

IMAGE BASED SPAM EMAIL DETECTION BY COMBINED APPROACH

Mallikka Rajalingam

School of Computer Sciences University Sains Malaysia,

Pulau Penang, 11800, Malaysia (India)

ABSTRACT

Electronic mail is one of the important communication channels of information technology which serves across the globe. Though the functionalities of e-mail have been very helpful in serving both individuals and institutions, it encounters a major issue called 'Spamming'. Spam mails are unwanted text or image-based messages, often sent without the consent of users so as to fill their mailboxes. In this paper, proposes two new novel methods-firstly the hybrid character segmentation method is proposed which uses Discrete Wavelet Transform and Hough Transform to segment the characters. Secondly, the hybrid character recognition is proposed which uses Template Matching and Contour Analysis towards recognizing the characters. The final phase of the paper is a complete spam detection system with the two proposed works built to detect spam messages. The overall efficiency of the proposed system reached above 90% which discerns the proposed work as a significant contribution to the research community.

Keywords-Text segmentation, Text recognition, Image spam detection.

I. INTRODUCTION

Email communication is one of the most efficient and most popular communication systems that enable people to communicate with each other. The total number of worldwide email accounts is expected to increase from 3.3 billion in 2012 to 4.3 billion accounts by the end of 2016 [4]. This represents an average annual growth rate of 6% over the next four years. In this regard with such an alarming usage of email communication, managing emails against fraudulent activities has become an important task. One such activity through emails is the impulsive posting of unwanted email to users known as spam messages. A spam mail is defined as an unsolicited/irrelevant/unwanted mail message received by users [2]. Spam mails usually contain commercial or profitable campaigns of uncertain products, dating services, get-rich-quick schemes and advertising. Spam emailing is also used to spread malicious or virus codes and is intended for fraudulence in financial transaction or phishing. Spamming is considered to regulate losses over the internet especially when they tend to turn malicious for business organisations. Several losses are mostly collateral damages not focusing a particular network or any organization. Spam mails occupy more network bandwidth during transmission. It also consumes user time in terms of searching. Statistical reports show, as of December 2014, spam messages accounted for 66.41 per cent of e-mail traffic worldwide and Asia constitutes 54% of the total percentage [5]. A recent study by [1] reveals the fact that most of the users receive more spam emails than non-spam emails.

Detection of spam email messages and quarantining it aside from the users are an important task. Spam detection consists of a series of steps- firstly, it starts with the tokenization phase in which the email content is parsed into a token. A token can be a word. The token is then transferred to the cleaning phase to process and form a single basic word without prefix and suffix. Then the processed tokens are sent to the spam detection phase to check whether the tokens are either spam or not. The clean token (not spam) is sent to the inbox folder and the infected known as 'spam' will be sent to the spam folder. The spam detection process requires understanding the message (token) (characters – alphabets, number, and symbols) written in the email. As a text email, the token is in ASCII character form for words and sentences therefore it is well understood and easily processed by the system for decision making.

Though text based spam emails are detected by most methods of email spam detection, spammers have identified new routes towards sending spam messages through images. Such a form of sending spam messages through images is called as image spamming and images embedded with spam characteristics are known as spam images or Image spam. Most algorithms find it easy to identify spam in text email. However, the same in image spam emails is a daunting task. A spam image carries a message which is intended to reach client systems and displays the same. One another complexity of spam detection techniques is though they are better methods to detect spam; they may also intend to block ham messages wherein the process is known as false positive [3]. However, detection of image spam is a difficult task as the messages or token (characters) is embedded within the images. The token or character embedded in image needs to be extracted and should be converted (also known as character recognition) into ASCII form. Character recognition within an image is indeed a challenging task as it involves image processing as the first process which involves character segmentation to mark the character in the image and the second process known as character recognition which is to convert the marked character into ASCII form. In the final process, ASCII forms are ready to be processed for identifying spam emails. Detecting spam mails especially image spam as shown in figure 1 is the focus of the present research which is a challenging task when compared with other conventional spam detection techniques.

