IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-6 E-ISSN: 2347-2693

An Effective Method of Image Mining using K-Medoid Clustering Technique

Ruchi Jayaswal[1*], Jaimala Jha[2], Ravi Devesh [3]

[1*] Dept. of CSE/IT, Madhav Institute of Technology and Science, Gwalior, India
[2] Dept. of CSE/IT, Madhav Institute of Technology and Science, Gwalior, India
[3] Dept. of CSE/IT, Madhav Institute of Technology and Science, Gwalior, India

*Corresponding Author: ruchi.jayaswal23@gmail.com, Tel.:+91- 8103274614

Available online at: www.ijcseonline.org

Received: 25/May/2017, Revised: 02/Jun/2017, Accepted: 20/Jun/2017, Published: 30/Jun/2017

***Abstract*—** The whole world is filled with a huge collection of digital data, digital images, and videos or can be anything that can be stored in a digitized manner. This data doesn't mean essentially anything. It is stored in an unorganized manner without any interpretation. Image Mining is an energetic concept for researchers. When there is a need to extract necessary information from the massive collection of image database through image mining techniques then this concept came into the picture. In this research paper, the proposed work is done through two steps. One is feature extraction, extract the features of images by RGBHist as a color feature and Edge Histogram Descriptor as a shape feature has taken to create feature dataset. While in second step K-Medoid clustering algorithm is applied to make good clusters and retrieval process is done from the clusters to increase the accuracy of the system. Manhattan similarity method is used a matching purpose from the query image. Three Database is used in this paper for testing the proposed image mining system.

Keywords*—Image Mining, RGB histogram descriptor, Edge Histogram Descriptor (EHD), Content Based Image Retrieval (CBIR), Clustering, K-Medoid Clustering Algorithm, Data Mining, Manhattan Similarity Measure,*

I. INTRODUCTION

Now, everything is digitized, a whole lot of data is available. This data can be anything it may be text, audio, video, images etc. But an image can interpret meaning more than words. These images and data are being extracted from marketing platforms like from online sites, governmental organization, business platforms, hospital records, forgery detection etc. So that images came into a picture to researches and to work on them and make it in a proper order. All the categories of images available are not necessarily useful to the user. To retrieve a useful information according to user's need or query image there is a need to extract the correct knowledge from the data. Basically, Image mining is the concept arise from two vast fields one is Data Mining and another is Image Processing. It deals with the extraction of image patterns from the gigantic collection of image database. The main focus of image mining on two most significant techniques. First one is to mine from the huge amount of images alone and the second technique is to mine from integrated collection of images and its related alphanumeric data. In this paper, image mining is applied to the first technique i.e. massive collection of images alone.

There are six techniques of image mining through which better knowledge can be obtained. It comprises of Object Recognition, Image Indexing, Image Retrieval, Classification, Clustering, and Neural Network. Clustering and Image retrieval techniques are used in this paper for better retrieval of images.

Image retrieval technique is just same as the CBIR. For a system, an image is nothing until significant features have extracted so it is important to extract the low-level features i.e. in numerical form from the image on the basis of color, texture, and shape. Color and shape descriptors are used and make a hybrid feature dataset in this proposed work. Clustering technique is used to make good clusters of images i.e. to generate the patterns of images. It is an unsupervised learning since a pre-defined data set is not available for learning.[22-23] Clustering is defined as the grouping the objects together into clusters where the objects within the cluster are similar to each other while the objects in two different clusters are dissimilar.[23]

The framework of Image mining process is shown in above figure 1. Image directory is select from the image repository. Pre-processing is done on next step i.e. resize the database or remove the noise further step is extract the features of the image on the basis of color, texture and shape descriptors. Subsequently, mine the dataset using mining techniques and evaluate the patterns. Finally, knowledge is obtained according to requirement.

The rest of the paper is described as below: Section 2 briefly summarizes the feature extraction techniques which are applied in this work. Section 3 brief about the clustering technique which is used in proposed work. Section 4 describes the similarity measure metrics. Section 5 discusses the literature survey in this field. Section 6 explains the proposed methodology of the system. Section 7 shows the experimental results of the system and Section 8 conclude the work and their limitations and brief about the future scope related this work.

