VOL. 3, NO. 4, April 2012 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

A Sobel Edge Detection Algorithm Based System for Analyzing and Classifying Image Based Spam

N. C. Woods, O.B. Longe, A.B.C. Roberts

Department of Computer Science, University of Ibadan

Ibadan, Nigeria

Chyn_woods@yahoo.com

ABSTRACT

Early spam mails were only text-based, however spammers have moved to more sophisticated spamming techniques that involve images now generally termed image based spam. In most image-based spam, the entire spam message, which could be sometimes text, is embedded in an image of any format. This type of spam emails creates another dimension to the spam filtering problem scenario. Extracting text from the image and filtering these text components is one method that has been used to deal with image spam with little success because Spammers modify their approaches to beat such filters even when such filters are based on Optical Character Recognition. In this work, we used employ the Sobel edge detection algorithm, which analyses a low level feature of an image as an alternative to the OCR only based filtering system. The low level feature resultant from the filtering activity is then used to calculate the global magnitude of the edge which aides in classifying the image as either spam or ham. Our system named WiSpaf can analyse images as well as photographic images and be able to tell them apart.

Keywords: *Image Spam, Edge detection, Image analysis, low level features*

1. INTRODUCTION

The availability of the Internet and its ease of access all over the world has brought with it, not only easy access to useful information, communication, but also the increase in the amount of useless information, especially via the electronic mailbox (email) of its users. The unwanted information sent via emails are generally called Spam. There is generally no single perfect definition for Spam, because there is a rough consensus as to what spam really is. This is partly because what one user considers as spam, another may consider as ham. However users of email immediately recognize it as soon as it gets into their mailboxes. Spam comes in a variety of ways and format. Some are to advertise / sell products while others aim at transmitting viruses to the computer system of the recipient. What is common to all the various definitions of spam is that they are unsolicited emails and the intention of the spammer is to flood people's computer with information most of which are useless to the recipient. Spam also known as Unsolicited Bulk E-mail (UBE) or Unsolicited Commercial Email (UCE) has been observed to be on the increase and so is the sophistication involved in generating the spam messages. Spamming is a frustrating aspect of technology because of the increase in Internet use/ availability, where once the spammers get hold of any email address, send unsolicited emails to that mailbox at an alarming level. Initially, most spam mails were text-based. Sometimes, disguising text spam usually involves misspelling key words or randomly inserting unrelated phrases to throw off spam filters [4]. However, since there are lots of ways to detect text based spam emails, usually from the mailbox server, the spammers have moved to higher or more sophisticated spam techniques that involve pictures or images known as image based spam. Some of these images are actually texts converted to picture in an attempt to beat the mail server text spam filter.

