Image Content-Based "Email Spam Image" Filtering

Jianyi Wang and Kazuki Katagishi

***Abstract*—With the population of Internet around the world, email has become one of the main methods of communication among people. Due to the flood of online information, great amount of the spam emails brings troubles to people. The number of technical approaches of spam filtering are increasing which are mostly based on the text spam filtering technologies. But it is not very effective for test messages imbedded into images which are developing rapidly in recent years. In this paper, we propose an approach of spam image filtering. Our approach combines the characteristics of spam images with the corner point density to detect spam images. The effectiveness of the proposed approach is experimentally evaluated.**

***Index Terms*—Spam email filtering, spam images, the corner point density.**

I. Introduction

A. The Background of the Email

With the popularization of Internet around the world, email has become one of the main methods of communication among people. However, because of the flood of online information in the recent years, a lot of people`s inbox space is plagued by spam emails. These spam emails not only waste internet users' network bandwidth and computing resources, but can also, on a much larger scale, interrupt enterprises' normal network operation.

For years, about 80% to 95% of emails on Internet are somehow composed of spams [1]. In 2012, the proportion of spam email was at its lowest level in 5 years, but it still composed 72% of all emails, among which spam image has grown to 17%. [2]

B. The Definition of Spam

The word "Spam" as applied to Email means "Unsolicited and Bulk Email". Unsolicited means that the recipient has not granted verifiable permission for the message to be received. Bulk means that the message is sent as part of a larger collection of messages, which has substantively identical content [3]. A message is a spam only if it is both Unsolicited and Bulk.

- Unsolicited Email is normal email

(Examples: first contact enquiries, job enquiries, sales enquiries)

- Bulk Email is normal email

(Examples: subscriber newsletters, customer communications, discussion lists).

However spam mail can also appear in the form of an image. Spam image is an obfuscating method in which the text of the message is stored as a GIF or JPEG image and displayed in the email [4], [5]. Fig. 1 shows us some examples of spam image. And Fig. 2 shows us an example of non-spam images.

(a) Image containing only text

(b) Image with photographic elements

Fig. 1. Examples of spam images.

Fig. 2. Example of non-spam images.

II. Existing Research

Although it is very subjective for users to judge whether a message is a spam or not. But according to the delivery address and contents of emails, it can be very beneficial for users to filter spams in advance.

Manuscript received September 12, 2013; revised December 18, 2013.

Jianyi Wang is with the University of Tsukuba Tennodai1-1-1, Tsukuba, Ibaraki 305-8577, Japan (e-mail: wjy816@gmail.com).

Kazuki Katagishi is with Academic Computing and Communications Center, Organization for Information Infrastructure, University of Tsukuba, Tennoudai 1-1-1, Tsukuba, Ibaraki 305-8577, Japan (e-mail:katagisi@cc.tsukuba.ac.jp).

DOI: 10.7763/JACN.2014.V2.92

We will make a brief introduction of existing filtering methods as follows:

- Word lists

Simple and complex lists of words that are known to be associated with spam. For example, “property sales”.

- Black lists and White lists

These lists contain known IP addresses of spam and non-spam senders, respectively.

- Hash-tables

These systems summarize emails into pseudo-unique values. Repeated sightings of hash values are symptomatic of a bulk mailing.

- Artificial Intelligence and Probabilistic systems

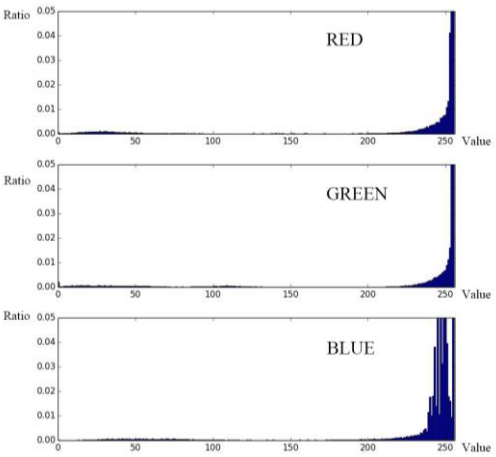
Systems such as SVM are used to identify word frequencies and patterns that usually are associated with both spam and non-spam messages.