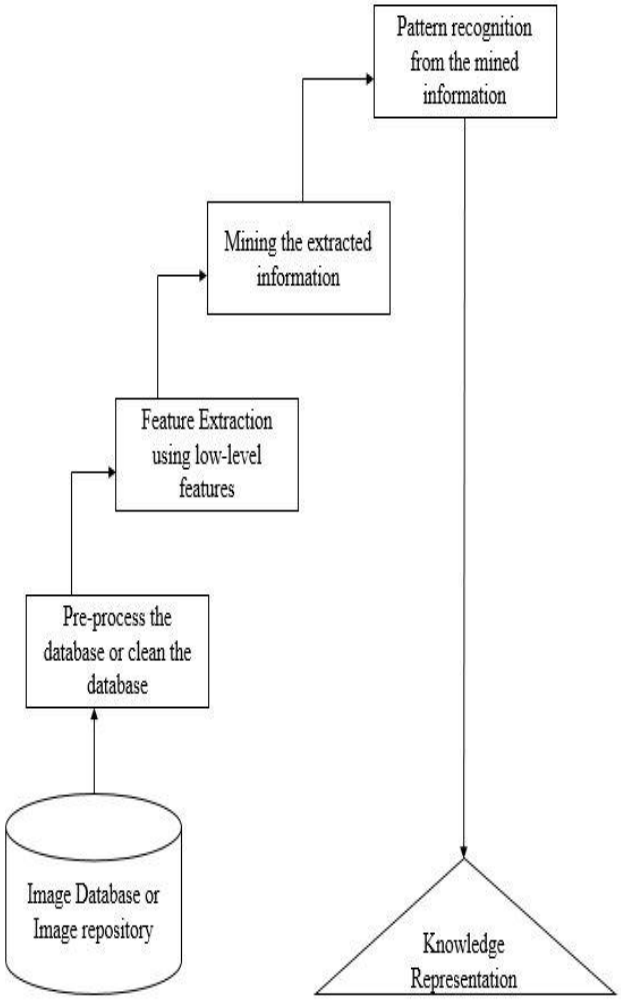

Figure 1 Framework for Image Mining Process

II. FEATURE DESCRIPTORS

There is a need to image is converted into the readable format or in numerical data so that comparison can be done and system retrieve the images. Low-level features are extracted from the image on the basis of color, texture, and shape. Color and shape descriptors will discuss here as they used in this work.

A. RGB Color Histogram

The RGB color model has three primary colors Red, Green, and Blue. RGB color histogram name arises from the RGB color model. The color histogram can be referred as "Three Color Histograms" i.e. RGB each of which shows the brightness distribution of each individual Red/Green/Blue color channel [2] [4].

A color histogram of an RGB image is computed as, a number of bins (for each color channel) is user input and is same for R, G and B. The output of RGB histogram can be un-normalized, L1-normalized or L2-normalized.If the number of bins is 8 then the output is 512(8*8*8) for each channel and get a dimension is 24[4].

B. Edge Histogram Descriptor (EHD)

The histogram is used to refer to the global feature composition of an image and Edge of the image is sensitive to human eyes for image perception. Edge distribution is a descriptor proposed for MPEG-7 [11] only consist of local edge distribution in the image. The size of the histogram should be small for the efficient storage. In MPEG-7, comprise only local edge distribution with 80 bins. Figure.2 illustrate that image is divided into 4*4 sub image further generate an edge histogram to represent edge distribution in the sub-image for each sub image then again divide into small square blocks known as image blocks[12].

