International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 10, No. 8 (2017), pp.135-144
http//dx.doi.org/10.14257/ijsip.2017.10.8.12

Mining of Images by K-Medoid Clustering Using Content Based Descriptors

Ruchi Jayaswal, Jaimala Jha and Manish Dixit

M.I.T.S.,M.P.(India)
ruchi.jayaswal23@gmail.com, jaimala.jha@gmail.com, dixitmits@gmail.com

Abstract

Image Mining is a challenging task in the data mining and image processing field. In Image mining, useful information is extracted from the enormous collection of image database. The expansion of images has been risen up in each areas medical field, business, forgery detection etc because images easily gain the attention to the users that's why more researchers are attracted towards this field. Lots of images are scattered in the database so to manage the database, clustering is applied which is one of the techniques of Image Mining. In this research work fusion of color and shape features are used for extracting the descriptors from the images through Color Moment (CM) and Edge histogram Descriptor (EDH). After that K-medoids Clustering algorithm is applied on the created dataset to obtain clusters. Finally, the output of clustered images will be shown. Three databases are used for testing this system Wang (1000), Coral (2000), Oliva (452). By using Precision, Recall, F-measure and Error rate metrics, we measure the performance of this proposed work and will also compare with other's conventional methods.

Keywords*: Clustering, Color Moment, Edge Histogram Descriptor, Kmedoids, Kmeans*

1. Introduction

With the advancement of digital storage and sensing technology in various utilization like surfing on the internet, in digital images and video surveillance, these all are generated high volume and high dimensional datasets. There are lots of data stored in digital form in an unorganized manner. At every second data is stored in digital form either it is text, audio, video or images on the internet and its size is increasing day by day. So with this increment, there is a need for analyzing the data, clustering into groups of data, classifying the data or by some other techniques used to manage these datasets. Digital images take a special attention rather than all other data. Mining of images from the enormous collection of images is a crucial task. Image mining is the advance field of Data mining [20]. It defines that the extraction of similar patterns from the large database of minimizes. There are many techniques of image mining. As shown in figure 1[7]. The goal of each technique is to extract similar patterns from the enormous collection of the database.

Here, we used clustering technique to organize the data into sensible manner. Organizing the images in a proper sequence is also a crucial task. In this work, the dataset of images is managed by using clustering. Clustering is a technique to the grouping of objects that contain similar patterns of a dataset in one cluster but dissimilar objects with other cluster group patterns [11]. For making the clusters we used the hybrid approach of color and shape feature extraction techniques. By using these techniques we get optimized clusters. Retrieval of Images from huge database size is a time-consuming process, and its accuracy of retrieving relevant images is also not fruitful. Instead of direct retrieving images from the database, we will form clusters of images so that efficiency of the system

ISSN: 2005-4254 IJSIP

would get improved. But in this work, we just make clusters of images and each cluster is different from other cluster group classes of images.

The rest of the paper is described as below: Section 2 briefly summarize as the literature survey of others research work. Section 3 explain about the feature extraction techniques of color and shape. Section 4 describe K–medoids clustering. Section 5 present the proposed work of this system. Section 6 shows the experimental results of the system. And last Section 7 conclude the work and brief about the future scope related this work.

