International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 4, Issue 3, May-June 2015 ISSN 2278-6856

Plagiarized Image Detection System based on CBIR

Prajakta Mahendra Ovhal[1], Prof. B.D. Phulpagar[2]

[1]Modern College Of Engineering, Shivajinagar Pune.

[2] Modern College Of Engineering, Shivajinagar Pune.

Abstract

Plagiarism detection is a well known phenomenon in the academic arena. Copying other people is considered as a serious offence that needs to be checked. For this many plagiarism detection systems are available. Many of such plagiarism systems donot include figures and diagrams for checking. They simply discard them which results in look holes that people can take advantage of. In short people can plagiarize diagrams and figures easily without the current plagiarism system detecting it. So there is a need to develop a system that will detect plagiarism in diagrams and figures as well. The main goal of this system is to develop a image plagiarism system which will be primarily based in CBIR. This system will focuses on image detection. There will be a database consisting of images. The user will give an image as input. The features of input image will be extracted and it will be compared with the features of the images present in database. The one which are most matching will be retrieved. The output image which is retrieved from database will be the similar image of the input image.

Keywords: Content Based Image Retrieval (CBIR), Red Green Blue model (RGB), Global Color Histogram (GCH), Local Color Histogram (LCH), Cyan Magenta Yellow model (CMYK), Hue Saturation Value (HSV), Hue Lightness Saturation (HSL).

1. INTRODUCTION

Plagiarism is nothing but using somebody's work without notifying or without giving them any credit. Plagiarism has become an important issue because of internet. Wide source of knowledge is available on internet today, because of which sharing of ideas and technology has become very easy, but due to this worldwide access people tend to copy someone's genuine work and necessary credit is not given to the actual researcher which is quite inappropriate. Plagiarism can be seen in many fields like research papers, art gallery, definitions etc. There are many plagiarism detection systems available for e.g, Plagiarisma.net, Viper, PlagScan etc which focus of plagiarism of text. When any research paper is being scanned by plagiarism detection system they scan the textual part but when any image is encountered it simply discards that image which is inappropriate because an image can also be plagiarized. An architecture of flow diagram of a project can be plagiarized, any snapshot of someone's result can also be plagiarized and so on. Hence there is a need to develop image plagiarism detection system.

1.1 Basics Of Image And Image Processing

1.1.1 Image Processing

Image processing is used to solve identification problems, such as in forensic medicine or in creating weather maps from satellite pictures. It deals with images in bitmapped graphics format that have been scanned in or captured with digital cameras. It can also be referred as improvement, such as refining a picture.

1.1.2 Image Retrieval

Image Retrieval has become a very active research topic, with two major research communities, database management and computer vision. One is text-based and another is visual-based. Text-based image retrieval involves annotating the image with keywords, and use text-based database management systems (DBMS) to retrieve the images. In text-based image retrieval system, keywords of semantic information are attached to the images. CBIR is an automatic process to search relevant images based on user input. A typical CBIR process first extracts the image features and store them efficiently. Then it compares with images from the database and returns the results. Feature extraction and similarity measure are very dependent on the features used. In each feature, there would be more than one representation. Among these representations, histogram is the most commonly used technique to describe features.

2. LITERATURE REVIEW

Arrish, et al.[1] in February 2014, published a paper in which they have developed a image plagarism detection system which works for flow chart images only. Here they are mostly concentrating on shape feature. Firstly there is a query image which is given as input, then preprocessing is performed on it like thinning. removing connected lines and removing text. Further they are detecting edges by using Canny edge detection technique, after this to detect shape they are using Chain code. Finally cosine similarity comparison is done and results are displayed .

Bhattacharjee, et al.[2] published a paper in which they have developed a plagarism detection system for equations. In this firstly they are extracting each line from document. In each line they are checking the presence of "=" symbol. If an "=" symbol is encountered that means some equation is present, then it converts that equation

line image to character by using character recognition they those generated characters are matched to equations from database and finally results are displayed.

Ait-Aoudia, et al.[3] published a paper in which they have developed a image retrieval system. In the beginning a query image is given as a input. Its features such as color, shape and point of inaterest are calculated, in which for color feature extraction they have used HSV color space, for texture feature extraction they are using contrast, energy and entropy, and for point of interest feature they are using Harris detector. Comparison is done and finally result is displayed.

