An Improved Anti Spam Filter Based on Content, Low Level Features and Noise

Anand Gupta[1], Chhavi Singhal[2], and Somya Aggarwal[1]

[1] Department of Computer Engineering
[2] Department of Electronic and Communication Engineering,
Netaji Subhas Institute of Technology, New Delhi, India
{Omaranand,chhavisinghal28,somya3322}@gmail.com

Abstract. Spammers are constantly evolving new spam technologies, the latest of which is image spam. Till now research in spam image identification has been addressed by considering properties like colour, size, compressibility, entropy, content etc. However, we feel the methods of identification so evolved have certain limitations due to embedded obfuscation like complex backgrounds, compression artifacts and wide variety of fonts and formats .To overcome these limitations, we have proposed a 4-stage methodology which uses the information of low level features and content of the spam images. The method works on images with and without noise separately. Also colour properties of the images are altered so that OCR (Optical Character Recognition) can easily read the text embedded in the image. The proposed method is tested on a dataset of 1984 spam images and is found to be effective in identifying all types of spam images having (1) only text, (2) only images or (3) both text and images. The encouraging experimental results show that the technique achieves an accuracy of 92%.

Keywords: Low level feature, anti obfuscation technique, noise.

1 Introduction

Image spam is a kind of spam in e-mail where the message text of the spam is presented as an image file. Anti spam filters label an e-mail (with image attached) as spam if they find suspicious text embedded in that image. For that, the filters employ OCR that reads text embedded in images. It works by measuring the geometry in images, searching for shapes that match the shapes of letters, then translating a matched geometric shape into real text. To defeat OCR, spammers upset the geometry of letters enough—by altering colours, for example—so that OCR can't "see" a letter, even though the human eye easily recognize it. To overcome this falsity, low level features of images are extracted as they are effective against randomly added noises and simple translational shift of the images. We now review the prior significant work in the area of image spam identification.

N. Meghanathan et al. (Eds.): CCSIT 2012, Part I, LNICST 84, pp. 563–572, 2012.

2 Prior Work

Till now spam identification has been carried out by considering the following spam image properties.

1. Content (C) 2. Metadata features (M) 3. Low level features (L) 4. Text region (T)
The following matrix shows the properties as used in the previous works. Whereas the left most column shows the reference numbers, the top most row 1 shows the properties employed as given above. Numerals '1' and '0' mean that the given property is used and not used respectively.

	C	M	L	T
[1]	1	0	0	0
[2]	1	0	0	0
[3]	0	1	0	0
[4]	0	0	1	0
[5]	0	1	1	0
[6]	1	1	1	0
[7]	0	0	0	1

In [1], a scheme is proposed which implements a spam filter based on both the text in the subject and body fields of e-mails, and the text embedded into attached images. The traditional document processing steps (tokenization, indexing and classification) are improved upon in [1] by employing text extraction using OCR from attached images. SpamAssassin (SA)[2] is a widely deployed filter program that uses OCR software to pull words out of images and then uses the traditional text based methods to filter spam. This happens to be an improvement of the earlier.

In the year 2008, a spam filtering technique has been proposed in [3] that uses image information (metadata features) such as file size, area, compressibility etc., and states a characteristics that appears for each information entity.

On the contrary, [4] identifies spam using a probabilistic boosting tree based on global image features (low level features), i.e. colour and gradient orientation histograms. In the year 2010, a feature extraction scheme that concentrates on both low-level and metadata features is proposed in [5].It does not rely on extracting the text.

In [6], a mechanism is proposed to ascertain spam embedded main body e-mail file, called as Partial Image Spam Inspector (PIMSI). The significant feature of this method is that it evaluates both low level features and metadata features to confirm whether a mail is spam or not. It analyses spam images by dividing it into 2 databases. Database of object image spam consists of images and its properties such as RGB colours, contrast, brightness. In the database of Vocal Spam, all keywords of the advertised spam images are recorded and are compared with the text extracted using OCR.

