Crossbreed Thresholding Text extraction Procedure for Images using DWT Domain and SVM Classifier

Manisha Bansal
M.tech Scholar
GZS PTU Campus Bathinda, India

Naresh Kumar Garg
Associate Prof.
GZS PTU Campus Bathinda, India

ABSTRACT
This paper proposes a novel crossbreed technique to vigorously extract the texts in images based on Discreet Wavelet Transform (DWT) and Support Vector Machine (SVM). Images on which experimentation has been done are taken from various book covers, newspapers, magazines and commercial products. Database of proposed technique includes 25 images. In addition to that the proposed technique is robust to language selection of the text that is embedded in an image. Experimental database includes images that contain English, Punjabi as well as Hindi font. The proposed technique can be used in the applications such as; keyword-based searching, document retrieving, database collection in an organized manner etc. The projected work is estimated using ICDAR 2013 competition metrics specification and the performance is good as well as results are promising for 3 languages as well.

Keywords
Support Vector Machines, Gradient Difference, Discreet Wavelet Transform.

1. INTRODUCTION
Text Extraction from image is concerned with extracting the relevant text data from a collection of images [1]. Rapid development of digital technology has resulted in digitization of all categories of materials [2, 3]. Lot of resources are available in electronic medium. Many existing paper-based collections, historical manuscripts books, journals, scanned document, video images, maps, posters, broadsides, newspapers, micro facsimile, microfilms, university archives, book plates, graphic materials, coins, currency, stamps, business cards, advertisements, web pages are converted to images and these images present many challenging research issues in text extraction and recognition [4,5]. Text extraction from images have many useful applications in document analysis, detection of vehicle license plate, analysis of article with tables, maps, charts, diagrams, keyword based image search, identification of parts in industrial automation, content based retrieval, object identification, street signs, text based video indexing, page segmentation, document retrieving, address block location [6,7]. Due to growing requirement for information many research work has been done on text extraction in images.
Several techniques have been developed for extracting the text from an image. The existing methods were based on morphological operators, wavelet transform, artificial neural network, skeletonization operation, edge detection algorithm, histogram technique. All these techniques have their benefits and restrictions. So in proposed work, hybridization will be done.

2. BACKGROUND
Text Extraction plays a major role in finding vital and valuable information. Text extraction involves detection, localization, tracking, binarization, extraction, enhancement and recognition of the text from the given image [12, 13.14]. These text characters are difficult to be detected and recognized due to their deviation of size, font, style, orientation, alignment, contrast, complex colored, textured background. Due to rapid growth of available multimedia documents and growing requirement for information, identification, indexing and retrieval, many researches have been done on text extraction in images. Several techniques have been developed for extracting the text from an image. The proposed methods were based on morphological operators [1], wavelet transform [19], artificial neural network [5], skeletonization operation [20], edge detection algorithm [15], histogram technique, connected Components [13] etc. [9]. All these techniques have their benefits and restrictions. Also, many techniques has come into existence that use fusion of wavelets, morphological operators, classifiers, histogram techniques etc. [8, 9, 10, 11, 16, 17, 18]. But till now, none of any researcher has used the concept of using wavelet, Gradient Method, SVM Classifier , Canny edge Detectors along with some pre and post processing steps, in which images containing text of both Hindi, Punjabi and English font.

3. PROPOSED METHODOLOGY
The various steps that are being followed to get the text extraction from images. These steps are applied to images on which English, Hindi and Punjabi text.

3.1 Image Uploading
Image acquisition/capturing of image is the first step of our proposed technique which is done through Nokia Phone based camera of 5 Mega Pixel. Captured image is of size 10-12 kb and of any format like bmp, png, jpej etc.

