International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

A Comparative Analysis of Distance Measurement Techniques for Image Retriveal Using Clustering Techniques

Dr. S. Hari Ganesh[1], A. Philip Arun Kennedy[2]
[1]Assistant professor, H.H The Rajah's College, Pudukkottai- 622 001
[2]Research scholar, H.H The Rajah's College, Pudukkottai-622 001
Email address: [1]hariganesh17[AT]gmail.com, [2]philip.arun23[AT]gmail.com

***Abstract*—** *Image retrieval is the dealing out of looking and retrieving photographs from a massive dataset. As the images grow complex and diverse, retrieval the right images becomes a difficult challenge. For centuries, most of the images retrieval is text-based which means searching is based on those keyword and text generated by human's creation. On the whole a content based image retrieval (CBIR) system will retrieve some of the aspects image like shape, texture, color and spatial information of each image which is placed in the database and then stores the feature details in a different database called the feature database. The characteristic database carries the function facts of all the snap shots present inside the major database. The feature records are very small in length whilst in comparison with the original picture. The feature database holds the description of the main image in a Compact format. It holds information about the coloration, shape, texture and spatial statistics in a set period actual-valued multi thing function vectors or signature. We can retrieve the feature vectors based on similarity measurements. In this paper, we can survey various similarity measurements in K-means clustering to retrieve the images from image database and a comparison study is also made on the various measurements using precision, recall and F1 score.*

***Keywords*—** *Content based image retrieval, feature vectors, similarity measurements, clustering approach.*

I. INTRODUCTION

Content based image retrieval is well known technology being used for the retrieval of images from large database. This image retrieval is a challenging topic that has been a research focus from many years. This has proven very much important because of its applications like face recognition, fingerprint recognition, pattern matching, verification and validation of images. The image retrieval is also called image classification in large database systems. In the past few years, there has been tremendous growth in database technology to store and retrieve large number of images. This requirement creates a demand for software systems for effective fast image retrieval from large database systems .The demand and use of multimedia applications in present world creates the need of content based image search and retrieval. The term content based image retrieval(CBIR) is originated by Kato from his work to retrieve images from database based on color and shape .Since then onwards the term CBIR is used for the process of searching and retrieving desired image from collection of database based on synthetically image features like color, texture and shape. The content based totally photograph retrieval is an essential software in scientific subject that is used to permit radiologist to retrieve images of comparable functions for input photo that result in similar analysis. Every CBIR system needs to have a module for feature extraction. This module is applied on query image and as well as on database of images. This module converts photo into binary shape and reveals its capabilities like shape, coloration, texture then it includes some other module referred to as similarity matching which is used to examine the input photo functions with functions of database pix. To find the features of image, the image is converted into wavelet histogram which is in binary matrix. For this we have a technique called Haar wavelet.

A. Color spaces

Color spaces are important component in representing images in digital form. Color is usually represented by color histogram, color coherence vector which are combined called as color space. The color histogram feature is most important in image retrieval. Here color histogram is a vector that's series of detail where every element represents no of pixels in a bin of image.

B. Texture

It is another important feature generally used in similarity matching of images. Texture means smoothness, coarseness and regularity. The texture features are analyzed using statistical approach.

In our proposed system we proposed to enforce content based totally photograph retrieval machine with determining assist Here we are using K-means clustering technique of data mining to create index of images in database once index is created then implements retrieval and search process using index. Generally in conventional image search and retrieval technique whenever if searches a question photograph in opposition to database then the query photo is proven with each photo in database. During this process feature extraction is performed for query images and database images in every search process. Similarly, similarity matching is also performed in every search process. Thus the search process is costly and time consuming every time. In case of proposed system, first K-means clustering creates index and images are organized in clusters in memory. Every cluster is collection of similar images based on content. Now each time seek system

Dr. S. Hari Ganesh and A. Philip Arun Kennedy, "A comparative analysis of distance measurement techniques for image retriveal using clustering techniques," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 209-214, 2017.

is implemented the feature extraction and similarity matching are achieved on clusters rather than man or woman photographs. This proves the process is improved and less cost than traditional one. The basic layout of CBIR is shown in fig. 1.