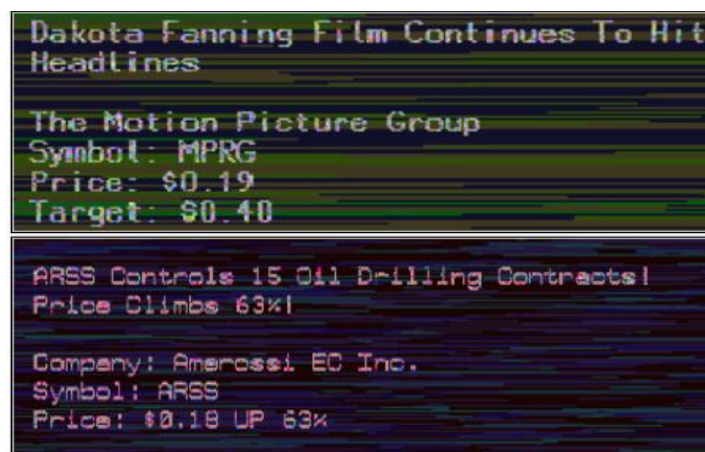


Figure 1 Sample spam email

This paper is organized as follows. Section II presents in detail an enhanced character segmentation algorithm that improves the detection efficiency of image based emails. Section III describes in detail the processes involved in character recognition. The components of this algorithm and their functions are discussed. Section IV presents in detail a detection algorithm for image based ham/ spam emails using the shape based feature

extraction technique. Section V discusses the entire approach with the discussion of the different algorithms used followed by testing the entire system based on various parameters. Section VI concludes the investigation and gives recommendations for upcoming work with esteem to this work.

II. CHARACTER SEGMENTATION

This section presents the details of the proposed work called Hybrid Based Character Segmentation. The hybrid refers to the combination of DWT and Hough transform techniques which are hypothesised to enhance the character segmentation algorithm thereby improving the detection efficiency of the image based emails. Furthermore, the accuracy of character segmentation is improved using the pixel count analysis method.

The proposed digital image character segmentation is shown in figure 2. The proposed method consists of two main components- (1) pre-processing component and (2) segmentation component. The pre-processing component prepares the image into an easy and simplified form for the next segmentation component. The pre-processing component performs three main tasks which include RGB to Grey-scale conversion, binary conversion and removal of connected components. The character segmentation component performs three main tasks: application of DWT, line segmentation and finally character segmentation. In this proposed work, the hybrid algorithm combines two methods namely DWT and Hough transform techniques which are applied in the segmentation component wherein pixel count analysis is performed to improve the accuracy of segmentation. The combination of these two techniques enables good segmentation of the characters from the image.

