Image Content Based Retrieval System using Cosine Similarity for Skin Disease Images

Sukhdeep Kaur[1], Deepak Aggarwal[2]

[1] **Computer Science and Engineering, PTU Jalandhar, RBGI,**
Mohali Campus, Kharar, Punjab 140301, India
Sukh.kaur.rb@gmail.com

[2] **Computer Science and Engineering, PTU Jalandhar,BBSBEC,**
Fatehgarh Sahib, Punjab 140301, India
Deepak.aggarwal@bbsbec.ac.in

Abstract

A content based image retrieval system (CBIR) is proposed to assist the dermatologist for diagnosis of skin diseases. First, after collecting the various skin disease images and their text information (disease name, symptoms and cure etc), a test database (for query image) and a train database of 460 images approximately (for image matching) are prepared. Second, features are extracted by calculating the descriptive statistics. Third, similarity matching using cosine similarity and Euclidian distance based on the extracted features is discussed. Fourth, for better results first four images are selected during indexing and their related text information is shown in the text file. Last, the results shown are compared according to doctor's description and according to image content in terms of precision and recall and also in terms of a self developed scoring system.

Keyword: Cosine similarity, Euclidian distance, Precision, Recall, Query image**.**

1. Basic introduction to cbir

CBIR differs from classical information retrieval in that image databases are essentially unstructured, since digitized images consist purely of arrays of pixel intensities, with no inherent meaning. One of the key issues with any kind of image processing is the need to extract useful information from the raw data (such as recognizing the presence of particular shapes or textures) before any kind of reasoning about the image's contents is possible. An example may make this clear. Many police forces now use automatic face recognition systems. Such systems may be used in one of two ways. Firstly, the image in front of the camera may be compared with a single individual's database record to verify• his or her identity. In this case, only two images are matched, a process few observers would call CBIR[15]. Secondly, the entire database may be searched to find the most closely matching images. This is a genuine example of CBIR.

2. Structure of CBIR model

Basic modules and their brief discussion of a CBIR modal is described in the following Figure 1.Content based image retrieval system consists of following modules:

Feature Extraction: In this module the features of interest are calculated for image database.

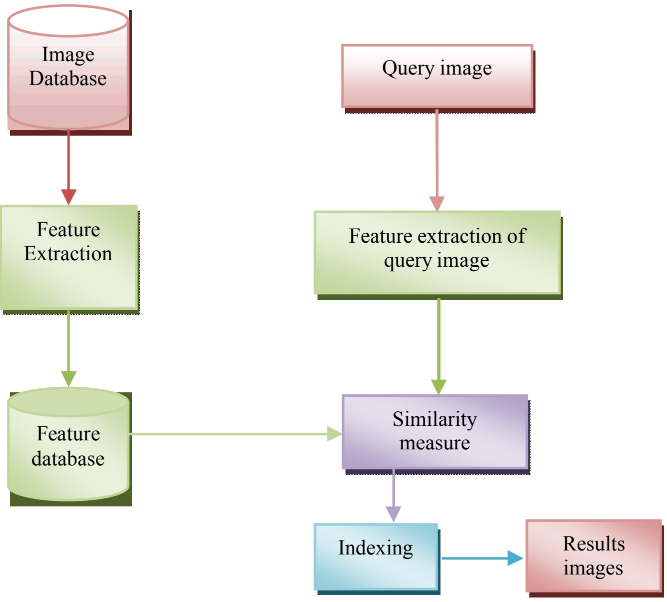


Fig.1 Modules of CBIR modal

Feature extraction of query image: This module calculates the feature of the query image. Query image can be a part of image database or it may not be a part of image database.

Similarity measure: This module compares the feature database of the existing images with the query image on basis of the similarity measure of the interest[2].

Indexing: This module performs filtering of images based on their content would provide better indexing and return more accurate results.

Retrieval and Result: This module will display the matching images to the user based on indexing of similarity measure.

Basic Components of the CBIR system are:

Image Database: Database which stores images. It can be normal drive storage or database storage.

Feature database: The entire extracted feature are stored in database like mat file, excel sheets etc.