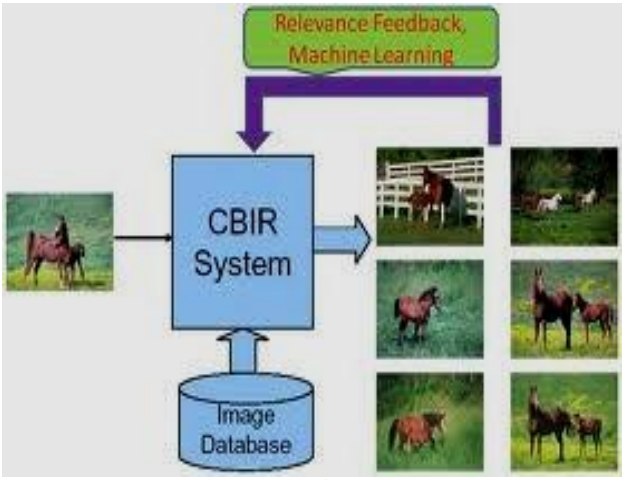

Fig. 1. CBIR framework.

II. RELATED WORK

Ying Liu, et al. [1] analyzed the fundamental difference between content material-based and text-primarily based retrieval structures is that the human interaction is an essential part of the latter device. Humans tend to use high-level structures (concepts), such as text descriptors, keywords to interpret images and measure their similarity. While the functions mechanically extracted the use of computer vision strategies are typically low-stage functions (color, texture, shape, spatial format, and many others). In general, there is no direct link between the high-level concepts and the low-level features. Low-level image feature extraction is the basis of CBIR systems. To overall performance CBIR, photo functions can be both extracted from the entire photograph or from areas. As it has been found that users are usually more interested in specific regions rather than the entire image, most current CBIR systems are region-based. Global feature based retrieval is comparatively simpler. To perform RBIR, the first step is to implement image separation. Then, low-level structures such as color, shape, texture or spatial location can be extracted from the segmented areas. Comparison between two images is defined based on region features. In this set of rules, the spatial proximity of every vicinity is taken into account by using defining a new center for the ok-approach algorithm and by using integrating the okay-method with a issue labeling manner.

M. Banerjee, et al. [2] proposed human visual system is highly efficient in sorting and selecting similar images from a very large collection. In this process of selection, a limited number of visually prominent features may be used to evaluate similarity between images. High curvature points play a significant role in characterizing an object with limited number of pixels compared to the total number of pixels. Designing a system using such features for retrieval mechanism will be both fast and cost effective proposition. The local features (color, texture etc.) are computed on a window of regular geometrical shape surrounding the corner points. General purpose corner detectors are also used for this purpose. However the curvature points may be of different types (sharp, medium, weak). The characteristics of sharp curvature points will be confined within a small region but for that of medium and weak type the region will be larger. These facts indicate that extracting the possible high curvature region of interest (roi) where shape and size of the extracted (roi) varies adaptively according to the nature of curvature type could be a better solution. It may be considered an alternative to segmenting an image and using it in overall scene matching applications. An efficient CBIR system should be able to handle imprecise image data, to some extent the difference arising due individual perception in evaluating similarity between images. A Fuzzy set theoretic based approach may be considered a good choice for handling uncertainties arising at different stages of processing and analysis of a CBIR system.

Hui Yu, et al. [3] propose a novel low-level feature, named color texture moments, for representing image contents. It is able to integrate the color and texture characteristics of an image in one compact form. Preliminary experimental results show that the new feature achieves better performance than many existing low-level features. More importantly, the dimension of this new feature is only 48, much lower than that of many features with good performance. Furthermore, the function extraction algorithm could be very smooth to put into effect. It is precious for the improvement and alertness of the CBIR systems. The experimental outcomes certify the effectiveness of the (SVcosH, SVsinH, V) colour area.The performance of the proposed feature is compared with that of RGB based and HSV-based features as well as grey-scale texture features. On the average, the color features are better than the grayscale features in terms of retrieval precision. The consideration of the color information enhances the representation ability of color texture. The comparison between RGB and HSV demonstrates that the HSV color space is more suitable for simulating and analyzing visual perception.

N. Jhanwar, et al. [4] presented a technique for content based image retrieval which effectively retrieved images using MCM. MCM are computationally inexpensive but were sensitive to translation. A heuristic method defined as sensitivity test was useful in diminishing the effect of limited translation. We also found that MCM is useful in encoding the dependency of both color and texture. In future MCM could be extending in multi-resolution since the effect of translation diminishes into lower resolution. One can try to find out the optimal resolution where the translation effect is minimized and use feature corresponding to these layer for retrieval purpose thereby eliminating the need for sensitivity test. The information which the motif contains is how pixel intensities vary in a local neighborhood i.e. a grid. Consider two grids at a distance k from another grid which are used to update the probabilities in MCM. Consider both the grids could individually be traversed optimally using all the six scans. This implies that all the values local to the two grids are same.