Chaudhari, et al.[4] published a paper in which they have developed a image retrieval system. In this paper they have developed an image retrieval system in which a query image is given as input, its features like color and shape is extracted. For color they are using Color Coherence Vector, for shape they are using mass, centroid and dispersion. Finally Euclidean distance is used for similarity measure.

Sakhare, et al.[5] published a paper in which they have developed a image retrieval system in which they have concentrated on two features color and texture. For color feature extraction they are using RGB color madel. For similarity measure they are using Euclidian distance.

3. SCENARIOS FOR IMAGE PLAGIARISM

An image can be plagiarized by doing some minute changes in it so that the plagiarism test will not be able to detect it easily. Following are some scenarios that can be considered while plagiarizing an image:

1. Rotating an image.

An image can be plagiarized by rotating it by some angle. If such rotation is performed then in plagiarism check it will be treated as two different images which is actually inappropriate.

2. Changing background color of an image.

An image can be plagiarized by simply changing its background. In this way those images will be treated as totally different images and it will easy pass the plagiarism test.

3. Changing size of image.

An image can be plagiarized by changing its size, which when compared with original image having a standard size will differ totally. Hence plagiarism test will fail here.

4. Shearing an image.

An image can be plagiarized by shearing it, which when compared with original image will differ totally. Hence plagiarism test will fail here.

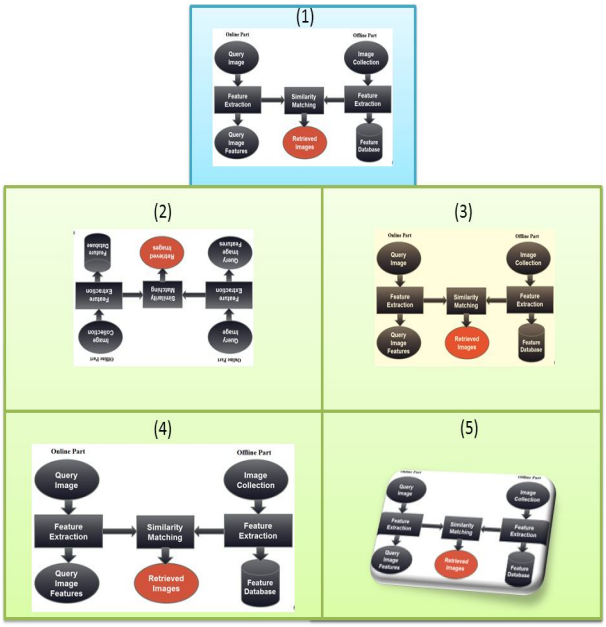

Figure 1. Different Scenarios For Image Plagiarism.

4. Proposed System

The main goal is to develop a Plagarized image detector System that if based on feature extraction techniques. It focuses on all kind of images like flowchart, snapshots, tables, graphs etc. The database contains images stored in a single folder. The retrieval system works by comparing query image with the images present in database. Proposed System is designed in such a way that the problems metioned in section three is resolved properly. Following figure shows the architecture of proposed system. In the architecture image is preprocessed in which image is enhanced in such a way that the time and space required for processing is reduced. In preprocessing image is threshold so that object in image is clearly visible then its boundary is detected and cropping of image is done, after cropping, image is resized. Then feature extraction is done which include color feature, shape feature and greyscale feature. After that comparison of images is performed through Euclidean distance and final image matching result is performed.

4.1 Architecture

Following is the flow of proposed system.

4.1.1 Image to verify

In this a query image is given as input. The image whose plagarism has to be checked is specified here. In this image can be of any size and any resolution any image format(.png, .jpg, .bmp etc) will work. By using thresholding we can convert a grey scale image to a binary image which is nothing but pure black and white image. Its main purpose is to identify the features of an image, it sets a threshold value and whatever features are use those are converted to black and everything else to white or vice- versa. Before that first find out threshold value.

For threshold value estimation Iterative algorithm is used which is as follows:
1. Select an initial estimate for global threshold, T.

