Image Spam Clustering – An Unsupervised Approach

Wei-Bang Chen and Chengcui Zhang
Department of Computer and Information Sciences
The University of Alabama at Birmingham, Birmingham, AL 35294, USA
{wbc0522, zhang}@cis.uab.edu

ABSTRACT

We propose an unsupervised image clustering framework for revealing the common origins, i.e. the spam gangs, of unsolicited emails. In particular, we target email spam with image attachments because spam information is harder to extract due to information hiding enabled by various image obfuscation techniques. To identify spam gangs, we observe that spam images from the same source are usually composed of visually similar elements which are arranged and altered in many different ways in order to trick the spam filter. We propose to infer spam images originated from the same spam gang by investigating spam email similarity in terms of their visual appearance and editing style. In particular, a data mining technique based on unsupervised image clustering is proposed in this paper to solve this problem. This is achieved by first dividing a spam image into different areas/segments, including texts, foreground graphic illustrations, and background areas. The proposed framework then extracts characteristic visual features from segmented areas, including text layout, visual features of foreground graphic illustrations and its spatial layout, and background texture features. In the clustering stage, all spam images are first categorized as illustrated images and text mainly images according to the existence of foreground illustration objects. Then illustrated images are clustered based on the color and/or foreground layout, while text mainly images are clustered based on the text layouts and/or background textures. A novel unsupervised ranked clustering algorithm is proposed for feature fusion, which is used in combination with the traditional hierarchical clustering algorithm for clustering. We test the proposed approach using different settings and combinations of features and measure the overall performance with V-measure.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering – *algorithms*, *similarity measures*.

General Terms: Algorithms, Experimentation, Security, Legal Aspects.

Keywords: Image spam, Clustering, Computer Forensics.

1. INTRODUCTION

Spam is unsolicited emails that adversely affect the regular email communications on the Internet. Billions of dollars of counterfeit software, electronics, as well as shoes, etc., are being sold through spam advertisements. Spam emails, claiming to be from banks, might lure users to give out their identifications. In order to survive the spam filtering, spam usually falsifies the sender information. Spammers often utilize botnets to keep sending automatic spam. Nowadays, botnets are the main choices for cyber criminals who seek to conceal their true identities by using third-party computers as media for their crimes [1].

The most effective way of controlling spam emails at the moment is spam filtering [2-5]. However, filters can only differentiate spam emails from non-spam ones but cannot track their origins. In order to hide the origins, escape spam detection and penetrate filters, and to conceal the fact that there are relatively few organizations creating the vast majority of these unsolicited emails, criminals use a variety of intentional obscuring techniques. For example, one of the techniques is to present text primarily as an image, to avoid traditional text-based filtering.

When creating image spam, spammers use certain tricks to defeat traditional anti-spam technologies like fingerprinting, OCR, and URL block-list. Spammers vary the space between words and lines and also add random speckles to make messages look different to different recipients, though all of them have the same text. By this way, they evade fingerprinting technology by making the images appear unique to standard spam analysis. The use of different colors and varying font size makes it impossible for OCR techniques to find out spam. Botnets are also becoming efficient and they can produce a large number of random images within a short time.

In order to stop spam, it is essential to trace the spam origins and bring down the spamming servers. In this process, law enforcement officers shall be actively involved as spam propagating is also a legal issue. The goal of this paper is to facilitate spam gang tracking by providing scientific proof to the common sources of spam. This paper is dedicated to the analysis and clustering of image spam based on their visual characteristics. Through clustering, spam images are grouped together. Each cluster contains spam images whose visual effects resemble each other in the cluster, indicating common origins/sources of those images, i.e., they are created from the same template hence by the same spam gang.

There are relatively few works in image spam identification [6-8]. All these works address the image spam filtering problem. For example, Byun et al. [6] report a classification method to model and identify image spam. McAfee, an Internet security vendor, also provides image spam filtering functions. The main purpose of these works is to correctly identify image spam from non-spam ones so that the filter can selectively pass or block a particular email.