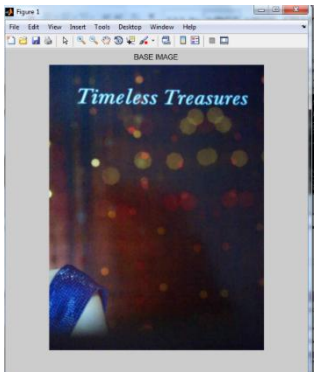

Figure.1 Base Image

Image uploading in proposed model is done using following fnction:

[File path] = uigetfile('*.jpg','Image Files');

img = imread([path file]);

3.2 Pre-Processing

3.2.1 Gray Scale Conversion

If the input image is gray-scale image, then there is no need of conversion, it is directly fetched to next step. But if the image is RGB image, then it needs to be converted into gray – scale image. Intensity image I is given as:

I= R (.29) + G (.58) + B (.11) (1)

The gray scale coversion has been done using following function:

img = rgb2gray(img);

3.2.2 Canny Edge Detection

Canny edge detector is applied to get the edges of the image. The purpose of edge detection in general is to significantly reduce the amount of data in an image, while preserving the structural properties to be used for further image processing. Following function has been used for edge detection

img_edge = edge (img,'canny');

3.2.3 Feature extraction

In MATLAB function "Region props" is then used to get the properties of image like Area, Filled Image, and Pixel list.

Area- It gives the total number of "ON" pixels; Filled Image- The on pixels correspond to the region, with all holes filled in.; Pixel list- It specifies each row of the matrix has the form [x y z ...] and specifies the coordinates of one pixel in the region.

3.2.4 Calculate Connected Components

Connected components are used to get the connected component values. Following function has been used:

B=conv2(double(BW),double(msk))

3.2.5 Median Filtering

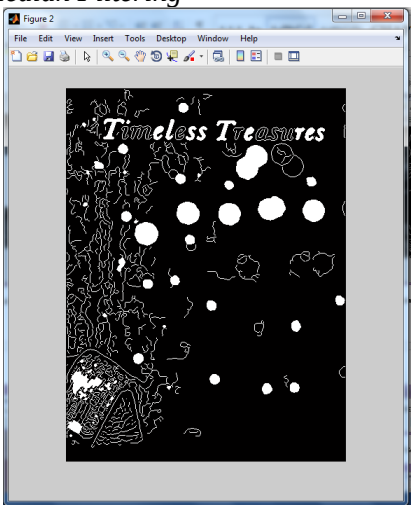

Figure.2 Pre-Processing Image

Median filtering is then done to make the gray scale image noisy free. Following function has been used for median filtering:

im1=medfilt2 (im1,[3 3]);

3.3 DWT for Edge Highlight

The processed image is then again inputted to Daubechies Db4 Wavelet Transform to get the three kinds of edges and texture as missed by the traditional edge detectors. With Daubechies Db4 DWT, the detected edges become more precise and obvious. The main reason of applying wavelet transform for edge detection is that wavelet transform can remove the noise whereas conventional edge operators identifies noisy pixels as edge pixels. It gives the values into 4 components LH, LL, HL, HH [19, 20].

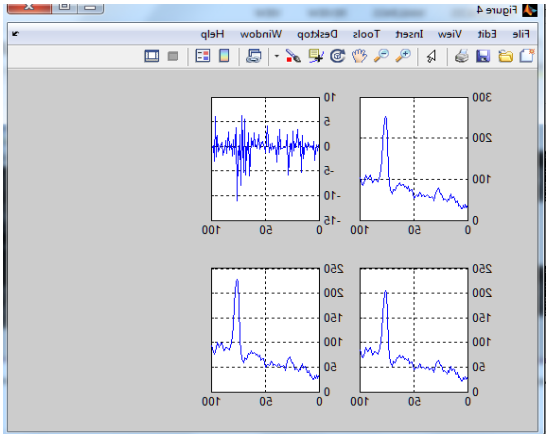

Figure.3 DWT components in terms of Frequency values

3.4 Application of gradient method

Use of Gradient method has been done to get pixel values of an image.

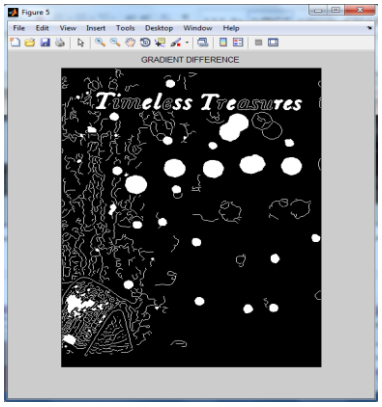

Figure.4 Gradient Method

3.5 Application of SVM for non-text Removal

Although several stages are employed algorithm to reject non-text areas, these are removed completely. Therefore a text verification employed in this stage of algorithm to verify blocks. The algorithm is based on SVM following features:

a. No. of Occurrences in database.

b. No. of attributes in database.

c. Generating two items dataset.

d. Develop Minimum support and minimum Confidence function.

e. Optimise fitness function.

f. Application of SVM rule.