3. Scope of CBIR for skin disease images

Skin diseases are well known to be a large family. The identification of a certain skin disease is a complex and demanding task for dermatologist. A computer aided system can reduce the work load of the dermatologists, especially when the image database is immense. However, most contemporary work on computer aided analysis skin disease focuses on the detection of malignant melanoma. Thus, the features they used are very limited. The goal of our work is to build a retrieval algorithm for the more general diagnosis of various types of skin diseases. It can be very complex to define the features that can best distinguish between classes and yet be consistent within the same class of skin disease. Image and related Text Database is collected from a demonologist's websites [17, 18].

There are mainly two kinds of methods for the application of a computer assistant. One is text query. A universally accepted and comprehensive dermatological terminology is created, and then example images are located and viewed using dermatological diagnostic concepts using a partial or complete word search. But the use of only descriptive annotation is too coarse and it is easy to make different types of disease fall into same category. The other method is to use visual features derived from color images of the diseased skin. The ability to perform reliable and consistent clinical research in dermatology hinges not only on the ability to accurately describe and codify diagnostic information, but also complex visual data. Visual patterns and images are at the core of dermatology education, research and practice. Visual features are broadly used in melanoma research, skin classification and segmentation. But there is a lack of tools using content-based skin image retrieval.

4. Problem formulation

However, with the emergence of massive image databases, the traditional manual and text based search suffers from the following limitations:

- Manual annotations require too much time and are expensive to implement. As the number of images in a database grows, the difficulty in finding desired images increases. It is not feasible to manually annotate all attributes of the image content for large number of images.

- Manual annotations fail to deal with the discrepancy of subjective perception. The phrase, "an image says more than a thousand words," implies a Content-Based Approach to Medical Image Database Retrieval that the textual description is not sufficient for depicting subjective perception. Typically, a medical image usually contains several objects, which convey specific information. Nevertheless, different interpretations for a pathological area can be made by different radiologists. To capture all knowledge, concepts, thoughts, and feelings for the content of any images is almost impossible.

5. Methodology of work

5.1 *General approach*

The general approach of image retrieval systems is based on query by image content. Figure 2 illustrate an overview of the image retrieval modal of skin disease images of proposed work.

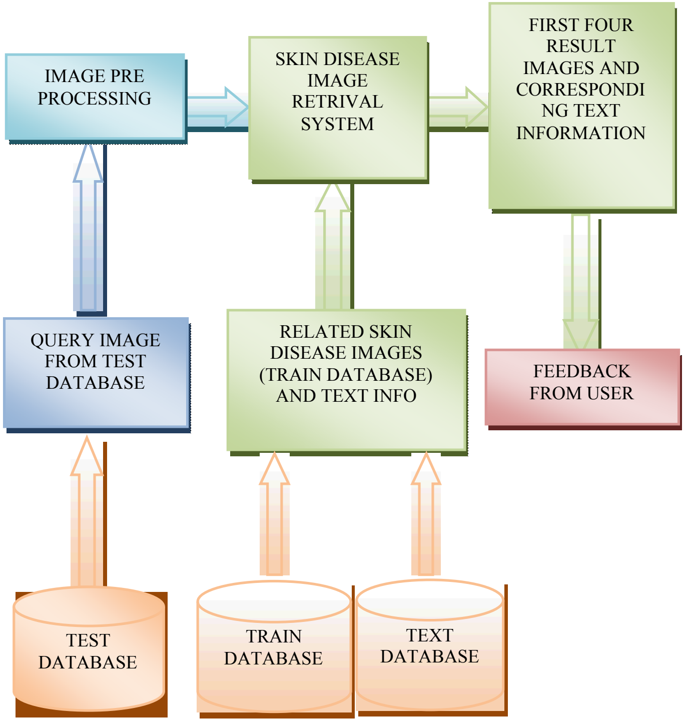


IMAGE PRE PROCESSING
SKIN DISEASE IMAGE RETRIVAL SYSTEM
FIRST FOUR RESULT IMAGES AND CORRESPONDI NG TEXT INFORMATION
QUERY IMAGE FROM TEST DATABASE
RELATED SKIN DISEASE IMAGES (TRAIN DATABASE) AND TEXT INFO
FEEDBACK FROM USER
TEST DATABASE
TRAIN DATABASE
TEXT DATABASE

Fig.2 Overview of the Image query based skin disease image retrieval process

5.2 Database details :

Our train database contains total 460 images (approximately) which are divided into twenty eight classes of skin disease, collected from reputed websites of medical images [17,18].