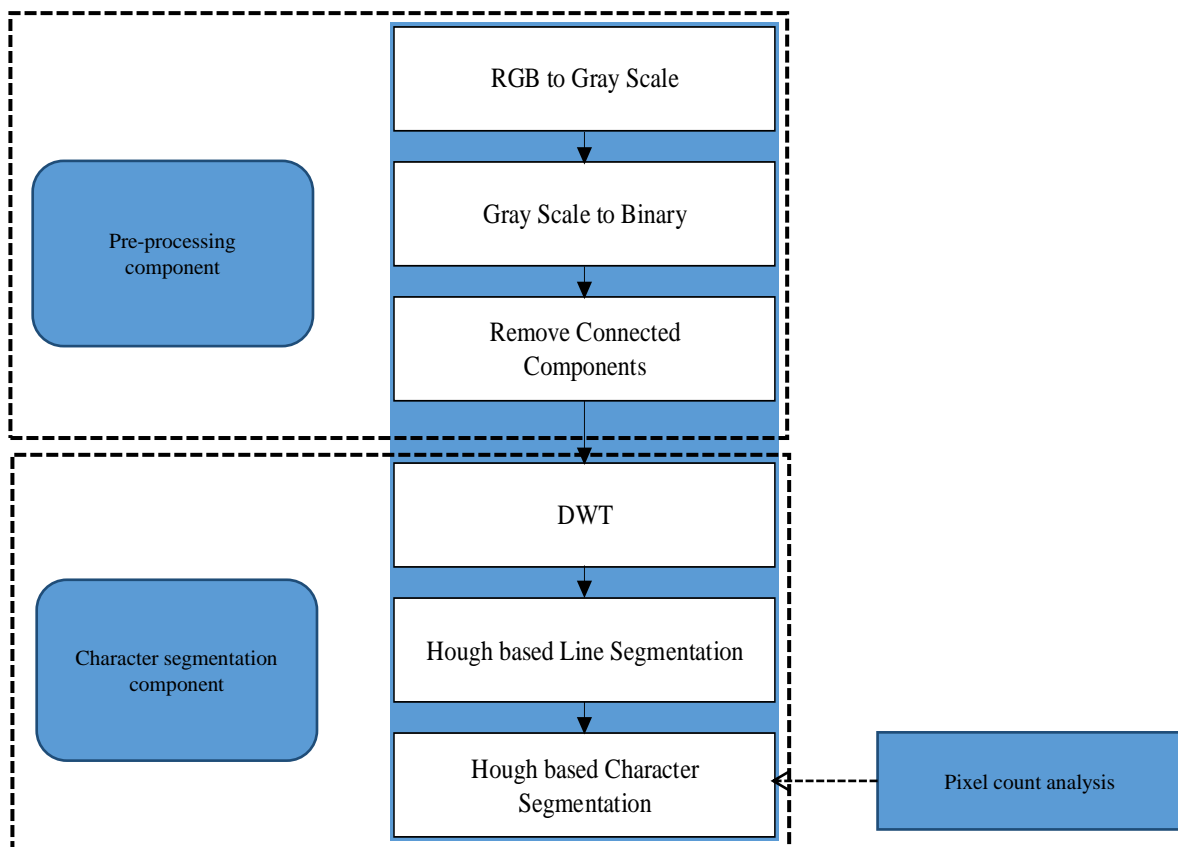


Figure 2 Detailed flowchart of character segmentation

A. Overall Hybrid Algorithm

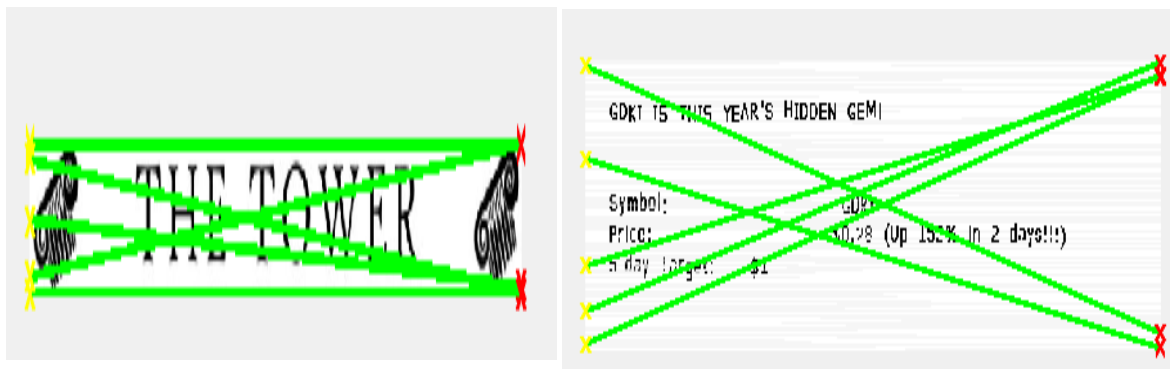
Character segmentation from image is a daunting task since it suffers from characteristics such as occluded character and low resolution image. The difficulty is to segment a character when detecting white space between words. When the characters are touched or overlapped, projection profile method does not give good results therefore the following hybrid algorithm is used. At each input image point, a number of lines are plotted at different angles. The vertical projection is used to separate the base characters using the white space between them. The hybrid method of DWT and Hough Transform to segment characters from image email improves the performance of character segmentation.

