IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Feature Extraction Approach for Content Based Image Retrieval

Komal Ramteke[1], Ashwini Vinayak Bhad[2]

Assistant Professor, Dept. of IT, RGCER, Nagpur University, Nagpur, Maharashtra, India[1]

Student, Dept. of C.S.E, RGCER, Nagpur University, Nagpur, Maharashtra, India[2]

Abstract: Content Based Image Retrieval (CBIR) is a significant and increasingly popular approach that helps in the retrieval of image data from a huge collection. Image representation based on certain features helps in retrieval process. Three important visual features of an image include Color, Texture and Histogram. Here image retrieval techniques used are color dominant, texture and histogram features. Using that technique, as a first step an image can be uniformly divided into coarse partitions. GLCM (Gray Level Co-occurrence Matrix) is used here for texture representation for image retrieval based. Although a precise definition of texture is untraceable, the notion of texture generally refers to the presence of a spatial pattern that has some properties of homogeneity. Color histogram is the most important color representation factor used in image processing. Color histogram yields better retrieval accuracy. Histogram finds out the number of pixels in gray level. After that we are applying Euclidean distance, Neural Network, Target search methods algorithm and K-means clustering algorithm for retrieval of images from the database and making a comparison based approach between them to see which method helps in fast retrieval of images in terms of distance and time.

Keywords: Color feature extraction, Texture feature extraction, Histogram based extraction, image database, Euclidean distance, neural network, Neighbouring Divide-and-Conquer Method and Global Divide-and-Conquer Method, K-means clustering, Threshold=15000.

I. INTRODUCTION

CBIR has become a prominent research topic since there is a huge demand of video and image in digital form. Therefore fast retrieval of images from large database is the need of people. Finding images that are nearly or totally similar to a query image, retrieval of images searches images from large database. CBIR enhance the accuracy of the information being found and is an effective option and match to conventional text-based image searching. For describing image content, color, texture and histogram features have been used. Color is the most effective feature in image processing which completely describe an image. Texture describes how the object differs from its surroundings and structural arrangement with other objects.

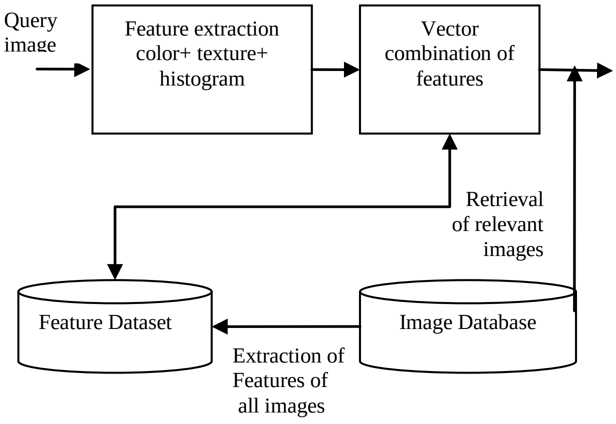

Query image
Feature extraction color+ texture+ histogram
Vector combination of features
Retrieval of relevant images
Feature Dataset
Image Database
Extraction of Features of all images

Fig.1. Block Diagram of CBIR

Histogram feature can also be extensively used for retrieval systems. CBIR system that is based on dominant color, texture and histogram can be implanted.

A. Aim and Objective:

CBIR had been a very important and effective research area in many fields. Searching images from large database and giving appropriate results, increased bandwidth availability will help to increase the use of internet by user in future. Therefore a significant difficulty that needs to be look after is fast retrieval of images from large databases. Image retrieval system searches for images from large database and tries to find exact or nearly same images. CBIR can greatly enhance the precision of the data being returned and is an effective to conventional textual-based image searching. Color, texture and histogram features are used to differentiate an image from other images. Research and development issues in CBIR cover a range of topics, many shared with mainstream image processing and information retrieval. Some of the objective can be: Extracting color, texture and histogram features from images, providing compact storage for large image databases, matching query and stored images in a way that reflects human similarity judgments.