Test database contains images which are selected as query image. In the present work size of train database and test database is same.

All the images are in .JPEG format. Images pixel dimension is set 300X300 by preprocessing. The illumination condition was also unknown for each image. Also, the images were collected with various backgrounds.

Text database corresponding to each image contains skin disease name, symptoms, cure, and description of the disease.

5.3 Use Of Descriptive Statistics Parameters for Feature Extraction

Statistical texture measures are calculated directly from the original image values, like mean, standard deviation, variance, kurtosis and Skewness [13], which do not consider pixel neighborhood relationships. Statistical measure of randomness that can be used to characterize the texture of the input image. Standard deviation is pixel value analysis feature [11].

First order statistics of the gray level allocation for each image matrix I(x, y) were examined through five commonly used metrics, namely, mean, variance, standard deviation, skewness and kurtosis as descriptive measurements of the overall gray level distribution of an image. Descriptive statistics refers to properties of distributions, such as location, dispersion, and shape [15].

5.3.1 Location Measure:

Location statistics describe where the data is located.

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

$$mean(x) = SUM_i \, x(i)/N$$

if x is a matrix , compute the mean of each column and return them into a row vector[16].

5.3.2 Dispersion Measures:

Dispersion statistics summarize the scatter or spread of the data. Most of these functions describe deviation from a particular location. For instance, variance is a measure of deviation from the mean, and standard deviation is just the square root of the variance.

Variance : For calculating the variance of element of vector x.

$$var(x) = 1/((N-1)SUM\_i \, x(i) - mean(x)\hat{}2)$$

If x is a matrix , compute the variance of each column and return them into a row vector [16].

Standard Deviation: For calculating the Standard Deviation of element of vector x.

$$std(x) = sqrt(1/(N-1)SUM\_i \, (x(i) - mean(x))\hat{}2)$$

If x is a matrix , compute the Standard Deviation of each column and return them into a row vector[16].

5.3.3 Shape Measures:

For getting some information about the shape of a distribution using shape statistics. Skewness describes the amount of asymmetry. Kurtosis measures the concentration of data around the peak and in the tails versus the concentration in the flanks.

Skewness: For calculating the skewness of element of vector x.

$$skewness \quad (x) \; = \; 1/N \, std \, (x) \hat{} \, (-3) \, SUM \, ((x - mean(x).\hat{}3)$$

If x is a matrix, return the skewness along the first non-singleton dimension of the matrix [16].

Kurtosis : For calculating the Kurtosis of element of vector x.

$$kurtosis \quad (x) \; = \; 1/N \, std(x) \hat{}(-4) \, SUM((x - mean(x).\hat{}4) - 3$$

If x is a matrix, return the Kurtosis over the first non-singleton dimension of the matrix [16].

5.4 Distance metrics:

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

5.4.1 Equlidian Distance(EU):

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

$$EU(u, v) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

Instead of two dimensions, if the points have n- dimensions, such as $a = (x_1, x_2 \ldots\ldots, x_n)$ and $b = (y_1, y_2, \ldots\ldots, y_n)$ then, Eq. (1) can be generalized by defining the Euclidean distance [14] between a and b as:

$$EU(a, b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots\ldots + (x_n - y_n)^2}$$

$$EU(a, b) = \sqrt{\sum_{i=1}^{n} (x_{i-} y_i)^2} \tag{2}$$

5.4.2 Cosine Angle Distance:

If we consider two vectors X and Y where $X = (x_1, x_2, \ldots., x_n)$ and $Y = (y_1, y_2, \ldots., y_n)$ then $\cos\theta$ may be considered as the Cosine of the vector angle between X and Y in n dimension [14]. Formally, we define CAD as follows

$$CAD(X, Y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

$$CAD(X, Y) = \frac{X.Y}{\|X\|\|Y\|} \tag{3}$$

One important property of vector cosine angle is that it gives a metric of similarity between two vectors unlike Euclidean distance, which give metrics of dissimilarities

5.5 Basic performance measures of CBIR system

5.5.1 Precision and Recall method

Precision and recall are widely used parameters in evaluating the CBIR systems. Precision is a measure of fidelity whereas recall is a measure of completeness. Precision basically is a measure of the number of retrieved images that are relevant to the search. Recall as mentioned earlier is a measure of completeness i.e. it is basically the ratio of relevant retrieved images to the total relevant images present in the database. The mathematical computation of precision and recall can be understood by Eq. (4) and Eq. (5).