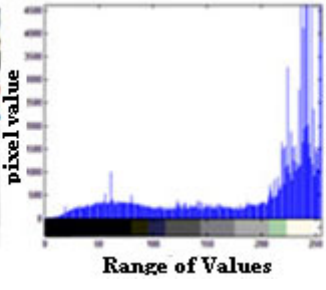
To overcome the shortcomings of the methods mentioned above, a spam identification model has been proposed in [7] which does not exploit low level features and OCR to extract text from images. Instead, it identifies spams by using the visual-BOW (VBOW) based duplicate image detection and statistical language model. Computation-efficient edge-detection method is used to locate possible text regions, and then text coverage rate in an image is calculated. Text region in a large majority of normal image is less than 15%, while text region in most spam is larger than such a threshold.

2.1 Motivation

The following drawbacks in prior related works have motivated us to develop a method that mitigates them.

Nowadays, spammers use different image processing technologies to vary the properties of individual messages 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.

[4][6][5][8] use colour histograms to distinguish spam images from normal images. Colour histograms of natural images 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 not likely to be satisfactory, since colour distribution is solely dependent on the format of the image. Figures 2 and 3 illustrate the difference between the colour histograms of a single image shown in Figure 1 but saved with different formats (gif and jpeg).



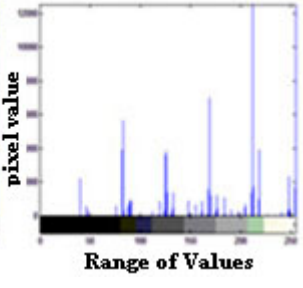


Fig. 1. Image **Fig. 2.** Colour histogram in jpeg **Fig. 3.** Colour histogram in gif

[3][5][6] use metadata features. Different images can have similar (even same) metadata features. This technology of image spam detection may be wrong and has low accuracy rate. After carrying out experiments we have found that 58.65% of spam images and 44.28% of normal images are smaller than 10KB. It implies that metadata features can be similar for both kind of images. Hence it is not a reliable method to distinguish between spam images from normal images.

The method mentioned in [7] has helped achieve significant results in identifying spam images which contain only text. However, few spam images contain both text and images. Figures 4 and 5 show that edge detection is not able to distinguish between text and images. Also edge detection will detect noise and treat it as text.

Fig. 4. Original Image **Fig. 5.** Edge Detection of Figure 4

3 System Architecture

Figure 6 depicts the System Architecture of the proposed approach.

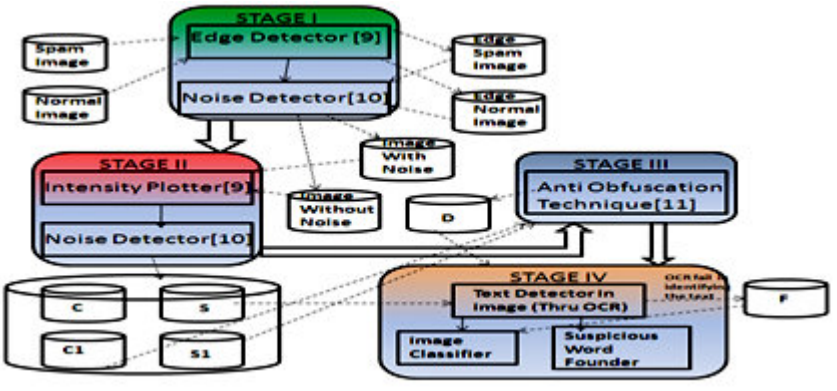


Fig. 6. System Architecture

3.1 Stage I (Identification of Noise)

Canny's edge detection [9] is used to identify noise in images. Images with noise are stored in set A and images without noise are stored in set B. Edge detection highlights even the slightest of noise added in an image. Images in Set A are more likely to be spam images. Set A is further classified into databases, namely -(A1) Dots & Dashes (A2) Lines. This classification is done on the basis of type of noise usually found in spam images. Figures 8 and 10 show a sample image for each kind of noise found in Figures 7 and 9 respectively. They also display the difficulty to identify noise without edge detection.