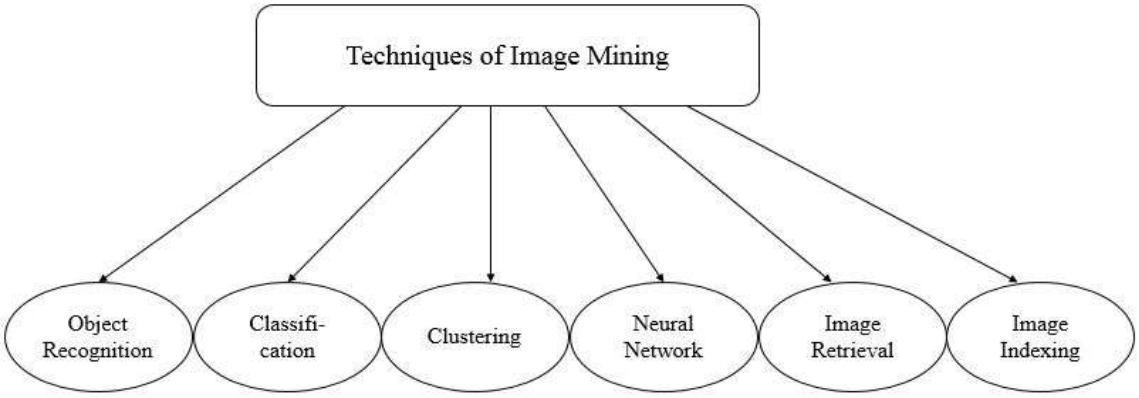

Figure 1. Techniques of Image Mining

2. Literatures Survey

Table 1. Survey of Some Research Work in this Field

S.NO	Year	Author	Proposed work	Database used	Result Analysis
1	2016	Shubhangi P. Meshram [13]	ACO and PSO hybrid optimization used by shape feature extraction for clustering	CORAL Database, 100 images use for testing having 3 categories.	0.98 accuracy achieved to make clusters.
2	2009	Manish Maheshwari [11]	Color Moment and Gabor Filter used as color, texture features, then applied Kmean and Hierarchical clustering algorithm	Wang database used having 3 classes of image.	The performance of Kmeans algorithm is better than Hierarchical clustering.
3	2015	Amit Khatami [9]	Swam intelligence algorithm based on kmedoids is proposed to solve problems in the fire detection field. FCS is used to extract color features	Forest Images	The proposed system is fast to detect fire in the forest.
4	2013	Manish Maheshwari [3]	HVS color space is used for feature extraction and Kmeans clustering algorithm is applied to created dataset.	Wang Database used having 1000 images	Recall is 51.5% and precision is 50.7%

5	2013	Annesha Malakar [2]	Color histogram, Color moment as the color feature used, Canny edge detection as a shape feature and finally kmeans is applied over this dataset.	Up to 40 images takes for testing of 3 classes	Overall accuracy 90.5% for 10,20,30,40 images
6	2015	Padmavati Shrivastava [12]	Two feature sets,1st is obtained by color moment, Gabor filter, edge direction histogram then applied kmeans clustering on this set after that 2nd feature vector extract by using Hue, Tamura's ad edge direction classification is done by second feature set	Natural scenes from the Oliva database	Total classification accuracy achieved 83.4%
7	2008	Nor Hafizah Abd. Razak [14]	Segment image into objects, texture feature and shape feature extracts by GLCM and 2-D moments invariants then applied hybrid kmeans and Hierarchical algorithm	PASCAL Database 2006 collection and Google images.	Performance achieved by hybrid system is 66%
8	2016	Maria Fayez [15]	Two methods are proposed:In 1st method GLCM texture feature extract by Hiralick statistical then kmeans applied. In 2nd method 2D wavelet transform feature extract then dataset pass to K-means code.	Medical database: 300 x-rays and 200 CT scans	From first proposed method overall accuracy is 67.2% and from Second proposed system accuracy is 86%

3. Visual Features Descriptors

For representing an object, visual features extracts from the images and make a dataset so that the relevant information can be obtained. To finding out the perfect combinations of the descriptors is still a major task. Color and shape are the primitive features for an image.

3.1 Color Moment

The color is the fundamental feature descriptors that make for human eyes perception easier. Color moment basically used to measure the color similarity amid the images. Distribution of color in an image can be explained as probability distribution [1][4]. Mean, Standard Deviation, Skewness is used as central moment [2] for an image's color distribution. These moments are calculated for each three primary colors Red, Green, Blue in an image. But in this work, we calculated only two moments for each channel of an image. We obtained 6 moments from 2 moments for each 3 color channels [1][2][4].
Mathematically, these two moments can be expressed as: P_{ij} is the image pixel of j^{th} pixel value in the i^{th} color channel.[4]

