IJCTA, 9(26) 2016, pp. 171-179
© International Science Press

An Enhanced Character Segmentation and Extraction Method in Image-based Email Detection

Mallikka Rajalingam[1] and Putra Sumari[2]

ABSTRACT

As the number of internet users increases, email has been an efficient and popular communication mechanism. The integral of classification of text and image based email by machine learning approaches were used. Image email is one of the most recent tricks introduced by spammers. The unwanted text embedding into images which are sent as email attachments. This paper proposed an enhanced way to segment characters in image based email classification domain. A hybrid approach of Discrete Wavelet Transform (DWT) and Hough Transform based character segmentation is used. This algorithm can enhance the segmentation accuracy. The experimental results show that proposed method produces less number of false negatives when compared with existing machine learning techniques.

Keywords: character segmentation, discrete wavelet transform, email detection, Hough transform.

1. INTRODUCTION

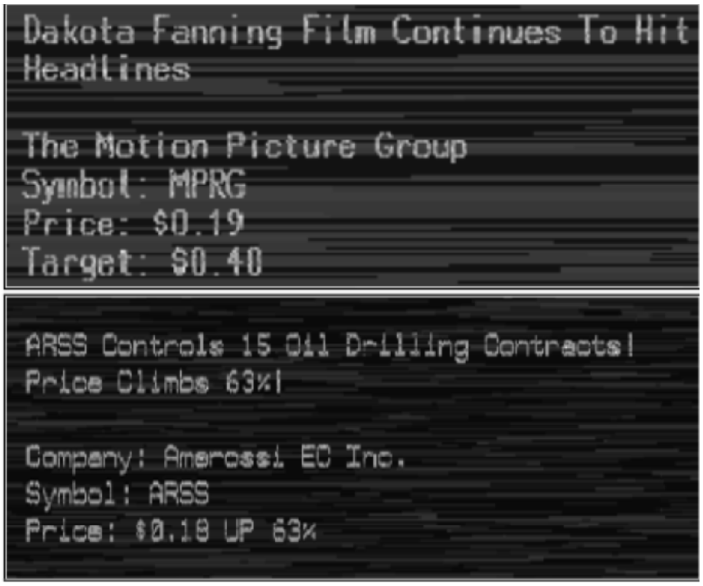
Segmentation of character from an image is an vigorous area in machine learning approach. In large and complex images, it is dynamic to segment the image and then recognize the characters using character segmentation method. Email message bodies consists of header information, subject, and content of the email. Frequently spam images are built by familiarizing random changes to a given prototype image, to make signature-based detection techniques useless, and are complicated to prevent optical character recognition (OCR) tools from analysing the embedded text.

Connected Component (CC) based methods begin with extraction and localize the text regions by processing only connected component information. There are three types of difficulties to be tackle in connected components, 1) To infer text blocks from connected component 2) To filter out non-text objects 3) to extract text. After the connected component extraction, CC-based approaches filter out non-text object. Fig 1 shows sample example of real spam image email [1].

Text line segmentation methods have been developed within diverse projects which perform RGB to gray scale, gray scale to binary conversion, applied hybrid approach and finally, input images are segmented as lines and characters. Otsu's thresholding method involves iterating through all the possible threshold values and calculating a measure of spread for the pixel levels each side of the threshold, i.e. the pixels that fall in either background or foreground.

The rest of the research paper is organized as follows. Section 2 presents a survey of existing character segmentation algorithms in related work. We discuss character segmentation, extraction algorithm and methodology in section 3. We present and analyze the results of the experiment in section 4 and the conclusion with future work is described in section 5.

[1] School of Computer Sciences, Universiti Sains Malaysia, Pulau Penang, 11800, Malaysia, *E-mails: mallikka2002@gmail.com; putras@cs.usm.my*



Dakota Fanning Film Continues To Hit
Headlines

The Motion Picture Group
Symbol: MPRG
Price: $0.19
Target: $0.40

ARSS Controls 15 Oil Drilling Contracts!
Price Climbs 63%!

