Detecting Image Spam Based on File Properties, Histogram and Hough Transform

Zin Mar Win and Nyein Aye

***Abstract*—Spammers are constantly creating sophisticated new weapons in their arms race with anti-spam technology, the latest of which is image-based spam. In general words, image spam is a type of email in which the text message is presented as a picture in an image file. This prevents text based spam filters from detecting and blocking such spam messages. There are several techniques available for detecting image spam (DNSBL, GrayListing, Spamtraps, etc,…). Each one has its own advantages and disadvantages. On behalf of their weakness, they become controversial to one another. This paper includes a general study on image spam detection using file properties, histogram and hough transform, which are explained in the following sections. The proposed methods are tested on a spam archive dataset and are found to be effective in identifying all types of spam images having (1) only images (2) both text and images. The goal is to automatically classify an image directly as being spam or ham. The proposed method is able to identify a large amount of malicious images while being computationally inexpensive.**

***Index Terms*—File properties, histogram, hough transform, spam archive dataset.**

I. Introduction

As the use of email for the communication is increasing, the number of unwanted 'spam' is also increasing [1]. For example, there's the occasional joke sent in mass from friend to friend and back again, or that all-important virus alert, or the occasional inspiration, etc. [2]. Spam message volumes have doubled over the past year and now account for about 80% of the total messages on the Internet. A major reason for the increased prevalence of spam is the recent emergence of image spam (i.e. Spam embedded in images). Image spam volumes nearly quadrupled in 2006, increasing from 10% to 35% of the overall volume of spam; worse, the volume of image spam continues to rise. The situation has significantly frustrated end-users as many image spam messages are able to defeat the commonly deployed anti-spam systems. In order to reduce the impact of spam, it is crucial to understand how to effectively and efficiently filter out image spam messages. Spammers have recently begun developing image-based spam methods to circumvent current anti-spam technologies since existing anti-spam methods have proved quite successful in filtering text-based spam email messages. Early image-based spam simply embedded advertising text in images that linked to HTML formatted email so that its content could be automatically displayed to end-users while being shielded from text-based spam filters. As spam filters started to use simple methods such as comparing the hashes of image data and performing optical character recognition (OCR) on images, spammers have quickly adapted their techniques. To combat computer vision techniques such as OCR, spammers have begun applying CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Human Apart) techniques. These techniques distort the original image or add colorful or noisy background so that only humans can identify the intended message [3]. Fig. 1 shows the recipient behavior of spam e-mail.

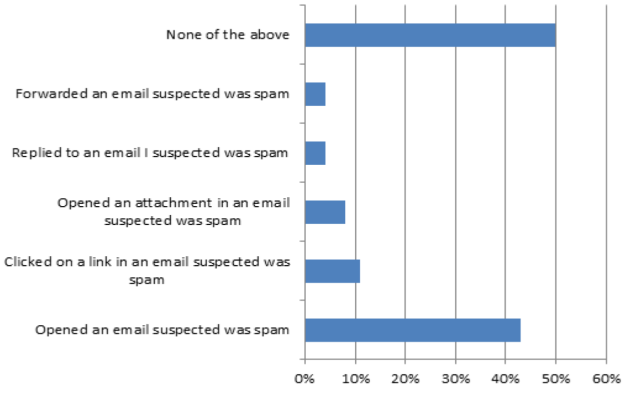


Fig. 1. Spam survey.

II. Related Works

Congfu Xu *et al.* [4], proposed approach based on Base64 encoding of image files and n-gram technique for feature extraction. It transformed normal images into Base64 presentation, and then it used n-gram technique to extract the feature. Using SVM, spam images were detected from legitimate images. This approach shows time efficient performance. Our proposed method detects spam images from normal images using SVM.

Yan Gao [5] proposed supervised detection method builds its training dataset based on two image features ie. colour and gradient orientation histograms and utilizes this data on probabilistic boosting tree (PBT) to distinguish spam images from ham images. Each node of PBT contains colour or gradient orientation histogram data of corresponding part of images inside training dataset. New incoming images are compared with PBT nodes to detect spam. In our proposed method, colour histograms are used to identify spam because it can easily differentiate normal images from spam images.