Moment 1: Mean

$$\mu_i = \left(\frac{1}{n}\sum_{j=1}^{n} P_{ij}\right) \quad (1)$$

Moment 2: Standard Deviation

$$\sigma = \left[\frac{1}{n}\sum_{j=1}^{n} (P_{ij} - \mu_i)^2\right]^{1/2} \quad (2)$$

3.2 Edge Histogram Descriptor (EHD)

To recognize the image, shape descriptor provide significant information. The histogram is used to characterize the global feature composition of an image. Edge of the image is considered as a sensitive feature for image perception [6]. This EHD suggested for MPEG-7[5][26] comprises only of local edge orientations in the image. Global EHD and Local EHD are used to extract the features of this descriptor. To represent the spatial distribution, there are five edge types in EHD. Four are directional edges named as Vertical, Horizontal, 45 Degree, 135 Degree, these are extracted from image blocks and the fifth one is the non-directional edge which described as if image block contains random edge in the absence of any directionality. Mean values are obtained from the four sub-blocks and obtained edge magnitudes as they are convolved with edge filter coefficients shown in figure 2 [5].

1	-1
1	-1

1	1
-1	-1

$\sqrt{2}$	0
0	$-\sqrt{2}$

0	$\sqrt{2}$
$-\sqrt{2}$	0

2	-2
-2	2

Figure 2. Edge Filter Coefficients

4. Clustering

In order to manage the large database, some techniques are required to organize them. Among from those techniques, Clustering is the finest method for an unsupervised class whose labels are not predetermined.

There are some methods for clustering such as Partitioning Based Method, Hierarchical Based Clustering, and Density Based Method. In Partitioning Based method, K-means, K-medoids clustering algorithm comes and in Hierarchical Based Method, there are two approaches used one is Agglomerative approach and another is Divisive approach and Density Based Method, DBSCAN algorithm comes, As we know that K-means is one of the convenient algorithms to implement but it severe from a major drawback. The distribution of data get irregular outcomes in improper clustering in the case of extreme valued data items, so that this algorithm very sensitive for noisy data and to outliers and makes the performance is low[8].

To overcome such drawbacks of K-means algorithm, there is another algorithm namely K-medoid clustering algorithm that is quite similar to K-means algorithm. K-medoid algorithm is more robust and minimizes the sensitivity to noisy data and to outliers which are bound to occur in the realistic abandoned environment [10]. Here**,** we applied K-medoid algorithm in this dataset.

4.1. K-Medoids Clustering Algorithm

This algorithm also called as Partition Around Medoids(PAM) is suggested in 1987 by Kaufman and Rousseuw[8][9][10].There are various approaches of K-medoids algorithm

such as PAM, SMALL, CLARA(for large dataset), and CLARANS(randomized CLARA) selection of algorithm is depend on the size the dataset. Kmedoid uses actual objects to clusters rather than mean values/centroid as in K-means [10].
K-medoids algorithm Procedure [8][10]

Input set: a) Dataset file(.mat) containing n images
b) Give the number of clusters

Output: Obtained the set of clusters

Procedure: The steps of K-medoids(PAM) follow-

a. Begin with initial medoids of k objects through random selection.
b. Assign each rest data objects to a cluster with the most nearby medoid;
c. Arbitrarily choice a non medoid data objects(O);
d. Evaluate the total cost of swapping (S) old medoid data objects with newly chosen non-medoids data objects(O)
e. If (S>0) then, the swap operation is performed with the new medoids.
f. Repeat steps b, c, d, and e until medoids stabilize their positions.