In this study, we go one step further to track the source of the spam distributors based on image spam clustering, i.e., if two spam emails have similar composition, i.e., similar content, layout, and/or editing styles, they are likely related. This can be used as a strong evidence to identify and validate spam clusters or phishing groups for the purposes of cyber-crime investigation. Chun et al. [9] proposed an approach that uses clustering techniques to form relationships between email messages and group them into spam clusters. Clusters were evaluated using a visual inspection method. A routine was developed to fetch a thumbnail of the appearance of each destination website. Where

MiFor09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-755-4/09/10...$10.00.

the resultant collection of website images from a single cluster was visually confirmed to be the same, a high confidence was placed upon the integrity of the cluster. Our proposed method can not only automate this visual validation process, but also link visual similarity directly to the presence of spam clusters.

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

2. The PROPOSED APPROACH

The overview of the proposed image spam clustering framework is illustrated in Figure 1.

We first detect the main composition type of a spam image. A spam image may contain all or a subset of the following contents: background, text, and illustration. By proper image segmentation, we can classify image spam based on their composition into two types: illustrated images (mainly composed of graphic illustrations) and text mainly images. Subsequently, we extract different visual features in order to cluster these two types of image spam separately since their characteristics in term of the visual composition are quite different.

The details of the proposed method are illustrated step-by-step based on the flow of the framework shown in Figure 1.

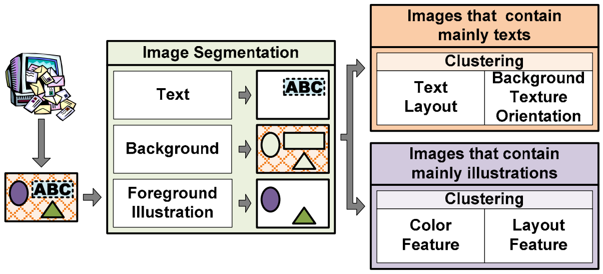


Figure 1. The overview of the proposed framework

2.1 Image Spam Segmentation

Spam images are usually composed of foreground objects and background. The foreground carries the target text message and/or illustrations (e.g., product pictures, logos, etc.) while the background contains various color and/or textures. Spam images from a common source are often based on a common template which is permuted (e.g., change the layout and/or color schema) to form large number of similar looking but distinct spam images. To separate the concerns, it is essential to distinguish foreground objects (texts and illustrations) from the background.

Two spam images are said to visually resemble each other if they have similar text layout, and/or similar foreground illustration, and/or similar background textures. To recognize foreground objects in the spam image, we first extract text areas through Optical Character Recognition (OCR). The coordinates of the bounding rectangles of each recognized word in the image is used in the subsequent text layout analysis.

After text areas in the image have been identified by OCR, the rest of the areas in that image should ideally contain the foreground illustrations and the background. Hence, our next step is to separate the foreground illustrations from the background. The foreground illustrations are actually sub-images in a spam image. Typically, these sub-images are full of variety in their visual appearance, and thus, difficult to characterize with any fixed set of visual features. On the contrary, the background is generally composed of a pure color base or computer-generated textures, and has relatively more uniformity than illustrations. In order to separate these two parts, our strategy is to obtain the illustration areas by removing the background in that image. However, this is not trivial since random noise and textures were often added on purpose to increase the background randomness and variations. Therefore, it is not suitable to use a single threshold value on visual homogeneity to separate the foreground objects from the background.

Instead, the proposed method is based on the following two assumptions: 1) Spam images must have sufficient foreground/background contrast to ease the reading of their recipients, which is usually the case as indicated by Byun et al. in [6]. More specifically, the intensity values of foreground and background must have significant difference. 2) The background area occupies a significant portion of an image, which is often the largest or at least comparable to foreground illustrations. Also because background usually demonstrates more uniformity than foreground, background pixels tend to be clustered together in the pixel intensity histogram while foreground pixels demonstrate a larger range of intensities.

