# Revealing Common Sources of Image Spam by Unsupervised Clustering with Visual Features

Chengcui Zhang, Wei-Bang Chen, Xin Chen, Gary Warner
Dept. of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, AL 35294, USA
{zhang, wbc0522, chenxin, gar}@cis.uab.edu

## ABSTRACT
In this paper, we investigate image spam with data mining techniques in order to reveal the common sources of unsolicited emails. To identify the origins, a two-stage clustering method groups visually similar spam images by exploring their visual features, including color feature, layout feature, text layout, and background textures. We test the proposed approach under different settings and combinations of features and measure the performance with a modified F-measure.

## Categories and Subject Descriptors
I.5.3 [**Pattern Recognition**]: Clustering – *algorithms, similarity measures.*

## General Terms
Algorithms, Experimentation, Security.

## Keywords
Image spam, Clustering, Computer Forensics, Botnet

## 1. INTRODUCTION
Spam, unsolicited emails, adversely affects the regular email communications on the Internet. Spammers use botnets, a malicious program hidden in a group of computers which are remotely controlled by spammers, to distribute spam and conceal their identities [1]. The most effective way of controlling spam currently is spam filtering which, however, can only differentiate spam emails from non-spam emails but cannot tell the origins of spam [2-3]. In order to hide their origins, escape spam detection and penetrate filters, criminals use various obscuring techniques.

Image spam, a commonly used obscuring technique, presents text primarily as an image to avoid text-based filtering. Spammers use various style editing tricks to generate images which look visually similar but appear unique to standard spam analysis to evade anti-spam technologies, e.g. fingerprinting. There are relatively few works in image spam identification [4-5]. All these works address the image spam filtering problem passively, and hence, it is essential to actively trace the origins of spam and bring down the botnets in order to stop spam.

This paper provides scientific evidence for identifying the common sources of spam, and is dedicated to the analysis and clustering of image spam based on the visual characteristics. Each resultant cluster contains visually resemble images, which indicates common origins, the spammers, of those images. The proposed method can further combine with other approach in [6] to identify and validate spam clusters or phishing groups for the purposes of cyber-crime investigation.


*SAC'09*, March 8-12, 2009, Honolulu, Hawaii, U.S.A.


In the rest of this paper, Section 2 details the proposed methods. Section 3 presents the results, and Section 4 concludes the paper.

## 2. METHODS

### 2.1 Image Spam Segmentation
Spam images generally comprise foreground and background. The foreground carries the text content and/or illustrations while the background contains various color and/or textures. Spam images are said to visually resemble if they have similar text layouts, illustrations, and/or background textures. Therefore, it is essential to separate these areas for similarity assessment.

First, we use Optical Character Recognition (OCR) to recognize texts whose bounding boxes represent the text layout. Second, we separate illustrations from the background by detecting the background. A color code histogram is then built for detecting the dominant color(s), i.e. the top 2% of high frequency bin(s), which are considered as background since the background area demonstrates more uniformity than foreground and occupies significant portions of an image. After background has been detected, the rest parts are illustrations (as shown in Figure 1).

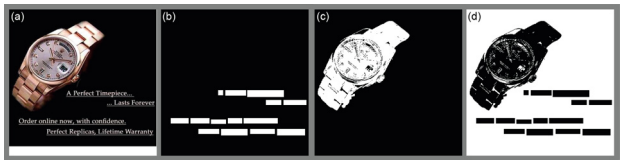


Figure 1. (a): The original image; (b): the text area mask; (c): the foreground illustration mask; and (d): the background mask.

### 2.2 Visual Feature Extraction
Four visual features are extracted for measuring image similarity. We adopt the color feature and the spatial layout feature to cluster illustrated images, and use the background texture and the text layout features in order to group text mainly images.

**Color features**: The foreground illustrations usually have more vivid color features to engage the viewers. Therefore, in this study, we use the color-code histogram to describe the color composition of the foreground illustrations.

**Layout features:** Spammers often obscure color-based filtering by adjusting the color scheme, which motivates us to use illustration layout for similarity measurement. The layout difference D between two images is defined in Equation 1.

$$D = \left(\#\text{of } \textit{trues} \text{ in } XOR\left(\text{mask}_1, \text{mask}_2\right)\right) / \text{mask size} \quad (1)$$

**Background texture features:** Another way to easily generate unique spam images with the same template is to change the background color. However, for the same template, the background texture often remains the same. Hence, we use two texture features, i.e., the homogeneity and orientation in the background and find that the "orientation" better distinguish among different templates than "homogeneity" in our dataset. The orientation feature is actually a normalized edge-orientation histogram, an adapted version of Tamura's directionality [7].

**Text layout analysis:** Spammers often use the same text layout template to generate different advertisements by only changing

the wordings for different products. To analyze text layout, we extract the minimum bounding box of the whole text area, which is then dilated to connect words in the same line. The dilated text area is then scaled and normalized for text layout comparison. The similarity of two text layouts is defined in Equation 2.

$$layout(I_1, I_2) = \sum_{i,j} (I_1(i,j) - I_2(i,j)) / (l_{small} \times w_{small}) \quad (2)$$

where $I_1(i, j)$ and $I_2(i, j)$ are the corresponding values at $(i, j)$ of the two text area masks. The pixel value is either 1 (text) or 0 (non-text). $l_{small}$ and $w_{small}$ are the length and width of the smaller text area. A series of distances are thus calculated by sliding the smaller text area over the larger one. The minimum distance value is used as the distance between the two text layouts.