5. Proposed Methodology

The proposed work is based on the fusion of CBIR and Data clustering techniques. For feature extraction, we used color (Color Moment) and shape (Edge Histogram Descriptor) combination which yield the perfect information of the images. K-medoids Clustering is used as a Data Clustering Techniques as described above. This methodology works well to produce the optimal clusters for the image database. The steps of the system are as follows:

Steps:-

1. Select the folder of image repository contains a number of images.
2. Pre-Process is done, resize our image database into 384*256, 256*256.
3. Create the dataset of images by visual feature descriptors through Color Moment and Edge Histogram Descriptor techniques.
4. Load the created dataset and specify the number of clusters k for the k-medoids algorithm.
5. Obtained cluster group index numbers for all the images.
6. Finally, show some images of each cluster group

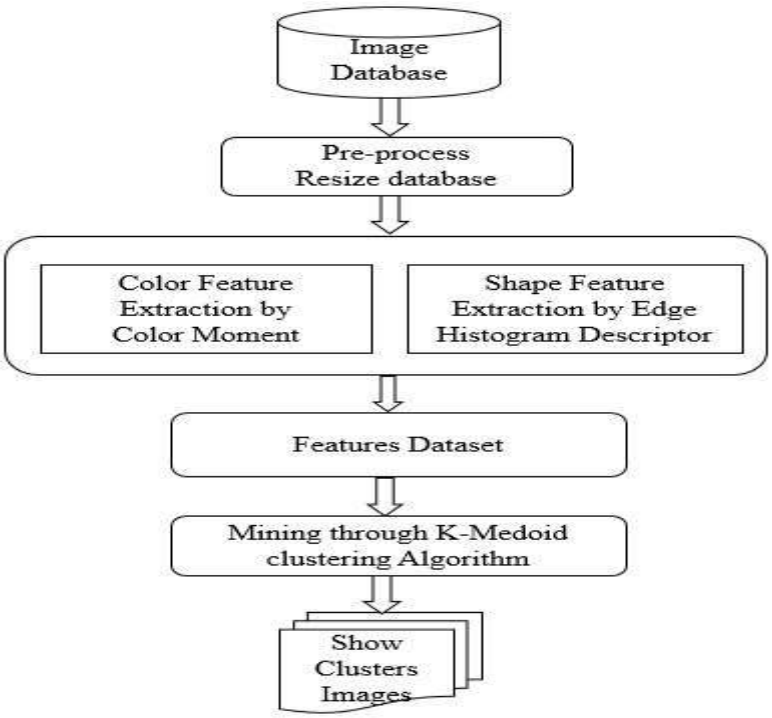

Figure 3. Flow Chart of Proposed System

6. Experimental Analysis and Outcomes

6.1 Tools and Dataset

This proposed system is implemented in MATLAB R2013b version. Coral[25](2000), Wang[3](1000) and Oliva[25] (452) Databases are used to testing this proposed system.

6.2 Performance Measures Analysis

The performance of projected work is analyzed by Precision, Recall metrics and F-measure. Precision measures about the relevance of the proposed system and recall measures about the accuracy of the proposed system[3]. F-measure is used to measure the accuracy of the same kind of images in a cluster and also belong to that class. F-measure is calculated as:[13]

$$P_R = \frac{M_R}{N_R} \quad (1)$$

$$R_C = \frac{M_R}{N_T} \quad (2)$$

$$FM = \left[\frac{(2) * P_R * R_C}{1 * (P_R + R_C)}\right] \quad (3)$$

$$E_R = \frac{K_N}{T_R} \quad (4)$$

Where, P_R = Average Precision Value
R_C = Average Recall Value
M_R = Number of Relevant Images Retrieved

N_R = Number of images retrieved in output window
N_T = Total number of images present in database
FM = Fmeasure Values
E_R= Error rate
K_N = number of non-relevant images retrieved
T_R = total number of images retrieved

Performance metrics is shown in Table 2 in which Precision is computed as the total number of germane images in a cluster from the number of cluster images shown in output window here, the window size is 30. The recall is computed as the average of a total number of germane images in each cluster from the whole database. Fmeasures computed by the amalgamation of precision and recall values and calculate the accuracy of the clustered images. Error rate should be minimum and is calculated by a number of irrelevant images in each cluster from the total number of images in the database of each category.