II. PROPOSED WORK

Describing image in order to obtain better search results and to express more image information, here it can be considered the dominant color, texture and histogram features combined. The work of this paper is based on dominant color, texture and histogram features of image.

DOI 10.17148/IJARCCE.2015.41115

After extraction of features from database and saving all the features in feature dataset we apply a query input image where features are extracted from the image and combined in a vector form. Then we have used fast retrieval algorithms such as Euclidean distance, Neural Network, Target methods and K-means clustering. After extraction of features a comparisons is made between the three algorithms on the basis of time and distance. Time here in terms means retrieving the images from database and distance means the distance between query image and target images.

Only simple features of image information cannot give the exact description of an image. We are using the three retrieval features color, texture and histograms to get images that are similar to query images. The working is based for dominant color, texture and a histogram feature of image.
Retrieval algorithm is as follows:

Step1: Uniformly divide every image as shown in Fig.2 in the database and the Query image into 8-common partitions.
Step2: For each partition, the center portion is selected as its main color.
Step3: Texture features (Energy, Contrast, Entropy and homogeneity) from GLCM are now obtained.
Step4: Obtain histogram features of the images in the database as well as for query images.
Step5: Now after obtaining features we will combine the features in vector form.
Step6: Using weighted and normalized Euclidean distance we will now find distances of target images.
Step7: A euclidean distance is sorted after sorting apply bubble sort to get the most relevant images.
Step8: Apply neural network approach and train the data and then again apply Euclidean distance approach on train data to sort the images**.**
Step9: Now apply Target methods on the vector of query image with all features and the feature vectors of image database.
Step10: After that we are applying k-means algorithm to retrieve images.
Step11: Now we will do a comparative based analysis on these algorithms and see which algorithms retrieves fast relevant images and which algorithm takes less time to retrieves the images.

A. Color Feature Extraction:

Color feature is very important for describing any image in surrounding, image is made up of number of colors. But in general all colors are combination of blue, red and green. Dominant Color Descriptive [1, 2, 4, 6] mainly consists of two factors to describe an image they are specific color percentage and the specific color. DCD helps in finding the exact distribution of color in an image .Colorspace selection is not a critical issue in DCD [7]. So, for making work more effective and easier Red, Green and Blue colorspaces are used. As shown in Fig. 2 firstly the images are divided in 8 partitions and there is separate partition for R, G and B. Suppose there are multi color located at same position they are assume to be same or similar. After coarse partition of the R, G and B slices we are taking average values of each block. After getting the averages we will combine the averages.

B. Extraction of texture of an image:

No more Texture feature is again important feature to describe an image because it differentiates the objects in an image. Texture recognition is beneficial in many fields of computer vision. Here For texture representation we have worked on Gray Level Co-occurrence Matrix [16, 17]. Gray Level Co-occurrence Matrix which is created in four directions i.e. distance between the selected pixel and its neighboring pixel. Statistics are used to take out texture features [8, 9, 10].Here we have using four statistics. It uses the probability value, it is defined by the distance between the pixel of interest and its neighboring pixel and when the distance is determined, is showed by Pi,j. GLCM is a matrix which is symmetric and the image gray-level is used to determine its level of co-occurrence. Elements in matrix are calculated by the equation shown below:

$$P\,(i, j\xi d, \text{❧}) = \frac{P(i\,, j\xi d, \theta)}{\sum_i \sum_j P\,(i, j\xi d, \theta)} \qquad \text{eq. (1)}$$

GLCM expresses the texture feature according to the correlation of the pixels with its neighbouring pixels. Four statistics used here are Energy, Contrast, Homogeneity and Entropy to describe the texture feature.

$$\text{Energy } E = \sum_x \sum_y P(x, y)^2 \qquad \text{eq.(2)}$$

Energy is a texture measure of gray-scale image which represents homogeneity changing and occurrence of similar changes in images. Gray-scale uniformity of weight and texture.

$$\text{Contrast } I = \sum \sum (x-y)^2\, P(x, y) \qquad \text{eq.(3)}$$

Contrast this feature shows the clarity of image and how the values of matrix are dispersed in the given images. As the value of contrast goes increased the clearer image looks.