3.6 Post Processing

Sometimes some small part of background color in candidate region occurs that are similar to text region. S to remove this global thresholding value has been used with SVM.

Figure.5 Text Extracted Image

3.7 Result values

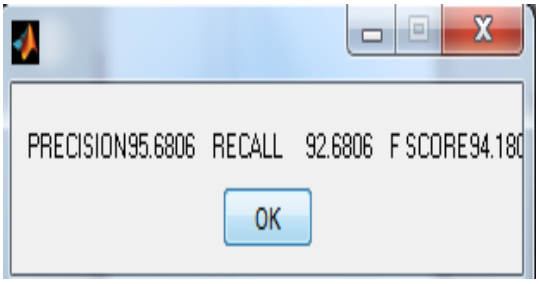

Figure.6 Result Values

Above window shows the result values for English font only.

3.8 ROC Curve

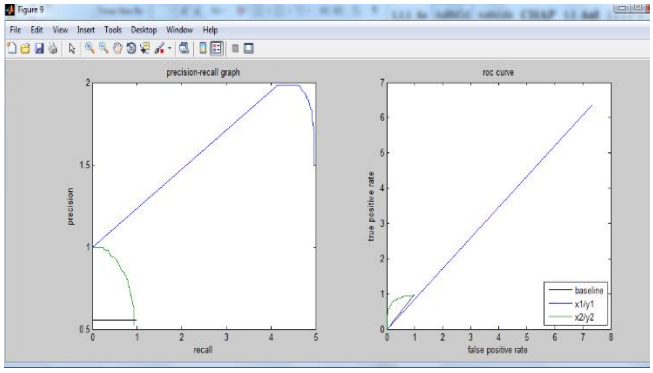

Figure.7 ROC Curve

Above figure shows the ROC curve obtained for English font image.

4. PROPOSED ALGORITHM

[file path] = uigetfile('*.jpg','Image Files');

img = imread([path file]);

img = rgb2gray(img);

img_edge = edge(img,'canny');

img_filled = imfill(img_edge,'holes');

props = regionprops(img_filled,'Area','PixelList'); max_area = props(1).Area;

pixels = props(max_count).PixelList;

conf_region = uint8(double(img) .* double(out_img));

text_region = edge(conf_region);

im1=medfilt2(im1,[3 3]); %Median filtering the image to remove noise%

B=conv2(double(BW),double(msk)); %Smoothing image to reduce the number of connected components

L = bwlabel(B,8);% Calculating connected components

[dwta,dwtb]=dwt(iman,'db4');

load rest.mat

diff_mat=imanb-new_img2;

SVM, siz=M*N;

max_length = size(dataset,2);

no_f_attributes=repval(att);

[num_of_occurance,first_itemset,min_no] =single_itemset(dataset,no_f_attributes,siz);

[sec_itemset,no_two_item,sup_2_item,conf_2_item,simp_2_item,compl_2_item] = two_itemsets(dataset,first_itemset,no_f_attributes,num_of_occurance);

min_supp=0;

min_confi=0.40;

M=size(dataset,1);

two_item=find(sup_2_item>min_supp);

for h=1:length(two_item)

sec_itemset(h,:)=sec_itemset(two_item(h),:);end;

thresh = multithresh(img_edge,2);

seg_I = imquantize(I,thresh);

Find Recall, Precision and accuracy.

5. RESULTS AND DISCUSSION

The whole implementation has been done in MATLB 7.10. Above result snapshots are shown for only English font image. In proposed work Punjabi, Hindi images are also taken as well. Below Table shows the result values for 25 images including Hindi, Punjabi and English font images.