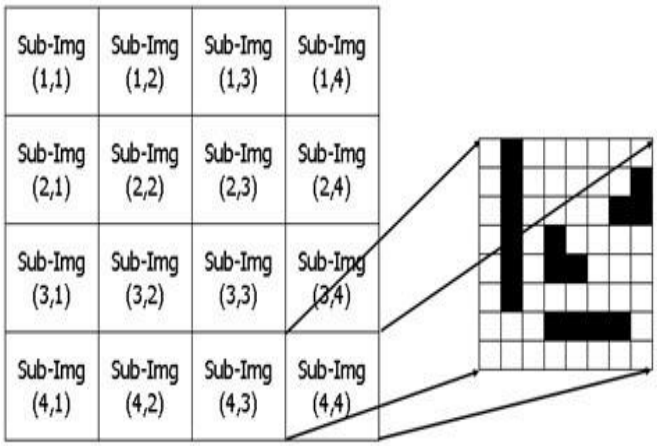

Figure 2. Example of partition of image with image block

There are five kinds of edges in this descriptor of which four are directional edges (vertical, horizontal, 45 degrees and 135 degree) these extracted from image blocks and other is non-directional edge whose contains an arbitrary edge with an absence of directionality [11].

III. MINE THE IMAGES USING CLUSTERING

Clustering technique is the finest method used to organize the data into sensible manner. Organizing the images in a proper sequence or to make good clusters is also a crucial task. It helps in outlier detection and also clustering is adaptable to any change. The distance measure is generally used in estimating the similarity, which implies that the objects belong to the same cluster are separated by minimum distance. There are some methods for clustering such as Partitioning Based Method, Hierarchical Based Clustering, and Density Based Method. K-means, K-Medoids are under the category of Partitioning Based Method. But K-means is sensitive to outliers and K-medoid is less sensitive to outliers. In this methodology, K-Medoid clustering algorithm is used to produce best clusters.

A. K-Medoids Clustering Method

Elect the actual objects to signify clusters in its place of mean values. Each rest object is clustered with the representative medoid to which is the most analogous. The algorithm diminish the sum of the dissimilarities between each object and its corresponding representative object [13][14]

$$E = \sum_{k-1}^{m} \sum_{p \in C_k} (p - o_k) \quad (1)$$

Where,

E is the sum of absolute error for all objects feature dataset.

p is the data point in the space representing an object

O_k is the representative object of cluster C_k.

Euclidean distance as a similarity distance in the K-medoid clustering so that time takes less to execute this code. The cost function is calculated as the difference in absolute error value if a current representative medoid is swapped by a non-representative object.

In Figure 3 the picture illustrates that the dataset is divided into a number of clusters like C1, C2, C3… Cn and in each clusters number of instances or images are residing such as I1, I2, I3… In and each cluster having its own properties that must be different from other cluster properties. In this figure Red cluster have its own properties, Yellow has its own and Green have its own cluster characteristics.

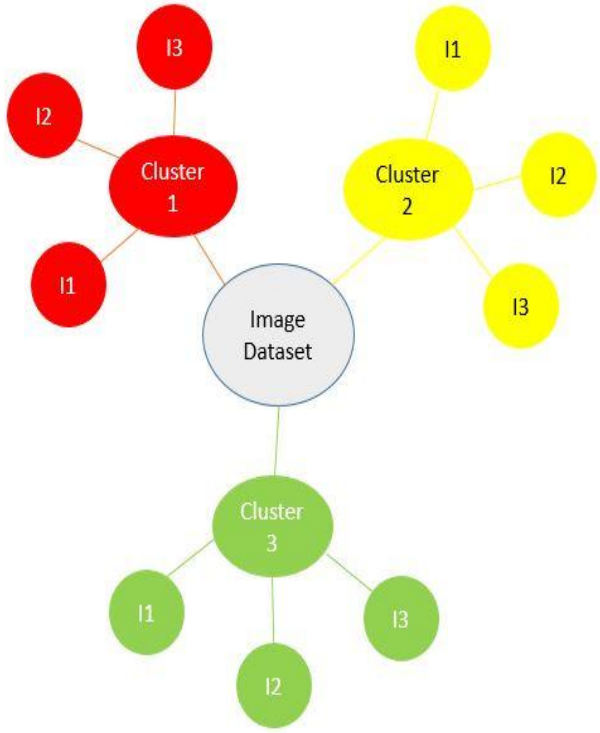

Figure 3. Example of Clustering the Image Dataset

IV. SIMILARITY MEASURE

The similarity measure is defined as the comparison the closest distance between the query image and image database. Selection of similarity metrics has a direct impact on the performance of the system. There are many similarity measures such as Euclidean, Cityblock, Spearman, Mahalanobis, Minkowski distances all have their own specialty. Manhattan Distance is applied in this work [25]. The distance between two images X and Y with n-dimensional features vectors is calculated as follows:

$$D(I,T) = \left(\sum_{i=1 \text{to } n} [I_i - T_i] \right) \quad (2)$$

Where ,

D (I, T) is the Manhattan distance between the feature vector of query image I and every image T in the database. Ii is the feature vector of the query image and Ti is the feature vector of database ages.