Most of the above methods are based on the text spam filtering technologies. But they are not very effective for spam images which are growing rapidly in recent years.

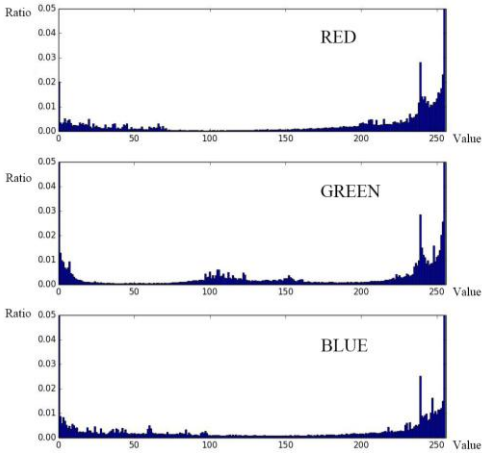
III. Experiment

A. Preparation

By collecting spam mails, corpus is established. And then we pick 999 pieces of typical spam to analyze the color characteristics. Here we used the 8-bit RGB mode to analyze. The histograms are shown in Fig. 3 as follows:



(a) The histogram of the spam image in Fig.1 (a).



(b) The histogram of the spam image in Fig.1 (b).

Fig. 3. The histograms of two spam images in Fig.1.

In order to reflect the overall characteristics of spam images in corpus, we worked out the average histogram of all. The result is Fig. 4.

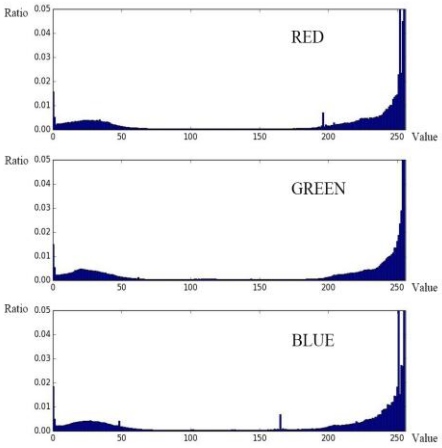


Fig. 4. The average histogram with 999 pieces of spam images.

At the same time we also have carried out contrast experiments on non-spam images. We have collected a total of 453 various images. The histogram of a non-spam image is in Fig. 5.

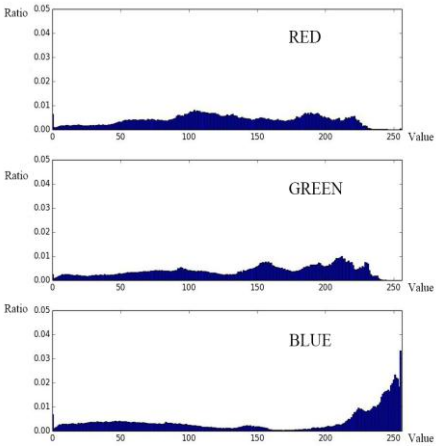


Fig. 5. The histogram of non-spam image in Fig. 2.

Fig. 6 shows us the average histogram of all non-spam images.

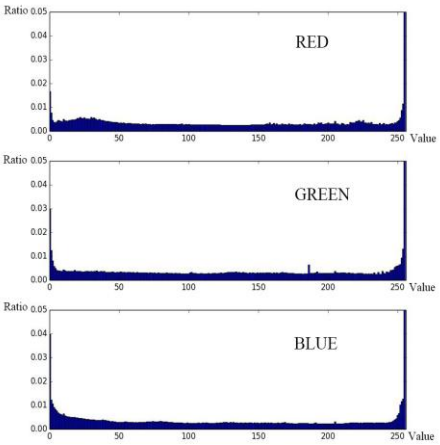


Fig. 6. The average histogram with 453 pieces of non-spam images.

By comparing Fig. 4 with Fig. 6, we find that spams have no rich color information, which is more monotonous. But non-spam images have more uniform color distribution.