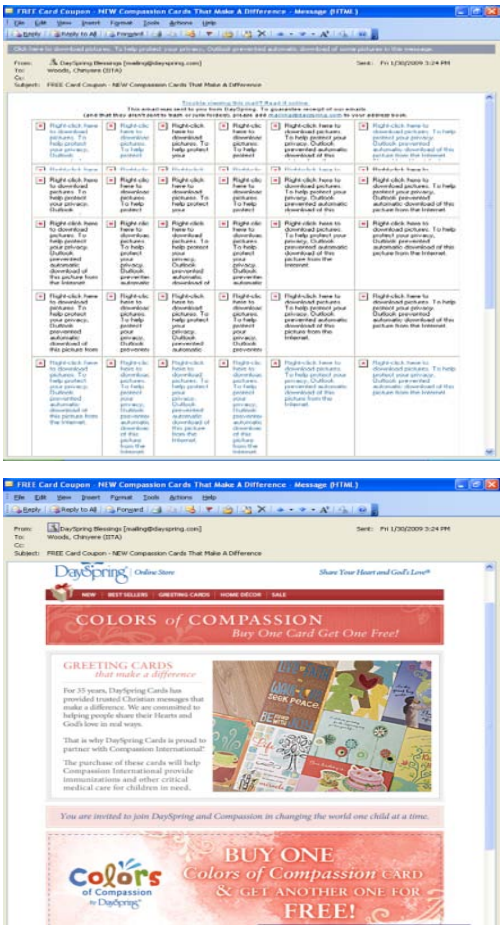

Figure 1: A spam message before and after the pictures are downloaded

In some image-based spam, the entire spam message is carried as an embedded JPEG or GIF image with minimum amount of text. While for others, the image that appears as one are actually a set of images arranged side by side to give the impression that it is just one image. When the image is fully downloaded, the viewer then gets to see the actual content of the image part of the message. For people that have 'automatically turn of pictures' set in their mailboxes these kind of spam messages are easy to spot. Figure 1 shows an example of what such a spam message looks like before and after downloading the pictures.

Sometimes, the spammers generate the spam images quickly by varying the image slightly each time. Image spam is so annoying because of the fact that spammers have learned to represent words in an image that are recognizable to the human eye but not to the computer via any OCR software. Spammers can also circumvent fingerprinting by changing some few pixels in the image. This way it will appear as a different image and may not be captured by the filter. Because of the increase in image-spam, companies have to bear the expenses with respect to bandwidth and computing power for mail server. Individual's mailboxes are also filled with so many of these unsolicited email that sometimes one misses out the wanted emails. Although anti-spam companies are trying various techniques to filter the image based spam, this type of spam still poses a serious problem to all email users and their servers.

2. ANTISPAM TECHNIQUES

Various techniques have been used as weapons by different anti-spam companies to try and trap as well as filter spam e-mails. These techniques/ methods can be broadly classified into Content based filter, origin based filter and rule based filter depending on the part of the e-mail message that is examined to determine whether an e-mail is spam or ham. Origin or address based filters generally examine the e-mail senders address making use of network information for spam classification these include Whitelisting and Blacklisting. While content based filters as the name implies, examine the actual content of the e-mail message. Optical Character Recognition (OCR) techniques, Keyword-based, Bayesian and Fingerprint methods are some that fall under this category. Each technique has its advantages and disadvantages as well as their level of successes. In some cases, two methods are combined to make the filter more effective. However as a new technique is discovered to filter the unsolicited massages, the spammers also modify their approach. For instance, to combat computer vision techniques such as Optical Character Recognition (OCR), spammers began applying CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) techniques. These techniques distort the original image or add colourful or noisy backgrounds so that only humans can identify the intended message while making it almost impossible for the computer. Once spammers have applied an image creation algorithm to make a message difficult to detect with computer vision algorithms, they apply further randomization to construct a batch of images for delivery. The additional randomization defeats fingerprinting detection mechanisms. Another way that spammers try to evade the filter is to fake the senders address. This is possible because the SMTP standard has not defined authentication of the sender's e-mail address. SMTP only authenticates the receivers email address. So it is possible to receive an email where your email address appears as the sender's address. Despite the efforts of anti-spam companies, e-mail spam is still on the increase and still manages to pass through anti-spam filters.

3. EFFORTS AT FIGHTING SPAM

One of the biggest challenges with fighting spam is that spammers become more creative as spam-fighting tools evolve. Various companies have employed different ways to fight spam. The most common way to tackle spam is at the receiving end. That is to say that most anti-spam filters work on inbound email messages to filter them. However, some researchers have proposed outbound spam filtering [4], usually to protect any company's image because if it is known that spammers operate from the company's domain, it's network address could be blacklisted.

Research has shown that content-based-filter (CBF) appears to be the best approach to filter image based spam. In this approach, the actual content of the image is analyzed and the resultant information used for classification. One of the first CBF is Optical Character Recognition, OCR. OCR attempts to recognise and convert the text within the suspect image to meaningful words and then filter these words using the traditional Bayesian and Heuristics methodologies.