Dr. S. Hari Ganesh and A. Philip Arun Kennedy, "A comparative analysis of distance measurement techniques for image retriveal using clustering techniques," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 209-214, 2017.

Therefore there is no variation in the intensities of pixels in these two grids. It implies that both of these grids belong to a homogeneous region and there is no texture local to both of these grids which could be captured. Updating probabilities using these two grids would give us wrong information since our feature would suggest that these two grids are actually traversed optimally by this particular motif which would be an error. Moreover the error is not equally distributed in our feature vector. The change in probabilities would be more for the pairs of motifs which occur less frequently in the image then the pairs which occur more frequently.

III. Image Retrieval using Clustering Algorithm

Image mining refers to set of tools and strategies to explore pix in an automatic method to extract semantically meaningful data. The retrieval process represents a visual query to the system and extracts the images based on the user request such mechanism referred to as query-by-example and It requires the definition of an photograph representation a fixed of descriptive functions and of some similarity metrics to examine question and goal photographs. The additional mechanisms have been introduced to achieve better performance and relevance feedback proved to be a powerful tool to iteratively collect information from the user and transform it into a semantic bias in the retrieval process. RF increases the retrieval overall performance and it enables the system to examine what's applicable or inappropriate to the user throughout successive retrieval-remarks cycles. RF approaches critical issues yet unexplained. And user interface is time overriding and tiring, and it is needed to reduce as much as possible the number of iterations to convergence This is particularly difficult when only a few new images are retrieved during the first RF steps and no positive examples are available for successive retrieval.

Image Mining

Image mining is an extension of data mining to image domain and it is an interdisciplinary endeavor that draws upon expertise in Data mining, Computer vision, Image retrieval, Image processing, Database, Machine learning, and artificial intelligence. The development of image acquisition and storage technology have led to tremendous growth in very large and detailed image databases and it can reveal useful information to the human workers. Image mining contracts with the extraction of implied knowledge, image statistics relationship, and other patterns not explicitly stored in the images. This one is differs from data mining in respect of the data and the nature of the data. Image mining has led to tremendous growth in significantly large and detailed image databases. The most important areas belonging to image mining are the image knowledge extraction; content based image retrieval, video sequence analysis, video retrieval, change detection, object recognition as well as model learning,.

Image Clustering and Techniques

Clustering is a way of grouping statistics objects into special organizations, such that similar statistics objects belong to the identical group and distinct facts objects to special clusters. Current research increasing interest in digital image searching, identification, classification, storage and management. Some common but important applications of are person identification in movie clips and recognition in biometric system, festive home videos, commercials filtering, natural scene classification for robot vision, segmentation of important topics in lectures and meetings. The image clustering, an important technology for image processing, takes been aggressively researched for a long period of time and explosive development of the Web, image grouping has even been a critical technology to help users digest the large amount of online visual information. The most common clustering algorithm is K-Means Clustering algorithm. K-means (Macqueen) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K Means algorithm in CBIR systems is used to initialize the clusters. The most important idea is to define okay centroids, one for every cluster. These centroids need to be positioned in a cunning way due to distinctive place causes distinct end result. So, the better choice is to place them as much as possible far away from each other. The subsequent step is to take every point belonging to a given facts set and partner it to the nearest centroid. After no point is pending, the first step is completed and an early group age is done. At this point we want to re-calculate okay new centroids as bary center of the clusters as a result of the previous step. After we've these k new centroids, a new binding must be achieved among the same facts set points and the nearest new centroid. A hoop has been generated. As a result of this loop we may additionally observe that the ok centroids trade their region little by little till no more changes are executed. In other words centroids do now not circulate any more. Finally, this set of rules goals at minimizing a goal function, in this situation squared blunders function.

IV. Distance and Its Significations

Clustering is an important data mining technique that has a wide range of applications in many areas like biology, medicine, image analysis and market research etc. which is the process of partitioning a set of objects into different subsets such that the data in each subset are similar to each other. In Cluster analysis Distance measure and clustering algorithm plays an important role. In order to measure the similarity or regularity among the data-sets, distance metrics plays a very important role. It is necessary to identify, in what manner the data are interrelated, how various data dissimilar or similar with each other and what measures are considered for their comparison. The main purpose of metric calculation in specific problem is to obtain an appropriate distance /similarity function. Many distance measures have been proposed in K-means clustering. Some measurements are listed below in fig. 2.