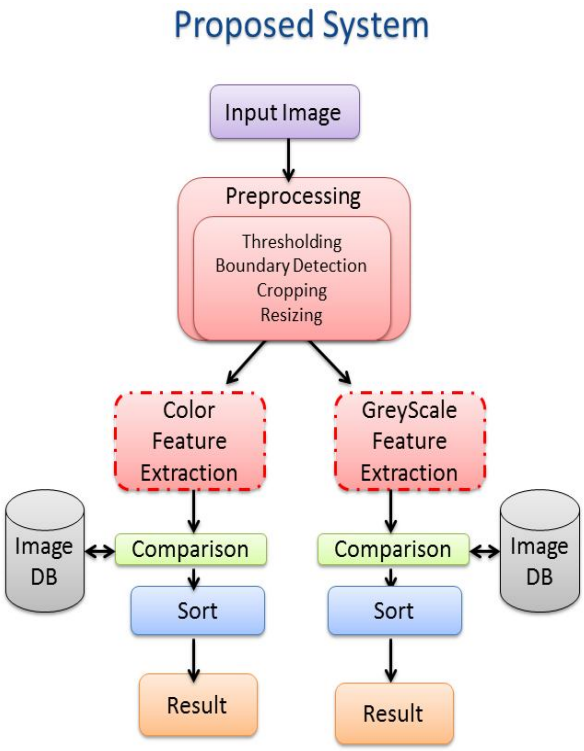

Figure 2. Flow of Proposed System.

2. Segment the image using T. Two groups of pixels will be produced G1 consisting of all pixels with intensity values greater T and G2 consisting of all pixels with intensity values less than or equal to T
3. Compute the average intensity values m1 and m2 for the pixels in G1 and G2,

$$m_1 k = \frac{1}{p_1 k} \sum i * p_i$$

$$m_2 k = \frac{1}{p_2 k} \sum i * p_i$$

4. Compute a new threshold value,

$$T = \frac{1}{2}(m1 + m2)$$

Repeat step 2 to 4 until difference between values of T in successive iterations is smaller than a predefined parameter ΔT.
Steps of thresholding an image is as follows:
1) Check each element in image array.
2) Compare it with threshold value.
3) If element value is less than threshold value then store zero at that location in array.
4) If element value is greater than threshold value store one at that location in array.

4.1.2 Boundary Detection

In this after thresholding an image its boundary has to be detected so that further processing can be done on it.The points at which image brightness changes sharply are typically organized into a set of curved line segments termed boundaries.
There are many methods for edge detection, out of which few are listed below:
a) Sobel Operator
b) Perwitt Operator
c) Laplacian of Gaussian
d) Canny Edge Detection
Advantages of Sobel Operator are Simplicity, Detection of edges and their orientations but disadvantage is sensitivity to noise and Inaccuracy.Advantages of Perwitt Operator is simplicity, Detection of edges and their orientations and disadvantage is sensitivity to noise and Inaccuracy. Advantage of Laplacian of Gaussian is finding the correct places of edges, Testing wider area around the pixel and its disadvantage is malfunctioning at the corners, curves and where the gray level intensity function varies. Not finding the orientation of edge because of using the Laplacian filter. Advantage of Canny Edge detector is using probability for finding error rate, localization and response, improving signal to noise ratio, better detection specially in noise conditions. Disadvantage is complex computations, false zero crossing, time consuming. For boundary detection we are following an algorithm as follows:

1. Calculate height and width of image.
2. Set MaxX=1 and MaxY=1
3. while(MaxX<wt)
4. while(MaxY<ht)
5. if(image[x][y]==0)
store MaxX value in array1
6. store MaxY value in array2
7. endif
8. Identify max and min value from array1
9. Identify max and min value from array2
10. cropW=1+maxX-MinX
11. cropH=1+MaxY-MinY
12. end while
13. end while

4.1.3 Cropping

After an image's boundary is found out, orignal input image is taken and the boundary extracted image is subtracted from orignal input image. By doing so we will get an image with only shape present. Other than boundary of shape everything else is discarded. The main purpose of cropping is to make image smaller and compact so that if we consider overall system performance then processing time required will be less.