In the proposed detection method, authors in [6] postulated that spammers use the same template to send a lot of spam images and they add random noises to an image template in order to bypass filters. Authors classify random noises into 17 categories and utilize three feature vectors in order to analyze

Manuscript received December 2, 2013; revised April 26, 2014.

The authors are with University of Computer Studies, Mandalay, Vietnam (e-mail: zinmarwinn19@gmail.com).

DOI: 10.7763/JACN.2014.V2.127

them. By extracting these features from images, the system builds training dataset, compares new images with dataset and labels them as spam or ham images. Our proposed method use spam archive dataset as training dataset.

Authors on this paper [7] propose fast and robust image spam detection method for dealing with image spam in emails. They extract 9 features from images for feeding the maximum entropy model (i.e, logistic regression based on binary case) to detect spam. They also use Just in Time (JIT) feature extraction to speed up process of spam detection that dramatically reduces processing time. JIT is a feature extraction method, which only focuses and extracts features based on each image. Hough transform detection method is used in our system to reduce time.

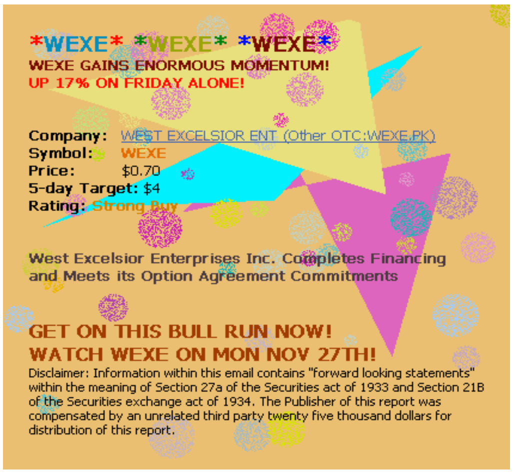
Pattarapom Klangpraphat *et al.* [8] verity image with content-based image retrieval. It also considers the partial similarity of e-mail spam from the normal e-mail. Our system uses file properties to detect spam as most legitimate images are Jpeg format and small file size.

III. Image Spam Detection

Nowadays, spammers use different image processing technologies to vary the properties of individual message e.g. by changing the foreground colours, background colours, font types or even rotating and adding artifacts to the images. Thus, they pose great challenges to conventional spam filters. To get rid of anti-spam filters in email spam, some spammers put their spam content into the images and attach these images to emails .Those anti-spam filters that analyses content of email cannot detect spam text in images.

Image spam is junk email that replaces text with images as means of fooling spam filters. If the recipient's email program downloads the image automatically, the image appears when the message is opened. The image itself may be a picture or drawing of alphanumeric characters that appears as text to the viewer, although it is processed as an image by the user's computer. The increase in more complex email spam attacks has caused spam capture rates across the email security industry to decline, resulting in wasted productivity and end-user frustration as more spam gets delivered to their inboxes. The root cause behind this sharp increase in spam volume is money. The more messages that are delivered to inboxes, the better the chances recipients take action on the messages, resulting in more income for spammers [9]. Fig. 2 shows the example of normal image and Fig. 3 shows the example of spam image.

Fig. 2. Normal image.



WEXE *WEXE* *WEXE*
WEXE GAINS ENORMOUS MOMENTUM!
UP 17% ON FRIDAY ALONE!

Company: WEST EXCELSIOR ENT (Other OTC:WEXE.PK)
Symbol: WEXE
Price: $0.70
5-day Target: $4
Rating: Strong Buy

West Excelsior Enterprises Inc. Completes Financing and Meets its Option Agreement Commitments

GET ON THIS BULL RUN NOW!
WATCH WEXE ON MON NOV 27TH!

Disclaimer: Information within this email contains "forward looking statements" within the meaning of Section 27a of the Securities act of 1933 and Section 21B of the Securities exchange act of 1934. The Publisher of this report was compensated by an unrelated third party twenty five thousand dollars for distribution of this report.

Fig. 3. Spam image.

