapes, fish, trees and mixed etc) is 20 images. Images are in JPEG, PNG and GIF format. The illumination and pixel dimensions are unknown for each image. It could be from medical or any universal photographs. During indexing, the feature o images are extracted and inserted into the database. When a query image is provided to the system, the feature extracted from the query images are compared with the stored feature values of the database images.

The data collection is going on, with the aim of extending the categories of images and number of images in each category is extended as part of future work.

7. EXPERIMENTAL RESULTS

The performance of the system is analyzed by taking tests over our database. Each test is evaluated as one-versus-the-rest basis, by querying each image in the database against the remainder images. Unless otherwise indicated, the full database is used in the tests.

The proposed system has been implemented using JAVA and tested on general purpose images of five different categories like cloud, landscape, fish, tree and mixed categories containing total 100 images in JPEG, PNG and GIF formats. The search is usually based on similarity rather than the exact match. We have followed the image retrieval technique described in section-4. The starting interface of the system is shown in Fig-2, the first step is indexing. We select the folder that contains image database, from this selection we want to search an image similar to query image and start the indexing process by clicking on start button.



Content-based Image Retrieval System

Image Indexing:

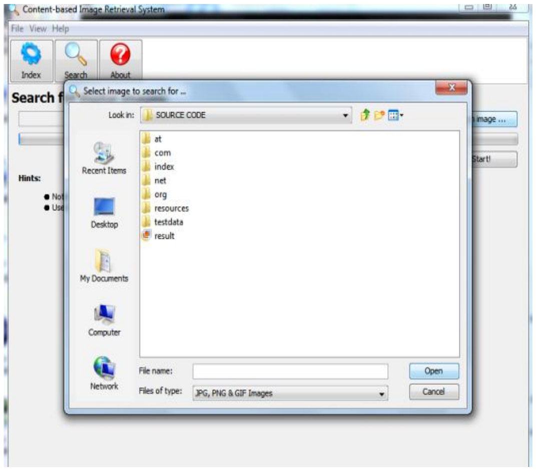
□ add to existing index

Hints:

- Only images in the selected directory but also all *subdirectories* will be indexed.
- If you don't specify a directory images are downloaded from *Flickr*. Configure in options panel the actual number

Figure-2 Starting Interface of Hybrid CBIR system

The second step of the proposed system is inputting a query image to the system is shown in Fig.-3. After selecting the image the searching process is going on. This is the process in which inputted query image is compared with the all the images in the database based on the RGB color feature space using hierarchical clustering technique.



Content-based Image Retrieval System

Select image to search for ...

Look in: SOURCE CODE

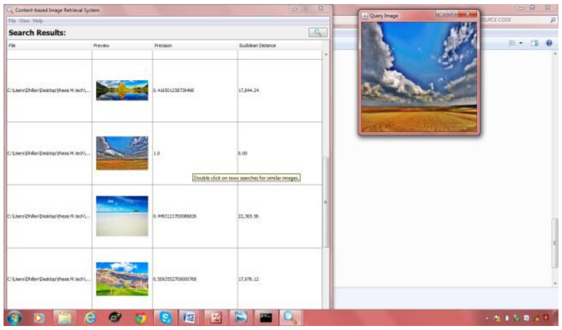
File name:

Files of type: JPG, PNG & GIF Images

Figure-3 Select a query image

After comparing query image with all database images the resultant set of images are processed by k-mean clustering technique. In this process the images that are most similar are clustered together. In this way the images those are most similar finding together. Finally all the images retrieved through k-mean search are displayed.

Below Fig.-4 and 5 shows the result of the system for searching from a cloud category and landscape category of images, the query image is also shown in right side of screen. The results was displayed with less than 1 second and also show the precision and Euclidean distance of each database image with respect to the query image.



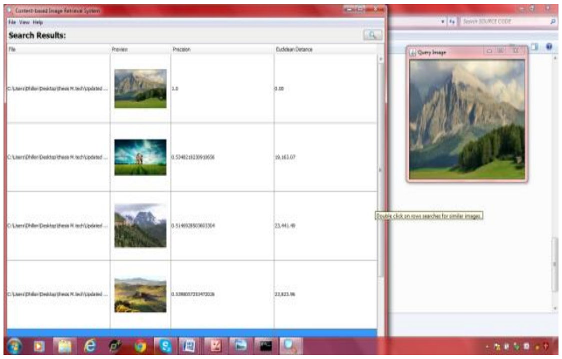
Content-based Image Retrieval System

Search Results:

File	Preview	Precision	Euclidean Distance
C:\Users\Dhillon\Desktop\thesis M.tech\...		0.416501238736498	17,844.24
C:\Users\Dhillon\Desktop\thesis M.tech\...		1.0	0.00
C:\Users\Dhillon\Desktop\thesis M.tech\...		0.449312170008826	22,365.56
C:\Users\Dhillon\Desktop\thesis M.tech\...		0.509355270900768	17,976.12

Double click on rows searches for similar images.