4.1.4 Segmentation

After an image's boundary is found out, orignal input image is taken and that boundary extracted image is subtracted from orignal input image. By doing so we will get an image with only shape present. Other than boundary of shape everything else is discarded. The main purpose of cropping is to make image smaller and compact so that if we consider overall system performance then processing time required will be less. In Non-contextual techniques ignores the relationships that exists between features in an image, pixels are simply grouped together on the basis of some global attribute, such as grey level. In Contextual techniques, additionally exploit the relationships between image features. Thus, a contextual technique might group together pixels that have similar grey levels and are close to one another.

4.1.4 Feature Extraction

Image is best represented by its features. The following features are used to extract from an image.

1. **Color Feature Extraction**
   Color feature extraction has to be done in two ways, defining color space and color descriptors. A color is represented using color space and it is described using color descriptors. In Color space represents color of an image. Many color spaces are available like RGB color space, HSV color space, HSB color space, CMYK color space etc. When we consider RGB color space it is an additive color space i.e. new colors are made by adding two primary colors from R,G,B. It is widely used and mostly in computer displays. Then next one is HSV color space. It is mainly used in Computer Graphics. It represents more colors than RGB. Then next is HSB, it represents color same as how humans do. CMYK color model is basically used in printers. In proposed system we are using HSV color space.
2. **Shape Feature Extraction**
   Shape is the characteristic surface configuration that outlines an object giving it a definite distinctive form. It is a fairly well-defined concept. Shape extraction techniques are Mass, Centroid, Mean, Variance, Aspect ratio, Circularity, Moment invariants etc. A histogram H for a given image is defined as a Vector,

$$H = h[1], h[2], h[3], \ldots h[i], \ldots h[N]$$

where i represents a color in the histogram, h[i] is the number of pixels in color i in that image, and N is the number of bins in the histogram, i.e., the number of colors in the adopted color model.

4.1.5 Comparison

Features of the database and the query images are extracted and the distance between them are obtained by the Euclidean distance which is given as

$$d(u,v) = \sqrt{\sum_{i=1}^{N} (h_i^v - h_i^u)^2}$$

4.1.6 Ranking and Result

When comparison is done the image which is matching accurately will be displayed first. The images which are partially matching will be displayed and finally images which are less matching will be displayed. In this way images are ranked according to matching criteria and then displayed in such fashion.

5. Mathematical Model

Mathematical Model for propsed approach is as follows:
Let S be a system where,

$$S = \{I, I_{gs}, I_{th}, I_{bd}, I_{cr}, I_{rz}, I_{fe}, I_{sim}, R\}$$

Where,
I → *input image*
I_{gs} → *greyscaled image*
I_{th} → *threshold image*
I_{bd} → *Boundary detection of image*
I_{cr} → *cropped image*
I_{rz} → *resized image*
I_{fe} → *feature extraction*
I_{sim} → *similarity of images*
R → *result*

Activation Function,

$F = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8\}$

$f1 \rightarrow I_{gs} = GreyScale(I)$

$GreyScale = R + G + B/3$

$f2 \rightarrow I_{th} = Thresholding(p, Th)$

p_0 *and* p_1 *are two classes,* p *is each pixel*

$If\ I(p) < Th, \quad p_0 = 1$

$If\ I(p) > Th, \quad p_0 = 0$

$f3 \rightarrow I_{bd} = Detect\ Boundary(I_{th})$

$f4 \rightarrow I_{cr} = Subimage(I_{bd}, I)$

$f5 \rightarrow I_{rz} = Resize(I_{cp}, nw.nh)$

$where, nw = 200\ and\ nh = 200$

$f6 \rightarrow I_{fe} = Extract\ Image\ Feature(f_a, f_b)$

$f7 \rightarrow I_{sim} = Calculate\ Similarity$

$I_{sim} = Compute\ Score(S_f)$

Let pixel in each query image be, $Q_i(x_s, y_s)$

Let pixel in each database image be, $D_i(x_t, y_t)$

$$d = \sqrt{\sum (QI_f - DI_f)^2}$$

$f8 \rightarrow I_r = ResultGeneration$

$I_r = \min(d)$

I_r *will be displayed in ascending order*

6. PARAMETERS USED FOR EXPERIMENTAL ANALYSIS

The performance of retrieval of the system can be measured in terms of its recall and precision. Recall measures the ability of the system to retrieve all the models that are relevant, while precision measures the ability of the system to retrieve only the models that are relevant. It has been reported that the histogram gives the best performance through recall and precision value
They are defined as:

$$Precision = \frac{A}{A+B}$$

$$Precision = \frac{A}{A+C}$$

Where, A represent number of relevant images that are retrieved, B, number of irrelevant items and the C, number of relevant items those were not retrieved. The number of relevant items retrieved is the number of the returned images that are similar to the query image in this case. The total number of items retrieved is the number of images that are returned by the search engine. The average precision for the images that belongs to the q_{th} category (A_q) has been computed by:

$$p^i = \sum \frac{p(i)_k}{|A_q|}$$

A) Experimental Results

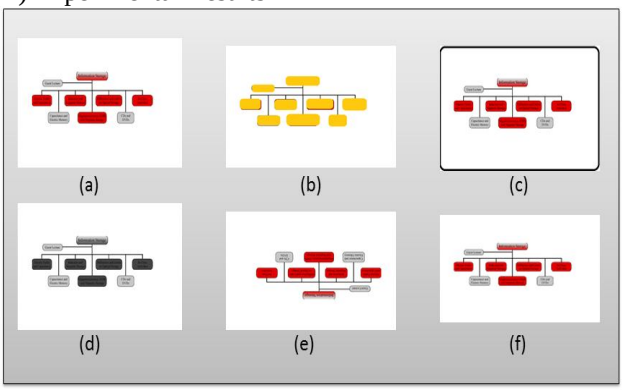

Figure 3. Sample query images.

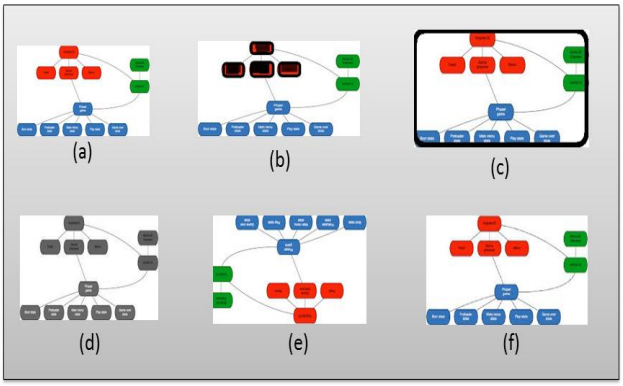

Figure 4. Sample query Images.

Above are the images listed which are used as query images. An image can get plagiarized by altering it to sme extend. Some alteration are given in this Fig 4. And Fig 5. (a)Firstly an exact image is used. (b) Color is changed of query image. (c) Border is given to the query image. (d) Black and white image is used as query image. (e) An inverted image is used as query image. (f) Size of image is changed of query image.

Shown in graph given in Figure 8 and Figure 9, where the red bar represents proposed approach and the blue bar represents the existing approach. The existing approach doesnot contain the extra step which is nothing but the preprocessing steps in it. The graph shows precision and recall values for existing and proposed approach. By observing all values in detail we can come to a conclusion that the precision and recall values of proposed approach gives slightly better performance than the existing system.

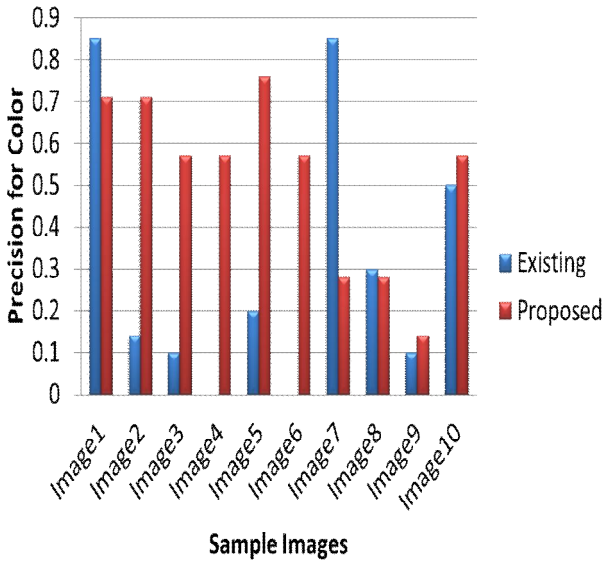

Figure 5. Graph showing precision values for color feaure results.