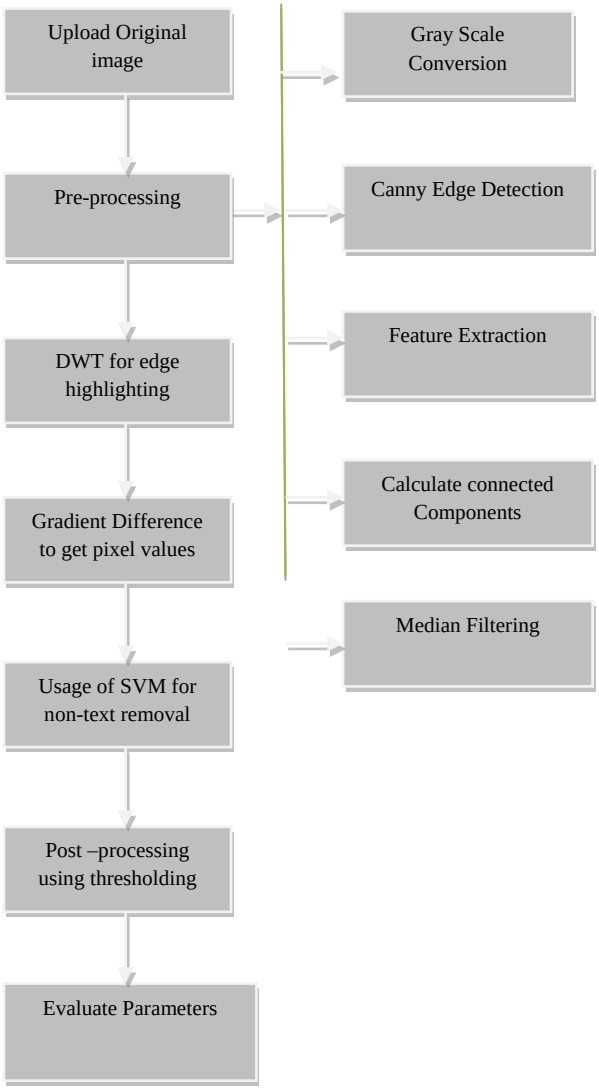

Figure. 8 Proposed Flowchart

5.1 Computation Parameters [21]

5.1.1 False Negatives (FN)

False Negatives (FN)/ Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

5.1.2 False Positives (FP)

False Positives (FP) / False alarms are those regions in the image which are actually not characters of a text, but have been detected by the algorithm as text.

5.1.3 Recall rate (r)

Recall rate (r) is defined as the ratio of the correctly detected characters to sum of correctly detected characters plus false negatives.

r = correctly detected characters / [correctly detected characters + FN] (2)

5.1.4 Precision rate (p)

Precision rate (p) is defined as the ratio of correctly detected characters to the sum of correctly detected characters plus false positives.

p = correctly detected characters / [correctly detected characters + FP] (3)

5.1.5 F-score

F-score is the harmonic mean of the recall and precision rates.

5.1.6 ROC

ROC is really a graphical method for com- paring two empirical distributions.

5.2 Result values

The following Table 1 shows the precision value, recall rate and f-score measures that is being calculated automatically at the end of each image after the extraction of text.

Table. 1 Performance Comparison

Image.no	Precision value (%)	Recall rate (%)	F-Score (%)
1	95.716	92.71	94.21
2	95.713	92.71	94.21
3	95.69	92.69	94.19
4	95.68	92.68	94.18
5	95.70	92.70	94.20
6	95.70	92.70	94.20
7	95.71	92.71	94.21
8	95.71	92.71	94.21
9	95.71	92.70	94.20
10	95.70	92.68	94.18
11	95.68	92.71	94.21
12	95.71	92.72	94.22
13	95.72	92.72	94.22
14	95.72	92.71	94.21
For Punjabi Font Images			
15	95.71	91.70	93.20
16	93.70	91.69	92.19
17	93.69	91.68	93.18
18	94.68	91.69	92.19
19	92.69	91.72	9.22
For Hindi Font Images			
20	95.72	92.69	94.19
21	95.69	92.66	94.16
22	95.72	92.72	94.22