IV. File Properties

Image spam e-mails will mostly contain images in JPEG or GIF file types. The basic features (Table I) that can be derived from an image at an extremely low computational cost are the width and the height denoted in the header of the image file, the image file type and the file size. In this study, we focus on all file formats that are commonly seen in emails, which are the Graphics Interchange Format (GIF), and the Joint Photographic Experts Group (JPEG) format, Bitmap (BMP) and Portable Network Graphic (PNG). By parsing the image headers of the image files using a minimal parse, a general idea of the image dimensions can be obtained; as this does not involve any decompression or decoding on any actual image data, the dimensions can be obtained rather faster.

In the case of GIF files there will be presence of virtual frames [10], which may be either larger or smaller than the actual image width. And this issue can be detected by decoding the image data. The problem imposed in the case of corrupted images is that the lines near to the bottom of the image will not decode properly. Any further decoding of the image data from that point of corruption will be decisive.

TABLE I: Image Features

Features	Description
f1	Image width denoted in header
f2	Image height denoted in header
f3	Aspect Ratio: f1/f2
f4	File Size
f5	File Area: f1.f2
f6	Compression: f5/f4
f7	File Format

This feature analysis reflects the percentage of images in the respective formats. Most legitimate images ("ham") in emails are JPEG images. The f3 is the aspect ratio of the image (i.e.) f1/f2. The feature f6 captures the amount of compression achieved by calculating the ratio of pixels in an image to an actual image size. The compression is better if more number of pixels is stored per byte.

V. Color Histograms

The color histogram is a simple feature and can be calculated very efficiently by one simple pass of the whole image. We have used 64-dimensional color histogram based in the RGB color space. Values in each of the three color channels (R, G, B) are divided into 4 bins of equal size, resulting in 4×4×4 =64 bins in total. For each bin, the amount of color pixels that falls into that particular bin is counted. Finally it is normalized so that the sum equals to one [11]. We use L1 distance to calculate the distance between two color histogram features. For image represented by D-dimensional real-valued feature vectors, the L1 distance of the pair of points $X=(X_1,\ldots\ldots,X_D)$ and $Y=(Y_1,\ldots\ldots\ldots,Y_D)$ has the form:

$$d(X,Y)=\sum_{i=1}^{D}|X_i-Y_i| \qquad (1)$$

We adopt color histogram in our system for its simplicity and efficiency. The color histogram is effective against randomly added noises and simple translation shift of the images. For spam randomization techniques, the color histogram is designed to handle shift size, dots, bar, frame, font type, font size, line, rotate, bits, content, fuzzy, url. We use color histograms to distinguish spam images from normal images. Color histograms of natural images tend to be continuous, while the color histograms of artificial spam images tend to have some isolated peaks. We point out that the discriminating capability of the above feature is likely to be satisfactory, since color distribution is solely dependent on the format of the image. Fig. 4 shows a sample image. The differences of colour histograms are illustrated in Fig. 5-Fig. 8 when the images are saved in different formats (jpeg, gif, png and bmp).

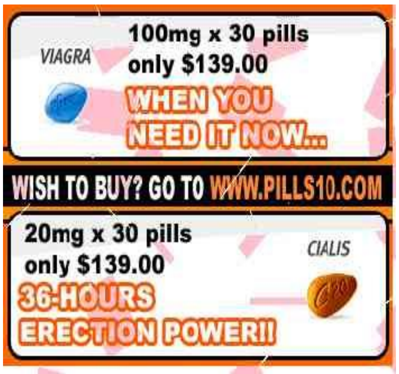


Fig. 4. Original image.

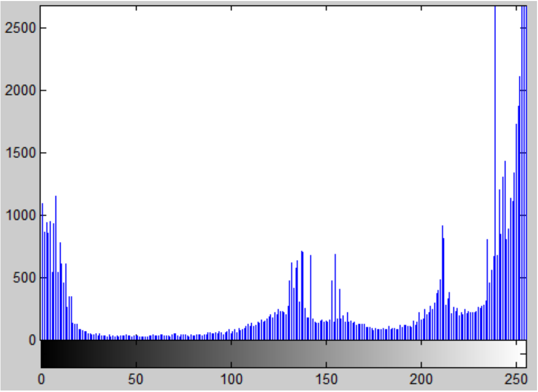


Fig. 5. Color histogram of Fig. 4 in Jpeg format.