[5, 6] proposed to carry out the semantic analysis of text embedded into images using text categorisation techniques like the ones applied to the body of the e-mail. One thing to note in their approach is that since text extracted from attached images via OCR may contain noise, they are not used in generating the vocabulary but at the indexing level where tokens from both the header and the extracted images are combined. At the classifier module, they used Support Vector Machines as the text classifier after training the machines. The method based on text extraction from images and analysis works well under stable conditions like plain black text on a white background, that way it is easy for OCR to recognize the text. However, once there is a little distortion as illustrated in figures 2a & 2b, then the OCR algorithm becomes confused.

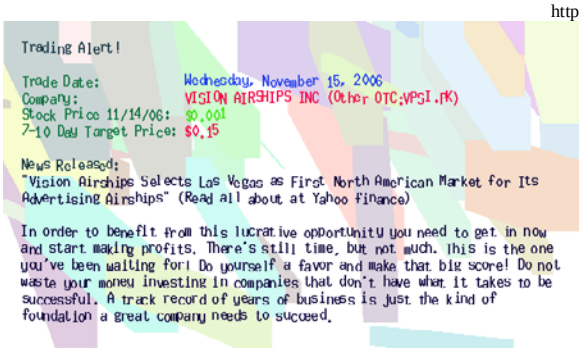

Figure 2a

Figure 2b: Randomized images designed to circumvent OCR

Another approach by [6, 9], stated that image spam filtering is a pattern recognition task in adversarial environment. They showed by experiments that filtering of adversarial obscured images can be an extremely difficult task, if spammers' actions for evading classifiers are not taken into account explicitly. Their results showed that spammers can evade OCR tools quite easily using obscured text images without compromising human readability. They then proposed an approach to filter obscured spam images based on the detection of obfuscated text, namely, an approach which takes into account explicitly the adversarial environment.

The authors in [2] also proposed an approach based on low-level image processing techniques to detect one of the main characteristics of most image spam, namely the use of content obscuring techniques to defeat OCR tools. Some of these obscuring techniques include; adding background noise interfering with text, distorting text lines or single characters, methods developed for building CAPTCHAs. Their approach measured three functions for the calculation of how obscure the text in the image is with one function aimed at detecting the presence of large background noise components overlapping with characters. This occurs when text is placed over non-uniform background. To achieve this, they extracted the edges from the original image using the Canny edge algorithm, and computed the average number of edge pixels which lie inside each component of the binarized image.

Another approach is to improve the filtering of image based spam by using image text features [3]. This approach was based on using image classifiers aimed at discriminating between ham and spam images. The researchers argued that though OCR modules could be used to filter image based spam, the OCR approach requires high processing time and can only be effective for clean images. For this reason, spammers often obfuscate the text embedded into images. Thus, a spam filter equipped with an OCR based module as the unique countermeasure against image spam is vulnerable to image spam with obfuscated text. Their approach was based on the idea of detecting the presence of obfuscation techniques into an image containing embedded text, which could be considered as an evidence of 'spamminess' of the email to which the image is attached. They developed some measures aimed at detecting and quantifying the amount of image text defects which are typical consequences of known obfuscation techniques used by spammers, like the presence of small fragments around characters (due for instance to characters broken by random lines of the same colour as the background, to characters filled with different colours, or to small background components, like random dots, around characters); the presence of large fragments around characters (due for instance to characters interfering each other, or interfering with noise components like random segments of the same colour as the text); large background shapes overlapping with characters (due to placing text over non-uniform background).