Figure-4 Result after search for cloud category



Content-based Image Retrieval System

Search Results:

File	Preview	Precision	Euclidean Distance
C:\Users\Dhillon\Desktop\thesis M.tech\Updated ...		1.0	0.00
C:\Users\Dhillon\Desktop\thesis M.tech\Updated ...		0.534821923091065	19,163.07
C:\Users\Dhillon\Desktop\thesis M.tech\Updated ...		0.514692850360330	23,441.49
C:\Users\Dhillon\Desktop\thesis M.tech\Updated ...		0.539800723347202	23,823.96

Double click on rows searches for similar images.

Figure-5 Result after search for landscape category

The above results show that the resultant image which has Euclidean distance is 0.0 and precision is 1.0 is best matched result to the query image.

7.1 PERFORMANCE MEASURES

To measure the performance of system we use:-

7.1.1 Precision and Recall method

Precision and recall are widely used parameters in evaluating the CBIR systems. Precision is a measure of fidelity whereas recall is a measure of completeness. Precision basically is a measure of the number of retrieved images that are relevant to the search. Recall is a measure of completeness i.e. it is basically ratio of relevant retrieved image to the total relevant images present in the database [6].

Formula to find precision is:

Precision for Red Pixel, $R = \frac{\min\left(R(i_q),R(i_d)\right)}{\max\left(R(i_q),R(i_d)\right)}$ (5)

Precision for Green pixel, $G= \frac{\min\left(G(i_q),G(i_d)\right)}{\max\left(G(i_q),G(i_d)\right)}$ (6)

Precision for Blue pixel, $B= \frac{\min\left(B(i_q),B(i_d)\right)}{\max\left(B(i_q),B(i_d)\right)}$ (7)

Total precision for image= $\frac{(R+G+B)}{3}$ (8)

Where $\min(R(i_q),R(i_d))$ means select the red cluster of query image or image from database which one has min number of red pixel. And $\max(R(i_q), R(i_d)$means select the red cluster of query image or image from database which has max number of red pixel.

- $R(i_q)$ = Number of red pixels in query image.
- $R(i_d)$ =Number of red pixels in database image.
- $G(i_q)$=Number of green pixels in query image.
- $G(i_d)$=Number of green pixels in database image.
- $B(i_q)$ =Number of blue pixel in query image.
- $B(i_d)$=Number of blue pixel in database image.

Formula for Recall.

$$\textbf{Recall}=\frac{(\text{Number of images reterieved})}{\text{Total number of images}} \qquad (9)$$

The Hybrid method, as discussed in section -4 has been used to study image retrieval using RGB color matching scheme using Hierarchical and K-mean clustering. The performance of the proposed system has been evaluated by analyzing results with the different author's results [5], [7], [9] .Each query image returns the top relevant images from the database, and the calculated average precision values using the formula described in equation (8) are shown in Table 1. From this table the effectiveness of hybrid method is evaluated by selecting each query image under every category of different semantics. For each query, we examined precision of the retrieval, based on the relevance of the image semantics. The precision of the selective category of image database is higher than the searching from mixed category of image database. These results clearly show that the performance of the proposed method is better than the other methods. The search for particular category provides us a better and efficient result than searching from the universal or mixed category of images. The time taken by system to search an image from the database is less than one second so it will show this system provide us fast and better results.

Table-1 Average precision of different category of images.