Company: Amerossi EC Inc.
Symbol: ARSS
Price: $0.18 UP 63%

Figure 1: Example of spam image

2. RELATED WORK

Researchers [2] presented a robust segmentation method for text extraction from the historical document images. The method is based on Markovian-Bayesian clustering on local graphs on both pixel and regional scales. It consists of three steps. In the first step, an over-segmented map of the input image is created. The resulting map provides a rich and accurate semi-mosaic fragments. The map is processed in the second step, similar and adjoining sub-regions are merged together to form accurate text shapes. The output of the second step, which contains accurate shapes, is processed in the final step in which, using clustering with fixed number of classes, the segmentation will be obtained. The method employs significantly the local and spatial correlation and coherence on both the image and between the stroke parts, and therefore is very robust with respect to the degradation. The resulting segmented text is smooth, and weak connections and loops are preserved thanks to robust nature of the method.

In the proposed work [3] presented a methodology for extracting text from images such as document images, scene images etc. Discrete wavelet transform (DWT) is used for extracting text information from complex images. For extracting text edges, the sobel edge detector is used. There are two different approaches have been used for text extraction from complex images namely region based approach and texture based approach. The region based method uses the properties of the color or gray scale in a text region or their differences regarding the background. This method is basically divided in two sub categories: edge based and connected component (CC) based methods. The edge based method is mainly focus on the high contrast between text and background. Connected component based method considers text as a set of separate connected components, each having distinct intensity and color distribution. The texture approach is based on the concept of textural properties. In this method, Fourier transforms. Discrete cosine transform and wavelet decomposition are generally used.

A novel [5] hybrid approach of text segmentation using edge and texture feature information. The texture features such as homogeneity, contrast, energy for texts are different from non-text. The texture features are used to detect the text region from image. The edge based textures have many desired properties. The gradient magnitudes usually have higher values in the edge of the characters, even when the text is embedded in pictures.

Classification of [6] image spam and legitimate email using feed forward back propagation neural network (BPNN) model was proposed. In this technique the features of an image is extracted using histogram.

In this method gradient histogram is extracted from an image. The obtained feature point of an image is then processed using Artificial Neural Network (ANN). In particular, this paper utilize feed forward back propagation neural network for classifying the image spam from those of legitimate mail (“ham”).

Methodology for extracting text [11] from images such as document images, scene image etc. Discrete Wavelet transform (DWT) is used for extracting text information from complex images. For extracting text edges, the sobel edge detector is used.

3. PROPOSED CHARACTER SEGMENTATION AND EXTRACTION ALGORITHM

Segmentation is a process of partitioning a digital image into multiple segments. The objective of segmentation is to simplify and/or change the representation of an image into something that is more expressive and easier to analyze. This section explains how the image email characters are segmented. Fig 2 shows the detailed flow chart of character segmentation phase. It contains five stages. The stage 2 and 3 will be combined and also stage 4 and 5 are combined into single stage.