Hybrid Algorithm

Input: After the process of binarization, the pre-processed image is provided as input to the hybrid algorithm. The following steps depict how the Discrete Wavelet Transform and Hough Transform operates with the selected Image:

1. Apply DWT – Discrete Wavelet Transform is applied to the pre-processed binary image
2. The binary image is decomposed into single level through DWT
3. The LL, LH, HL, HH components are identified
4. For each components in the decomposed image
 - 4.1 Find the Edges use Canny Edge Detector Algorithm [6]
5. End For
6. Hough Transform technique is applied to segment lines
7. Find the horizontal lines in image
8. Extract all the Lines from image
9. For each line in the image
 - 9.1 Identify the character location
 - 9.2 Segment the Character
10. End For

Output: Line and Character Segmented Image



(a) Ham email

(b) Spam email

Figure 3 Results of proposed segmentation algorithm

The images decomposed using DWT is then subjected to Hough transforms wherein the lines are detected in the image. When the line is detected it marks horizontal line to the image. The output images after the application of Hough transforms with the lines are provided in Figure 3.

III. CHARACTER RECOGNITION

This section discussed in detail the proposed work called combined approach (Template matching and Contour analysis) of Character Recognition. This chapter also presents the details of the algorithms that would enhance character recognition algorithm which improves the identification competence of the image attachment based emails with the aid of Template Matching and Contour analysis.

The proposed digital image character recognition flowchart is shown in figure 4. The proposed method consists of two main components- (1) pre-processing component and (2) character recognition component. The pre-processing component prepares the projection parallel to the true alignment of the lines which will likely have the extreme variance; since when it is parallel, each given ray projected through the image will hit either almost no black pixels or many black pixels. After skew detection, skew correction is performed by rotating the digitized image by skewed angle which is performed through simple rotation. If an image is skewed in clockwise direction then the image has to be rotated in the anti-clock wise direction. If image is skewed in anti-clockwise, then image has to be rotated in clock wise direction. The recognition component has three main tasks: template matching, contour analysis and finally character recognition. Template matching is one of the techniques for recognition of characters. It is the method of finding the position of a sub image termed as template inside an image. After skew correctness, characters are compared with the template. The matched word is given to contour which helps to mark the edges, shapes of the characters for the exact match of the individual character.

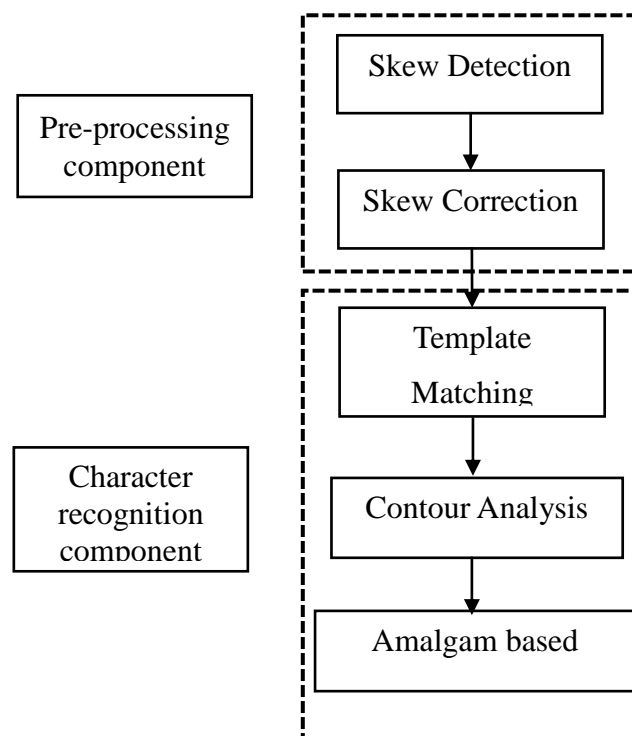


Figure 4 Detailed flowchart of character recognition

B. Overall Hybrid Algorithm

Character recognition from image is slightly difficult because of poor quality document, smaller font size, and different font types. The proposed algorithm is hence used to refine the result for better performance.

