Identification of Image Spam by Using Histogram and Hough Transform

Zin Mar Win[1], **Nyein Aye**[2]

[1]University of Computer Studies, Mandalay, Myanmar

[2]University of Computer Studies, Mandalay, Myanmar

Abstract: ***Today, the internet is the most powerful tools throughout the world. But the explosive growth of unsolicited emails has prompted the development of numerous spam filtering techniques. It needlessly obstruct the entire system. Spammers are creating new ways against anti-spam technology. By the end of 2006, the nature of spam had totally shifted. The newest 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 histogram and hough transform, which are explaind 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.***

Keywords: histogram, hough transform, anti-spam technology, image spam detection, spam archive dataset

1. 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 friends 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 at 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 using 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]. Once spammer have applied an image creation algorithm to make a message difficult to detect with computer vision algorithm, they apply further randomization to construct a batch of images for delivery. The additional randomization is that current image spam methods present serious challenges for anti-spam systems.

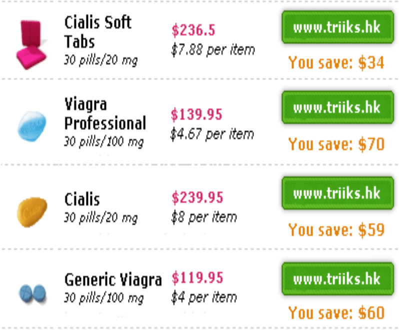
Figure 1: Spam survey

2. Image Spam Detection

Nowadays, Spammers use spammers use different image processing technologies to vary the properties of individual message e.g. by changing the foreground colours, backgrounds, 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 currently some spammers put their spam content into the images (i.e. they embed text such as advertisement text in the images) and attach these images to emails .Those anti-spam filters that analyse content of email cannot detect spam text in images [4].

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 [5].

Figure 2: Natural image



Cialis Soft Tabs 30 pills/20 mg	$236.5 $7.88 per item	www.triiks.hk You save: $34
Viagra Professional 30 pills/100 mg	$139.95 $4.67 per item	www.triiks.hk You save: $70
Cialis 30 pills/20 mg	$239.95 $8 per item	www.triiks.hk You save: $59
Generic Viagra 30 pills/100 mg	$119.95 $4 per item	www.triiks.hk You save: $60

Figure 3: Spam image

3. Related Works

Congfu Xu et. al [6], 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.

Yan Gao [7] 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 the proposed detection method, authors in [8] 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 utilized three feature vectors in order to analyse 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.

Authors on this paper [10] 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 extract features needed based on each image. Pattarapom Klangpraphat et. al [13] verity image with content-bases image retrieval. It also considers the partial similarity of e-mail spam from the normal e-mail.

4. Colour Histograms

Nowadays, spammers use different image processing technologies to vary the properties of individual message e.g. by changing the foreground colours, backgrounds, font types or even rotating and adding artifacts to the images. Thus, they pose great challenges to conventional spam filters. 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 4x4x4 =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,. . . . . . .,X_D) and Y =(Y_1,. . . . . , 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. Use colour histograms to distinguish spam images from normal images. Colour histograms of natural image tend to be continuous, while the colour histograms of artificial spam images tend to have some isolated peaks. We point out however that the discriminating capability of the above feature is likely to be satisfactory, since colour distribution is solely dependent on the format of the image. Figure 2 shows a sample image and the difference between its colour histograms when saved with different formats (jpeg, gif, png and bmp) is illustrated in Figures 5,6,7,8.

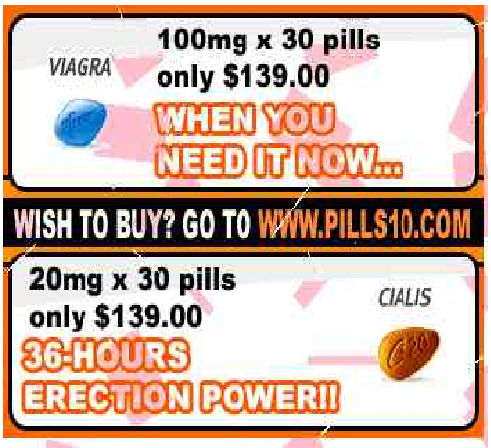


Figure 4: Original image

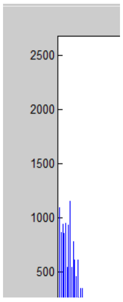


Figure 5: Color histogram of fig 4 in jpeg format

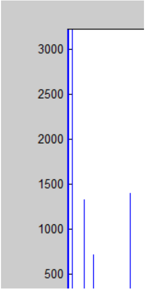


Figure 6: Color histogram of fig 4 in gif format

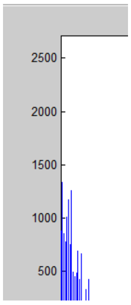


Figure 7: Color histogram of fig 4 in png format

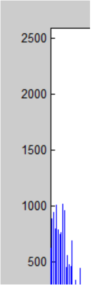


Figure 8: Color histogram of fig 4 in bmp format

5. Hough Transform

The Hough transform is a technique which can be used to isolate features of a particular shape within an image. Because it requires that the desired features be specified in some parametric form, 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 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:

$$y = a * x + b \qquad (2)$$

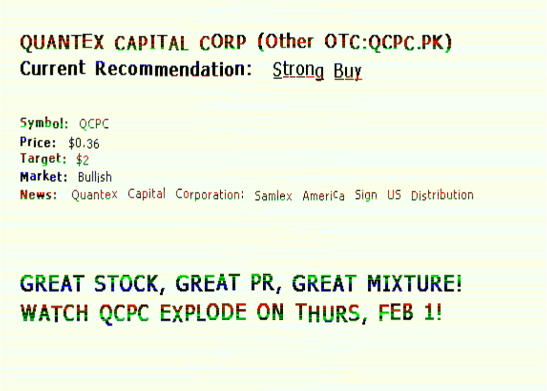
Where a is the slope, and b is where the line intersects the y-axis. These parameters, a and b, can be used to represent a straight line as single point (a, b) in the parameter-space spanned by the two parameters a and b. The problem by represent a line as a point in the (a, b) parameter-space, is that both a and b goes toward infinity when the line becomes more and more vertical, and thereby the parameterspace 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 distance ρ (rho) is the distance from the origin to the line along a vector perpendicular to the line, and the angle θ (theta) is the angle between the x-axis and the ρ vector, can be written as:

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

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

$$\rho = x * \cos(\theta) + y * \sin(\theta) \qquad (4)$$

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 $\theta \in [0, p]$ 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 3 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!

Figure 9: Original image

Figure 10: Edge detection of fig 9

Figure 11: Line detection of fig 9

6. 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 [9] partly. The Spam Archive images were taken from the Spam Archive data provided by Giorgio Fumera's group and used in this paper. This spam archive data set contains combination of personal image ham and personal image spam. In total, the images considered to this proposed work is with 5087 images combined of 3209 spam and 1878 ham images, which consist of JPEG, GIF, PNG and BMP images.

The system performance is measured in terms of accuracy. The Accuracy tells the ratio of the number of spam which are identified accurately to the total number of images in the database. The objective is to reduce the false positive rate of classifier and to classify the images correctly into the actual class. The number of correctly identified spam is termed as true positive, number of correctly identified ham denoted as true negative, number of spam images misidentified as ham is false negative and ham images misidentified as spam represents false positive. As false positives are generally considered to be more harmful than false negatives, the goal is to ensure that low false alarm rate is the first priority, while at the same time minimizing the rate of false negatives as much as possible. The system also evaluates the performance in terms of Accuracy (A), Precision (P) and Recall (R).

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. Color histogram features and hough transform for line detection 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.

Table 1: Classification Results

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

Table 2: Comparison of Accuracy, Precision and Recall

Approach	*Accuracy(A)*		*Precision (P)*		*Recall (R)*	
	Ham	Spam	Ham	Spam	Ham	Spam
Color histogram	94.60%	92.10%	88.70%	84.10%	90.50%	89.60%
Hough Transform	96.50%	95.40%	90.50%	88.70%	92.00%	91.40%

7. 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 color histogram and hough transform method. Detection rate depends on the type of spam images, i.e. only images or both text and images. Both methods have their advantages and disadvantages. According to the experimental result, the approach 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. The later, 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.

References

[1] Bhasakr Mehta, Saurabh Nangia, Manish Gupta, Wolfgang Nejdl, Detecting image spam using visual features and near duplicate detection, Proceeding of the 17th international conference on World Wide Web, April 21-25, 2008, Beijing, China [doi>10.1145/1367497.1367565]

[2] Biggio, B, Fumera, G, Pillai, I , Roli, F, 2007, Image spam Filtering by content obscuring detection, Fourth conference on email and antispam, CEAS 2007, Mountain View, California, August 2-3, 2007.

[3] The CAPTCHA Project, 2000. Http:// captcha.net

[4] M. Dredze, R. Gevaryahu, and A. E. BAchreach, “Learning Fast Classifier for Image Spam”, in Fourth conference on Email and Anti-spam (CEAS 2007) Mountain View. California, 2007

[5] Minal Kamble, Chhaya Dule, “Detecting Image Spam based on Image Features using Maximum Likelihood Technique” IJCST, March, 2012

[6] Congfu Xu, Yafang Chen, Kevin Chiew, “An Approach to Image Spam Filtering Based on Base64 Encoding and N-gram Feature Extraction”, pp. 171-177, 2010.

[7] G. Yan, Y. Ming, Z. Xiaonan, B. Pardo, W. Ying, T. N. Pappas, and A. Choudhary, “Image Spam Hunter,” in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference, Las Vegas, Nevada, U.S.A. 2008, pp. 1765-1768

[8] Z. Wang, W. Josephson, Q. Lv, M. Charikar, and K. Li, “Filtering Image Spam with Near-Duplicate Detection,” in Proceeding of the Fourth Conference on Email and Anti-Spam, CEAS, Mountain View, California, 2007

[9] G.Fumera, I, Pillai, and F. Roli, “Spam Filtering based on the Analysis of Text Information Embedded into Images,” Journal of Machine Learning Research (special issue on Machine Learning in Computer Security), vol. 7, pp.2699-2702, 2006

[10] M. Dredeze, R. Gevaryahu, and A. E. Bachrach, “Learning Fast Classifiers for Image Spam,” in Fourth Conference on Email and Anti-Spam (CEAS 2007) Mountain view, California, 2007

[11] D- Q Zhang and S-F Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In MULTIMEDIA’ 04: Proceedings of the 12th annual ACM international conference on Multimedia, pages 877-884, 2004.

[12] Xinguo Yu, Hoe Chee Lai, Sophie, Hon Wai Leong, “A Gridding hough Transform for detecting the staright lines in the sport videos,” IEEE, 2005

[13] Pattarapom Klangpraphant, Pattarsinee Bhattarakosol, (2010), “ Detect Image Spam with Content Base Information Retrieval,” pp.505-509

Author Profile

Zin Mar Win received the B. C. Sc. (Hons) degree from Computer University (Mandalay) in 2005 and M. C, 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.