To identify the background in a spam image, we first convert each pixel in a color image to a 6-bit color code by taking the 2 most significant bits of each R, G, and B color components. The goal of this process is to transform a RGB image to an index image such that similar colors can be grouped together. After the transformation, we build a histogram for the index image. Based on our second assumption, background pixel intensities usually have a relatively smaller range than that of the foreground and thus correspond to high frequency bin(s) in the histogram. To identify the high frequency bin(s), we first calculate the average frequency (m) of all bins as well as their standard deviation (σ). Any bin with its frequency above the threshold $m+2\sigma$ is considered as a high frequency bin. This threshold can remove 98% of bins and keep the top 2% high frequency bin(s). These high frequency bins indicate the dominant color(s) in the image, and thus are considered as background, based on the second assumption.

Up to this point, each spam image is segmented into text, foreground illustrations, and the background (see Figure 2 (a)-(d)). These segmentation masks will be used in the subsequent visual features extraction steps.

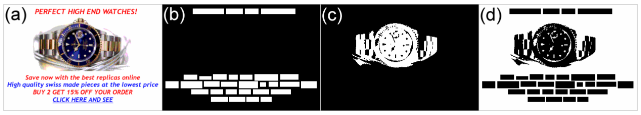

(a)
PERFECT HIGH END WATCHES!
Save now with the best replicas online
High quality swiss made pieces at the lowest price
BUY 2 GET 15% OFF YOUR ORDER
CLICK HERE AND SEE
(b)
(c)
(d)

Figure 2. (a): The original image; (b): the text area segmentation mask; (c): the foreground illustration segmentation mask; and (d): the background segmentation mask.

2.2 Visual Feature Extraction

Recall that our goal is to identify the spam images produced from the same origin through image clustering. To achieve this goal, our strategy is to group visually resembled images by using the similarity of their visual features. Therefore, proper visual features must be extracted in order to establish the connection.

For illustrated image spam, we observe that spam gangs attempt to obscure anti-spam filtering with tricks such as adding, replacing or relocating the foreground graphic illustrations. These minor alterations may change the color scheme and/or the spatial layout of the foreground objects. This indicates that we can associate two images if their foreground illustrations are highly similar in terms of the color composition and/or the spatial layout.

For text mainly images, spam gangs typically create image spam of this kind by using a template. The same template can take different text messages, generate similar text layout, and produce images with various text colors and background colors. We also observe that the images produced from the same template often have similar background texture despite the difference in their background colors. Based on these observations, we may assume that if two images are created from the same template, their background texture and text layout must be similar.

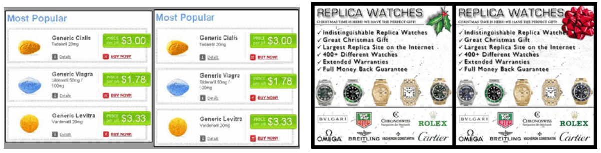
In this study, we adopt the color features and the spatial layout features to cluster illustrated images, and use the background texture and the text layout features to cluster text mainly images. The following subsections detail the feature extraction process.

2.2.1 Color Features

A large number of spam images contain foreground illustrations in them, such as artworks, pictures, and tables. We assume that these images with similar foreground sub-images advertise the same kind of product/service and therefore may originate from the same source. In addition, relocating the foreground illustrations does not change the color composition of the foreground illustrations. Moreover, we can also assume that these foreground sub-images have vivid color features to engage the viewers. In this study, the color code histogram is extracted from the foreground illustrations and used as the color features to describe the color composition of the foreground illustrations of an image.

2.2.2 Layout Features

Spam gangs may alter image spam by slightly adjusting the color composition of the foreground illustrations without significantly affecting the visual composition perceived by human eyes. This obfuscation may defeat the color histogram-based detection, and therefore, motivates the integration of shape and layout information of foreground graphic illustrations. To compare the layout difference of foreground illustrations between two images, we should first normalize their size. However, spam gangs often produce mutated images by shifting the entire foreground content, and/or rescaling the entire image or even changing the canvas size, which is exemplified in Figure 3(a).