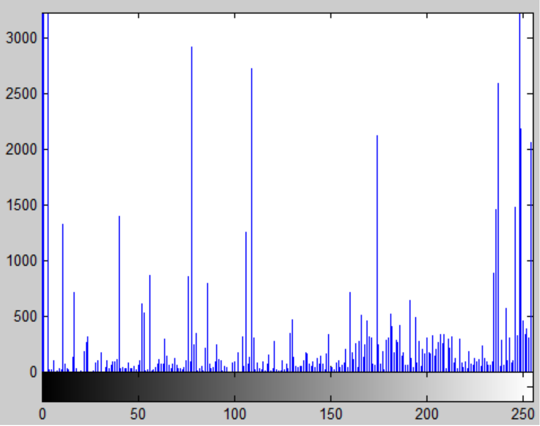


Fig. 6. Color histogram of Fig. 4 in gif format.

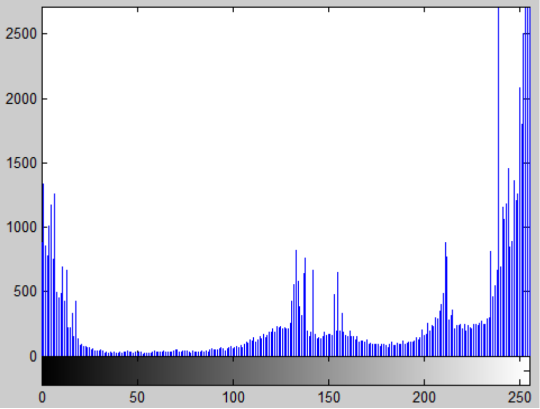


Fig. 7. Color histogram of Fig. 4 in png format.

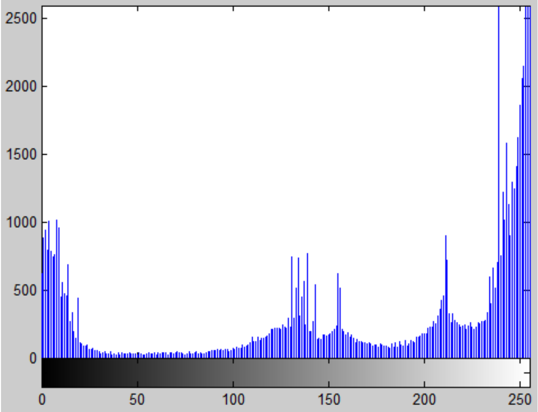


Fig. 8. Color histogram of Fig. 4 in bmp format.

VI. Hough Transform

The Hough transform is a feature extraction technique used in image analysis, computer vision and digital image processing. The purpose of the technique is to find imperfect instances of within a certain class of shapes by a voting procedure. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by an algorithm for computing the Hough transform.

The *classical* Hough transform is most commonly used for the detection of regular curves such as lines, circles, ellipses, *etc.* The Hough transform is a technique used to find shapes in a binary digital image. By using Hough Transform it is possible to find all kind of shapes that can be mathematical expressed, for instance lines, circles and ellipses, but only straight lines will be considered in this paper. If having a white pixel in a binary image, infinity many straight lines can go through that single pixel, and each of these lines can go through other white pixels in the same image, and the more white pixels on the same line the more is this line represented in the image. This is the principle of the Hough transform for straight lines. As mentioned above a shape can be found if a mathematical expression can be set for the shape, and in this case where the shape is a straight line, an expression can be set as:

$$y = a * x + b \tag{2}$$

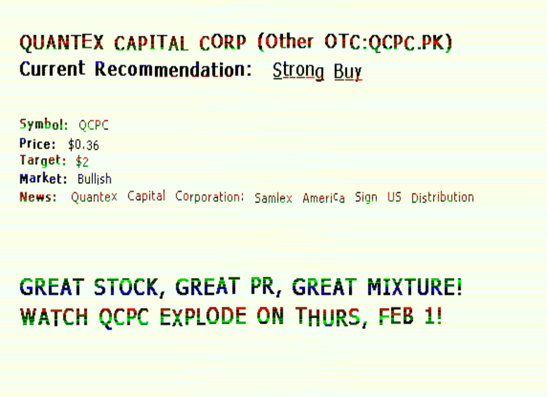
where *a* is the slope, and *b* is where the line intersects the *y*-axis. These parameters, *a* and *b*, can be used to represent a straight line as single point (*a*, *b*) in the parameter-space spanned by the two parameters *a* and *b*. The problem of representing a line as a point in the (*a*, *b*) parameter-space, is that both *a* and *b* goes toward infinity when the line becomes more and more vertical, and thereby the parameter space becomes infinity large. Therefore it is desirable to find another expression of the line with some parameters that have limited boundaries. It is done by using an angle and a distance as parameters, instead of a slope and an intersection. If the ρ (rho) is the distance from the origin to the line along a vector perpendicular to the line, and θ (theta) is the angle between the *x*-axis and the ρ vector, can be written as:

$$y = -\frac{\cos(\theta)}{\sin(\theta)} * x + \frac{\rho}{\sin(\theta)} \tag{3}$$

The expressions, instead of *a* and *b*, is found by trigonometrical calculations. To get an expression of ρ

$$\rho = x * \cos(\theta) + y * \sin(\theta) \tag{4}$$

Contrary to when the parameters is a and b, the values that ρ and θ can have are limited to: $\theta \in [0, 180]$ in degrees or $\theta \in [0,\pi]$ in radians, and $\rho \in [-D, D]$ where *D* is the diagonal of the image. A line can then be transformed into a single point in the parameter space with the parameters θ and ρ, this is also called the Hough space. If, instead of a line, having a pixel in an image with the position (*x*, *y*), infinity many lines can go through that single pixel. By using equation (4), all these lines can be transformed into the Hough space, which gives a sinusoidal curve that is unique for that pixel. Doing the same for another pixel, gives another curve that intersect the first curve in one point, in the Hough space. This point represents the line, in the image space, that goes through both pixels. This can be repeated for all the pixels on the edges, in an edge detected image. When the Hough transform is made on the image for all the white pixels (edges) the lines that have most pixels lie on can be found. The result of the Hough transform is stored in a matrix that often called an accumulator. One dimension of this matrix is the theta values (angles) and the other dimension is the rho values (distances), and each element has a value telling how many points/pixel that lie on the line with the parameters (rho, theta). So the element with the highest value tells what line that is most represented in the input image [12].



QUANTEX CAPITAL CORP (Other OTC:QCPC.PK)
Current Recommendation: Strong Buy

Symbol: QCPC
Price: $0.36
Target: $2
Market: Bullish
News: Quantex Capital Corporation: Samlex America Sign US Distribution

GREAT STOCK, GREAT PR, GREAT MIXTURE!
WATCH QCPC EXPLODE ON THURS, FEB 1!

Fig. 9. Original spam image.

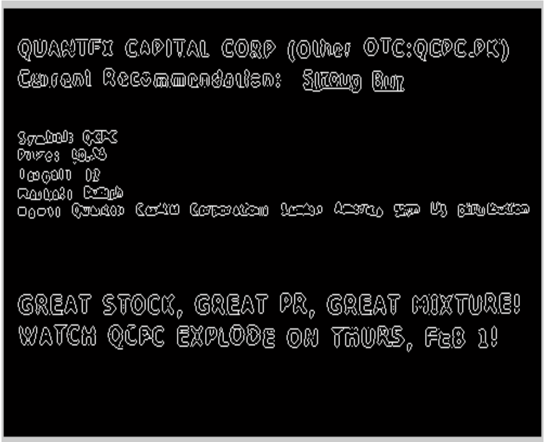


Fig. 10. Edge detection of Fig. 9.

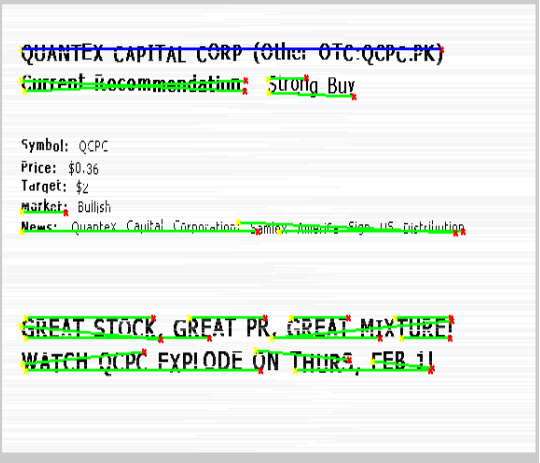


Fig. 11. Line detection of Fig. 9.