Fig. 7. Sample Image 1 **Fig. 8.** Image 1 noise **Fig. 9.** Sample Image 2 **Fig. 10.** Image 2 noise

3.2 Stage II (Extraction of Low Level Features)

In this stage, all the input images pass through Intensity Plotter [9], which plots the variation of intensity along a line segment or a multiline path of an image. Since spam images are artificially generated, we expect their low level features to be different

from those of images typically included as attachments to personal e-mails. The plots thus obtained do not change on varying the format of the image (from gif to jpeg or vice versa). Stage II has two sub stages.

3.2.1 Sub-Stage II (a)

Images without noise are classified into two sets, S and C. The classification is based on the difference in the shape of the plots obtained. Figures 11 and 12 show the intensity plots of normal and spam images respectively. Herein X and Y are two element vectors specifying X and Y data of the image. Images common to both Set B and Set C are labelled as normal images and are not processed further. Images in set S are directly passed to stage IV.

3.3 Sub-Stage II (b)

Images with noise are classified (as mentioned above) on the basis of plots obtained in two sets, S1 and C1. Images in set S1 and C1 are passed to stage III.

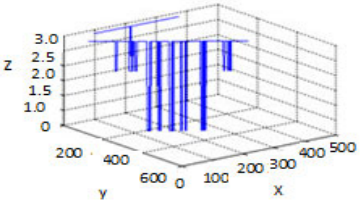


Fig. 11. Intensity plot for normal images

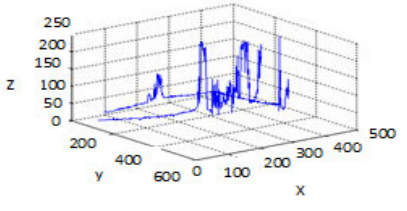


Fig. 12. Intensity plot for spam images

3.4 Stage III (Removal of Noise)

Spammers add noise in images so that it becomes difficult for OCR to read the embedded text. To overcome this difficulty, obfuscation techniques are applied. Therefore, only images with noise are passed through this stage. In this technique we alter the RGB properties of the images. Images thus obtained are stored in set D. Table 1 shows a comparison between the words identified by OCR before and after applying stage III on Figure 13.

Fig. 13. Original Image

Table 1. Comparison between the words identified by OCR before and after stage III

Text identified in Fig. 13 before applying anti obfuscation technique	Text identified in Figure 13 after applying anti-obfuscation technique
DEMDEAG ERANDE ING ClTC:DMGB,PI(—~,I,._- Teda•,r's Breaking news sent shares up +122% in just a few minutes, revelutienaryr new pre-duet in eenstruetien. r___..F-—— u_\ Huge PR campaign is under wa•,r, eemhined with teda•,r's news there's ne telling where this stuck is geing te end up, ——	DEMOBAG BRANDS INC OTC:DMGB.PK Today's Breaking news sent shares up +122% in Just a few minutes, revolutionary new product in construction. Huge PR campaign is under way, combined with today's news there's no telling where this stock is going to end up,

3.5 Stage IV (Content Extraction Using OCR)

Input to this stage comprises of images in set S and D. Images are passed through OCR, which identifies embedded text and compares it with a list of keywords. If text identified matches with the list of spam words, then the image is labelled as spam image else it is a normal image. If OCR fails, then the graphs plotted in stage II are considered. Images in sets S1 and S are labelled as spam images. Images in set C1 are labelled as normal images.

4 Experiments and Results

We have collected two sets of images to test the filter: spam images and normal images. The dataset consist of spam images in gif and jpeg format. A set of 1984 images are taken out of which 802 are normal images and 1182 are spam images from [12]. Experiments are performed on a system with the following specifications: 32- bit operating system, 2.40 Ghz processor and 4 Gb RAM. Matlab version 7.7 is used. It employs image processing techniques like Canny's edge detection and intensity plotter. For applying anti obfuscation techniques, we make use of an online editor [11]. Free OCR V3 is used to extract the content of the spam image. The experiment has shown that our technique is effective in identifying spam images in any format. According to the experimental results, detection rate of the new system is 0.92, false positive rate is 0.0064 and false negative rate is 0.059 by calculation. The following are the results.