(a) **(b)**

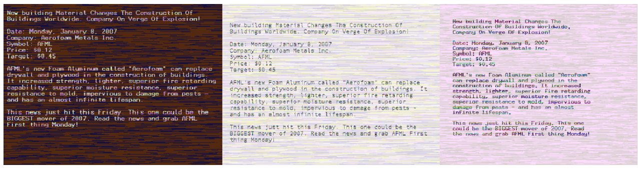
Figure 3. a) Spam images in different scale. b) Spam images with illustration substitution

These tricks change the absolute or relative position of foreground content in an image, but the relative distances between foreground objects are still retained. For this reason, the comparison of layout difference is performed on the foreground illustration region instead of on the entire image region. To extract the foreground illustration region, the minimum and maximum *x*-coordinate and *y*-coordinate of all foreground objects in an image are found, which forms a minimum bounding rectangle of all the foreground objects. The foreground illustration segmentation masks of two images (see Figure 2 (c)) are further cropped according to their foreground bounding rectangles. Subsequently, two cropped masks are normalized. Finally, XOR operation is applied on the two normalized foreground masks for comparing two images. This operation measures the difference in the foreground illustration layout between two images. Small difference values indicate high similarity in terms of the foreground illustration layout. Figure 3(b) shows two spam images with similar foreground illustration layout but different illustrations in their upper right corners. The formal definition of the layout similarity matrix *L* is defined in (1).

$$L(I_i, I_j) = 1 - \frac{\#\text{of 1s(trues) in } XOR(\text{mask}_i, \text{mask}_j)}{\text{The size of the normalized mask}} \tag{1}$$

2.2.3 Background Texture Analysis

One of the easiest ways to generate unique spam images based on a common template is probably to change the background color. Thus, color similarity alone cannot be treated as an important indication of common templates for background analysis. Instead, we first convert the image background into grayscale. We further find that, although different in color, the background texture features of images created from the same template tend to have less variation. As shown in Figure 4, three images that have exactly the same text content have different background colors but very similar background textures. The first and the second images in Figure 4 also share a similar text layout.



New building Material Changes The Construction Of
Buildings Worldwide. Company On Verge Of Explosion!

Date: Monday, January 8, 2007
Company: Aerofoam Metals Inc.
Symbol: AFML
Price: $0.12
Target: $0.45

AFML's new Foam Aluminum called "Aerofoam" can replace
drywall and plywood in the construction of buildings.
It increased strength, lighter, superior fire retarding
capability, superior moisture resistance, superior
resistance to mold, impervious to damage from pests -
and has an almost infinite lifespan.

This news just hit this Friday. This one could be the
BIGGEST mover of 2007. Read the news and grab AFML
First thing Monday!

New building Material Changes The Construction Of
Buildings Worldwide. Company On Verge Of Explosion!

Date: Monday, January 8, 2007
Company: Aerofoam Metals Inc.
Symbol: AFML
Price: $0.12
Target: $0.45

AFML's new Foam Aluminum called "Aerofoam" can replace
drywall and plywood in the construction of buildings. It
increased strength, lighter, superior fire retarding
capability, superior moisture resistance, superior
resistance to mold, impervious to damage from pests -
and has an almost infinite lifespan.

This news just hit this Friday. This one could be the
BIGGEST mover of 2007. Read the news and grab AFML First
thing Monday!

New building Material Changes The
Construction Of Buildings Worldwide,
Company On Verge Of Explosion!

Date: Monday, January 8, 2007
Company: Aerofoam Metals Inc.
Symbol: AFML
Price: $0.12
Target: $0.45

AFML's new Foam Aluminum called "Aerofoam"
can replace drywall and plywood in the
construction of buildings. It increased
strength, lighter, superior fire retarding
capability, superior moisture resistance,
superior resistance to mold, impervious to
damage from pests - and has an almost
infinite lifespan.

This news just hit this Friday. This one
could be the BIGGEST mover of 2007. Read
the news and grab AFML First thing Monday!

Figure 4. Background texture similarity based on the edge orientation

Therefore, we analyze the “homogeneity” and the “orientation” texture features of image backgrounds and found that with our current spam image collection, “orientation” can better distinguish among different templates than “homogeneity”. The orientation feature used in this study is an adapted version of Tamura’s directionality feature [10], which measures the local direction of the edge in the background textures by first applying the Prewitt edge operators, and then, computing the local orientation angle θ with the following formula.

$$\theta(u,v) = \tan^{-1}\left(\Delta_y(u,v) / \Delta_x(u,v)\right) \tag{2}$$

where (u, v) are the coordinates of an edge pixel; Δx and Δy are the filter results obtained from the corresponding Prewitt edge operators. The obtained edge orientation values are then quantized into a 16-bin histogram H_{dir}. In Tamura’s paper, the directionality feature is the sum of the second moments around each peak in H_{dir} from valley to valley. However, this measurement may cause problems since we may obtain the same directionality feature from two totally different H_{dir}. Thus, the normalized H_{dir} (divided by the total number of edge pixels) is adopted to represent the orientation feature of the background texture.