Table 2. Average Performance Metrics of Three Dataset

Database used	***Coral (2000 images)***	***Wang (1000 images)***	***Oliva (452 images)***
Categories	20	10	3
Average Precision of clustered images	1	0.99	0.95
Average Recall of clustered images	0.99	0.98	0.96
Average FMeasure of clustered images	0.99	0.98	0.95
Average Error rate of clustered images	0.01	0.18	0.4

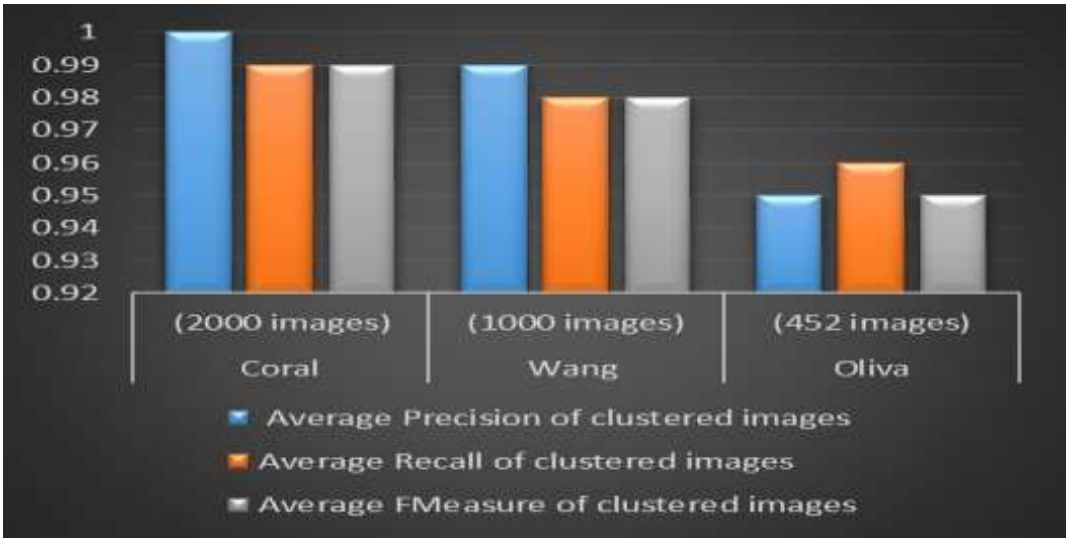

Figure 4. Bar Graph of Proposed Work on Different Dataset

Figure 5. Sample of Cluster 6 images of Coral Database

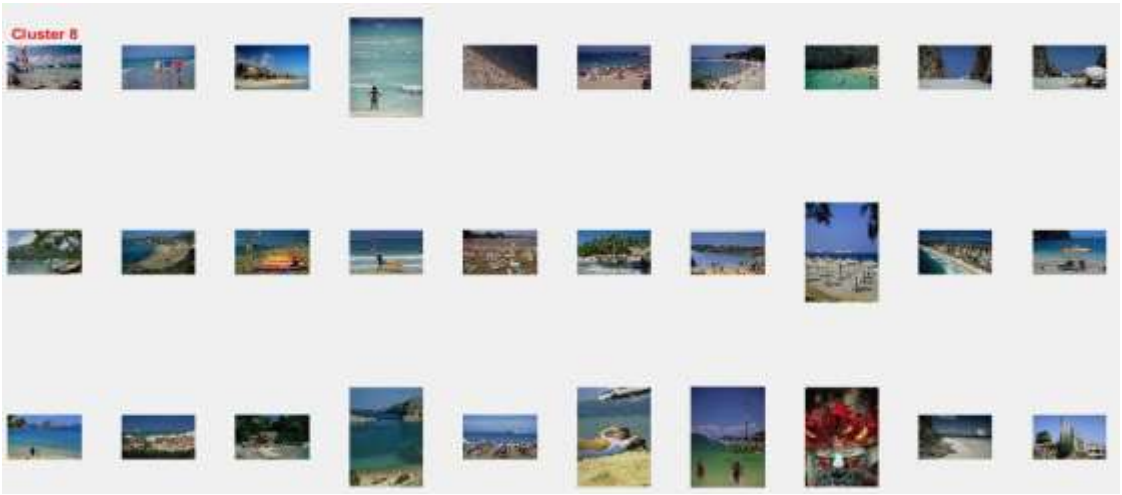

Figure 6. Sample of Cluster 8 images of Wang Database

Figure 7. Sample of Cluster 2 images of Oliva Database

Table 3. Comparison with other Conventional Methods

Database/Methods	*Database*	*Performance Metrics*
Shubhangi's Method[13]	Coral Database	Accuracy is 0.98 for 3 categories -
Maheshwari's Method[3]	Wang Database	Precision is 50.7% for 10 categories
Manish's Method[11]	Wang Database	Precision is 92.26 for 3 categories
Proposed Method	**Coral Database and Wang Database**	**Accuracy is 0.99 for Coral Dataset and Precision is 0.98 for Wang dataset**

7. Conclusion and Future Work

This research work presents the "Mining of images by K-medoid Clustering using content based descriptors" to obtain the clusters, partitioning based method i.e. K-medoids clustering algorithm is used. In this proposed system, color and shape features are used as a color moment and edge histogram descriptor is extracted from the images and created image dataset. By this combination, we obtained fine dataset for clustering. After applied clustering, we got optimized clusters i.e. similar images belong to one cluster or make one group.

The accuracy of proposed system is achieved 0.99 for Coral database is better than other conventional system but still, there is the need for the future scope to get more accuracy for others database.

References