After analyzing the characteristics, we did the following several steps ahead of image processing:

Step1: Color Edge Detection

In our researches, for the purpose of image preprocessing, we transformed from RGB mode to gray level and the outline of the image is obtained (in Fig. 7) [6]. Because of the sensitivity of the human`s eye to color is different, so the According to statistics, the best value is

$$Gray=0.3R+0.59G+0.11B$$

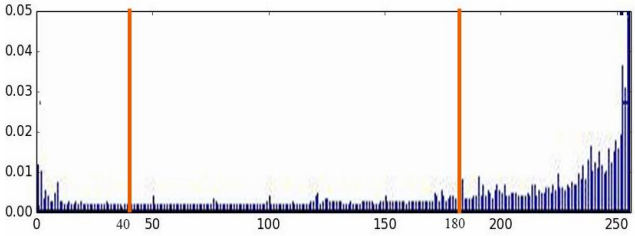


Fig. 7. The histogram of the gray level in Fig. 1. (b).

After obtaining gray level's value, through trial and error, we get the best outline effect which ranges between 40 and 180.

Step 2: Image Binarization

This study uses the simplest constant threshold segmentation. After repeatedly testing the system the threshold value is set to 120 [7], [8]. (See Fig. 8)

Fig. 8. The result of spam image by Image Binarization in Fig. 1 (b).

Step 3: Corner Point Detection

By improving gradient operator from [-2 -1 0 1 2] which is used in Harris corner detection to [5 0 -5; 8 0 -8; 5 0 -5], 8 neighborhoods is taken into account to expand the scope of the detection in the x-axis. Similarly in the y-axis operator [-2,-1, 0, 1, 2] also changed to [5 8 5; 0 0 0;-5 -8 -5]. By the maximum obtained within the operator, we can get the positions and the number of corner points. As the result of improving the operator, corner points increased from 489 to 573. (In Fig. 9 and Fig. 10)

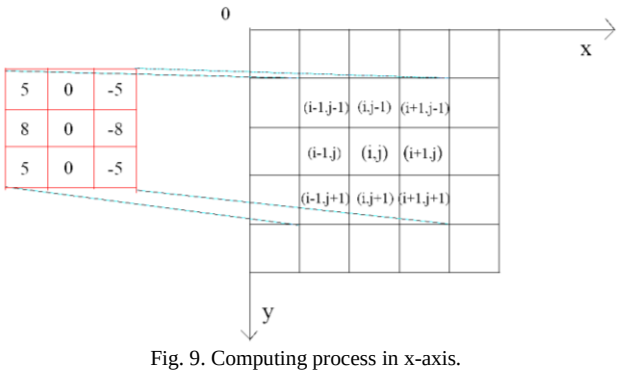


Fig. 9. Computing process in x-axis.

Fig. 10. The result of corner detection.

B. Spam Image Discrimination

The general idea of the algorithm is based on the corner proportion of the images to judge if it is a spam or not. Since the information must be expressed in words, and it`s difficult to detect out whether the words are in the images or not. Spammers disguise spam information by graphical words. In this way, the spam images should have much more graphical text than non-spam images. At the same time, words have larger number of corner points than number of curves, so we can judge if it is a text area according to the corner proportion of the area. [4] Furthermore, we can judge whether it is a spam or not.

- The marked matrix is established in accordance with the calculated corner point coordinate.

1) The whole zero matrix "mk" which has the same size as the original image is generated which the length and the width which are marked as *H* and *W* respectively.

$$mk(i,\ j)=0$$

$$(i=0,1,2...H-1;\ j=0,1,2...W-1)$$

2) In accordance with the corner point coordinate the place of the corner point is set as 1 and then the matrix represents the distribution of the corner points in original images.

$$mk(i,\ j)=\begin{cases}0 & (i,\ j)\ is\ not\ a\ corner\\ 1 & (i,\ j)\ is\ \ a\ corner\end{cases}$$

- The calculation concept is to count the total number of all pixel of 1 in "mk" in the area of model matrix "mask".