Hybrid Algorithm

Input: Segmented Character Image

Output: Recognized Text Character

Steps:

1. Read Segmented Character Image (The characters are segmented using hybrid approach of DWT and Hough Transform)
2. Read Template Training
3. For each image in Segmented Character
 - 3.1 Read the character image
 - 3.2 For each template in Training
 - 3.2.1 Compute Similarity between character image and template training image
 - 3.2.2 If the character is matched
 - 3.2.2.1 Return Corresponding Text Character
 - 3.2.2.2 Else
 - 3.2.2.3 Apply Contour Analysis technique
 - 3.2.2.4 Find the character Shape
 - 3.2.2.5 Match with Training Templates
 - 3.2.2.6 Return Corresponding Text Character
 - 3.2.3 End If
 - 3.3 End For
4. End For

The work flow of character recognition by combined approach is the process to read segmented character image using hybrid approach of DWT and Hough Transform. For each segmented characters, template training words are read. For each template in training set, the character image is read. The similarity between the character image and template in training image should be computed further. If the character is matched, then the text character is returned wherein the resultant text character is refined using Contour Analysis to attain character recognition improvements. Contour analysis allows storage, comparison and recognition of characters presented in the form of the exterior outlines. It is supposed that the contour contains the sufficient information on the character shape. Interior points of the fonts or characters are not accepted in the system. The contour is the boundary of characters, a population of points (pixels), and separating character from a background.



Figure 5 Results of proposed recognition algorithm

Figure 5 shows the results of the segmented and recognised characters from the images.

IV. IMAGE SPAM EMAIL DETECTION

This section presents in detail the use of Shape based feature extraction which enables the identification of Spam/ Ham from Image emails. In this regard, this part at first elaborates on the methods of visual feature extraction (Text layout analysis) and the details of the algorithms used for Image Ham/ Spam detection. Furthermore, the score and performance metrics of the identified images are provided which is the result of the experiments.

Since spammers generally send image spam in the form of batches which consists of similar features, image based spam detection method can filter those images effectively on the basis of known image spams that are collected, stored, trained and classified. The underlying principle for the spam detection system is as follows: firstly, the features of the detected image such as low-level features (visual) and high-level features (semantics). Secondly, the features are compared with the features in two feature databases (DB with spam features and DB with ham-features). Finally, the image is judged whether it is spam or ham. The architecture for image spam detection is shown in figure 6.

Examining the various spam images from the image spam dataset, it is revealed that spammers generally utilise the same text layout template for the generation of different advertisements wherein only the use of words/ text in the images change based on the different products they attempt to advertise. For the analysis of the text layout, the minimum bounding box technique is used for the whole area, which is again dilated for connecting words that are in the same line. Scaling is then performed for the text area which is dilated and is then normalised for the comparison of text layout [7].

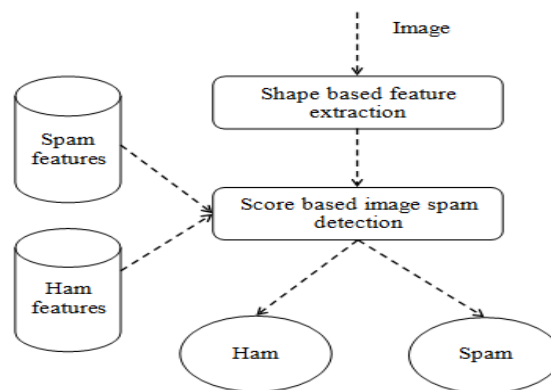


Figure 6 Architecture for image spam detection

V. EXPERIMENTATION AND RESULTS

The proposed algorithms that will improve the detection accuracy of text and image based email classification are experimented to achieve the research objectives. All the proposed algorithms are fulfilled using MATLAB (version R 2013a), and the experimentations 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. For experimentation, images were taken from image spam dataset. The image spam dataset acts as a fast classifier and hence is utilised in the study [8]. The proposed approaches (Character segmentation using DWT and Hough transforms, Template matching and Contour analysis, and Shape based feature extraction) are evaluated for performance wherein the samples of images would be 23 for testing and 560 images to measure accuracy by true positive, true negative, false positive, false negative, precision, recall and F-measure.

This step inspects the performance efficiency of the proposed algorithms by assessing the results with respect to detection accuracy. Furthermore, comparisons among the proposed algorithms and current state-of-the-art methods are shown in the experimentations. The accuracy of proposed algorithms is measured using True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Recall, and Precision.