In the proposed work total number of relevant result images against the query image shown is fixed i.e. four images in case of query by image. So for the proposed work modified precision and recall formulas are given below:

$$\text{precision} = \frac{\text{Number of relevant images from the class of QUERY IMAGE}}{\text{Total Number of retrived images=4}} \tag{4}$$

$$\text{recall} = \frac{\text{Number of relevant images from the class of QUERY IMAGE}}{\text{Size of the class to which the query image belongs}} \tag{5}$$

5.5.2 Self Developed Scoring System

In previous section the standard evaluation method for content-based image retrieval was introduced. However, the Precision vs. Recall graph is drawn while adjusting the number of images retrieved from 1 to the size of the whole image library. So, the Precision vs. Recall graph can be regarded as an evaluation for the image retrieval system, but not an evaluation based on a given fixed number of retrieved images. In present case, the number of images retrieved is fixed at four, so a scoring system particular to this preset range without changing the number of images retrieved is developed.

$$W_1 = (W_{i1} * W_{d1}) + (W_{i2} * W_{d2}) + (W_{i3} * W_{d3}) + (W_{i4} * W_{d4})$$

$$W_2 = (W_{i1} * W_{c1}) + (W_{i2} * W_{c2}) + (W_{i3} * W_{c3}) + (W_{i4} * W_{c4})$$

Where: W_i = weight of image according to indexing

Weight of Result image 1 W_{i1} = 4/4 =1.00

Weight of Result image 2 W_{i2} = 3/4 =0.75

Weight of Result image 3 W_{i3} = 2/4 =0.50

Weight of Result image 4 W_{i4} = 1/4 =0.25

W_d = weight of image according to doctor description i.e. if query image and result image is of same class.

If both belongs to same class then weight $W_{di} = 1$

If both belongs to different classes then weight $W_{di} = 0$

(Value of i varies from 1-4)

W_c = weight of image according to image content i.e. if query image and result image is seems to be of same class.

If both seems to be of same class then weight $W_{ci} = 1$

If both belongs to different classes then weight $W_{ci} = 0$

$$W_{i\,max} = (1.00 * 1) + (0.27 * 1) + (0.50 * 1) + (0.25 * 1) = 2.5$$

$$W_{i\,min} = (1.00 * 0) + (0.27 * 0) + (0.50 * 0) + (0.25 * 0) = 0$$

(Value of i varies from 1-4)

$$\text{performance score of each disease dataset} = \frac{\text{sum of weights of all input query images of a particular disease class}}{\text{size of paricular disease class dataset}} \quad (6)$$

6. Experimental results

6.1 Results for self developed scoring system

Table 1: Scores of first five skin diseases

S.No.	Disease Class	Score (c)= Average weight of particular Disease dataset(Acc.to image content)	Score(d)= Average weight of particular Disease dataset(Acc. to doctor description)	Size of data set
1	Acne	1.90	1.05	15
2	Alopecia	1.78	1.00	15
3	Atopic Eczema	2.00	1.08	16
4	Basal Cell Carcinom	1.98	1.05	15
5	Bowen's Disease	1.88	1.11	19