1) A model matrix "mask" which is composed of 1 will generate the length and the width which are marked as *m* and *n*. And at the same time also a marked matrix *R* is set as all values are 0.

$$mask(m,n)=1$$

$$(m=0,1,2...M-1;n=0,1,2...N-1)$$

$$R(i,\ j)=0$$

$$(i=0,1,2...H-1;\ j=0,1,2...W-1)$$

2) The pixel coincided by "mask" and "mk" should be multiplied, then comes the summation. The value is taken as the central location of "mask" matrix. Then

"mask" will be changed in pixel and the calculation will be made again to get the value of another point in *R*. The process is two-dimensional convolution. The value of the point got from *R* matrix represents that the point is taken as the center. The value of the corner points in the area corresponding to "mask" matrix area will be summed up.

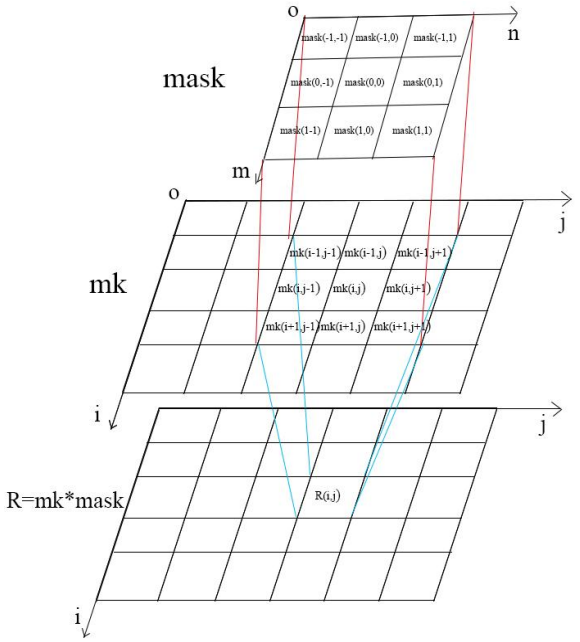


Fig. 11. Process diagram.

- When the value of a point in *R* matrix is *k*, it represents that the corner point density of the area with center point and the same size with "mask" is *k*/ (*M*×*N*). Here we use "denr" to express the density of *R*.

$$denr = R(i, j) / (M \times N) = k / (M \times N)$$

A threshold value of corner point density will be determined as "rate", when the value of "denr" is lower than the threshold value "rate", it means that the density is low and the denr here is set as 0. Through this method, the corner point area which distributes scatteredly will be removed and the values of *R* represents that the threshold value of corner point is larger than rate. (See Fig. 12)

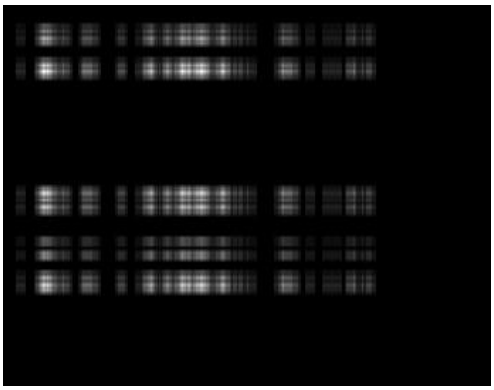


Fig. 12. The result of text corner point recognition.

- We judge a spam image by getting the ratio of non-zero elements in the matrix *R* where the value is larger than rate.

A threshold value "rat-spam" will be used to distinguish whether the image is a spam picture or not. When the ratio is larger than rat-spam, it means that there are many areas with larger corner point density and it is also correspondent to larger text area. Then it will be regarded as spam image. Otherwise, it will be regarded as normal pictures.

$$denr \begin{cases} \geq rat - spam & spam \\ < rat - spam & non - spam \end{cases}$$

According to experimental results from collected corpus, the value of the rat-spam 0.01 is given here. The value "rate" of the image in Fig. 1 (b) is 0.0368. So the image is judged as a spam image.

IV. Evaluation

We collected and extracted a corpus of spam mails with the images, and then pick up 999 pieces of typical spam images to analyze. At the same time, we also select 453 pieces of non-spam images mixed into spam images in the experiment.