Image spam dataset is downloaded from [9]. The image spam data set contains 2173 images in SpamArchive corpus, 2359 images in personal ham corpus, and 1248 images in personal spam. The additional data will be given the name non-benchmark data in this thesis. For the assessment of the proposed algorithms with respect to the detection accuracy, the proposed algorithms are compared with existing individual methods.

Furthermore, to examine the performance of the proposed approaches (Hybrid Character segmentation, Template matching and contour analysis, and shape based feature extraction), the values of false negative, false positive, true negative, true positive, precision, recall and F-measure are measured and are compared with the values of the factors acquired in previous researches. Following is the description of the performance analysis indicators used in the present research:

False Positive Rate (FP)

When a test falsely or incorrectly reports a positive result,

$$FP = \frac{b}{b + d}$$

False Negative Rate (FN)

When a test falsely or incorrectly reports a negative result,

$$FN = \frac{c}{c + a}$$

Precision Rate (P)

Precision rate (P) is calculated by the ratio of correctly segmented characters to the sum of correctly segmented characters and false positive.

$$P = \frac{\text{Correctly segmented characters}}{\text{Correctly segmented characters} + \text{False Positive}}$$

Recall (R)

Recall (R) is calculated by the ratio of correctly predicted characters to sum of correctly identified characters plus false negative.

$$R = \frac{\text{Correctly segmented characters}}{\text{Correctly segmented characters} + \text{False Negative}}$$

F-Measure (F)

F-Measure (F) is calculated using precision and recall.

$$F = 2 * \frac{P * R}{P + R}$$


Accuracy (A)

Overall accuracy is calculated by,

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN}$$

With the characters segmented and recognized, shape based feature extraction is performed. The shape features for each character segmented and recognized are extracted based on the region properties wherein several features are examined which included- Area, Bounding box, Centroid, Eccentricity, Euler’s number, Extent, Extremer, Major Axis length, Minor Axis length, Orientation and Perimeter. Of all these features, the threshold average fit best for the feature ‘Area’ which is hence used to detect whether an image is HAM or SPAM. The feature value of a testing input image is compared with the trained feature values of multiple images using Multi-SVM algorithm which classifies and produces the result whether an image is HAM/ SPAM. However, the performance analysis of the proposed system is measured using metrics such as Total positive rate/Sensitivity, Total Negative Rate/Specificity and Accuracy. Table I and Table II provides the information of the performance metrics for each image identified using the proposed Ham/ Spam detection system.

Table I: Performance metrics of images detected as HAM using the proposed approach

No.	Images	Performance metrics
1		THIS IS SPAM-FREE AND HAM IMAGE Correct Rate is: 82.2581% Error Rate is: 17.7419% True Positive is: 100 False Positive is: 10 True Negative is: 90 False Negative is: 0 True Positive Rate (TPR)/ Sensitivity is: 100%

	<p>True Negative Rate (TNR)/Specificity is: 78%</p> <p>False Positive Rate (FPR) is: 22%</p> <p>False Negative Rate (FNR) is: 0%</p> <p>Precision is: 0.909091</p> <p>Recall is: 1</p> <p>Fmeasure is: 0.952381</p> <p>Accuracy of Linear Kernel SVM is: 96.7742%</p>
--	---

Table II: Performance metrics of images detected as SPAM using the proposed approach

No.	Images	Performance metrics
1	<p>GOX1 IS THIS YEAR'S HIDDEN GEM!</p> <p>Symbol: GOX1</p> <p>Price: \$0.28 (Up 152% in 2 days!!!)</p> <p>5-day Target: \$1</p>	<p>SPAM IS DETECTED</p> <p>Correct Rate is: 82.2581%</p> <p>Error Rate is: 17.7419%</p> <p>True Positive is: 100</p> <p>False Positive is: 10</p> <p>True Negative is: 90</p> <p>False Negative is: 0</p> <p>True Positive Rate (TPR)/ Sensitivity is: 100%</p> <p>True Negative Rate (TNR)/Specificity is: 78%</p> <p>False Positive Rate (FPR) is: 22%</p> <p>False Negative Rate (FNR) is: 0%</p> <p>Precision is: 0.909091</p> <p>Recall is: 1</p> <p>Fmeasure is: 0.952381</p>