V. RELATED WORK

TABLE I The literature survey on some mining techniques and their result analysis

S.No.	Year	Author	Proposed Work	Database Used	Result Analysis
1.	2016	Maria Fayez [1]	Two methods are implemented: In first method GLCM texture feature extract by Hiralick statistical then Kmeans clustering is applied. In the second 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 work overall accuracy is 67.2% and from secondly proposed work accuracy is 86%
2	2014	Snehal Mahajan [2]	Feature extract by RGB color Histogram, RGB Histogram with the canny edge and CLBP descriptors are used then Contribution Based clustering is applied.	Washington University database having 771 images.	From this work precision is 0.80, recall is 0.79 and f-measure is 0.59
3	2012	Ruziana Mohamad Rasli [3]	This paper is done the comparative study between the color histogram with GLCM feature, Color histogram with Gabor and color histogram using K-means and Euclidean distance is used as a similarity measure.	9960 image database has been used.	Precision is color histogram with K-means is 0.90 and Recall is 0.546 both are higher than color hist with GLCM
4.	2014	Y. Poornima [4]	A novel Partitioning clustering algorithm based on the idea of contribution of data points is applied on RGB histogram feature dataset	804 visual art images are taken for testing from Yale University Art Gallery.	From the proposed work 94.86 precision and 92.05 recall are achieved and show it is better than other three K-means and BTC pruning features.
5.	2016	Geethu Varghese [5]	In this proposed work advanced technique of BTC is used known Ordered Dither Block Truncation Coding(ODBTC) & K-means clustering is used then retrieve images repeat three times for better result	Wang image database having 1000 images	Average accuracy is 92.88% after the second iteration of K-means.
6.	2014	Rupinder Kaur[6]	Kmeans is applied on the dataset then the average RGB value is calculated then group the images using hierarchical clustering further images are retrieved.	Microscopic, Landscape and Fish image database are taken	Precision of Microscopic image is 0.71, Landscape images 0.69 and Fish images 0.65 achieved
7.	2016	Ankita Tripathi [7]	MATLAB and WEKA tool is used to perform this method. GLCM feature is used then applied classification in WEKA using J48 algorithm further clustering is applied and visualize the instances.	Textual images are taken having 60 different images	Accuracy from classification images is 66.66% and by clustering is 73.33%
8.	2015	Amit Khatami [8]	Swam intelligence algorithm based on K medoids is proposed to resolve 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.
9.	2016	Katta Sugamya [9]	Classification is done by SVM classifier by color, texture and shape features.	200 images are taken having 2 categories	Accuracy and Error rate is found.
10	2012	Monika Sahu[10]	Texture Feature is taken. Energy, Entropy and contrast features are extracted from images.	Texture based images like bricks, clouds, leaves are used for testing.	Precision is 0.6 and Recall are 0.06.

VI. PROPOSED METHODOLOGY

The proposed system is based two image mining techniques one with image retrieval and other is clustering. For feature extraction, we used color (RGB color histogram) and shape (Edge Histogram Descriptor) combination which yields the perfect information of the images. K-medoids Clustering is used as a Clustering Techniques as described previously. This methodology works well to produce the optimal clusters for the image database and retrieved images as per query image.

The steps of the system are as follows:

Step 1 Select the directory of images for the input.

Step 2 Pre-process is done, resize the database 384*256, 256*256.

Step 3 Apply RGB color histogram to extract the features of the database, no of bins is 3 and Edge Histogram Descriptor feature as shape feature is also applied threshold value is 0.5 is given.

Step 4 Combine the features and make a dataset of the database.

Step 5 Load the created dataset and give the number of clusters K for the k-medoids clustering algorithm.