$$\text{Entropy } S = \sum_x \sum_y P(x, y) \log P(x, y) \qquad \text{eq. (4)}$$

Entropy measures unpredictability in the image texture. If the value of the GCM is equal then the entropy is minimum and if value GCM is not equal then its value is greater. Therefore, entropy is always maximum given by GCM.

$$\text{Homogeneity } H=\sum_{i,j}(p(i,j)/1+\xi i-j\xi) \qquad \text{eq.(5)}$$

Local changes whatever comes or occurs in an image is found by Homogeneity. Here p (i, j) is the gray level value at the Coordinate (i, j) shown in equation (5). Features of texture are computed when distance =1 and direction =0°, 45°, 90°, 135°. Four texture features are calculated In each direction. After computing the features vector form of both color and texture are formed**.**

C. Image database and Feature dataset:

The image database contains 1000 images in each of 10 categories as shown in Fig 1. Feature Database contains all

DOI 10.17148/IJARCCE.2015.41115

the extracted features of all the images present in the image database. Features here are in number format and all algorithms are applied here for fetching images. Here also extracted features of 1000 images are stored.

III.PROPOSED ALGORITHM

After feature extraction of query image following algorithms are applied to retrieve relevant images from the database. All the features of images in the database are extracted and stored in feature database where algorithms are one by one applied to extracts relevant images.

A. Euclidean distance:

It is used for fast retrieval of target images from the database. The Euclidean distance is the straight-line distance between two pixels. Euclidean distance here is used to match extracted features of query image with the feature database and then finds the images where features are matching with feature database images after match it sorts out that images which are having shortest distance from the query image and gives us the relevant images. We have used pdist that is pairwise distance between pair of objects. The direct Euclidean distance between an image P and query image Q can be given by the equation:

$$ED = \sum_{i=1}^{n} \sqrt{(Vpi - Vqi).(Vpi-Vqi)} \qquad eq.(6)$$

B. Neural network approach:

In neural network we have both inputs and outputs given and we have to train the neurons to get the exact outputs we required. Here we have given inputs all the extracted features of the images an output is given in the form of 10, 20, 30……n. Now, we have to train the neurons here. The work flow for the neural network design process has six primary steps:

- Collect data
- Create the network
- Configure the network
- Initialize the weights and biases
- Train the network
- Validate the network and use the network

Neural network training can be more successful and efficient if we perform certain pre-processing steps. For the hidden layers Sigmoid transfer functions are generally used. Functions become saturated if input is greater so here in this case gradient will be small and network training will be slow. The net input is nothing but the product of the input times, the weight and the bias for prevention of transfer function from saturation the input should be very large and the weight must be very small. For getting better images first normalize the input. As normalization step is applied to both the input vectors and the target vectors in the data set the network output always falls into a normalized range. The output can be achieved by reversing the data got by normalization method i.e., by renormalising the result.

C. Target Search methods:

Target Search methods consist of two retrieval methods to fetch out images from the database.

1) Neighboring Divide-and-Conquer Method: To find out good results and to speed convergence, we propose to use Voronoi to reduce search space. The nearest neighbors are found by the Voronoi diagram approach finds. NDC searches for the target images. From the starting query Mi, j points are randomly retrieved (line 2). Then the Voronoi region VRk is initially set to the minimum bounding box of M (line 3). In the while loop, NDC determines the Voronoi Seed set Mj+1 (lines 6 to 10) and qi, the most relevant point in Mj+1 according to the user's relevance feedback (line 11). Next, it constructs a Voronoi diagram VD inside VRk using Mj+1 (line 12). The Voronoi cell region containing qi in VD is now the new VRk (line 13). Because only VRk can contain the target, we can safely prune out the other Voronoi cell regions. To continue the searching VRk, NDC constructs a k-NN query using pi as the anchor point (line 15), and evaluates it (line 16). The procedure is repeated until the target pt is found. When NDC encounters a local maximum trap, it employs Voronoi diagrams to aggressively prune the search space and move towards the target image, thus significantly speeding up the convergence. Therefore, NDC can overcome local maximum traps and achieve fast convergence.

```
Neighboring Divide Conquer (M, j)
Input:
Set of images M
Number of retrieved images at each iteration j
Output:
Target image pt
•    Mi ← h0; PQ; WQ; DQ; M; ij
•    Mj ←Evaluate query (Mi) /* randomly retrieve j points in M */
•    VRk ←the minimum bounding box of M
•    iter ← 1(while user does not find pt in Mj do
if iter 6= 1 then)
•    Mj+1← { Mj + qi}
else
•    Mj+1← Mj
•    Endif
•    qi← most relevant point € Mj+1
(construct a Voronoi diagram VD inside VRk using points in Mj+1 as Voronoi seeds)
•    VRk ← the Voronoi cell region associated with the
Voronoi seed pi in VD.
•    M0 ← such points € M that are inside VRk except qi
•    Qr ← h{1: }qi ; WQ; DQ; M0; ij
•    Mj ← CALCULATE QUERY /* perform a constrained k-NN query */
•    iter ←iter + 1
•    Enddo
•    Return pt
```