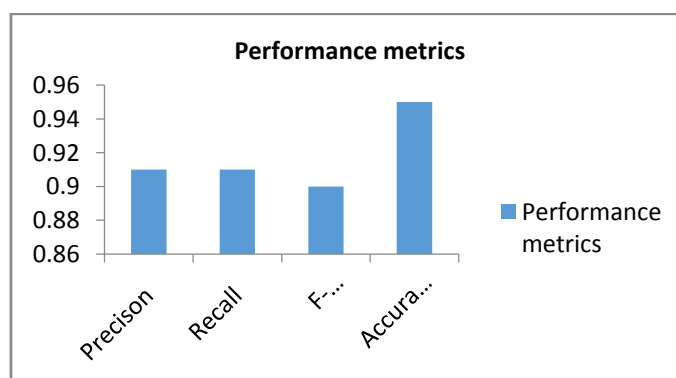


Figure 7 Performance metrics of the proposed Image Ham/ Spam detection approach

Of all the 23 images tested from 560 images in the training dataset, an average F-measure value of 0.90 is obtained with the recall value of 0.91. The accuracy reached 0.95 and the precision value recorded is 0.91 as shown in figure 7.

The performance of the proposed system with respect to character segmentation and recognition accuracy is examined through comparison with previous researchers. A research by [10] proposed a character recognition system using OCR which could detect three different types of fonts wherein the accuracy reached 0.92 for Californian, 0.94 for Georgia and 0.97 for Tibook antique. A comparison with similar researches in feature extraction based character detection method revealed that the proposed method operates with great precision and accuracy. A research by [11] revealed a precision value of 66 and recall value of 70. Similarly, researches by [12] [13] [14]; (Zhao et al. 2010) revealed precision and recall values of (59,55), (79,76) and (100,81). However, the present research could achieve 95 per cent accuracy and 87 per cent recall. Furthermore, the values of accuracy in comparison with other methods of image ham/ spam detection are examined which revealed that the proposed system achieved an accuracy of 0.95. Figure 8 compares the results achieved by previous researchers and the proposed system.

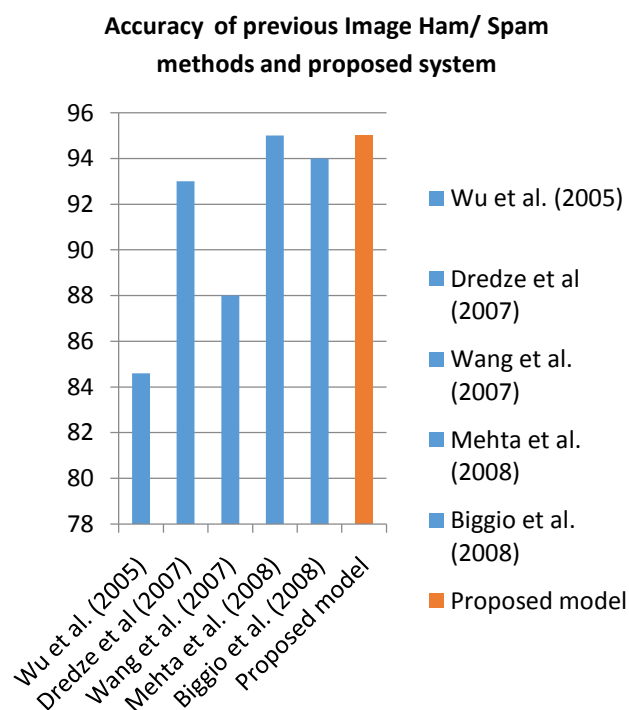


Figure 8 Comparison of other image ham spam systems’ accuracy with the proposed system