Some authors tried to improve on fingerprinting method as well as try to counter the gimmick by spammers of altering just a few pixels of an image, by filtering image spam with Near-Duplicate detection [10]. They observed that image spam emails were often sent in large batches that consisted of visually similar images that differ only in a few pixels, due to the application of randomization algorithms. They also noted that traditional spam detection methods such as honeypots, message header analysis or human reporting mechanisms can detect some image spam. Their own basic idea was to use traditional anti-spam methods to detect some image-based spam messages and then use fast near-duplicate detection filters to detect the variations of known spam images. Rather than studying the image itself to determine whether the particular image is a spam image or not, their system uses very efficient near duplicate detection techniques to find spam images that are variations of other spam images caught by traditional anti-spam methods.

Another set of researchers took a different approach by combining the methods of the above two researchers [7]. Their approach detected image spam using visual features as well as near duplicate detection. Their near duplicate detection algorithm is based on the intuition that they could recognize a lot of images similar to an identified spam image; since image based spam is usually generated from a template, near duplicates should be easy to detect. They also employed the use of low level global visual features of images like texture, shape, edge and colour as well as learn classifiers using these selected

features. From their approach, they felt it was easy to extract the features of image spam because they had some standard properties:

1. They often contain text messages conveying the intent of the spammer.
2. The images differ from natural images because natural images tend to have smoother distribution in RGB colour-space than image based spam.
3. They are usually noisy and different from one another because spammers use algorithms to generate them in bulk.
4. They are usually template based, making them easily near-duplicates of one another.

Their approach showed that low level global visual features of images are highly indicative of spam images.

Some other authors [8] proposed the approach that utilizes the extraction of a low level feature of images, the edge and then using Support Vector Machines (SVM) combined with vector representation of images to distinguish between spam and ham image messages. Using simple edge-based features, the method computes a vector of similarity scores between an image and a set of templates.

4. IMAGE FEATURES

All file types including images have features and properties that can be used in analyzing and classifying them. These file features and properties can be classified as either global or local and at the same time as high level or low level features. Being global signifies that that feature pertains to the file as a whole and not just a part of the file. It is common knowledge that high level global features of any file type is much easier to extract than low level global features. Some examples of high level properties of a file, which can be easily extracted are it's size usually measured in kilobytes, date of creation/ modification and its type which tells the computer system what application to use in opening that file. The low level features on the other hand, have to do with the actual content of the file and these features are more difficult to extract. In the case of images which are our main concern for this work, the high level global features include file format, size while the low level global visual features include the colour distribution, texture, shape, edge and the area of text regions (if any). These features could represents a property of the entire image or part of the image. Feature extraction plays a significant role in computer vision (e.g. in the area of CAPTCHA) where it aids in the analysis of images for security purposes. Low level feature extraction in images deals with algorithms that analyze the whole image, pixel by pixel and bringing out or marking any useful information depending on the feature being extracted. The next section briefly explains the edge feature and its extraction.

5. VISUAL TEXTURE FEATURES

Image texture, which is defined as a function of the spatial variation in pixel intensities, is useful in a variety of applications. The intuition behind choosing texture features for classification is that natural images have different quality of texture as compared to textures in computer generated images where most spam images fall into. The features that fall under texture are autocorrelation, edges and *primitive length* which is a continuous set of maximum number of pixels in the same direction that have the same grey level. The visual feature, edge, of an image is a typical example of one feature that can differentiate between spam and ham images. An edge is a property attached to an individual pixel of an image and is calculated from the image function behaviour in the neighbourhood of that pixel. These edges are areas with strong intensity contrasts, which is a jump in intensity from one pixel to the next or are pixels where brightness changes abruptly. Edges can be extracted by simply calculating the difference D(C1,C2) between the RGB colour of the pixel being studied C1= (R1,G1,B1) and the next neighbour C2 = (R2,B2,G2), using the equation:

$$D(C1,C2) = \sqrt{(R1-R2)^2 + (G1-G2)^2 + (B1-B2)^2}$$

If the difference is too big, then it is considered an edge. The aim of edge detection is to determine the edge of shapes in a picture and to be able to draw a result bitmap where edges are in white on black background (for example). The idea is very simple; we go through the image pixel by pixel and compare the colour of each pixel to those of its neighbours. If these comparison results in a too big difference the pixel being studied is part of an edge and should be turned to white, otherwise it is kept in black.