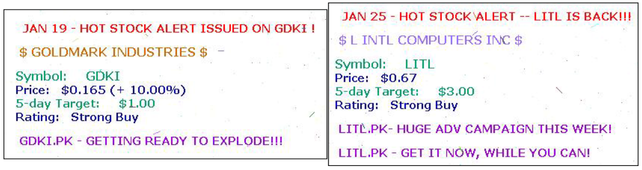
2.2.4 Text Layout Analysis

The texts in two advertising spam images are probably not the same when they are trying to sell different things. However, it is highly possible that a spammer uses the same text layout template in generating different advertisements by only changing the wordings for different products. For example, Figure 5 shows two spam images that have very similar text layout but advertise different things, which indicates a possible common origin.

The proposed text layout analysis method consists of the following steps:

1) **Bounding Box Extraction** – The first step is to extract the minimum bounding box of the whole text area in each image.

2) **Dilation** – We notice that two text layouts may look similar in their general layout yet their word and space distributions are very likely to be dissimilar. If we directly compare the text layout masks, noises will be introduced by different word length, line spacing, and word positions in each text line. We alleviate this problem by coarsening the text area. In doing so, we try to connect words in one line if they are only separated by a small space. The method used is called dilation. For each pixel in the bounding box, if it is "1", the m pixels on its right and m pixels on its left are also set to 1. In this way, small spaces are "closed" and therefore ignored. Only the general layout of the whole text area will be considered in the analysis. Examples of dilated text areas are shown in Figure 6.



JAN 19 - HOT STOCK ALERT ISSUED ON GDKI !

$ GOLDMARK INDUSTRIES $

Symbol: GDKI
Price: $0.165 (+ 10.00%)
5-day Target: $1.00
Rating: Strong Buy

GDKI.PK - GETTING READY TO EXPLODE!!!

JAN 25 - HOT STOCK ALERT -- LITL IS BACK!!!

$ L INTL COMPUTERS INC $

Symbol: LITL
Price: $0.67
5-day Target: $3.00
Rating: Strong Buy

LITL.PK - HUGE ADV CAMPAIGN THIS WEEK!

LITL.PK - GET IT NOW, WHILE YOU CAN!

Figure 5. An example of spam images that advertise different things but have a very similar text layout

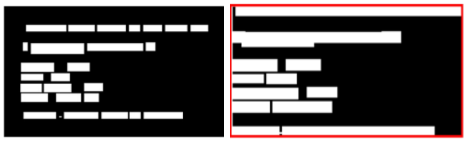


Figure 6. An example of text area dilation. The left is the original text mask and the right one is after dilation.

3) **Scaling** – Text areas from different spam images are usually not of equal size. We cannot directly compare them without normalization. A common way is to down-sample the larger text area, bounded by its minimal bounding box, to the same size of the smaller text area. However, this method may cause the larger text area to be skewed since the aspect ratio of the two images can be different. Therefore we only resize the larger text area so that its length is the same as that of the smaller text area, with its original aspect ratio being preserved.

4) **Similarity Calculation** – After resizing, we superimpose the text area with the shorter width on the one with the longer width and conduct the pixel-wise comparison. Then we slide the smaller text area one step at a time and repeat the comparison. At each time of compare, their distances are calculated as follow:

$$layout(I_1, I_2) = \left(\sum\nolimits_{i,j} (I_1(i,j) - I_2(i,j))\right) \Big/ (l_{small} \times w_{small}) \qquad (3)$$

where $I_1(i, j)$ and $I_2(i, j)$ are the corresponding pixel values at the position (i, j) of the two text area masks. Here the value of a pixel is either 1 (white: text pixel) or 0 (black: non-text pixel). 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 to represent the distance between the two text areas, i.e., the distance of the two text layouts.