VI. CONCLUSION

The conclusion of this paper is to propose a new text and image based email filtering method. This work consists of three main phases- Character Segmentation, Character Recognition and Email Classification. The

algorithms in Character Segmentation and Character Recognition are all some common image processing algorithms with some modifications. The characters are segmented using hybrid approach combining DWT and Hough Transform. This algorithm is robust with respect to different languages, font size, style, orientation, colour and alignment of text and can be used in large variety of application fields. The Characters are recognized using combined approach of Template Matching and Contour Analysis. These methods work faster to recognize characters. The combination of several techniques together for the character segmentation, recognition and shape based feature extraction methods provided better results. This combination is indeed a novelty since no research has combined such extensive algorithms together. The proposed method has been evaluated based on performance of false positive (FP), false negative (FN), true positive (TP), true negative (TN) and accuracy. Experiments and results show that, this application yields above 90% efficiency for character segmentation, character recognition and email filtering which shows better efficiency of the proposed method. One extensive recommendation would be to propose a system which could detect websites. Though such systems are outside the scope of the present research, future researches could be conducted which will be a significant contribution to the research community.

REFERENCES

- [1] Biggio, B., Fumera, G., Pillai, I. & Roli, F. (2007). Image Spam Filtering by Content Obscuring Detection. In: *CEAS 2007 - Fourth Conference on Email and Anti-Spam*. 2007, Mountain View, California USA.
- [2] Kamboj, R. (2010). *A Rule Based Approach for Spam Detection*. Thapar University.
- [3] Mehta, B., Nangia, S., Gupta, M. & Nejdil, W. (2008). Detecting image spam using visual features and near duplicate detection. In: *WWW '08 Proceedings of the 17th international conference on World Wide Web*. [Online]. 2008, Beijing, China: ACM, pp. 497–506. Available from: <http://www.www2008.org/papers/pdf/p497-mehta.pdf>.
- [4] Radicati, S. & Hoang, Q. (2012). *Email Statistics Report*. PALO ALTO.
- [5] statista (2017). *Global spam volume as percentage of total e-mail traffic from January 2014 to September 2016, by month*. 2017. The Statistics Portal.
- [6] Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions of Pattern Analysis and Machine Intelligence*. pp. 679-698.
- [7] Zhang, C., Chen, W.-B., Chen, X. & Warner, G. (2009). Revealing common sources of image spam by unsupervised clustering with visual features. In: *Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09*. [Online]. 2009, New York, New York, USA: ACM Press, p. 891. Available from: <http://portal.acm.org/citation.cfm?doid=1529282.1529474>.
- [8] Das, M. & Prasad, V. (2014). Analysis of an Image Spam in Email Based on Content Analysis. *International Journal on Natural Language Computing*. 3 (3). p.pp. 129–140.
- [9] Dredze, M., Gevayahu, R. & Elias-Bachrach, A. (2007). Learning Fast Classifiers for Image Spa. In: *proceedings of the Conference on Email and Anti-Spam*. [Online]. 2007, CEAS. Available from: http://www.cs.jhu.edu/~mdredze/datasets/image_spam/
- [10] Singh, D., Khan, M.A., Bansal, A. & Bansal, N. (2015). An application of SVM in character recognition with chain code. In: *2015 Communication, Control and Intelligent Systems (CCIS)*. [Online]. November



2015, IEEE, pp. 167–171. Available from: <http://ieeexplore.ieee.org/document/7437901/>.

[11] Pan, Y.-F., Liu, C.-L. & Hou, X. (2010). Fast scene text localization by learning-based filtering and verification. In: *2010 IEEE International Conference on Image Processing*. [Online]. September 2010, IEEE, pp. 2269–2272. Available from: <http://ieeexplore.ieee.org/document/5651862/>.

[12] Neumann, L. & Matas, J. (2011). A Method for Text Localization and Recognition in Real-World Images. In: R. Kimmel, R. Klette, & A. Sugimoto (eds.). *Computer Vision – ACCV 2010. ACCV 2010. Lecture Notes in Computer Science*. [Online]. Berlin, Heidelberg: Springer, pp. 770–783. Available from: http://link.springer.com/10.1007/978-3-642-19318-7_60.

[13] Anthimopoulos, M., Gatos, B. & Pratikakis, I. (2010). A two-stage scheme for text detection in video images. *Image and Vision Computing*. [Online]. 28 (9). p.pp. 1413–1426. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0262885610000430>.