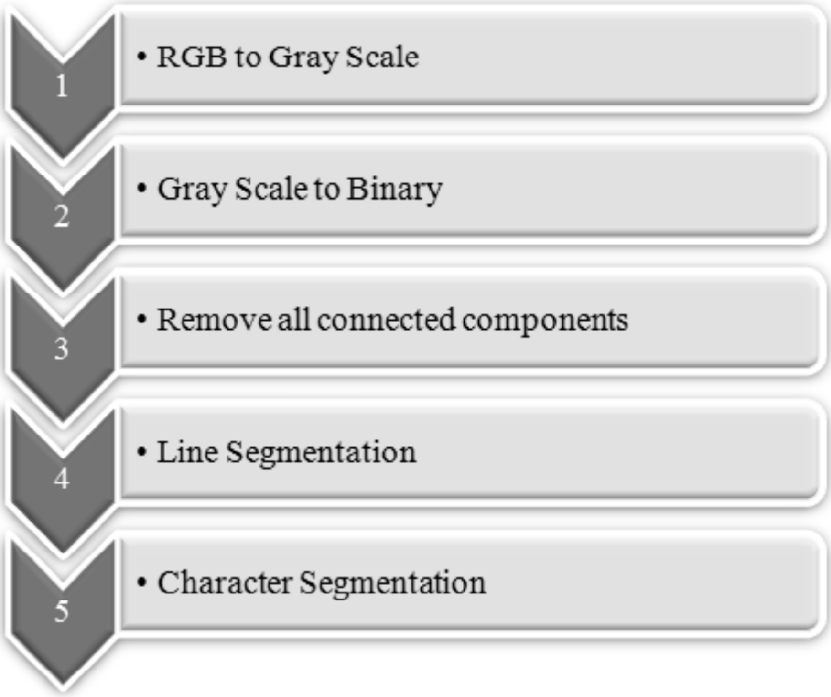


Figure 1: Detailed Flowchart of Character Segmentation

(A) RGB to Gray Scale

In this process the color image is converted into gray scale image. To convert any color to a grayscale demonstration of its luminance, initially obtain the values of its red, green, and blue (RGB) primaries in linear intensity encoding, by gamma expansion. Then, add together 30% of the red value, 59% of the green value, and 11% of the blue value (these weights depend on the exact choice of the RGB primaries, but are typical). Regardless of the scale employed (0.0 to 1.0, 0 to 255, 0% to 100%, etc.), the resultant number is the desired linear luminance value; it typically needs to be gamma compressed to get back to a conventional gray scale representation. By human eye method,

$$\text{Gray} = 0.299 * \text{Red} + 0.587 * \text{Green} + 0.114 * \text{Blue}$$

To convert a gray intensity value to RGB, simply set all the three primary color components red, green and blue to the gray value, correcting to a different gamma if necessary.

(B) Binarization

Binarization is performed so as to convert the RGB and gray scale images to the black and white pixel images. Only in a black and white image noise removal can be done efficiently without affecting the character pixels. The gray scale is converted into binary image using Otsu's method. The aim is to find the threshold value where the sum of foreground and background spreads is at its minimum.

After conversion of binarization removes all connected components (objects) that have fewer than 15 pixels from the binary image.

(C) Line and Character Segmentation

After binarization the lines and characters are segmented. The hybrid approach Discrete Wavelet Transform (DWT) and Hough Transform is used to segment the characters.

DWT decompose signal into different components in the frequency domain, and the 2-d DWT in which it decomposes input image into four components or sub bands. One average component (LL) and three detail components (LH, HL, HH) as shown in fig 3, Sub bands are used to detect candidate text edges in the original image. The LL band is more significant band it contains more information of the original image, so it is most important part of the algorithm process. The LL sub-band can be further decomposed into four sub-bands. This process can continue to the required number of levels. It is known multi-level decomposition.

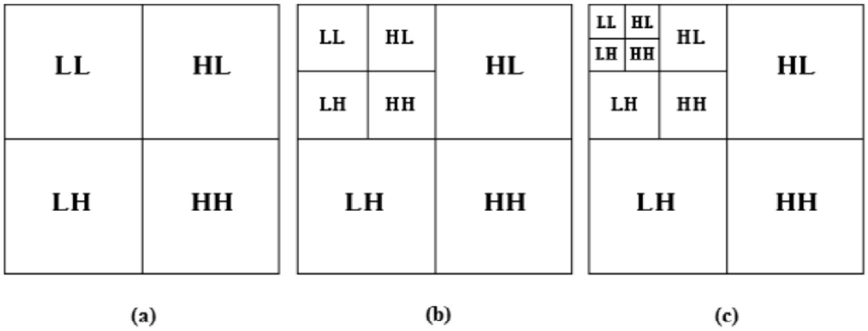


Figure 2: Different level of DWT Decomposition

The three level decomposition of the given digital image is as shown in fig 3. High pass and low pass filters are used to decompose the image first row-wise and then column wise. Hough transform is applied on the binarized edge map DWT image to generate the Hough image of it. For this purpose, the parameters of the Hough transform, like theta (line rotation angle), rho (distance from the coordinate origin), peaks (row and column coordinates of Hough transform bins) are initialized or tuned in such a way that the lines are extracted as a set of connected words.