Step 6 Obtained cluster groups with index numbers for all images like C1, C2, C3… Cn

Step 7 Find Cluster centroid of the clusters and match with the query image cluster number (feature vector).

Step 8 Define some threshold limit of similarity measurement based Manhattan distance so that we can find relevant images from the stored image database.

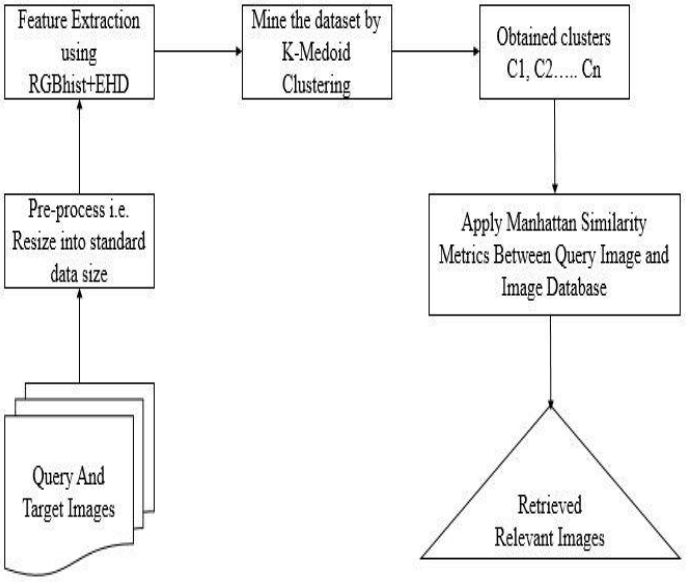

Figure 4. Flow Diagram of Proposed System

VII. EXPERIMENTAL ANALYSIS AND OUTCOMES

A. Tool and Image Database

The tool in which this proposed work is implemented in MATLAB R2013b version. There is three databases are used to test this system Coral [21] (2000), Wang [5] (1000) and Caltech [21] (468).

Figure 5. Coral Database

Figure 6. Wang Database

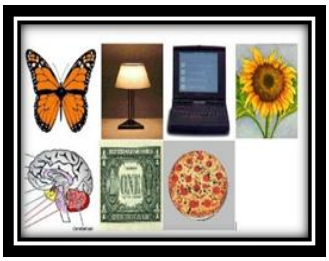

Figure 7. Caltech Database

B. Performance 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 [24].

$$P_C = {M_R}/{N_R} \quad (3)$$

$$R_C = {M_R}/{N_T} \quad (4)$$

$$FM = {(2 * P_C * R_C)}/{(P_C + R_C)} \quad (5)$$

Where P_C = 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

Performance metrics are shown in Table 2 in which Precision is computed as the total number of germane images from the total number of images retrieved shown in output window here, the window size is 50. The recall is computed as the average of a total number of germane images retrieved from the whole image database. FMeasures is computed by the amalgamation of precision and recall values and calculate the accuracy of the retrieved image [24].

Table 2 Performance measure on Wang Database

Categories	RGB+EHD descriptor N=50			RGB+EHD+K-Medoid N=50		
	Precision	Recall	F-measure	Precision	Recall	FM-easure
Tribes	0.35	0.17	0.22	0.9933	0.496	0.667
Beach	0.42	0.22	0.288	0.966	0.483	0.644
Monuments	0.47	0.23	0.308	0.966	0.483	0.644
Buses	0.46	0.24	0.315	0.977	0.499	0.66
Dinosaurs	0.95	0.47	0.628	1	0.5	0.66
Elephants	0.44	0.22	0.301	0.966	0.483	0.644
Flowers	0.95	0.47	0.628	1	0.5	0.66
Horses	0.66	0.34	0.448	0.988	0.493	0.66
Mountains	0.31	0.15	0.201	0.966	0.48	0.644
Food	0.33	0.16	0.215	1	0.5	0.66
Average	0.534	0.26	0.349	0.98223	0.491	0.65