2) Global Divide-and-Conquer Method: To reduce the number of iterations in the worst case in NDC, we propose the GDC method. To construct a Voronoi diagram GDC uses the query point and j points randomly sampled from VRk. Specifically, GDC replaces lines 15 and 16 in NDC

DOI 10.17148/IJARCCE.2015.41115

with:

- Qr ← h0; PQ; WQ; DQ; M0; ij
- Mj ← Calculate query /* randomly retrieve j points in M0 */

Similar to NDC, while encountering a local maximum trap, GDC employs Voronoi diagrams to aggressively prune the search space and move towards the target image, thus significantly speeding up the convergence. Therefore, GDC can overcome local maximum traps and achieve fast convergence. In the first iteration, Mj = p1; p2; pg is the result of k = 3 randomly sampled points, of which ps is picked as pi. Next, GDC constructs a Voronoi diagram and searches the VR enclosing ps. At the second iteration, Mj+1 = fps; p4; p5; p6g and p5 is the most relevant point pi. In the third and final iteration, the target point is located.GDC takes 3 iterations to reach the target point.

3) K-means clustering: Clustering is sampling of similar objects in space or gathering similar images. So, it is a method of data exploration where we can find out the data which we required and which we want to neglect. The algorithm has access only to the set of features describing each object; it is not given any information as to where each of the instances should be placed within the partition. K-means clustering is a method commonly used to automatically partition a data set into k groups. 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.

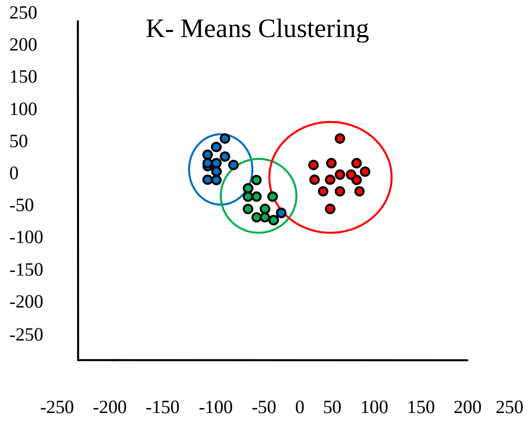


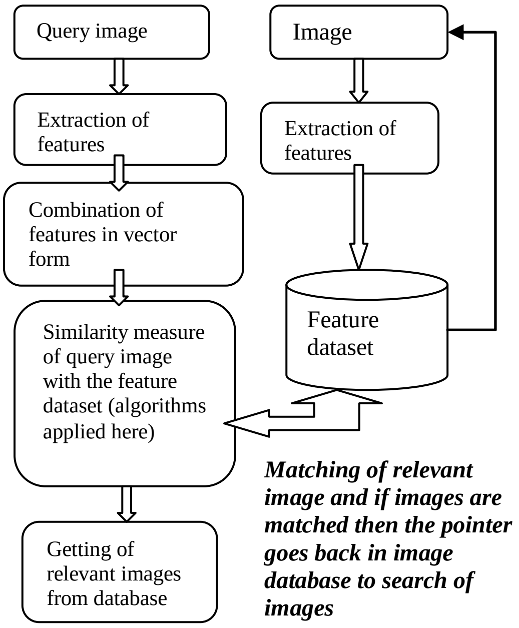
Fig.2. Cluster form for k=3 of red, green and blue clusters