The Hough Transform Algorithm [7]

1. Find all of the desired feature points in the image space
2. For each feature point in image space
3. For each possibility *i* in the accumulator that passes through the feature point
4. Increment that position in the accumulator
5. Find local maxima in the accumulator
6. On requirement map each maxima in the accumulator back to image space

First the image has to be segmented row-wise (line segmentation), then each rows have to be segmented column-wise (character segmentation). The binary image is segmented row-wise (line segmentation).

Skew Detection

The projection parallel to the true alignment of the lines will likely have the maximum variance, since when parallel, each given ray projected through the image will hit either almost no black pixels as it passes between text lines or many black pixels while passing through many characters in sequence

Skew Correction

It is done by rotating the digitized image by skewed angle. If image is skewed in clockwise then image has to be rotated in anti-clock wise direction and wise versa.

Character Recognition

The combined approach template matching and contour analysis is used to recognize the characters. For template matching, the segmented character is resized into 24?42 grid. Each segmented character is matched with the training set. If it is matched, the corresponding image character will be returned as a text character.

4. RESULTS AND DISCUSSION

The proposed method was implemented using MATLAB (version R2013a) and the experiments are performed on an Intel(R) Core (TM) i5 machine with a speed 2.60 GHz and 8.0 GB RAM using Windows 8.1 64-bit Operating System. The resulted image of the line segmentation is shown in fig 4 and the resulted image of the Hough transform is shown in fig 5. The accuracy of classification is measured by computing the False Positive (FP), False Negative (FN), True Positive (TP), True Negative (TN), Recall, and Precision.

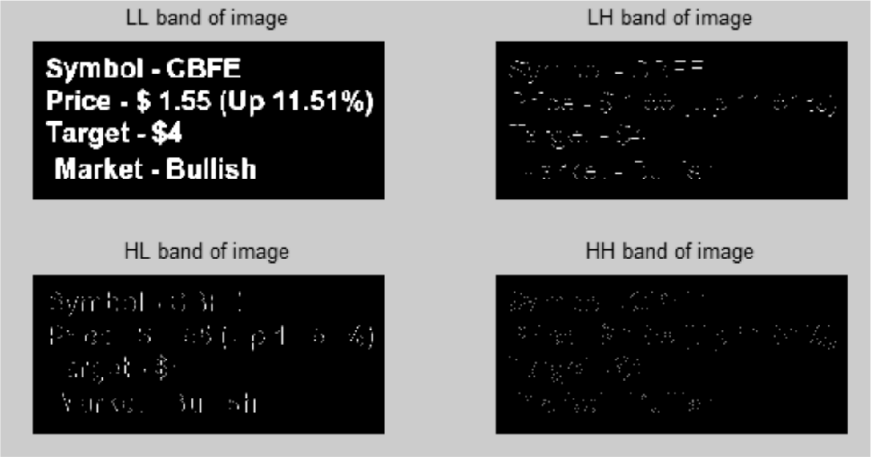


Figure 4: DWT of Image

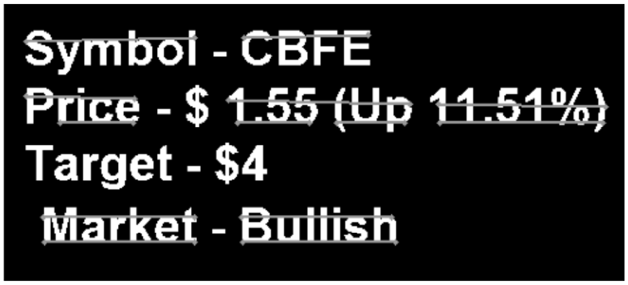


Figure 5: Hough Transform

The outputs of each step in Character segmentation are illustrated in following figures fig 6, fig 7 and fig 8.