Dr. S. Hari Ganesh and A. Philip Arun Kennedy, "A comparative analysis of distance measurement techniques for image retriveal using clustering techniques," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 209-214, 2017.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

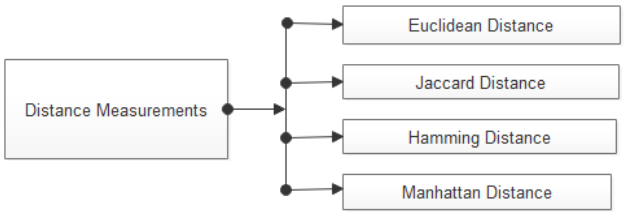

Fig. 2. Distance measurement techniques.

We can elaborate these distances in following sections.

Euclidean Distance:

The Euclidean distance function measures the distance. The formula for this distance between a Point X (x1, x2,…., xn) and a Point Y (y1, y2,…, yn) is:

$$\text{Dist}(X,Y)=\sqrt{\sum_{i,j=1}^{m,n}(x_{ij}-y_{ij})^2} \quad (1)$$

Developing the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values.

Jaccard Distance:

The Jaccard coefficient, which is every other similarity measure, additionally known as the Tanimoto coefficient, is recycled to degree the comparison in the intersection divided by the union of the objects. The Jaccard distance, which measures dissimilarity between sample groups, is matching to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1 or equivalently by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

$$\text{Dist}(X,Y)=1-J(X,Y) \quad (2)$$

X, Y is the image points and J is Jaccard Co-efficient
Where as

$$J(X,Y)=\frac{|(X\cup Y)|-|(X\cap Y)|}{|(X\cup Y)|} \quad (3)$$

For photo, the use of Jaccard coefficient is to make a assessment of the sum weight of shared terms and the sum weight of terms presented in either of the two documents but in condition that they are not the shared terms

Manhattan Distance:

Manhattan distance is also named as image block distance because it is a distance the image pixel would drive in a image put out in square blocks like Manhattan

$$Dist(X,Y)=\left\|x_{ij}-y_{ij}\right\| \quad (4)$$

X, Y are image points

Manhattan distance is also known as L1 distance. The distance between two points is the absolute difference between the points. Absolute value distance gives more robust result whereas Euclidean influenced by unusual values.

Hamming Distance:

The Hamming distance which refers to difference between strings of equal period is the quantity of positions for which the corresponding symbols are different

$$Dist(X,Y)=\sum_{i=1}^{n}|x_i-y_i| \quad (5)$$

$$x=y\rightarrow Dist=0$$

$$x\neq y\rightarrow Dist=1$$

X,Y are image points

The Hamming distance may be interpreted as the variety of bits which want to be changed (corrupted) to show one string into other. Sometimes the range of characters is used in location of the number of bits. Hamming distance can be visible as Manhattan distance among bit vectors. The merits and demerits of the distance measurement are shown in table I.

TABLE I. Merits and demerits of distance measures.

Distance measure	Merits	Demerits
Euclidean Distance	Flexible to support all data	Does not work on large datasets
Jaccard Distance	Reduce complexity in computation	Does not work for small values
Manhattan distance	Provide generalized model in distance measurements	Can't be implemented in image datasets
Hamming Distance	Easy to calculate distance values	Only support Document clustering

V. PROPOSED FRAMEWORK

The proposed work focus to provide the framework for retrieving images from datasets using K means clustering with various distance measurement. Each image has three feature Shape and Texture, color, for fast and improve Image retrieval performance we are using color feature extraction. The proposed framework is shown in fig. 3.

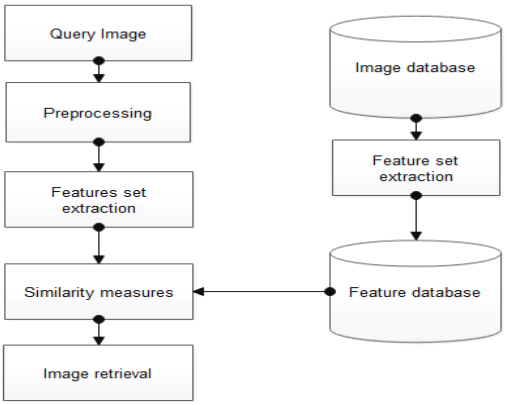

Fig. 3. Proposed framework.