4.1 Result I

Figures 14 and 15 show the efficiency of spam and normal images which were correctly identified by using our methodology. We have tabulated the results by classifying hams and spam images into the following categories.

Spam Images: Advertisement, URL (Uniform Resource Locator), Pornography, Stock.

Normal Images: Only Text, Text and Images, Only Images.

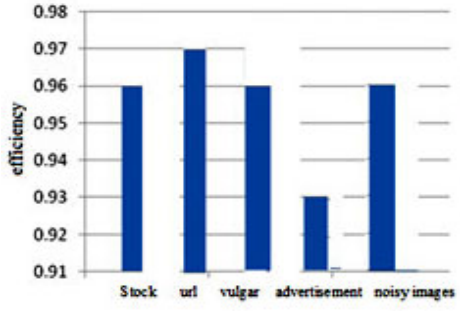


Fig. 14. Efficiency of spam images

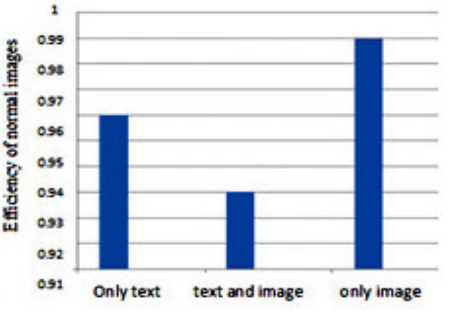


Fig. 15. Efficiency of normal images

4.1.1 Discussion

Hams with only images are easiest to identify because the range of colour components used is quite vast. Their intensity plots are curved and continuous. Their plots are easily distinguishable from those of spam images. Hams with only text consist mostly of survey forms, documents and newspaper articles. Due to the use of limited colours (mostly black and white) intensity plots of these images may be similar to that of spam images. Hence their efficiency is less than that of hams with only images. However hams with only text have efficiency higher than that of hams with both text and images because the former can be easily read by OCR .The presence of colourful images in the background of the latter makes it difficult for OCR to read the embedded text. Spam images concerning Advertisements are made attractive by adding colourful images and text so that they look like real advertisements. Hence, they are often confused with hams with both text and images and are toughest to identify.

4.2 Result II

We have taken five samples of a single image and increased its brightness from 20% to 80%. Table 2 shows that OCR depends on the brightness of images. Tick (✓) shows images that OCR could read, cross (×) shows images that OCR could not read and P shows images that OCR could partially read.

Table 2. Effectiveness of OCR on varying brightness of images

Type of images	20%	35%	50%	65%	80%
URL Spam	×	✓	✓	✓	×
Stock Spam	×	✓	✓	P	×
Vulgar Spam	×	✓	✓	P	×
Advertisement Spam	×	✓	P	P	×
Noise Spam	×	✓	✓	✓	×

4.2.1 Discussion

OCR reads successfully the embedded text in an image if its brightness is confined to a particular range. However, it fails if brightness of an image deviates from a particular threshold having both upper and lower limit.

4.3 Result III

We have varied the background colour of an image keeping its font colour constant. We have set the font colour at [R:255 G:0 B:0] and decreased the background colour from [R:255 G:240 B:240] to [R:255 G:0 B:0] in sets of 30 keeping R constant. Figure 16 shows that the detection rate of OCR decreases as background colour approaches font colour. Black dots ( ) show that OCR has failed. Red dots ( ) show that OCR has been successful. The largest circle is [R:255 G:0 B:0].

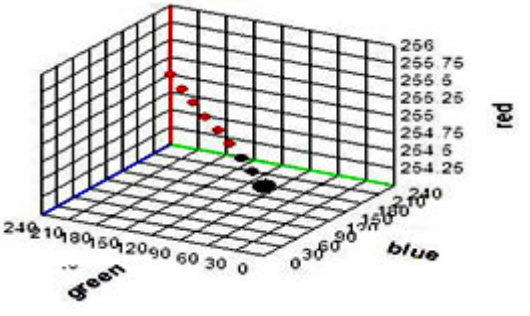


Fig. 16. Detection rate of OCR decreases as background colour approaches font colour

4.3.1 Discussion

OCR reads successfully the embedded text in an image if there exists a dissimilarity between font colour and background colour. As font colour approaches background colour it becomes difficult for OCR to read text.