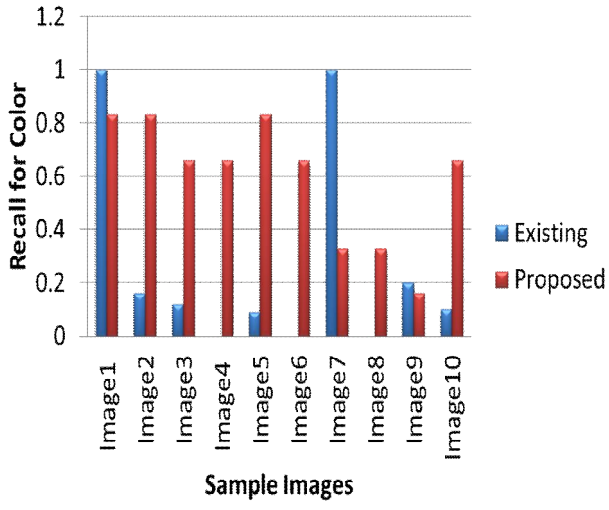

Figure 6. Graph showing recall values for color feature results.

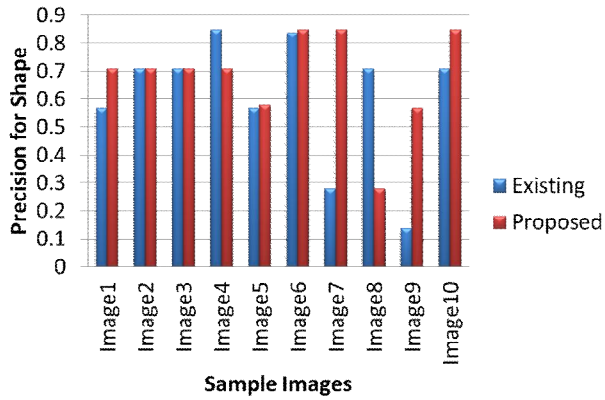

Figure 7. Graph showing precision values for shape feature results.

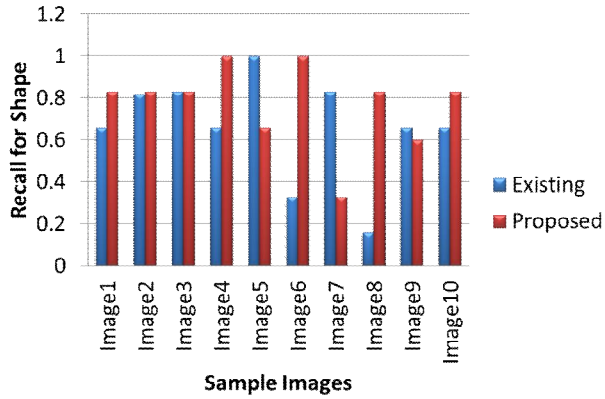

Figure 8. . Graph showing recall values for shape feature results.

In the Figure 9, graph of time required for system to fetch result is mentioned. From the graph we can see that average time required for proposed system is better as compared to existing system.

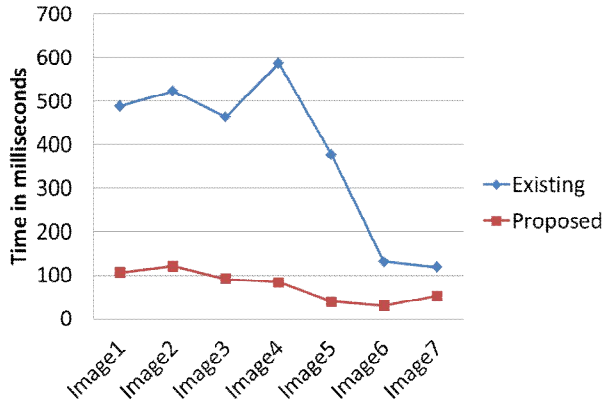

Figure 9. Graph showing processing time difference between Existing and proposed approach.