2.3 Image Spam Clustering

As mentioned earlier, image spam can be typically categorized into two major groups, including images that contain mostly illustrations (e.g., pictures and tables) and those mostly text-based. Since the visual clues that lead us to their origins are quite different, we deal with illustrated images and text mainly images separately. In practice, the entire image set is classified into illustrated images and text mainly images, which is based on the percent of foreground area. The reason is that foreground objects (excluding texts) usually occupy a significant portion in an illustrated image. If the foreground area is too small, it tends to be noise in the image. In this study, we test various foreground-image ratios and find that using a cutoff value of 1% produces the best classification results according to our dataset. To this point, illustrated images and text mainly images are separated and can be individually clustered.

To test the effectiveness of various features, including their combinations, the proposed ranked clustering algorithm and the traditional agglomerative hierarchical clustering algorithm are both adopted and tested in the proposed framework, and their best combination is selected as our final clustering method. It is essential for a traditional hierarchical clustering algorithm to have a predefined cutoff value to form clusters. On the contrary, the proposed ranked clustering algorithm, which will be detailed later in Section 2.3.1, is unsupervised in the sense that it does not require a preset cutoff value to produce clusters since it forms a cluster by comparing and fusing ranked lists. However, the constraint of this algorithm is that it needs at least two ranked lists, i.e., requiring at least two features to be used in the clustering. Therefore, we use traditional agglomerative hierarchical clustering algorithm when clustering spam images with one single feature and adopt the proposed ranked clustering when two or more features are used in clustering.

2.3.1 Ranked Clustering

In this paper, we propose a ranked clustering algorithm for clustering spam images based on the extracted features. These features have their own strength and complement each other in grouping similar spam images. The basic idea behind the proposed clustering algorithm is that different features capture image similarity in different ways. Therefore, no matter how images are differently ranked when different features are used for calculating similarity scores, similar images tend to be listed on the top of the ranked list.

Assume there are in total n images in a set $P=\{p_1, p_2, \ldots, p_n\}$. Given a query image p_x, according to some visual feature f_k we can generate a ranked list l_k by measuring the distance between p_x and $\forall\, p \in P$ with the first image of each list being the query image itself. Therefore, all images in the list are ranked according to their distance to the query image in a decreasing order according to that feature f_k. For the reason aforementioned, the image ranks in the lists of different visual features can be different. We then compare all the lists and find all the indices in the lists above which the sub-lists contain the same image set. Let all the images in l_k above index y be the set $S_k y$ ($S_k y \subset l_k$). We collect all the y values in a set Y where all the $S_1 y = S_2 y = \ldots = S_m y$ (m is the index of visual features used in the clustering). Suppose y_{max} is the maximum value in Y and $y_{max}<n$ ($y_{max} = n$ indicates the equality of the full lists; y_{max}=1 indicates the equality of only the top images which are actually the query image itself, from all the lists. y_{max} indicates that the first y_{max} images in all the lists are very similar in terms of their corresponding visual features. These images then form a cluster and are removed from the image set. A new query image is then selected from the remaining images before the next query starts. This process will be repeated until the entire image set is empty, resulting in a set of image clusters within each of which member images are highly similar in terms of all the visual

features. In this study, we cluster illustrated spam images with the proposed ranked clustering algorithm and cluster text mainly images with a combined approach of the ranked clustering and the traditional agglomerative hierarchical clustering algorithm. The proposed ranked clustering algorithm is formalized as follows:

```
Let $P=\{p_1, p_2, \ldots, p_n\}$ be the original image set.
Let $F=\{f_1, f_2, \ldots, f_m\}$ be the feature set.
Step1: Randomly select $p_x$ from $P$ as a query image.
Step2: Generate ranked lists $l_1, l_2, \ldots, l_m$ corresponding to features
       $f_1, f_2, \ldots, f_m$ so that images in a list are ranked according to
       their distance to $p_x$ based on the corresponding feature.
       Thus $l_k = \{q_{k1}, q_{k2}, \ldots, q_{kn}\}$ where $q_{ki} \in P$
Step3: For $y \in \{1, 2, \ldots, n\}$
           $S_k y = \{q_{k1}, q_{k2}, \ldots, q_{ky}\}$, $k=1, 2, \ldots, m$, $(S_k y \subset l_k)$
           set $Y=\{y \neq n \mid S_1 y = S_2 y = \ldots = S_m y\}$
       end
       $y_{\max} = \max \{Y\}$
Step 4: $P= P-S_k y_{\max}$ (remove the first $y_{\max}$ images from the current
        image set $P$)
Step 5: Go back to Step 1 and continue the process until the entire
        image set $P$ is empty.
```