Edges are often used in image analysis for finding region boundaries. The purpose of detecting sharp changes in image brightness is to capture important events and changes. Laplace, Sobel and Canny are some algorithms that can be used to detect edges.

6. WISpaF: DESIGN AND IMPLEMENTATION

WISpaF is a system designed for the analysis and classification of Images as spam or ham. It is particularly aimed at image-based spam that are mainly text hidden or converted to images. It is based on the extraction of a low level image feature, the edge. WISpaF takes as input an image; analyze all the pixels of that image to identify and mark the edges. The identification of edges is done using the ***Sobel*** edge detection algorithm. After the edges are identified and marked, the system calculates and stores what percentage of the total number of pixels in the whole image are edges. With this percentage number of edges, we were able to classify the image as a spam image or ham depending on whether this percentage is above or below a set threshold. This is a slight variation from the method proposed by [8]. Since what we hope to filter are

mainly text-hidden-under-image spam mails, it is expected that there will be higher percentage of edges per computer generated images compared to natural pictures as the later tend to be smooth in texture.

Two approaches were tried when it came to the edge detection section of the system. In the first approach (which we termed IGS), we converted the loaded image to greyscale first before passing that greyscale image as an input to the sobel edge detection module. Figure 3 shows a snapshot of the output of this approach on a spam and natural image.

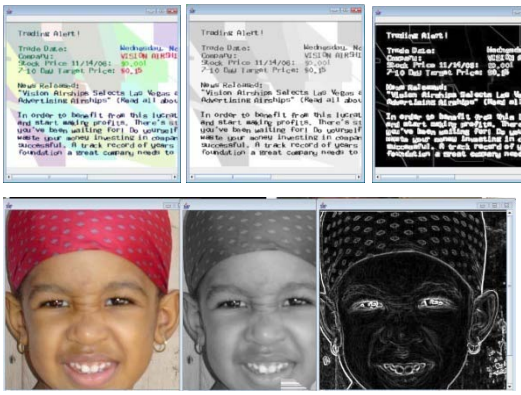

Figure 3: Output of first approach (image to greyscale to edge detection) on two different images

The second approach used (termed ISG) was to pass the original image to the sobel edge detection module first and then convert the resultant image to greyscale. Figure 4 shows a snapshot of this approach. This confirms that the edge detection method can be used on both greyscale as well as full colour images. It is interesting to note that the two approaches yielded slightly different final outputs.

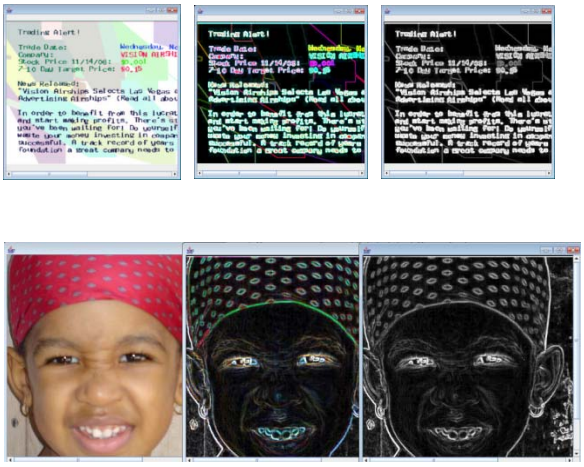

Figure 4: Output of second approach (image to sobel detection then greyscale) on two images

7. RESULTS

It took IGS 10 seconds and 5 seconds to analyze and classify the images called duck.jpg (figure 5) and Circum OCR.jpg (figure 6) respectively, while it took ISG 5 minutes 15 seconds and 9 seconds to analyze the same images respectively.