Algorithm for K-means

The summation of point-to-centroid distances is minimized by K-means, summed over all k clusters uses a two-phase iterative algorithm:

1) The first phase uses batch updates, where each updation consists of suggesting some points to the nearest cluster centroid, followed by recalculation of cluster centroids.
2) The second phase uses online updates, where points or number are separately given to the cluster to trim down the sum of distances, and cluster centroids are recomputed after each reassignment.
3) Decide on a value for k.
4) Decide the class memberships of the N objects by assigning them to the nearest cluster center.
5) Initialize the k cluster centers (randomly).
6) Re-estimate the k cluster centers, by assuming the memberships found above are correct.
7) If none of the N objects changed membership in the last iteration, exit. Otherwise go to 3.

In this way the clusters of similar kind are formed and it helps to reduce the elapsed time of the system.


Query image
Extraction of features
Combination of features in vector form
Similarity measure of query image with the feature dataset (algorithms applied here)
Getting of relevant images from database
Image
Extraction of features
Feature dataset

Matching of relevant image and if images are matched then the pointer goes back in image database to search of images

Fig.3. Working of CBIR

IV. COMPARISON BASED APPROACH

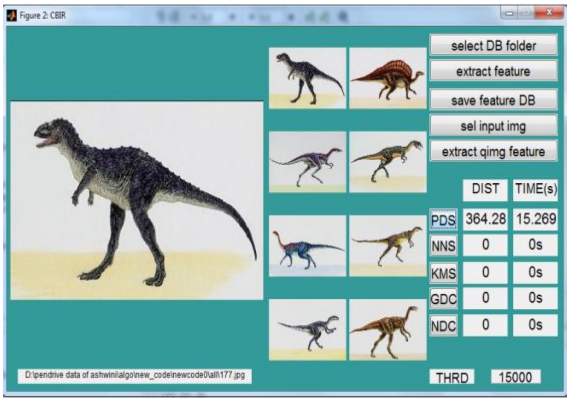
After we have applied all the algorithms we are going to do a comparative based analysis to see which algorithms retrieves fast relevant images from the database. We have made analysis by using distance and timing approach ,i.e., which algorithm will easily retrieval the images within short time and having smallest distance between query image and target images. After comparison based approach we have found Target search methods find out the relevant images in less time than other algorithms.

V. RESULT AND FUTURE WORK

The algorithm has been implemented in MATLAB-10 in Window 7 and run on CPU 2.80GHz PC. The input images are obtained from Internet. Database contains 1000 images in each of 10 categories. By using Target Methods we are getting good relevant images but we can add more measures such as output images got and the percentage matched with query images. In the future, we can develop a system that combines the texture, shape, and spatial features with the color feature to represent the image, which will give good results. Also, segmentation may be used as a method to extract regions and objects from the image that segmented regions are used for similarity matching