The goal of this study is to reveal the spam gangs through spam image clustering. Due to the great diversity of image spam, it is impossible to differentiate spam gangs by clustering image spam with one single feature. This manifests the need of feature integration. It is worth mentioning that the proposed ranked clustering algorithm can easily integrate multiple features in a single framework without providing a predefined threshold, and thus, provides a solution for clustering spam images with multiple features.

For illustrated images, we assume that spam images containing similar foreground illustrations may come from the same source. As aforementioned, two visual features, i.e., color features and foreground spatial layout features are extracted for clustering illustrated images. To better understand the effectiveness of those visual features in spam image clustering, we adopt the proposed ranked clustering algorithm when clustering with both features, and use the traditional hierarchical clustering algorithm with a predefined threshold when clustering spam image with either of the two features alone. The comparison result between single-feature based clustering and multi-feature clustering is presented in Section 3.

For text mainly images, commonly used modification tricks include changing the background color, background textures, and/or text layout. However, as mentioned earlier, judging from the background color alone is not sufficient in identifying common templates. Therefore, text layout features and background texture features are also adopted in the proposed framework for clustering text mainly images. Similarly, the proposed ranked clustering algorithm is used to cluster text mainly images when more than one feature is used, and the traditional hierarchical clustering algorithm with a predefined threshold is used in single-feature based clustering for text mainly images. The comparison results are summarized in Table 1.

3. EXPERIMENTAL RESULTS

In our experiments, the proposed two-stage clustering is performed on 1190 spam images in order to reveal the common sources – the spammers – from these images. The difficulty in this study is that there is indeed no “real” ground truth for evaluating the experimental results since we have no prior knowledge as to the actual sources of image spam. In order to evaluate the performance, we manually cluster the 1190 spam images into 61 clusters based on visual inspection and use this manual clustering result as the ground truth.

3.1 The Evaluation Method

To compare the resultant clusters with the ground truth, we use V-measure, a weighted harmonic mean of homogeneity (*hm*) and completeness (*cm*), proposed by Rosenberg and Hirschberg [13]. V-measure is a conditional entropy-based method to evaluate the clustering results and is independent of the clustering algorithm being used. The definition of V-measure is given in Equation 4.

$$V_\beta = \left((1+\beta^2) \times hm \times cm\right) \Big/ \left(\left(\beta^2 \times hm\right) + cm\right) \tag{4}$$

where β is a constant, which if greater than 1 would mean that *cm* is weighted β times more strongly; otherwise *hm* is weighted more in the calculation. In this study, we compare our clustering results with the ground truth using this measure with $\beta=1$. It is worth mentioning that different β values have also been tested and the same trend as shown in Table 1 (Section 3.2) has been observed.

3.2 Performance Evaluation

The proposed framework adopts both ranked clustering and hierarchical clustering algorithms. For hierarchical clustering algorithm, it requires a cutoff value in order to group images. In the first experiment, we test various cutoff values for each visual feature. According to the experimental results, the best cutoff values for color-code histogram (CCH), foreground layout (XOR), Scale Invariant Feature Transform (SIFT) [14], text layout (TxL), and background texture (BgTxO) features are 0.960, 0.900, 0.940, 0.600, and 0.925, respectively. These thresholds are used in testing single-feature based clustering.