[1] Naveena A K and N K Narayanan,"Image Retrieval using Combination of Color, Texture and Shape Descriptor," **(2016)** IEEE, pp-958-962

[2] Annesha Malakar, and Joydeep Mukherjee,"Image Clustering using Color Moment, Histogram, Edge and K-means Clustering," IJSR, **(2013),** Volume 2 Issue 1, ISSN: 2319-7064,pp-532-537

[3] Manish Maheshwari, Dr. Mahesh Motwani, and Dr. Sanjay Silkari, "New Feature Extraction Technique for Color Image Clustering," IJCSEE, **(2013)**, Volume 1, Issue 1 pp-131-135

[4] Muhsina Kaipravan and Rejiram R,"A Novel CBIR System Based on Combination of Color Moment and Gabor Filter", **(2016)**, IEEE

[5] Dong Kwon Park, Yoon Seok and Chee Sun Won," Efficient Use of Local Edge Histogram ", IEEE,2013.

[6] Swati Agarwal, A.K. Verma, Preetvanti Singh,"Content Based Image Retrieval using Discrete Wavelet Transform and Edge Histogram Descriptor", 978-1-4673-5986-3, **(2013)**, IEEE.

[7] Prabhjeet Kaur and Kamaljit kaur, "Review of Different Existing Image Mining Techniques," International Journal of Advanced Research in Computer Science and Software Engineering(IJARCSSE), Volume 4, Issue 6 June **(2014)**, pp 518-524

[8] Swarndeep Saket J. and. Sharnil Pandya,"Implementation of Extended K-Medoids Algorithms to Increase Efficiency and Scalability using Large Dataset", International Journal of Computer Applications(0975-8887) Volume 146- No. 5, July **(2016)**,pp-19-23

[9] Amit Khatami,and Saeed Mirghasemi,"A New Color Space Based on K-medoids Clusterings for Fire Detection", 978-1--4799-8697-2/15/$31.00 **(2015)**, IEEE,DOI 10.1109/SMC. 2015.481,pp 2755-2760

[10] Aruna Bhat, "K-Medoids Clustering Using Partitioning Around Medoids For Performing Face Recognition," IJSCMC, Vol 3,No 3, August 2014, DOI: 10.14810/ijscmc.**(2014).**3301,pp-1-12

[11] Manish Maheshwari, Dr. Mahesh Motwani, and Dr. Sanjay Silkari, "Image Clustering using Color and Texture," 978-0-7695-3743-6/09 $25.00 **(2009)** IEEE, DOI 10.1109/CICSYN.2009. pp-403-408.