23	95.69	92.69	94.19
24	95.66	92.66	94.16
25	95.72	92.72	94.22

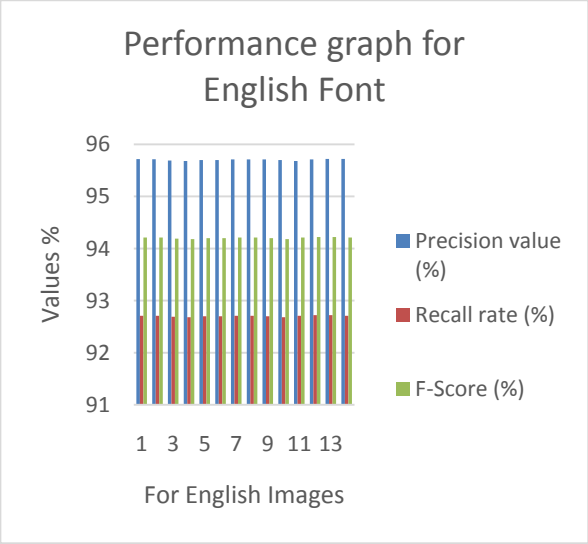

Figure .9 Performance Graph for English Font

Above graph shows the comparison of Precision rate, Recall rate and F-Score values for English font images.

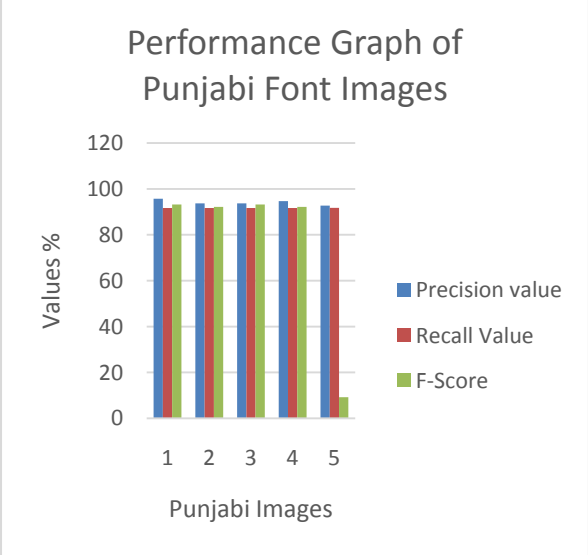

Figure .10 Performance Graph for Punjabi Font

Above graph shows the comparison of Precision rate, Recall rate and F-Score values for Punjabi font images, from the graphs it has been shown that it has medium performance in comparison to English font because each character of Punjabi font is connected to each other so there is no isolation of characters. Hence feature extraction process become difficult.

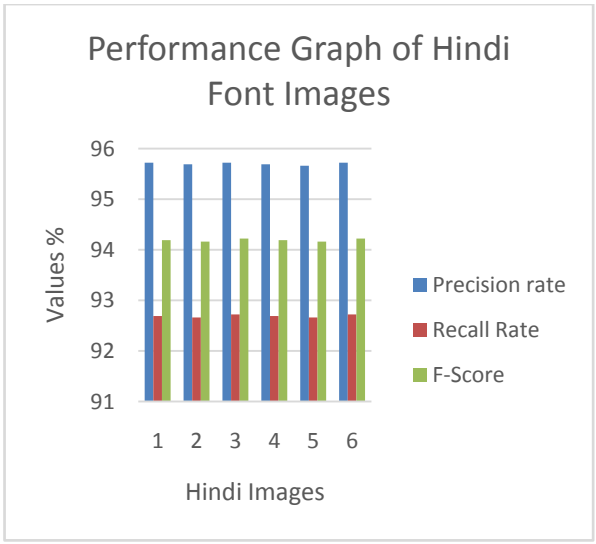

Figure.11 Performance Graph for Hindi Font

Above graph shows the comparison of Precision rate, Recall rate and F-Score values for Hindi font images, from the graphs it has been shown that it has good performance in comparison to Punjabi font but has low performance in comparison to English font images because of the less non-uniformity and less inter-spacing.