Figure 5: Duck.jpg

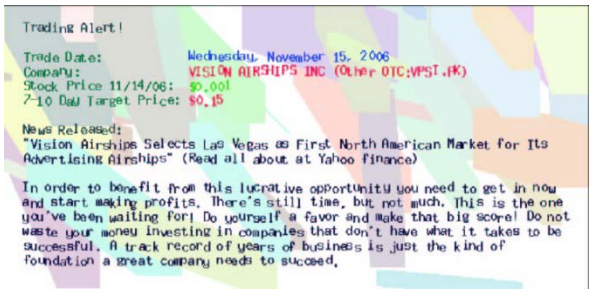

Figure 6: Circum OCR.jpg

We presume that this has to do with the approach as one works with more colours than the other. IGS was also much better in classifying the images, because of the magnitude of the edges calculated. The size of the image was not so much a constrain. Since in designing software, processing speed as well as good result is of paramount importance, the first approach, IGS was retained eventually because it gave a better result.

8. CONCLUSION

In this work, we have tried to show that using the low level image feature – edge, as well as the magnitude of the edges per image, it is possible to analyse and classify an image as spam or ham. The system is mainly targeted at text embedded in image in a bid to evade OCR filters. These kinds of images are usually computer generated graphic images. We have shown that our system can analyse images as well as photographic images and be able to tell them apart. WISpaF will assist in the reduction of image based spam emails especially those designed to evade OCR or finger printing filters. It is interesting to note that the more noise and obfuscation is

introduced to the image, the higher the chances of that image being classified as spam.

9. FUTURE WORK

One area that needs to be explored in image-based spam research is that of natural nude pictures. Our Future work will attempt to develop a system that can be used to filter such offensive images.

REFERENCES

[1] Bhaskar 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] Biggio, B., Fumera, G., Pillai, I., Roli, F., 2008, Improving Image spam filtering Using Image Text Features , Fifth conference on email and antispam, CEAS 2008,

[4] Chiemeke, S.C., Longe, O.B., Onifade, O.F.W, Longe, F.A., 2007, Text Manipulations and Spamicity Measures: Implications for Designing Effective Filtering Systems for Fraudulent 419 Scam Mails. ICASTdu Conference, Ghana, Dec 19-21 2007.

[5] Fumera, G., Pillai, I., Roli, F., 2006, Spam Filtering Based on the Analysis of Text Information Embedded Into Images, Journal of Machine Learning Research, Vol. 7, pp. 2699-2720.

[6] G.Fumera, I.Pillai, F.Roli, B.Biggio, Image spam filtering using textual and visual information, MIT Spam Conference 2007, Cambridge, USA, March 2007 (paper available at http://www.spamconference.org/).

[7] Mehta Bhaskar, Nangia Saurabh, Gupta Manish and Nejdl Wolfgang. 2008. 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. pp497-506.

[8] Nhung, Ngo Phuong and Phuong, Tu Minh, 2007, An Efficient Method for Filtering Image-Based Spam E-mail, In Computer Analysis of Images and Patterns, 12th International Conference, CAIP 2007, Vienna, Austria, August 27-29, 2007. Proceedings. W.G. Kropatsch, M. Kampel, and A. Hanbury (Eds.): pp. 945–953.

[9] Roli,F., Biggio, B, Fumera, G., Pillai, I., Satta, R., 2007, Image spam filtering by detection of adversarial obfuscated text, Workshop on Machine Learning in Adversarial Environments for Computer Secuirty, NIPS 2007, Whistler Canada, Dec 8, 2007,

[10] Wang, Zhe, William Josephson, Qin Lv, Moses Charikar, Kai Li, 2007, Filtering Image Spam with Near-Duplicate Detection. Proceedings of the 4th Conference on Email and Anti-Spam (CEAS).

[11] Youn Seongwook and McLeod Dennis, 2009, Improved Spam Filtering by Extraction of Information from Text Embedded Image E-mail, SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.