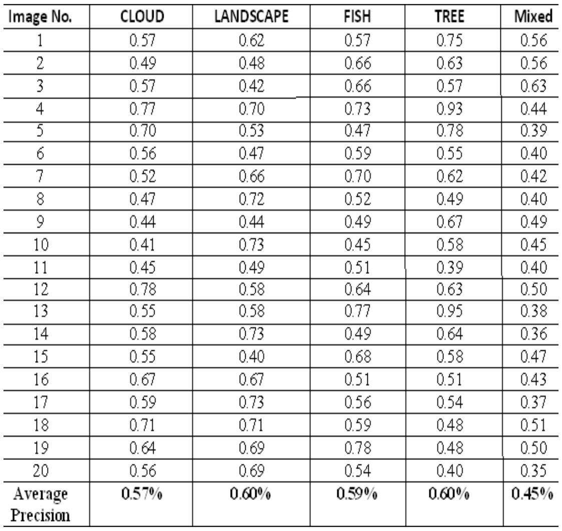
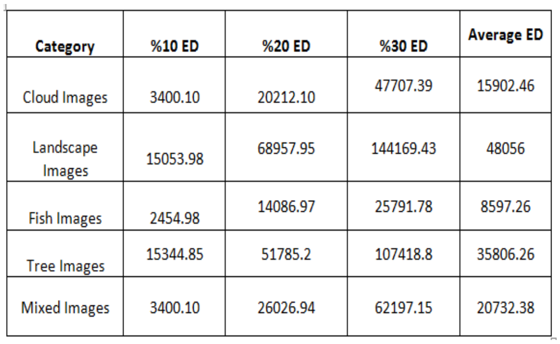


Image No.	CLOUD	LANDSCAPE	FISH	TREE	Mixed
1	0.57	0.62	0.57	0.75	0.56
2	0.49	0.48	0.66	0.63	0.56
3	0.57	0.42	0.66	0.57	0.63
4	0.77	0.70	0.73	0.93	0.44
5	0.70	0.53	0.47	0.78	0.39
6	0.56	0.47	0.59	0.55	0.40
7	0.52	0.66	0.70	0.62	0.42
8	0.47	0.72	0.52	0.49	0.40
9	0.44	0.44	0.49	0.67	0.49
10	0.41	0.73	0.45	0.58	0.45
11	0.45	0.49	0.51	0.39	0.40
12	0.78	0.58	0.64	0.63	0.50
13	0.55	0.58	0.77	0.95	0.38
14	0.58	0.73	0.49	0.64	0.36
15	0.55	0.40	0.68	0.58	0.47
16	0.67	0.67	0.51	0.51	0.43
17	0.59	0.73	0.56	0.54	0.37
18	0.71	0.71	0.59	0.48	0.51
19	0.64	0.69	0.78	0.48	0.50
20	0.56	0.69	0.54	0.40	0.35
Average Precision	0.57%	0.60%	0.59%	0.60%	0.45%

7.1.2 Euclidian distance Analysis

We can also find the Euclidian distance using method discussed in equation (4) for each category of images as shown in Table-2 which tells us the distance of query image to the database images and which image having 0 distances from the query image in the database that image is most relevant. The Table-2 show us the Euclidian distance of top 10%, 20%, 30% images of each category at the end it shows the average Euclidian distance of each category query image to the database.

Table-2 Euclidean Distance for different picture categories



Category	%10 ED	%20 ED	%30 ED	Average ED
Cloud Images	3400.10	20212.10	47707.39	15902.46
Landscape Images	15053.98	68957.95	144169.43	48056
Fish Images	2454.98	14086.97	25791.78	8597.26
Tree Images	15344.85	51785.2	107418.8	35806.26
Mixed Images	3400.10	26026.94	62197.15	20732.38

The results of Table-2 is better represented by bar chart in Fig.-6 the blue bar shows 10% Ed, red bar shows 20% ED, green bar shows 30% ED. Horizontal axis shows categories of images and vertical axis shows distance%. Those images who have Euclidian distance 0 is similar to the query image. The Table-2 shows that out of 10% of images the 3400.10 is distance of relevancy of query image to the database and so on.

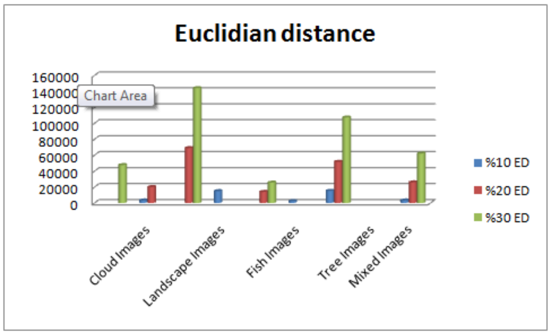


Figure-6: Euclidean Distance for different picture categories

8. SUMMARY AND CONCLUSION

In this paper, we present an innovative approach for content base image retrieval by combining the hierarchical and k-mean clustering approaches features. The invention of this hybrid technique provides us fast and better results. Similarity between the images is determined by means of distance function (e.g. Euclidean distance). The experimental result shows that the proposed method outperforms the other retrieval methods in terms of average precision. As a result, the average precision is high in category wise search than in the mixed category search.