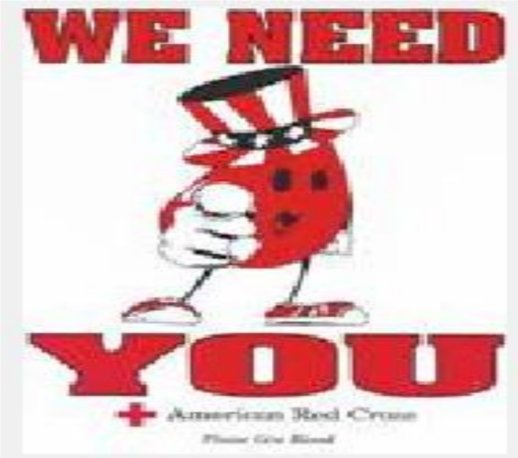


Fig. 12. Original ham image.

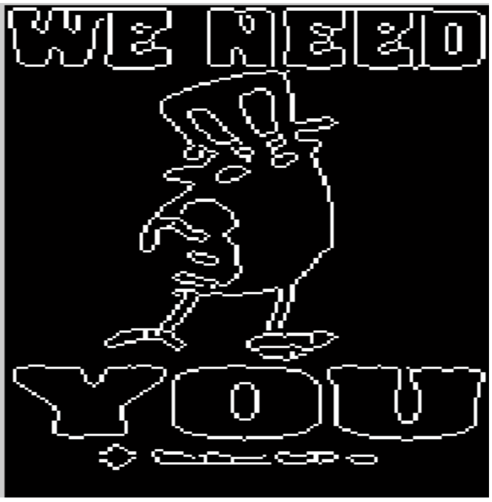


Fig. 13. Edge detection of Fig. 12.

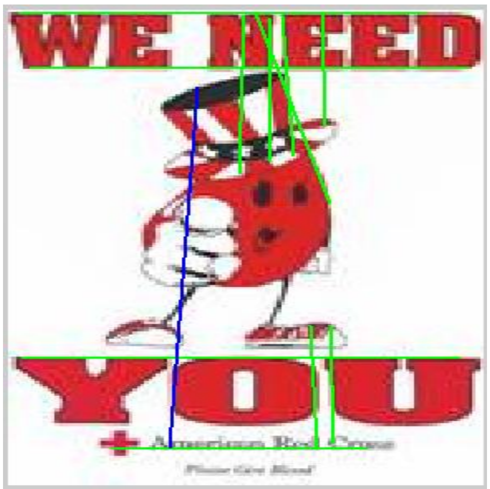


Fig. 14. Line detection of Fig. 12.

Fig. 9 is example of spam image. Fig. 10 is canny edge detection and Fig. 11 is line detection. Fig. 12 is example of ham image. Fig. 13 and Fig. 14 show edge detection and line detection respectively. According to Fig. 11 and Fig. 14, there are more horizontal lines in spam image.

VII. Experimental Results

Email corpora are difficult to construct due to the private nature of email communication. In many spam classification assessments, duplicate or highly analogous emails are included to imitate the real world nature of spam. To calculate the performance, the proposed approach used a spam archive data set [13] partly. The Spam Archive images were taken from the Spam Archive data provided by Giorgio Fumera's group. This spam archive data set contains a combination of personal image ham and personal image spam. In total, there are 5087 images combined of 3209 spam and 1878 ham images, which are JPEG, GIF, PNG and BMP images.

The aim is to develop a classifier that can distinguish legitimate from spam. The idea is to develop a method to filter spam based on image content, rather than text content. File Properties, Color histogram features and Hough transform method will be exploited. Finally the focus is to reduce the false positive rate of classifier i.e, if an image is spam, it should be detected as spam. (See Table II and Table III).