6. CONCLUSION AND FUTURE SCOPE

The proposed approach can handle images with varying background of multiple colors and texture; and foreground text in any color, font, size and orientation. From the implementation results it has been concluded that proposed algorithm works for English font images, Hindi Text as well as Punjabi Text images. In other words it can say that it is robust to type of language. In addition to this, in this method main stress is given on the removal of false-positives that is the most important drawback of previous techniques. Future scope lies in the use of the K-Means clustering algorithm for classification of texts.

7. REFERENCES

[1] Uddin, Sultana, M. Rahman, Busra, "Extraction of text from Scene Image using Morphological based approach", International Conference on Machine and Vision, IEEE, pp.876-880 (2012).

[2] Matko Saric, Hrvoje Dujmic, Mladen Russo, "scene text extraction in HIS color space using K-means and modified cylindrical distance", PRZEGLĄD ELEKTROTECHNICZNY, pp. 117-121 (2013).

[3] C. A. Bouman: Digital Image Processing - January 13 (2014).

[4] Fatma H. Elfouly, Mohamed I. Mahmoud, Moawad I. M. Dessouky, and Salah Deyab "Comparison between Daubechies wavelet and Haar transform using FGPA", World Academy of Science,Engineering and Technology, pp.395-400 (2008).

[5] Punam Patel, Shamik Tiwari, "Text segmentation From Images", International Journal of Computer Applications, pp.25-28 (2013).

[6] Neha Gupta, V.K Banga , "Localization of Text in Complex Images Using Haar Wavelet Transform", International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp.111-115 (2012).

[7] A.J.Jadhav Vaibhav Kolhe Sagar Peshwe, "Text Extraction from Images: A Survey", IJARCSSE, pp.333-337 (2013).

[8] R. Chandrasekaran, RM. Chandrasekaran, "Morphological Text Extraction in Images", IJSCT, vol-2, pp.103-107 (2011).

[9] R. Chandrasekaran, RM. Chandrasekaran, P Natrajan, "Text Localization and Extraction in Images Using Mathematical Morphology and SVM", International Conference on Computer Pattern and Recognition, IEEE, pp.55-60 (2012).

[10] Anubhav Kumar, "An Efficient Text Extraction Algorithm in Images", Conference on Contemporary Computing, IEEE, pp.6-12 (2013).

[11] B.H. Shekhar, Smitha M.L, P. Shivkumara, "Discrete Wavelet Transform and Gradient Difference based approach for text localization in videos", Fifth International Conference on Signals and Image Processing, IEEE, pp.280-284 (2013).

[12] Fixing Ye, Qingming Huang, Wen Gao and Debin Zhao, "Fast and Robust text detection in images and video frames", Image and Vision Computing 23 (2005).

[13] C. A. Bouman "Connected Component Analysis," Digital Image Processing, pp. 1-19, January 10 (2011).

[14] R.C. Gonzales and R.E. Woods, Digital Image Processing, Addison-Wesley, Reading (1992).

[15] Leon, M., Vilaplana, V., Gasull, A. and Marques, F., "Caption text extraction for indexing purposes using a hierarchical region-based image model", International Conference on Image Processing, IEEE, El Cairo, Egypt (2009).

[16] J.Sushma, M.Padmaja, "Text detection in colour Images", International Conference on Image Processing, IEEE (2009).

[17] S.Abirami, Dr. D.Manjula, "A Survey of Script Identification techniques for Multi-Script Document Images", International Journal of Recent Trends in Engineering, Vol. 1, No. 2 (2009).

[18] Davod Zaravi, Habib Rostami, Alireza Malahzaheh, S.S Mortazavi, "Journals Subheadlines Text Extraction UsingWavelet Thresholding and New Projection Profile World",World Academy of Science, Engineering and Technology (2011).

[19] J.Fabrizio, M. Cord, B. Marcotegui, "Text extraction from street level Images", CMRT09. IAPRS, Vol. XXXVIII, Part 3/ W4 3-4 (2009).

[20] C. Liu, C. Wang and R. Dai. "Text Detection in Images Based on Unsupervised Classification of Edge-based Features", pp. 610-614, ICDAR (2005).

[21] Niti Syal, Naresh Kumar Garg, "Text Extraction in Images Using DWT, Gradient method and SVM Classifier", IJEATE (2015).