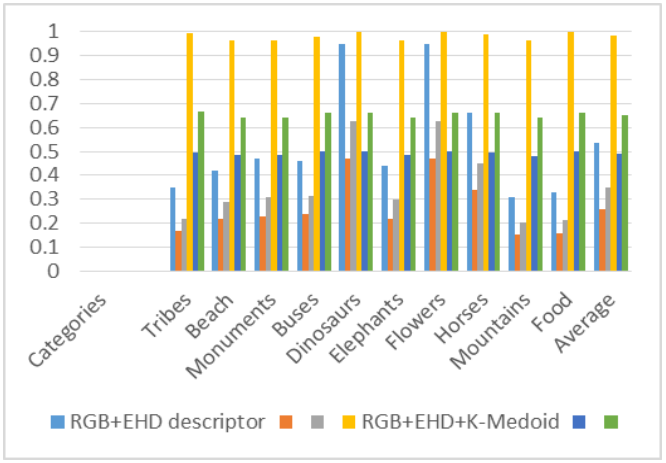

Figure 8. Bar graph shows the performance metrics on Wang database

Table 3 Comparison with other conventional systems from proposed system on Wang database

Categories	Yogita's Method [19]	Naveena's Method[20]	Proposed Method
Tribes	0.62	0.4407	**0.9933**
Beach	0.71	0.5745	**0.966**
Monuments	0.65	0.6	**0.966**
Buses	0.82	0.875	**0.977**
Dinosaurs	0.99	0.9804	**1**
Elephants	0.66	0.6552	**0.966**
Flowers	0.98	0.9302	**1**
Horses	0.58	0.902	**0.988**
Mountains	0.59	0.5556	**0.966**
Food	0.62	0.5345	**1**
Average Precision	0.828	0.70481	**0.98223**

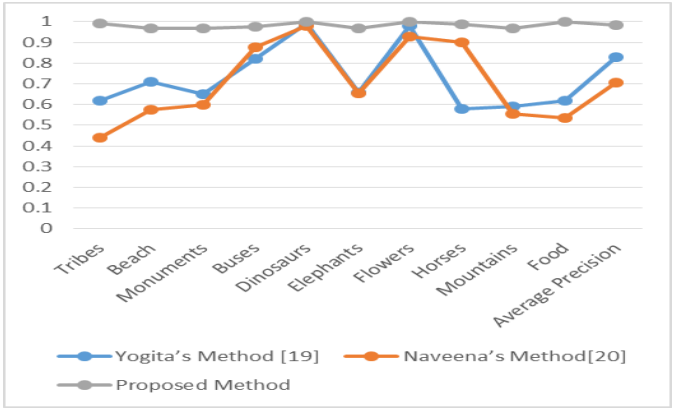

Figure 9. Graph shows the proposed system precision metrics is higher than another system