## 2.3 Image Spam Clustering

Image spam can be typically categorized to illustrated images and text mainly images, which are dealt with separately in this study.

### *2.3.1 Illustrated image clustering*

For clustering illustrated spam images, we propose a two-stage clustering algorithm. In the first stage, we cluster a set ($I$) of spam images by (1) selecting a query image, (2) measuring the visual similarities of the query image with all other images and generating two ranked lists that correspond to color and layout features, respectively, (3) collecting the first $n$ images in both lists as two sets A={$a_1$,…, $a_n$} and B={$b_1$, …, $b_n$}, (4) finding the maximal $n$ that satisfies A=B and $n<m$, where $m$ is the total number of images, (5) grouping all images in A (or B) as a cluster, (6) removing the clustered images from $I$, and (7) repeating (1)-(6) until $I$ is empty. The result is a set of image clusters that highly agree with histogram, shape, and/or location features. In the second stage, we use hierarchical clustering on the color feature with a very strict criterion (≥99% similarity) to repair the misclassified images due to significant rearrangement of foreground illustrations in the image.

### *2.3.2 Text mainly image clustering*

For text-mainly images, we combine the text layout and background texture features in a weighted manner. According to our experimental results for the best weight assignment, we weight the background texture twice as much as the text layout.

The clustering process begins with finding the pair of images that yields the minimum distance in the combined distance matrix. We then use one of the two images as a query image and retrieve those images that are at least 65% similar to the query image. The query image and the retrieved images forms a cluster and are removed from the image set. The above procedure is repeated until the image set becomes empty.

# 3. EXPERIMENTAL RESULTS

We use a database of 1190 spam images which are manually grouped into 61 clusters as the ground truth.

## 3.1 Performance Evaluation

The clustering performance is evaluated with a modified F-measure [8] which calculates the pairwise F-measure between clusters in the ground truth and those in the clustering results. The overall F-measure is defined as the mean of the maximum F-measures. Since our goal is to reveal the common source of spam images, recall is experimentally determined to weigh 4 times more than precision in this study. We test various cutoff values (99.5%, 99.0%, and 97.5%) in the second phase of illustrated image clustering. Among them, 99% is selected as the cutoff value because it has the highest F-measure value, and its number of clusters is the second closest to the ground truth.

We further compare the performance of other combinations of the two foreground features (color and layout features) for clustering, including color-code histogram only, foreground illustration layout feature only, and the combined use of the two features without the second-stage clustering. The experimental results show that the F-measure values are 0.748, 0.678, and 0.772, respectively, for the above three alternatives. The superiority of the proposed clustering method for illustrated images is obvious.

We also test various cutoff values (75%, 70%, 65%, 60%, 55%, 50%, and 40%) for clustering text images. Among them, we chose the cutoff value 65% for text image clustering, which result in an overall best F-measure 0.779.

In addition, we compare the F-measure of the proposed algorithm with the clustering methods reported in [9], which does not differentiate illustrated images from text images and combines several visual features with equal weights in the clustering process. The hierarchical clustering algorithm adopted in [9] was performed on the same dataset with various cutoff values, and its average F-measure value is are 0.599. It is obvious that the proposed method outperforms the one proposed in [9].

# 4. CONCLUSIONS

We report a clustering framework to identify the origins of spam images according to their visual similarity. This study goes beyond traditional means of spam filtering and is among a few recent efforts in identifying the origins of spam. Like the other approaches in this field, e.g., spam clustering according to common subjects, the image spam clustering is one key piece of a complex jigsaw puzzle. Those techniques, when put together, can help reveal the origins of spam.

# 5. ACKNOWLEDGEMENTS

This research of Dr. Zhang is supported in part by NSF DBI-0649894 and the UAB ADVANCE program.

# 6. REFERENCES

[1] www.cnn.com/2007/TECH/11/29/fbi.botnets

[2] Clark, J., Koprinska, I., and Poon, J. 2003. A neural network based approach to automated e-mail classification. In *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, pp. 702 – 705, Oct. 13-17, Beijing, China.

[3] Sanpakdee, U., Walairacht, A., and Walairacht, S. 2006. Adaptive spam mail filtering using genetic algorithm. In *Proceedings of the 8th International Conference on Advanced Communication Technology*, pp. 441-445.

[4] Byun, B., Lee, C.-H., Webb, S., and Pu, C. 2007. A Discriminative Classifier Learning Approach to Image Modeling and Spam Image Identification. In *Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS 2007)*, Mountain View, CA, USA.

[5] Mehta, B., Nangia, S., Gupta, M., and Nejdl, W. 2008. Detecting Image-based Email spam using visual features and Near Duplicate Detection. In *Proceedings of the WWW*.

[6] Chun, W., Sprague, A., Warner, G., and Skjellum, A. 2008. Mining Spam Email to Identify Common Origins for Forensic Application. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing*, Mar. 16-20, Fortaleza, Ceará, Brazil.

[7] H. Tamura, S. Mori, and T. Yamawaki. 1978. Textural Features Corresponding to Visual Perception. *IEEE Transaction on Systems, Man, and Cybernetics*, vol. SMC-8, pp. 460-472, 1978.

[8] Van Rijsbergen, C.J.: Information Retrieval. London; Boston. Butterworth, 2nd Edition 1979. ISBN 0-408-70929-4.

[9] Zhang, C., Chen, X., Chen, W-B., Yang, L., and Warner, G. 2008. Spam image clustering for identifying common sources of unsolicited emails. To appear in *International Journal of Digital Computer Forensics*.