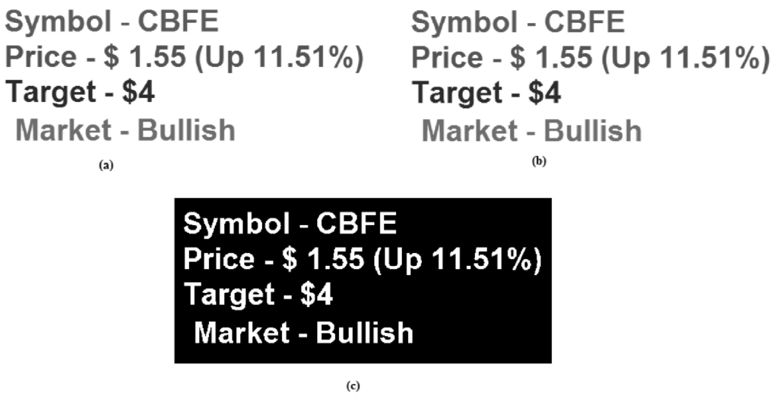

Symbol - CBFE
Price - $ 1.55 (Up 11.51%)
Target - $4
Market - Bullish
(a)
Symbol - CBFE
Price - $ 1.55 (Up 11.51%)
Target - $4
Market - Bullish
(b)
Symbol - CBFE
Price - $ 1.55 (Up 11.51%)
Target - $4
Market - Bullish
(c)

Figure 6: (a) Original Image, (b) Gray Scale Image, (c) Binary Image

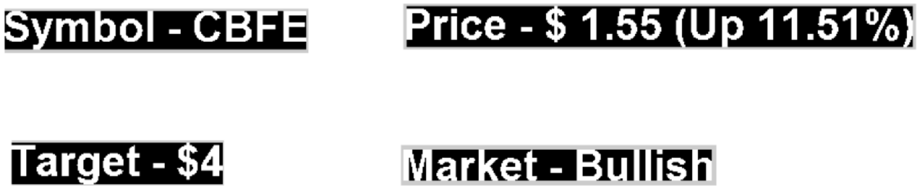

Symbol - CBFE
Price - $ 1.55 (Up 11.51%)
Target - $4
Market - Bullish

Figure 3: Segmented Lines

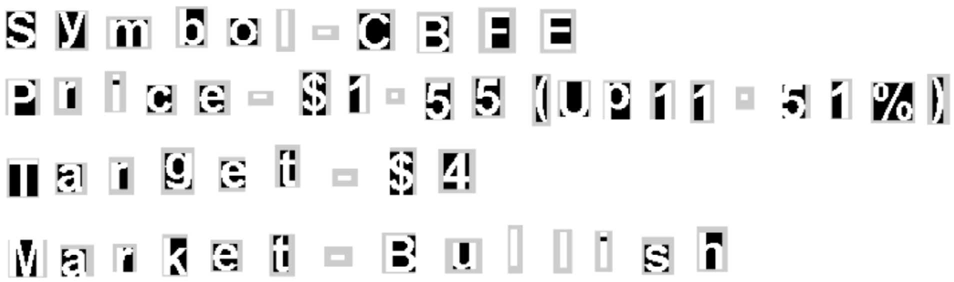
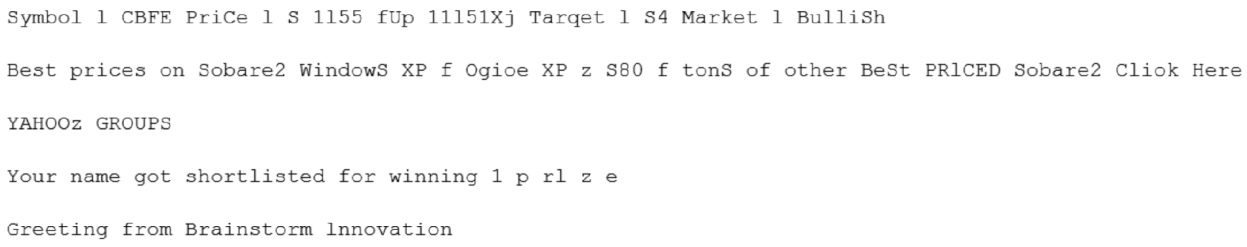