9. DISCUSSION AND FUTURE WORK

Since the lack of contemporary research in this specific area of image retrieval, this is preliminary research and could benefit from many improvements. Future studies will include following three aspects: image collection, feature extraction and distance metric.

9.1 Image collection

There might be some improvements in the image collection section. If the illumination condition for each image is given, color balancing may be performed in the pre-processing step, in order to reduce the impact of mismatched color balance between the query and database images. The images are recommended to be collected at the same physical scale level to eliminate the impact of spatial scale. This will more helpful in the pre-processing step and it will also decrease the image access time.

9.2 Feature extraction

In this only descriptive parameter (mean) were chosen to characterize the homogeneity property of images. In the future, many other parameters of descriptive statistics can be used. Along with this we can apply dimension reduction on extracted features to compensate the retrieval time as the size of the database is increased.

9.3 Distance metric

In the proposed methodology, we used Euclidean distance for the similarity measure of query image and database images which shows almost similar results. In further study, some other distance metric, such as the Cosine angle distance and Mahalanobis distance, could be explored. The distance metric combination scheme may be further investigated.

REFERENCES

i. *A.Kannan, D. M. (2010). Image clustering and retrieval using image mining techniques. IEEE International conference on computational intelligence and computing research*

ii. *B. Ramamurthy, K. C. (2011). CBMIR:Shape based Image Retrieval Using CANNY Edge Detection and K- Mean Clustering Algorithm for Medical Images. International Journal of Engineering Science and Technology (IJEST) , 3, No.3, 1870-1877.*

iii. *Devasena,]. C. (2011). A Hybrid Image Mining Technique Using LIM- based Data Mining Algorithm. International Journal of Computer Applications , 25, 11-15.*

iv. *Dubey, R. S. (2010). Image mining using content based image retrieval system. International Journal on computer science and engineering(IJCSE) , 02, No. 07, 2353-2356.*

v. *, D. X, et.al.(2009). Trend of content based images retrieval on the internet. IEEE 5th International conference on image and graphics , 733-738.*

vi. *GokberkCinbis, S. a. (2009). Image mining using directional spatial constraints. Geosciences and remote sensing letters, IEEE , 07, No. 1, 33-37.*

vii. *Guang-hai liu, L. Z.-K. (2010, Feb 11). Image retrieval based on multi-texton histogram. Elsevier Ltd. All rights reserved , 2380-2389.*

viii. *Hanife Kebapci, B. Y. (2010, April 9). Plant image retrieval using color, shape and text features. (H. Toroslu, Ed.) The computer journal advace access .*

ix. *Jan-Ming Ho, S.-Y. L.-W.-C.-I. (2012). A novel contnt based image retrieval system using k-means with feature extraction. International conference on systems and informatics , 785-790.*

x. *Madheswaran, P. R. (2009). An improved Image mining technique for brain tumour classification using efficient classifier. International Journal of computer science and information security , 06, No. 03, 107-116.*

xi. *Mahip M. Bartere, D. P. (2012). An efficient technique using text & content base image mining technique for image retrieval. International Journal of Engineering research and application(IJERA) , 2, 734-739.*

xii. *Murthy V.S.V.S, E. 1. (2010). Application of Hierarhical and K-means Techniques in Content based image retrieval. International Journal of engineering Science and Technology , 2, 749-755.*

xiii. *Nishchol Mishra, D. S. (2012). Image Mining in the Context of Content Based Image Retrieval: A Perspective. International Journal of Computer Science Issues , 9 (4), 98-107.*

xiv. *Peter, S. P. (2010). A novel minimum spanning tree based clustering algorithm for image mining. European Journal of scientific research , 540-546.*

xv. *S, H. P. (2011). Content Based Image Reterieval using Color Boosted Salient Points and shape features of an image. International Journal of Image processing , 2 (1), 45-47.*

xvi. *Shaikh Nikhat Fatma, M. a. (2011). Image mining using association rule. World congress on information and communication technologies,IEEE , 587-593.*

xvii. *Shanthi, J. a. (2007). Image mining techniques for classification and sgmentation of Brain MRI data. Journal of theoratical and applied information technology , 3, No. 4, 115-121.*

xviii. *Singha, M. (2012). Content base image retrieval using color and texture. Signal and image processing: An international Journal(SIPIJ) , 3.*

xix. *VanitaG.Tonge. (2011). Spatially related image mining on very large image collections. Proceedingd of ICETECT , 864-869*