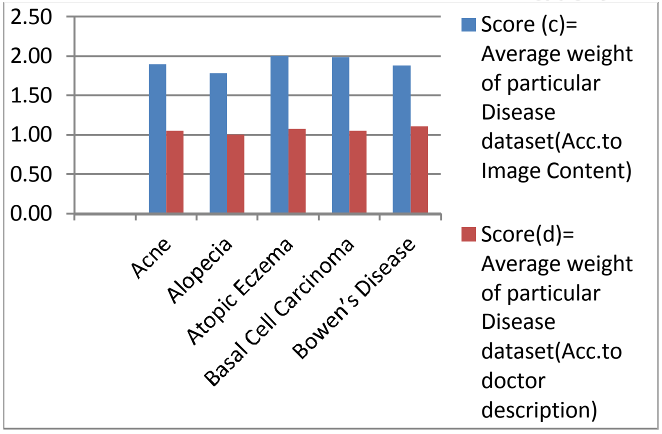


Fig..3 Chart of scores for individual disease

Query image Four result images

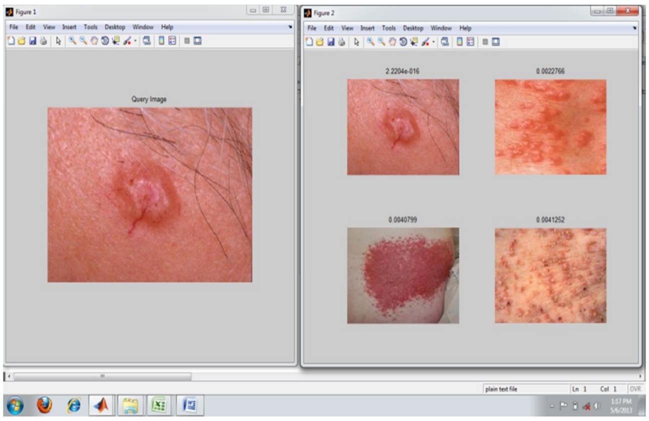
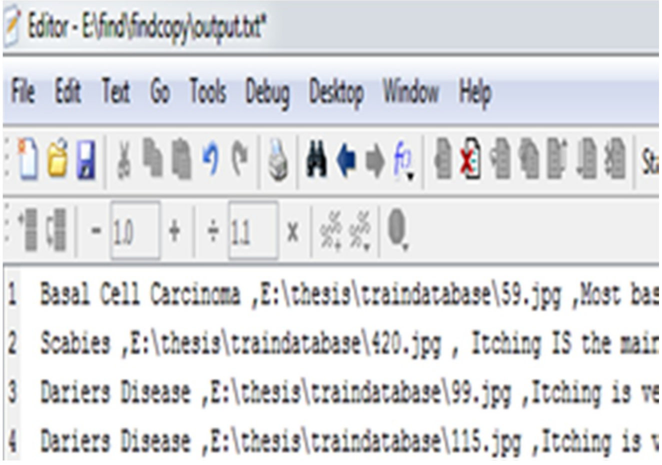


Fig.4 Results for "Basal Cell Carcinoma" from Train database


Editor - E:\find\findcopy\output.txt*
File Edit Text Go Tools Debug Desktop Window Help
- 1.0 + ÷ 1.1 x

```
1 Basal Cell Carcinoma ,E:\thesis\traindatabase\59.jpg ,Most bas
2 Scabies ,E:\thesis\traindatabase\420.jpg , Itching IS the mair
3 Dariers Disease ,E:\thesis\traindatabase\99.jpg ,Itching is ve
4 Dariers Disease ,E:\thesis\traindatabase\115.jpg ,Itching is v
```

Fig.5 Text Information Results for "Basal Cell Carcinoma" image from text database

6.2 Results for Modified Precision Vs recall Graphs

Table 2: Precision and Recall values for " Basal Cell Carcinoma"

S.No.	Relevant result images according to Doctor Description	Relevant result images according to Doctor Description	Precision according to Doctor Description (PD)	Recall according to Doctor Description (RD)	Precision according to image Content (PC)	Recall according to image Content (RC)
1	1	4	0.25	0.07	1.00	0.27
2	1	3	0.25	0.07	0.75	0.20
3	1	4	0.25	0.07	1.00	0.27
4	2	2	0.50	0.13	0.50	0.13
5	1	3	0.25	0.07	0.75	0.20
6	1	2	0.25	0.07	0.50	0.13
7	1	4	0.25	0.07	1.00	0.27
8	1	3	0.25	0.07	0.75	0.20
9	1	4	0.25	0.07	1.00	0.27
10	1	2	0.25	0.07	0.50	0.13
11	1	4	0.25	0.07	1.00	0.27
12	1	2	0.25	0.07	0.50	0.13
13	1	2	0.25	0.07	0.50	0.13
14	1	1	0.25	0.07	0.25	0.07
15	1	4	0.25	0.07	1.00	0.27