S y m b o l - C B F E
P r i c e - $ 1 . 5 5 (U p 1 1 . 5 1 %)
T a r g e t - $ 4
M a r k e t - B u l l i s h

Figure 4: Segmented Characters



```
Symbol l CBFE PriCe l S 1155 fUp 11151Xj Tarqet l S4 Market l BulliSh

Best prices on Sobare2 WindowS XP f Ogioe XP z S80 f tonS of other BeSt PRlCED Sobare2 Cliok Here

YAHOOz GROUPS

Your name got shortlisted for winning 1 p rl z e

Greeting from Brainstorm lnnovation
```

Figure 9: Extracted Text from Image Email

Fig 9 shows extracted text from image email. A combined-approach used for character segmentation i.e. Discrete Wavelet Transform (DWT) and Hough Transform techniques to achieve competitive segmentation accuracy. Morphological operations like erosion and dilations are used for better approach of refining text region segmentation. Morphological operations are helpful in the removal of no texted regions. The Hough transform technique is used for feature extraction in image analysis and most importantly designed for identification of lines in the image & identifying positions of arbitrary shapes. The purpose of Hough transform is to find imperfect or occurrence in the object within the shape. The output of DWT gives as input to Hough transform to get refined results. Table 1 shows the performance of image based email and result of total characters, extracted characters, correctly matched with given image email and incorrect match with given image emails.

Table I
Performance of Image Email

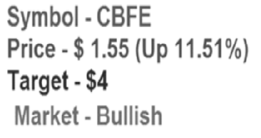
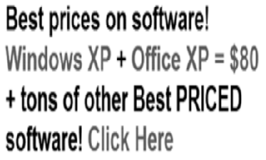
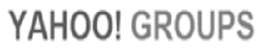
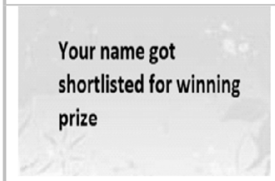
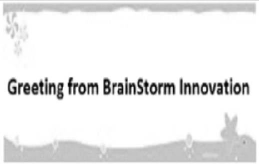
Image Email	Extracted Text	Result
 Symbol - CBFE Price - $ 1.55 (Up 11.51%) Target - $4 Market - Bullish	Symbol l CBFE Price l S 1155 fUp 11l51Xj Tarqet l S4 Market l Bullish	Total Char - 70 Extracted Char - 70 Correct - 58 Incorrect - 12
 Best prices on software! Windows XP + Office XP = $80 + tons of other Best PRICED software! Click Here	Best prices on Sobare2 Windows XP f Ogioe XP z S80 f tons of other Best PRlCED Sobare2 Cliok Here	Total Char - 103 Extracted Char- 98 Correct - 88 Incorrect - 10
 YAHOO! GROUPS	YAHOOz GROUPS	Total Char - 14 Extracted Char- 14 Correct Char - 13 Incorrect – 1
 Your name got shortlisted for winning prize	Your name got shortlisted for winning 1 p rl z e	Total Char – 44 Extracted Char - 49 Correct - 39 Incorrect - 5
 Greeting from BrainStorm Innovation	Greeting from Brainstorm lnnovation	Total Char - 36 Extracted Char- 36 Correct - 36 Incorrect - 0

Table II shows the image email accuracy from above table with image id for identification.