Using color function extraction first of all we converted color photo into gray degree that is containing values from 0 to 255. The images are kept in database called image database. After preprocessing, images are clustered by using the method K- Means Clustering. Clustering is a way of grouping together information samples which can be comparable in some way consistent with some standards that we pick its shape of unsupervised studying So, it's a method of data exploration – a way of looking for patterns or structure in the data that are of interest. Clustering algorithms are generally used in an unsupervised fashion. They are provided with a hard and fast of data instances that have to be grouped in keeping with a few notion of correspondence. The algorithm devises access only to the set of features describing each object; it is not given any information as to where each of the instances should

Dr. S. Hari Ganesh and A. Philip Arun Kennedy, "A comparative analysis of distance measurement techniques for image retriveal using clustering techniques," *International Research Journal of Advanced Engineering and Science*, Volume 2, Issue 3, pp. 209-214, 2017.

be placed within the partition. K-way clustering is a method generally used to mechanically partition a statistics set into okay organizations. It proceeds by selecting k initial cluster centers and then iteratively refining the results. The algorithm converges when there is no further change in assignment of instances to clusters. In this K means cluster we can implement two types of distance measurement such as Euclidean and Manhattan with K means clustering. The Algorithm Steps as follows:

Algorithm: Basic Euclidean Distance metric in K means

Input: X = {x1, x2, x3,…..,xn} be the set of data points, Y= {y1,y2,y3…yn} be the set of data points and V = {v1,v2,v3,….,vn} be the set of centers

Step 1: Select 'c' cluster centers arbitrarily

Step 2: Calculate the distance between each pixels and cluster centers using the Euclidean Distance metric as follows

$$Dist(X,Y) = \sqrt{\sum_{j=1}^{n}(X_{ij} - Y_{ij})^2} \quad (6)$$

X, Y are the set of data points

Step 3: Pixel is assigned to the cluster center whose distance from the cluster center is minimum of all cluster centers

Step 4: New cluster center is calculated using

$$V_i = \frac{1}{c_i}\sum_{1}^{ci} x_i \quad (7)$$

Where Vi denotes the cluster center, ci denotes the number of pixels in the cluster

Step 5: The distance among every pixel and new obtained cluster facilities is recalculated

Step 6: If no pixels were reassigned then stop otherwise repeat steps from 3 to 5

Algorithm: Manhattan Distance in K means clustering

Input: X = {x1, x2, x3,…..,xn} Y= {y1,y2,y3,…,yn}be the set of data points and V = {v1,v2,v3,….,vn} be the set of centers

Step 1: Select 'c' cluster centers arbitrarily

Step 2: Calculate the distance between each pixels and cluster centers using the Manhattan metric as follows

$$Dist(X,Y) = \left\| x_{ij} - y_{ij} \right\| \quad (8)$$

Step 3: Pixel is assigned to the cluster center whose distance from the cluster center is minimum of all cluster centers

Step 4: New cluster center is calculated using

$$V_i = \frac{1}{c_i}\sum_{1}^{ci} x_i \quad (9)$$

Where xi denotes as data points, Vi denotes as cluster centroids and ci denotes the number of pixels in the cluster

Step 5: The distance between each pixel and new obtained cluster centers is recalculated

Step 6: If no pixels were reassigned then stop otherwise repeat steps from 3 to 5

For similarity comparison between the query image and the database picture. Using a suitable threshold, images that are semantically closer are retrieved from the database and displayed as a thumbnail.

VI. Experimental Results

There are a number of evaluation matrices are used to evaluate the retrieval performance. For matric evaluation, we are using precision and recall. Where, precision measure the availability of relevant images from the retrieved image in CBIR system and recall measure the availability of relevant images from the retrieved image over the total number of relevant images in the database.

$$\text{Precision} = \frac{\text{No of relevant images extracted}}{\text{Total no of images extracted}}$$

$$\text{Recall} = \frac{\text{No of relevant images extracted}}{\text{Total no of images in database}}$$

To analyses the visual similarity of CBIR system, various types of distance measures are used. We took some random images from each class and applied these images one by one and retrieved top 40 images. Then calculate average precision and average recall for every class. Result shown that Manhattan distance measure provided the better result in comparison of Euclidian distance measure. The performance chart is shown in fig. 4.