4.4 Result IV

We have artificially added 'salt & pepper' noise in spam image, and have gradually increased the amount of noise added from 0.005 to 0.020. On increasing noise, distortion in the graphs obtained by intensity plotter also increases.

4.4.1. Discussion

Intensity plot of artificial spam images tends to be straight. Intensity plotter gives correct plot for spam images till a particular amount of added noise, above that it gives a distorted plot for spam images.

5 Conclusion and Future Work

In this paper we present a method for addressing image spam problems. It takes into account some of the recent evolutions of the spammers tricks in which obfuscation techniques are used to the extent that standard OCR tools become nearly ineffective. The paper proposes a method for identifying spam images by exploiting low level features and content of an image. Also anti-obfuscation techniques are applied to improve text filtering performance of the system.

According to the experimental results, we have concluded that the detection rate depends on the type of spam images, i.e. whether it contains only text, text and images or images only. Spammers add noise and changes brightness of an image till an extent only because then it hampers the clarity of the images and the user is unable to read the text.

Therefore, we aim to extend the proposed methodology by implementing a more efficient algorithm for Intensity Plotter so that it gives appropriate results for images with low signal to noise ratio.

References

1. Fumera, G., Pillai, I., Roli, F.: Spam Filtering Based On The Analysis Of Text Information Embedded Into Image. Journal of Machine Learning Research (JMLR) 7, 2699–2720 (2006)
2. Apache.org. The apache spamassassin project (2011), `http://spamassassin.apache.org/index.html` (last accessed May 3, 2011)
3. Uemura, M., Tabata, T.: Design and Evaluation of a Bayesian-filter-based Image Spam Filtering Method. In: Proceedings of the 2nd International Conference on Information Security and Assurance (ISA 2008), Busan, Korea, April 24-26, pp. 46–51 (2008)
4. Yan, G., Ming, Y., Xiaonan, Z., Pardo, B., Ying, W., Pappas, T.N., Choudhary, A.: Image Spam Hunter. In: Proceeding of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, Nevada, USA, March 30-April 4, pp. 1765–1768 (2008)
5. Wang, C., Zhang, F., Li, F., Liu, Q.: Image Spam Classification based on Low Level Image Features. In: Proceeding of the 8th International Conference on Communications, Circuits and Systems (ICCCAS 2010), Chengdu China, July 28-30, pp. 290–293 (2010)
6. Klangpraphant, P., Bhattarakosol, P.: PIMSI: A Partial Image Spam Inspector. In: Proceeding of the 5th International Conference on Future Information Technology (FutureTech), Busan, South Korea, May 21-23, pp. 1–6 (2010)
7. Hsia, J.H., Chen, M.S.: Language-Model-based Detection Cascade for Efficient Classification of Image-based Spam e-mail. In: Proceeding of the International Conference on Multimedia and Expo (ICME 2009), New York, USA, June 28-July 3, pp. 1182–1185 (2009)
8. Soranamageswari, M., Meena, C.: Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks. In: Proceeding of the 2nd International Conference on Machine Learning and Computing (ICMLC 2010), Bangalore, India, February 9-11, pp. 101–105 (2010)

9. Mathworks The Matlab image processing toolbox.M, http://www.mathworks.com/access/helpdesk/help/toolbox/images/ (downloaded on July 10)
10. Bag of Visual words Model: Recognizing Object Categories, http://www.robots.ox.ac.uk/~az/icvss08_az_bow.pdf
11. Image editor, http://www.lunapic.com/editor/?action=contrast (downloaded on July 10, 2011)
12. Image spam dataset, http://www.cs.jhu.edu/~mdredze/datasets/image_spam/ (downloaded on June 3, 2011)