[12] Padmavati Shrivastava, K.K. Bhoyar, and A.S. Zadgaonkar,"A New Approach to Scene Classification using K-means Clustering", International Journal of Computer Applications(0975-8887), Volume 125-No. 14,September **(2015)**

[13] Shubhangi P. Meshram, Dr. Anuradha D. Thakre,and Prof. Santwana Gudadhe "Hybrid Swarm Intelligence Method for Post Clustering Content Based image Retrieval,"7th International Conference on Communication Computing and Virtualization **(2016)**, Elsevier B.V., pp. 509–515.

[14] Nor Hafizah Abd. Razak, Noridayu Manshor, Mandava Rajeswari and Dhanesh Ramachandram,"Object Based Clustering using Hybrid Algorithm," IEEE Bombay Section Symposium (IBSS), 978-0-7695-3359-9/08 $25.00 **(2008)**, IEEE,pp-17

[15] Maria Fayez, Soha Safwat and Ehab Hassanein ,"Comparative Study of Clustering Medical Images 978-1-4673-8460-5/16/$31.00 **(2016)** IEEE, SAI Computing Conference , London , UK, pp. 312–318.

[16] Snehal Mahajan, Dharamaraj Patil,'Image Retrieval Using Contribution-Based Clustering Algorithm with Different Feature Extraction',978-4799-3064-7/14,IEEE,**(2014)**

[17] Ruziana Mohamad Rasli, T Zalizam, Yuhanis Yusof and Juhaida Abu Bakar", Comparitive Analysis of Content Based Image Retrieval using Color Histogram. A Case Study of GLCM and K-Means Clustering, 978-0-7695-4668-1/12 $26.00 © **(2012)** IEEE

[18] Geethu Varghese, Dr. Arun Kumar M N "Content Based Image Retrieval using Feature Extraction and K-Means Clustering" International Journal for Innovative Research in Science and Technology Vol.3, ISSN:04 September**(2016)**-4238, pp.291-302.

[19] Priti Maheshwary, and Namita Srivastav "Retrieving Similar Image Using Color Moment Feature Detector and K-means Clustering of Remote Sensing Image," 978-0-7695-3504-3/08 ©**(2008)**, IEEE

[20] Neethu Joseph.c, Aswathy Wilson ," Retrieval of Images using Data mining Techniques',978-1-4799-6629-5/14$31.00 IEEE,**(2014)**.

[21] Thamilselvan P and Dr. J. G. R Sathiseelan," Image Classification using Hybrid Data Mining Algorithms – A Review", 978-1-4799-6816-94. ICIIECS'**(2015)**

[22] Jaimala Jha Dr. Sarita Sign Bhaduaria "Review of Various Shape Measures for Image Content Based Retrieval" International Journal of Computer & Communication Engineering Research Nov.**(2015).**

[23] Jaimala Jha Dr. Sarita Singh Bhaduaria" Performance-based analysis of CBIR methods "International journal of 10 April **(2016)**.

[24] Jaimala Jha, Dr. Sarita Sign Bhaduaria " Review of Various Shape Measures for Image Content Based Retrieval " International Journal of Computer & Communication Engineering Research,Nov **(2015).**

[25] Ming Zhang, Ke Zhang,and Qinghe Feng, Jianzhong Wang,Jun Kong and Yinghua Lu," A novel image retrieval based on hybrid information desriptors,"1047-3203/ **(2014)**Elsevier.

[26] Sawet Somnupong,and Kanokwan Khiewwan,"Content –Based Image Retrieval using a Combination of Color Correlograms and Edge Direction Histogram,",JESSE,978-1-5090-2033-1/16©**(2016)** IEEE.

[27] Ruchi Jayaswal,Jaimala Jha,"A Hybrid Approach For Image Retrieval using Visual Descriptors", ISBN: 978-1-5090-6471-7/17/$31.00 (**2017)** IEEE, unpublished

[28] Ravi Devesh, Jaimala Jha,"An EEfficient approach for Monuments Image Retrieval using Multi-Visual Descriptors." In MCCS **(2017)**,unpublished