Outputs of algorithms

DOI 10.17148/IJARCCE.2015.41115

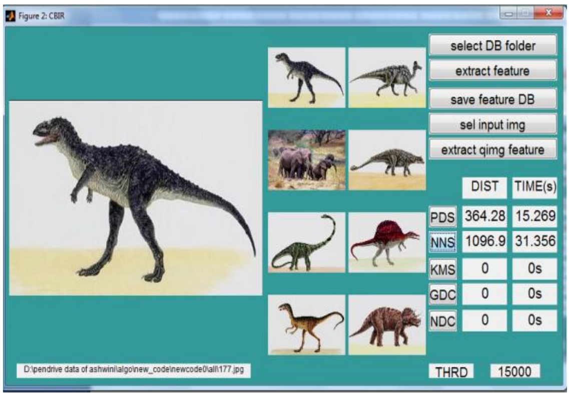


	DIST	TIME(s)
PDS	364.28	15.269
NNS	0	0s
KMS	0	0s
GDC	0	0s
NDC	0	0s

D:\pendrive data of ashwini\algo\new_code\newcode0\all\177.jpg

THRD 15000

Fig.4. Selecting query image and getting of images related to database

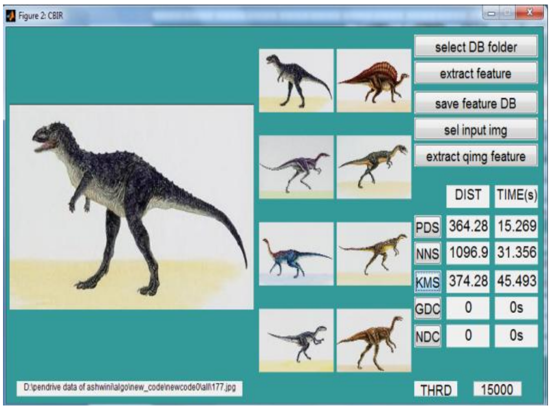


	DIST	TIME(s)
PDS	364.28	15.269
NNS	1096.9	31.356
KMS	0	0s
GDC	0	0s
NDC	0	0s

D:\pendrive data of ashwini\algo\new_code\newcode0\all\177.jpg

THRD 15000

Fig.5. Selecting query image and getting of relevant image by using by using Euclidean distance (Time =15.269 sec) neural network approach (Time= 31.356 sec)

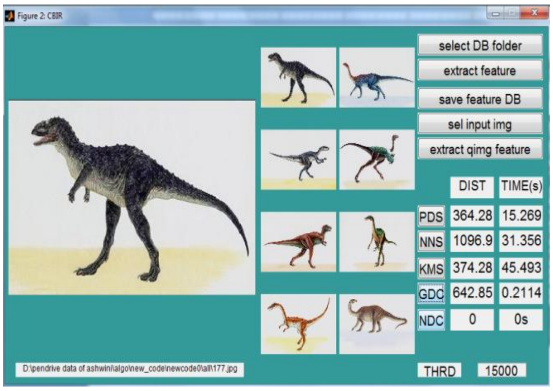


	DIST	TIME(s)
PDS	364.28	15.269
NNS	1096.9	31.356
KMS	374.28	45.493
GDC	0	0s
NDC	0	0s

D:\pendrive data of ashwini\algo\new_code\newcode0\all\177.jpg

THRD 15000

Fig.6. Selecting query image getting of relevant image by using

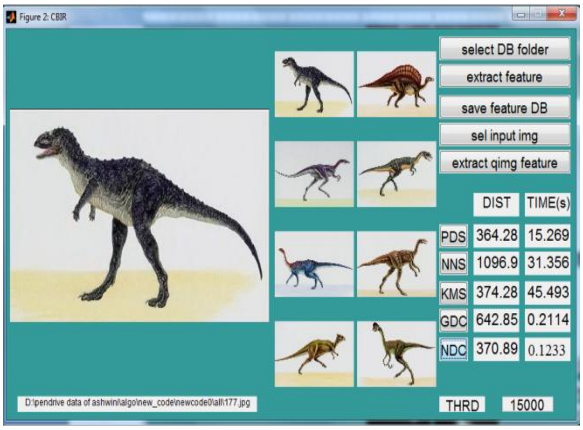


	DIST	TIME(s)
PDS	364.28	15.269
NNS	1096.9	31.356
KMS	374.28	45.493
GDC	642.85	0.2114
NDC	0	0s

D:\pendrive data of ashwini\algo\new_code\newcode0\all\177.jpg

THRD 15000

Fig.7. Selecting query image and getting of relevant image by using K-means clustering approach (Time= 45.493 sec) target methods approach gdc (Time= 0.2114 sec)



	DIST	TIME(s)
PDS	364.28	15.269
NNS	1096.9	31.356
KMS	374.28	45.493
GDC	642.85	0.2114
NDC	370.89	0.1233

D:\pendrive data of ashwini\algo\new_code\newcode0\all\177.jpg

THRD 15000

Fig. 8. Selecting query image and getting of relevant image by using target methods approach ndc (Time= 0.1233 sec)