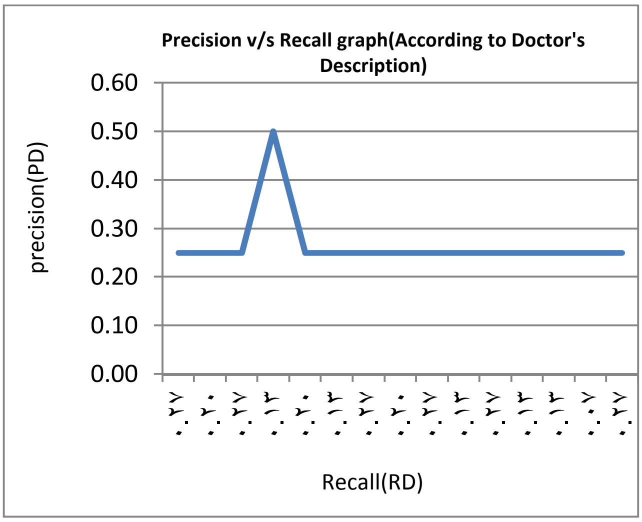


Fig.6 Precision Vs Recall graph for "Basal Cell Carcinoma" (According to Doctor's Description

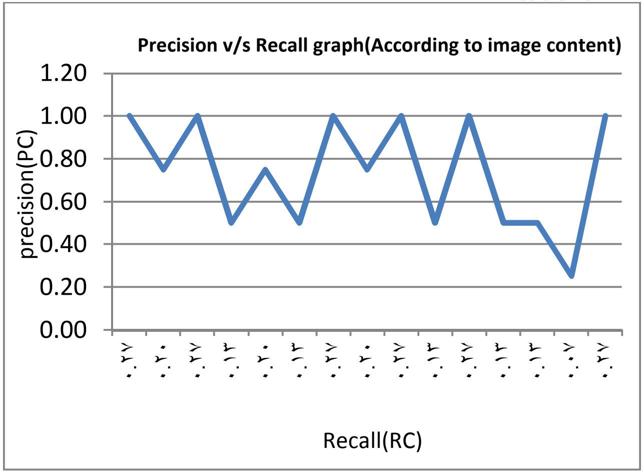


Fig.7 Precision Vs Recall graph for "Basal Cell Carcinoma" (According to image content)

6.3 Similarity between Cosine Angle Distance(CAD) and Euclidian distance(EUD)

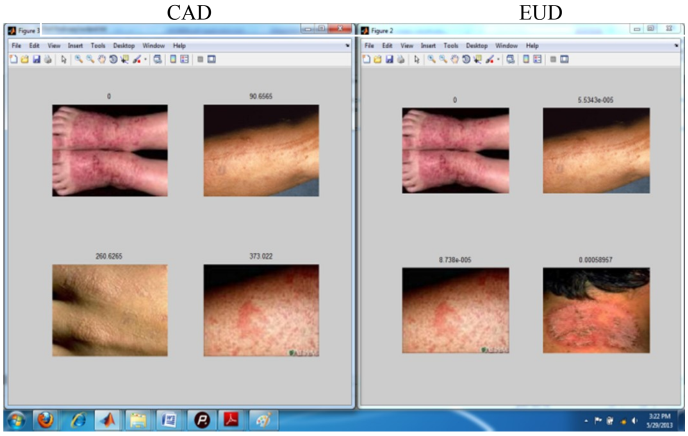


Fig.8 Result Images using CAD and EUD for same query image (Atopic Eczema)

7. Conclusion and future scope

7.1 Conclusion

It was observed descriptive statistics for feature extraction plays a dominating role in distinguishing different types of skin disease. The retrieved result images often contain the similar content against the query image when compared with the Train Database. Since the disease class of the result images may or may not be similar to the query image. Therefore the performance of the proposed system is evaluated according to doctor and also according to the image content.

The retrieval results of using Euclidian distance are almost similar to the results using Cosine Angle Distance.

As shown in Table 1 Atopic Eczema and Basal Cell Carcinoma diseases image dataset had the best accuracy according to self developed scoring system, since it is has the best consistency in color, texture and shape.

7.2 Future Scope

Since the lack of contemporary research in this specific area of skin disease retrieval, this is preliminary research and could benefit from many improvements. These include the following three aspects:

Image collection: There might be some improvements in the image collection section. If the illumination condition for each image is given, color balancing may be performed in the pre-processing step, in order to reduce the impact of mismatched color balance between the query and Train Database images

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

At last, if the feature identification and extraction can be associated with some medical knowledge of those skin diseases as a semantic feature, it could significantly improve the precision and recall of the disease description.