7. Conclusion

People easily plagiarizes images because many plagiarism detectors are based on text plagiarism detection, they simply discard images from the file. There is very less research in this area. In this project, a system will be developed to retrieve images having certain characteristics based on the given input image.. Some research has been performed on flowchart image but we are planning to develop a plagiarism system which will be applicable for all kind of images. For attaining accurate results, features such as color and greyscaling will be used. To improve the accuracy of system Iterative thresholding and some important preprocessing is done which will improve the accuracy of system. By including the extra steps i.e, the preprocessing steps there has been 30%-40% decrease in the processing time of existing system.

References

[1] Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim, “Shape-Based Plagiarism Detection for Flowchart Figures in Texts,” International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No 1, February 2014.

[2] Debotosh Bhattacharjee and Sandipan Dutta, “Plagiarism Detection by Identifying the Equations,”ELSEVIER, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013

[3] Samy Ait-Aoudia, Ramdane Mahiou, Billel Benzaid, “Yet Another Content Based Image Retrieval system,” 1550-6037/10 $26.00 © 2010 IEEE DOI 10.1109/IV.2010.83

[4] R. Chaudhari, A. M. Patil, Content Based Image Retrieval Using Color and Shape Features, “International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering”, Vol.1, Issue 5, November 2012.

[5] Swati V. Sakhare and Vrushali G. Nasre, “Design of Feature Extraction in Content Based Image Retrieval (CBIR) using Color and Texture”,” International Journal of Computer Science & Informatics”, Volume-I, Issue-II, 2011.

[6] 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.

[7] Guang Yang , Kexiong Chen, Maiyu Zhou, Zhonglin Xu and Yongtian Chen, “Study on Statistics Iterative Thresholding Segmentation Based on Aviation Image,” IEEE Computer Society 2007.

[8] Dr.N.Krishnan, M.Sheerin Banu and C.Callins Christiyana, “Content Based Image Retrieval using Dominant Color Identification Based on Foreground Objects,” IEEE International Conference on Computational Intelligence and Multimedia Applications 2007.

[9] Wang Wen and Wang Yanbo Li Bingbing, “Research on Plagiarism Identification of Digital Images,” IEEE Conference 2010.

[10] Aly S. Abdelrahim, Mostafa A. Abdelrahman, Ali Mahmoud, and Aly A. Farag, "Image Retrieval Based on Content and Image Compression", IEEE Conference 2011.
[11] Zhuo Zhang, Xiaodong Gu and Sunyuan Kung, "COLOR-FREQUENCY-ORIENTATION HISTOGRAM BASED IMAGE RETRIEVAL", IEEE Conference 2012.
[12] Kanwal Preet Kaur, "On Comparative Performance Analysis of Color, Edge and Texture based Histograms for Content Based Color Image Retrieval," IEEE Conference 2014.
[13] Dinakaran, J. Annapurna and Ch. Aswani Kumar, "Interactive Image Retrieval Using Text and Image Content", CYBERNETICS AND INFORMATION TECHNOLOGIES Volume 10, No 3, Sofia 2010
[14] Ms. K. Arthi and Mr. J. Vijayaraghavan, "Content Based Image Retrieval Algorithm Using Colour Models", "International Journal of Advanced Research in Computer and Communication Engineering", Vol. 2, Issue 3, March 2013.
[15] Swati V. Sakhare and Vrushali G. Nasre, "Design of Feature Extraction in Content Based Image Retrieval (CBIR) using Color and Texture"," International Journal of Computer Science & Informatics", Volume-I, Issue-II, 2011.
[16] Prof. C. S. Gode and Ms. A. N. Ganar, "Image Retrieval by Using Colour, Texture and Shape Features", "International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering", Vol. 3, Issue 4, April 2014.
[17] Senosy Arrish, Fadhil Noer, Ahmadu Maidorawa and Naomie Salim, "Shape-Based Plagarism Detection for Flowchrt Figures in Texts", "International Journal of Computer Science & Information Technology", Vol 6, No 1, February 2014.

AUTHOR

Prajakta Mahendra Ovhal received the Bachelor's degree (B.E.) in Computer Engineering in 2013 MIT AOE, Pune. She is now pursuing ME degree, in Computer Engineering at P.E.S Modern College Of Engineering. Her current research interests include Image Processing.