VI. CONCLUSION AND DISCUSSION

Currently available large images repositories require new powerful way to retrieve interesting information from these images database. This paper focused on developing a new matching strategy based on the idea of minimum area between a query image vector and each vector of images database in specific class in order to select the closest images to the query image. The proposed similarity measure played a major role in not only the relevant retrieval of images, but also in providing the faster mechanism. Performances of the six different similarity distances measures in comparison with the proposed similarity measure were evaluated. The results were promising and showed the effectiveness of the proposed approach. By using Target search algorithm we can better matched any kind of images and it takes less time to retrieve images than other algorithms. In Euclidean distance though the distance between query image and that of target image is less than other algorithms but the most important factor is time while retrieving images from database.

The above figures shows that having threshold 15000 and by using Euclidean Distance (PDS), Neural Network approach (NNS), Target Search methods and K-means clustering approach we got the relevant images but Target methods approach gives more perfect images and the time required to retrieve images from database for both NDC and GDC is also less than that of Euclidean distance, Neural Network approach and K-means clustering approach. In Euclidean distance though the distance between query image and that of target image is less than other algorithms but the most important factor is time while retrieving images from database.

REFERENCES

1. M.Rao, Dr.B.Rao, Dr.Govardhan, “Content based image retrieval using Dominant color and Texture features”, International Journal of Computer science and information security, Vol.9 issue No: 2, February 2011.pp:41-46.
2. X-Y Wang et al., “Effective image retrieval scheme using color, texture and histogram features”, Computer. Stand. Interfaces (2010), doi:10.1016/j.csi.2010.03.004.
3. Chia-Hung Wei, Yue Li, Wing-Yin Chau, Chang-Tsun Li, “Trademark image retrieval using synthetic features for describing

global histogram and interior structure", Pattern Recognition, 42 (3) (2009) 386–394.

4. FAN-HUI KONG, "Image Retrieval using both color and texture features", proceedings of the 8th international conference on Machine learning and Cybernetics, Baoding, 12-15 July 2009.
5. JI-QUAN MA, "Content-Based Image Retrieval with HSV Color Space and Texture Features", proceedings of the 2009 International Conference on Web Information Systems and Mining.
6. Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang, "Image retrieval: ideas, influences, and trends of the new age", ACM Computing Surveys, 40 (2) (2008) 1–60.
7. Nai-Chung Yang, Wei-Han Chang, Chung-Ming Kuo, Tsia-Hsing Li, "A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval", Journal of Visual Communication and Image Representation.
8. Young Deok Chun, Nam Chul Kim, Ick Hoon Jang, "Content-based image retrieval using multiresolution color and texture features", IEEE Transactions on Multimedia, 10(6) (2008) 1073–1084.
9. P. Howarth and S. Ruger, "Robust texture features for still-image retrieval", IEEE Proceedings of Visual Image Signal Processing, Vol.152, No. 6, December 2005.
10. S. Liapis, G. Tziritas, "Color and texture image retrieval using chromaticity histograms and wavelet frames", IEEE Transactions on Multimedia 6 (5) 676–686 (2004).
11. Subrahmanyam Murala, Anil Balaji Gonde, R. P. Maheshwari," Color and Texture Features for Image Indexing and Retrieval", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
12. L. Kotoulas and I. Andreadis, "Colour histogram content-based image retrieval and hardware implementation", IEEE Proc.-Circuits Devices Syst., Vol. 150, No. 5, October 2003.
13. X. Yang, Z. Wang, D. Li, J. Zhang, Color image retrieval with adaptive featureweight in Brushlet domain, in: IEEE 2nd Symposium on Web Society (SWS), pp. 97–102,16–17 August 2010.
14. H. Liu, L. Yu, Toward integrating feature selection algorithms for classificationand clustering, IEEE Transactions on Knowledge and Data Engineering 17 (4) 491–502 (2005).

DOI 10.17148/IJARCCE.2015.41115