Table II
Image Email Accuracy

Image Id	*1*	*2*	*3*	*4*	*5*
Accuracy	82.85 %	85.43%	92.85%	88.63%	100%

Image based spam is a breakthrough from the spammers' view point; it is a simple and effective way of deceiving spam filters since they can process only text. The embedded image carries the target message and most email clients display the message in their entirety. Since many ham emails also have similar properties using HTML, carrying embedded images, with normal text as image-based emails, existing spam filters can no longer distinguish between image-based spam and image ham. For experimentation, Ling-Spam Corpus database are used for training and testing in the ratio of 70/30. The number of emails taken from ham and spam dataset using confusion matrix is shown in table III.

Table III
Confusion Matrix of Email Classification

		Predicted	
		Ham	*Spam*
Actual	Ham	2310	102
	Spam	19	463

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.

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

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]

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]

F-score is the harmonic mean of the recall and precision rates. The false negative rate is reduced based on the performance analysis, the results of TP, FP, TN, FN, precision, recall, F-measure are 0.99, 0.18, 0.81, and 0.008 respectively.

5. CONCLUSION WITH FUTURE WORK

The main objective of this paper explains about proposed algorithms, methods, and techniques used to segment characters in image based email. For experimentation, Ling-Spam Corpus database are used for training and testing. Image-based email performance evaluation with results are discussed and outputs are shown. The overall accuracy 100% is obtained by proposed hybrid approaches of character segmentation. The performance of false negative rate is reduced. To increase the size of training and testing data set will be the further enhancement of current work.

REFERENCES

[1] Biggio, B., Fumera, G., Pillai, I. & Roli, F. (2011). "A Survey and Experimental Evaluation of Image Spam Filtering Techniques. " Elsevier. Pattern Recognition. pp. 1436-1446.

[2] Hedjam, R., Moghaddam, R.F. & Cheriet, M. (2010). "Text extraction from degraded document images." 2nd European Workshop on Visual Information Processing (EUVIP). 2010. vol.5. no. 6. pp.247- 252.

[3] Gupta, N. & Banga, V.K. (2012). "Image Segmentation for Text Extraction." 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE, 2012). Singapore. April 28-29.

[4] Shivananda, N. & Nagabhushan, P. (2009). "Separation of Foreground Text from Complex Background in Color Document Images." Seventh International Conference on Advances in Pattern Recognition. 2009. ICAPR '09. vol.4. no.6. pp.306-309.

[5] Patel, P. & Tiwari, S. (2013). "Text Segmentation from Images. "International Journal of Computer Applications. vol. 67. no.19.

[6] Soranamageswari, M. & Meena, C. (2011). "A novel approach towards image spam classification." International Journal of Computer Theory and Engineering. vol.3. no.1. pp. 84-88.

[7] Saha, S., Basu, S., Nasipuri, M. & Basu, D. (2010). "A Hough Transform Based Technique for Text Segmentation." Journal of Computing. vol. 2. no. 2.

[8] Jeffrey, Z., Ramalingam, S. & Bekooy, N. (2012). “Real-Time DSP-Based License Plate Character Segmentation Algorithm Using 2D Haar Wavelet Transform.” Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology. pp. 1-22.

[9] Singh, P.N. & Jain, A. (2014). “Text and Character Extraction of Colour Image using DWT in MATLAB Image Processing Tool.” “International Journal of Advanced Technology in Engineering and Science. vol.2, no. 6. pp. 152-158.

[10] Syal, N. & Garg, NK. (2014). “Text Extraction in Images Using DWT, Gradient Method and SVM Classifier.” International Journal of Emerging Technology and Advanced Engineering. vol. 4. no. 6. pp. 477-481.

[11] Gupta, N. & Banga, V.K. (2012). “Image segmentation for text extraction.” 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE’2012) Singapore. April 28-29.