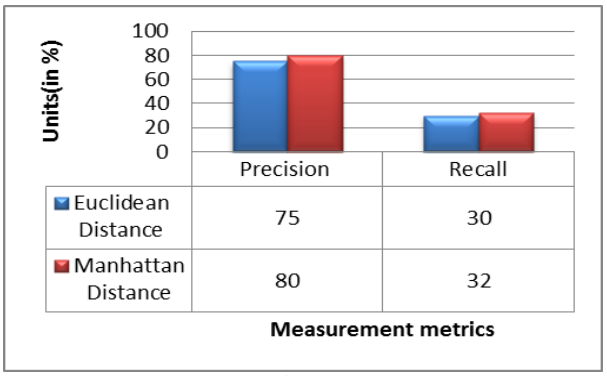

Fig. 4. Performance chart.

VII. Conclusion

In this paper, different types of distance measure techniques are describe. The main purpose of this paper to analyze the performance of Euclidean and Manhattan similarity measures with K-Means clustering. Here, Euclidian and Manhattan measure applied on texture and shape based features. We find that Manhattan Distance produced the better result in comparison with other distance measures in image retrieval system.

References

[1] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, issue 1, pp. 262-282, 2007.

[2] M. Banerjee, M. K. Kundu, and P. K. Das, "Image retrieval with visually prominent features using fuzzy set theoretic evaluation," *IET International Conference on Visual Information Engineering*, pp. 298-303, 2006.

[3] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," *IEEE Proceedings International Conference on Image Processing*, vol. 3, 2002.

[4] N. Jhanwar, S. Chaudhuri, G. Seetharaman, B. Zavidovique, "Content based image retrieval using motif cooccurrence matrix," *Image and Vision Computing*, vol. 22, issue 14, pp. 1211-1220, 2004.

[5] V. Vijaya Kumar, N. Gnaneswara Rao, A. L. Narsimha Rao, and V. Venkata Krishna, "IHBM: Integrated histogram bin matching for similarity measures of color image retrieval," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 2, no.3, 2009.
[6] R. Gonzales and R. E. Woods, *Digital Image Processing*, 2nd Edition, New Jersey Prentice Hall, 2002.
[7] S. Somnugpong and K. Khiewwan, "Content based image retrieval using a combination of color correlograms and edge direction histogram," *13th International Joint Conference on Computer Science and Software Engineering*, 2016.
[8] G. F. Ahmed and R. Barskar, "A study on different image retrieval techniques in image processing," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 1, issue 4, pp. 247-251, 2011.
[9] M. B. Rao and Dr. D. P. Rao, and Dr. A. Govardhan, "Content-based image retrieval system based on dominant color and texture features," *International Journal of Computer Applications*, vol. 18, no.6, pp. 40-46, 2011.
[10] N. Bagri and P. Kumar Johari, "A comparative study on feature extraction using texture and shape for content based image retrieval," *International Journal of Advanced Science and Technology*, vol. 80, pp. 41-52, 2015.
[11] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 610–621, 1973.
[12] S. Redrouthu and Annapurani. K, "Time comparison of various feature extraction of content based image retrieval," *International Journal of Computer Science and Information Technologies*, vol. 5, issue 2, pp. 2518-2523, 2014.
[13] M. A. Z. Chahooki and N. M. Charkari, "Shape retrieval based on manifold learning by fusion of dissimilarity measures," *IET Image Processing*, vol. 4, issue 6, pp. 327-336, 2012.
[14] T. Deselaers, D. Keysers and H. Ney, "Feature for image retrieval: An experimental comparision," *Springer*, pp. 1-22, 2007.
[15] B. Ramamurthy and K. R. Chandran, "Content-based medical image retrieval with texture content using gray level cooccurrence matrix and K-Means clustering algorithms," *Journal of Computer Science*, vol. 8, issue 7, pp. 1070-1076, 2012.
[16] C.-H. Lin, R.-T. Chen, and Y.-K. Chan, "A smart content-based image retrieval system based on color and texture feature," *Image and Vision Computing*, vol. 27, issue 6, pp. 658–665, 2009.
[17] P. Hiremath and J. Pujari, "Content based image retrieval using color, texture and shape features," *15th International Conference on Advanced Computing and Communication*, 2007.
[18] R. Gali, M. deval, and R. Anand, "Genetic algorithm for content-based image retrieval," *Forth Conference on Computational Intelligence, Communication System and Network*, 2012.
[19] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
[20] M. Narottambhai Patel and P. Tandel, "A survey on feature extraction techniques for shape-based object recognition," *International Journal of Computer Applications* (0975-8887), vol. 137, no. 6, 2016.