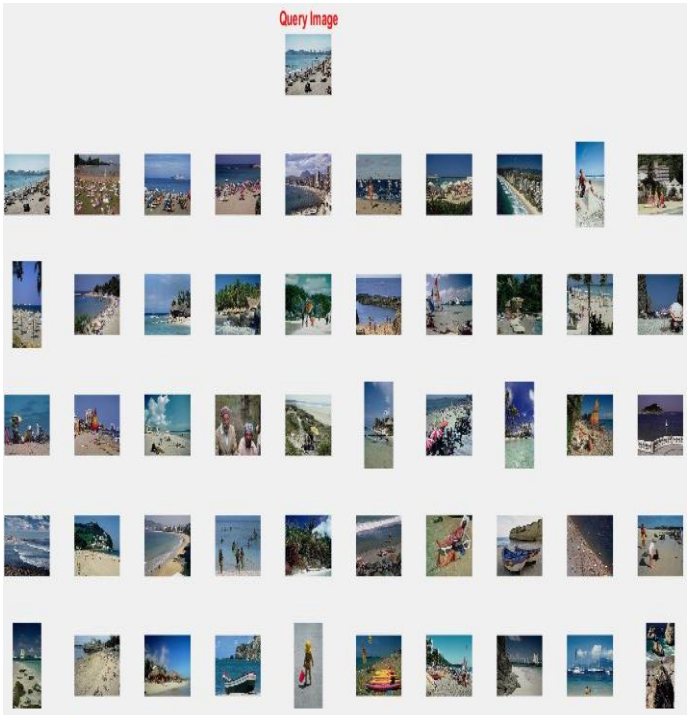

Figure 10. Outcomes of beach category from proposed method

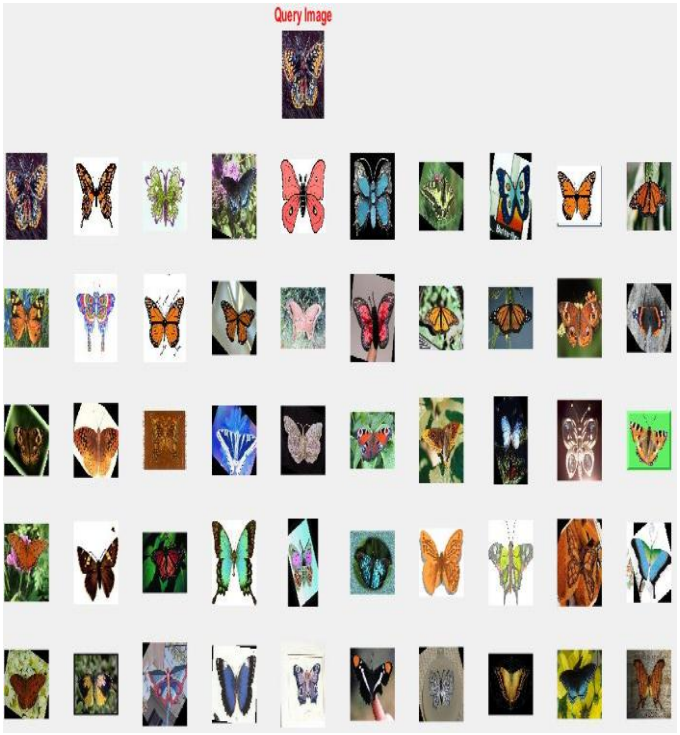

Figure 12. Outcomes of Caltech database category from proposed system

Table 4. Average performance metrics of Caltech and Coral database

Database / Performance Metrics (N=50)	CALTECH Database(468)	CORAL Database(2000)
Average Precision	0.945	0.985
Average Recall	0.77	0.492
Average F-Measure	0.848	0.656

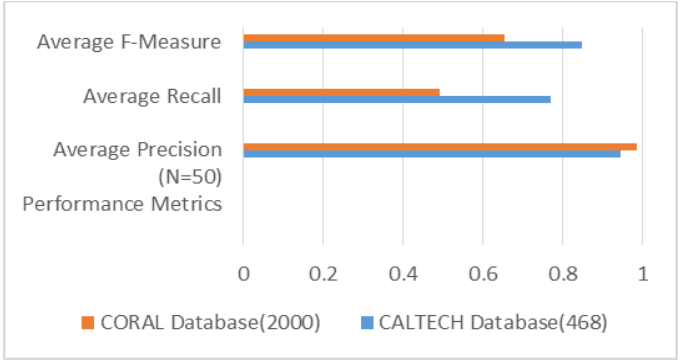

Figure 11. Bar Graph shows the performance of another database

Figure 13. Outcomes of Coral database category from proposed system

VIII. CONCLUSION AND FUTURE SCOPE

The conclusion of the entire discussion is that image mining system is the current research area for the researchers. The general standard image database has taken and implemented on it. Fused feature descriptors color RGB histogram and EHD are the best combination to create this dataset. Precision, Recall, F-measure are used to check the efficiency of the system.

Optimized Clustering can be applied by using optimization algorithm like PSO, Ant colony, Elephant Herding, Genetic algorithm, Cuckoo search etc. Another clustering algorithm can also be applied and reduce the drawbacks of K-Medoid algorithm.