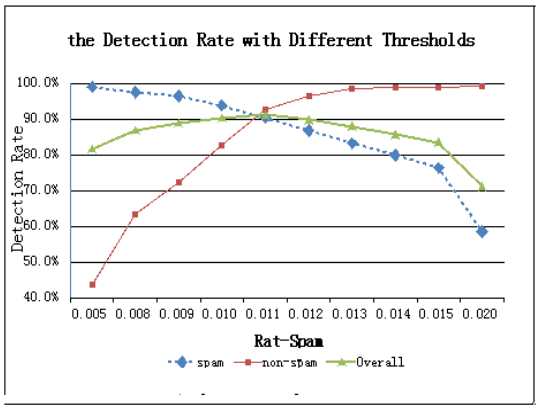


Fig. 13. The results by different thresholds.

From the results shown in Fig. 13 we find that by taking 0.01 as threshold we can obtain the best results. Based on the results, the detection rate of spam images is 904/999=90.5%. The non-spam's is 421/453=92.0%. Comprehensive detection rate is (421+904)/ (453+999) =91.3%.

V. Conclusion

On the whole, by the corner point density in images to filtering the spam images, the detection rate is relatively high. Our proposal method can be effective against spam images from spam mails. Combined with pre-existing methods which have been used into practical application, we can effectively filter spam mails.

Effectively identify the corner and conduct corner statistics is the key point in this experiment. In future researches, we strive to optimize the algorithm design of corner detection. We also hope to remove the noise around the outline of the binary image as much as possible in order to provide better image quality for corner point detection in next step.

Because spam image`s style changes fast, newer modifications of spam image are already found in the span of our research. Consequently, anti-spam technology should be improved continuously. In next step, our work will focus

on finding new algorism to cope with more crafty spams.

REFERENCES

[1] Symantec's quarterly report, (2009, September). [Online]. Available: http://www.symantec.com/zh/cn/about/news/release/article.jsp?prid=20090901_01

[2] *Chinese Anti-Spam Survey Report in the Fourth Quarter of 2012*, Internet Society of China, December 2012.

[3] The official website of The Spamhaus Project. http://www.spamhaus.org/consumer/definition/.

[4] S. Krasser *et al*. "Identifying image spam based on header and file properties using C4.5 decision trees and support vector machine learning," in *Proc the 2007 IEEE Workshop on Information Assurance*, 2007, pp. 255-261.

[5] D. G. Bachrachae, "Learning fast classifiers for image spam," presented at the 4th Conference on E-mail and Anti Spam, 2007.

[6] Y. Y. Xu and H. Yuan, "The advertisement spam image filtering method based on image content," *Journal of Shandong University (Natural Science)*, vol. 2006, no. 3, pp. 9, 2006.

[7] H. B. Aradhye, G. K. Myers, and J. A. Herson, "Image analysis for efficient categorization of image-based spam e-mail," in *Proc. Eighth International Conference on Document Analysis and Recognition*, 2005, vol. 2, pp. 914–918.

[8] D. Sagarmay and Y. C. Zhang, "An overview of content-based image retrieval techniques," in *Proc. 18th International Conference on Advanced Information Networking and Applications*, 2004, vol. 1, pp. 59-64.

Jianyi Wang had received the B. E. in computer science and technology from Tianjin University of Technology, China in 2011. Now he is pursuing his master of engineering at the University of Tsukuba, Japan. His current researches focus on network security especially spam mail filtering.

Kazuki Katagishi had received his B.E. in electronics engineering from Nagoya Institute of Technology, Japan in 1980. He had received two M.S. and Ph.D in information sciences and electronics from the University of Tsukuba, Japan in 1982, 1984, and 1987 respectively. In 1987, he joined in KDD R&D Laboratories. From 1990 to 1993, he worked in ATR Interpreting Telephony Research Laboratories and was mainly engaged in the development of speech recognition systems. From 1993 to 1999, he was engaged in the development of security mail gateway. He is now an associate professor of University of Tsukuba. His current researches include network security, network management and hyperfunctions-based signal processing and their applications to the new generation network.