TABLE II: Performance Parameter

Classifier	Natural Images	Spam Images
Natural images	Number of images correctly classified as natural images	Number of spam images is classified as natural image
Spam images	Number of natural images misclassified as spam image	Number of images correctly classified as spam images

TABLE III: Classification Results

Approach	Accuracy(A)		Precision (P)		Recall (R)	
	Ham	Spam	Ham	Spam	Ham	Spam
File Properties	90.5%	86.6%	84.5%	80.6%	88.3%	85.7%
Color histogram	94.6%	92.1%	88.7%	84.1%	90.5%	89.6%
Hough Transform	96.5%	95.4%	90.5%	88.7%	92.0%	91.4%

VIII. Conclusion

The spam images are growing continuously. They waste the storage on the network, also consumes the bandwidth. There is need for employing efficient method for differentiating spam and natural images. In this paper, the image is detected by using file properties, color histogram and hough transform method. Detection rate depends on the type of spam images, i.e. only images or both text and images. These methods have their advantages and disadvantages. According to the experimental result, the approach using file properties eliminates only 80% of the spam images. The method using histogram implements the distance measurements. This method eliminates only 84% of the spam messages and this makes the method not suitable for most of the cases. Hough transform method utilizes the edge detection and line detection to determine spam image. This method minimizes the low false positive rate to minimum. Thus, Hough transform method provides better performance result.

Acknowledgment

The researcher thanks to the professor, Dr. Mie Mie Thet Thwin, University of Computer Studies, Mandalay for her guidance. The researcher also thanks the Associate Professor, Dr. Nyein Aye, University of Computer Studies, Mandalay in providing constant assistance towards the accomplishment of this research project.

References

[1] B. Mehta, S. Nangia, M. Gupta, and W. Nejdl, "Detecting image spam using visual features and near duplicate detection," in *Proc. the 17th International Conference on World Wide Web*, April 21-25, 2008, Beijing, China.

[2] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "Image spam filtering by content obscuring detection," presented at Fourth Conference on Email and Antispam, August 2-3, 2007.

[3] The CAPTCHA Project. (2000). [Online]. Available: http:// captcha.net

[4] C. Xu, Y. Chen, and K. Chiew, "An approach to image spam filtering based on Base64 encoding and N-gram feature extraction," 2010.

[5] G. Yan, Y. Ming, Z. X. Nan, B. Pardo, W. Ying, T. N. Pappas, and A. Choudhary, "Image spam hunter," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, Las Vegas, Nevada, U.S.A., pp. 1765-1768.

[6] Z. Wang, W. Josephson, Q. Lv, M. Charikar, and K. Li, "Filtering image spam with near-duplicate detection," in *Proc. the Fourth Conference on Email and Anti-Spam*, California, 2007.

[7] M. Dredeze, R. Gevaryahu, and A. E. Bachrach, "Learning fast classifiers for image spam," in *Proc. Fourth Conference on Email and Anti-Spam*, California, 2007.

[8] P. Klangpraphant and P. Bhattarakosol, "Detect image spam with content base information retrieval," pp. 505-509, 2010.

[9] M. Kamble and C. Dule, "Detecting image spam based on image features using maximum likelihood technique," *IJCST*, March, 2012.

[10] S. Krasser, Y. Tang, J. Gould, D. Alperovitch, and P. Judge, "Identifying image spam based on header and file properties using C4.5 decision trees and support vector machine," presented at IEEE Workshop on Information Assurance 2007.

[11] D. Q. Zhang and S. F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *Proc. the 12th Annual ACM International Conference on Multimedia*, 2004, pp. 877-884.

[12] X. Yu, H. C. Lai, and H. W. Leong, "A gridding hough transform for detecting the staright lines in the sport videos," in *Proc. IEEE International Conference on Multimedia and Expo*, 2005.

[13] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded into images," *Journal of Machine Learning Research*, vol. 7, pp. 2699-2702, 2006.

Zin Mar Win received the B.Sc. (Hons) degree from Computer University (Mandalay) in 2005 and M.Sc. degree from Computer University (Banmaw) in 2007, respectively. Now, she is a PhD candidate at University of Computer Studies, Mandalay. Her interested fields are data mining, digital image processing and information security.

Nyein Aye received the PhD degree from Engineering Physics Institute, Moscow. He currently serves as an associate professor of Computer University, Mandalay. His current research interests include pattern recognition, image processing and information security.