7.3 Distance metric

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

REFERENCES

1. S.Antani, Z.Xue, L.R.Long, D.Bennett, S.Ward and R.Thoma, "Is there a need for biomedical CBIR systems in clinical practice? Outcomes from a usability study", SPIE Medical Imaging 2011: Advanced PACS-based Imaging Informatics and Therapeutic Applications. Orlando, 2011, pp.7967:7974,.
2. K.Bunte, M.Biehl, M.F.Jonkman and N.Petkov. "Learning effective color features for content based image retrieval in dermatology", Pattern Recognition 44, 2011, PP.1892–1902 ScienceDirect, journal homepage: www.elsevier.com/locate/pr.
3. R.Dobrescu, M.Dobrescu, S.Mocanu and D.Popescu. "Medical images classification for skin cancer diagnosis based on combined texture and fractal analysis", Wseas Transactions On Biology And Biomedicine Issue 3, Volume 7, 2010, pp.223-232.
4. H.B.Kekre, D.Mishra and A.Kariwala. "A Survey of CBIR Techniques and Semantics" International Journal of Engineering Science and Technology, Volume 3 No. 5, 2011, pp.4510-4517.
5. H.B.Kekre and K.Patil. "Standard Deviation of Mean and Variance of Rows and Columns of Images for CBIR" WASET International Journal of Computer, Information and System Science and Engineering (IJCISSE), Volume 3, Number 1, 2009, pp.8-11.
6. H.B.Kekre and K.Sonawane. "Bins approach to image retrieval using statistical Parameters based on histogram partitioning of R, g, b planes", International Journal of Advances in Engineering & Technology, ISSN: 2231-1963, Volume 2, Issue 1, , 2012, pp. 649-659.
7. Y.Liu, D.Zhang, G.Lu, and W.Y.Ma. "A survey of content-based image retrieval with high-level semantics", Pattern Recognition :The General of the Pattern Recognition Society 40, 2007, pp.262– 282
8. Mittra and Parekh, 2011. "Automated Detection of Skin Diseases Using Texture Features", International Journal of Engineering Science and Technology, Volume 3, pp. 4801- 4808.
9. Qian, Sural, and Pramanik, 2002. " A Comparative Analysis Of Two Distance Measures In Color Image Databases". Proc. of IEEE Int. Conf. on Image Processing, pp. 401–404, 2002.
10. G.Qian, S.Sural, Y.Gu and S.Pramanik. " Similarity between Euclidean and cosine angle distance for nearest neighbor queries" ACM, 2004 pp.14-17
11. P.Rane, P.Kulkarni, S.Patil and B.B.Meshram."Feature based image retrieval of images for CBIR". IJCEM International Journal of Computational Engineering & Management, Volume 14, October 2011 ISSN (Online): 2011, pp.2230-7893 www.IJCEM.org
12. N,Selvanathan, L.S.Yun,M.Sankupellay ,V.Purushotaman. and S.Jameelah. "Automatic retrieval of Microscopic blood cell images", Springer-Verlag Berlin Heidelberg, 2007pp. 245-249 ,www.springer .com
13. S.Selvarajah and S.R.Kodituwakku. "Analysis and Comparison of Texture Features for Content Based Image Retrieval", International Journal of Latest Trends in Computing (E-ISSN: 2045-5364) Volume 2, Issue 1, 2011pp.108-113
14. A.Vadivel, A.K.Majumdar and S.Sural. "Performance Comparison of Distance Metrics in Content-based Image Retrieval Applications", International Conference on Information Technology (CIT), Bhubaneswar, India, 2003, pp. 159-164.
15. "Descriptive Statistics" reference.wolfram.com. [online]. Available: http://reference.wolfram.com/mathematica/guide/DescriptiveStatistics. html/ [Accessed: Jan 2013].
16. "Descriptive Statistics" www.gnu.org. [online]. Available: http: // www.gnu.org / software / octave / doc / interpreter / Descriptive-Statistics.html [Accessed: Jan 2013].
17. "Skin Information" www.britishskinfoundation.org.uk.[online]. vailable: http://www.britishskinfoundation.org.uk/SkinInformation/AtoZofSkindisease/ [Accessed: Sep. 2012.].
18. "Skin Disease Pictures" www.dermnet.com . [online]. Available: http:// www.dermnet.com/dermatology-pictures-skin-disease-pictures/ [Accessed: Sep. 2012.]