REFERENCES

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

[2] Snehal Mahajan, Dharamaraj Patil, "*Image Retrieval Using Contribution-based Clustering", Algorithm with Different Feature Extraction Techniques*"978-1-4799-3064-7/14 IEEE

[3] Ruziana Mohamad Rasli, T Zalizam et al,"*Comparative Analysis of Content Based Image Retrieval Technique using Color Histogram. A Case Study of GLCM and K-Means Clustering*", 978-0-7695-4668-1/12 IEEE DOI 10.1109/ISMS.2012.111

[4] Y. Poornima, P.S. Hiremath,"*Efficient Modelling of Visual Art Color Image Clustering*", International Journal of Computer Applications (0975 – 8887) Volume 91 – No.11, April 2014

[5] 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

[6] Rupinder Kaur, Navleen Kaur,"*Content Based Image Mining Technique for Image Retrieval Using Optimized Hybrid Clustering*", International Journal of Computer Trends and Technology (IJCTT) – Volume 11 number 3 – May 2014

[7] Ankita Tripathi, Shivam Pandey,"*An Improved and Efficient Image Mining Technique for Classification of Textual Images Using Low-Level Image Features*", 10.1109@inventive.2016.7823220

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

[9] Katta Sugamya, Suresh Pabboju, Dr. A.VinayaBabu." *A CBIR Classification using support vector machines*", International Conference on Advances in Human Machine Interaction (HMI - 2016), 978-1-4673-8810-8/16 IEEE.

[10] Monika Sahu, Madhup Shrivastava "*Image Mining: A New Approach for Data Mining Based on Texture*", 978-0-7695-4872-2/12 IEEE DOI 10.1109/ICCCT.2012.11

[11] Dong Kwon Park, Yoon Seok Jeon, and Chee Sun Won, " *Efficient Use of Local Edge Histogram Descriptor*," ACM workshop on Multimedia, November 2000, pp. 51-55

[12] 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 IEEE.

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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, IEEE

[18] Neethu Joseph.c, Aswathy Wilson," *Retrieval of Images using Data Mining Techniques*",978-1-4799-6629-5/14 IEEE

[19] Yogita Mistry, D. T. Ingole, and M. D. Ingole," *Efficient Content Based Image Retrieval Using Transform and Spatial Feature Level Fusion*", The 2nd International Conference on Control, Automation and Robotics. 978-1-4673-9859-61/16 IEEE.

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

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

[22] Clustering- J. A. Hartigan, "*Clustering Algorithms*", JohnWiley & Sons, 1975.

[23] [9] P. Arabie, L. J. Hubert, and G. De Soete, "*Clustering and Classification",* World Scientific, 1996.

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

[25] Rajkumar Jain, and Punit Johari, "*An Improved Approach of CBIR using Color Based HSV Quantization and Shape Based Edge Detection Algorithm*," 978-1-5090-0774-5/16 ©2016 IEEE

[26] 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).

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

[28] 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).

[29] 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

[30] Rupali Sawant,"*Image Mining Using Texture and Shape Feature*", International Journal of Computer Sciences and Engineering, Vol 02, No. 09, 2010, 2808-2815

[31] Ansari Mohd Aquib, Dixit Manish*,"An Image Retrieval Framework: A Review*", International Journal of Advanced Research in Computer Science,[S.l.], v. 8, n. 5, p. 692-699, June 2017. ISSN 0976-5697.

Authors Profile

Ms. Ruchi Jayaswal has received the Bachelor of Engineering degree in Computer Science and Engineering from ITM Universe, RGPV University M.P. in 2015. She is currently pursuing her Master of Technology from MITS, Gwalior, M.P. Her research interest is included Digital Image Processing, Image Mining, Data Mining and Image Retrieval. She is currently a member of IEEE organization and has also published one research paper in International conference and other in the national conference.

Mrs. Jaimala Jha has received the Bachelor of Engineering degree in Computer Science and Engineering from MITS, Gwalior M.P. and Master of Technology from SATI, Vidisha, MP. She is currently working in MITS as an assistant professor in computer science and engineering department. She has 7-year experience of teaching experience and pursuing Ph.D. from Rajeev Gandhi Technical University, Bhopal. She has published 12 international journals, 08 national and 05 international conferences research papers.

Mr Ravi Devesh has received his B. Tech degree in Computer Science and Engineering from MANIT Bhopal, in 2014. He is currently pursuing M. Tech in Information Technology from MITS Gwalior, RGPV University, Bhopal. His research interests comprise of Content Based Image Retrieval, Image Mining and Image Processing.