It is worth noting that we compare SIFT feature with CCH and XOR features since SIFT is commonly used in image matching and measures the similarity from a different aspect than the visual features used in our framework. SIFT detects image features by building an octave of difference of Gaussian (DOG) images and finding local extrema in a scale space. For each feature point, a principle direction is computed and its neighborhood is rotated accordingly. Finally, a descriptor for each feature is formed by accumulating gradients in its neighborhood weighted by a Gaussian window. The extracted feature descriptors are not affected by image scaling or rotation, and are proved to be robust in matching objects across images. A match of a SIFT feature is defined as the nearest neighbor of its descriptor in Euclidean space. We organize all feature descriptors for each spam image into a kd-tree structure, so that the nearest neighbor search is reduced to logarithmic time.

In order to evaluate the effectiveness of visual features, we test various feature combinations. As aforementioned, we first test single-feature based clustering, i.e. CCH, XOR, SIFT, TxL, and BgTxO alone, with traditional hierarchical clustering algorithm. When using two or more features in clustering, i.e., CCH+XOR, and TxL+BgTxO, we adopt the proposed ranked clustering method. Table 1 demonstrates the experimental results in terms of V-measure under various combinations.

From Table 1, we observe that the best result (the highest V-measure (0.747)) is achieved when using the combined CCH+XOR features and the ranked clustering for clustering illustrated images and using the BgTxO feature and the hierarchical clustering algorithm for clustering text mainly images. We visually examine the resultant clusters and find that most similar spam images are grouped together. Figure 4 shows a subset of a resultant cluster. Three images have different background colors but exactly the same text content. It can be seen that all 3 images in the cluster have similar background texture as the cluster centroid, i.e., the leftmost image, despite the disparity in background colors and texture scales.

Table 1. Performance analysis on different combinations of features and clustering algorithms

Illustrated Images	Text Mainly Images	*hm*	*cm*	V_1
CCH	**TxL**	0.563	0.575	0.569
XOR	**TxL**	0.537	0.567	0.552
SIFT	**TxL**	0.542	0.558	0.550
CCH+XOR	**TxL**	0.563	0.579	**0.571**
CCH	**BgTxO**	0.772	0.720	0.745
XOR	**BgTxO**	0.747	0.717	0.732
SIFT	**BgTxO**	0.752	0.705	0.728
CCH+XOR	**BgTxO**	0.773	0.724	**0.747**
CCH	**TxL+BgTxO**	0.963	0.400	0.565
XOR	**TxL+BgTxO**	0.938	0.394	0.555
SIFT	**TxL+BgTxO**	0.943	0.392	0.554
CCH+XOR	**TxL+BgTxO**	0.963	0.401	**0.566**

From Table 1, we also observe that clustering illustrated images with CCH+XOR features and clustering text mainly images with TxL+BgTxO features produces high homogeneity, but relatively low completeness. By examining the resultant clusters, we find that illustrated images are mostly correctly clustered. However, the text mainly images forms many small clusters. This is due to the fact that many text-based spam images with same text content do not necessarily share the same text layout since the spam gangs may alter the page size, which also changes the text layout. Therefore, it would be beneficial to further incorporate textual clue into the proposed framework for clustering text mainly spam images.

The experimental results also show that when using the proposed ranked clustering algorithm to cluster illustrated images, the proposed color feature (CCH) and the foreground spatial layout feature (XOR) complement to each other. This claim can be justified by comparing the V_1 values contained in the rows 1-4, 5-8, or 9-12 in Table 1 where the proposed clustering algorithm produces the higher homogeneity, completeness, and V-measure values with the combined CCH+XOR features for illustrated images than when either feature is used alone. In addition, CCH+XOR outperforms SIFT feature in all cases.

4. CONCLUSIONS

In this paper, we propose an unsupervised image spam clustering framework in order to identify the common sources of unsolicited emails. The proposed framework extracts proper visual features from the spam images based on the image type (illustrated or text mainly images). The subsequent clustering associates similar images in terms of the similarity of the extracted features. Using this approach, clusters of spam used for spreading messages to encourage the purchase of a product or service through image attachments can be readily identified.

One of the major contributions of this paper is that we apply data mining and image visual analysis techniques to the field of computer forensics, which brings an innovative interdisciplinary perspective to this line of research. Our study on spam image clustering goes beyond traditional means of spam filtering and is among a few recent efforts in identifying the common source of spam. Just like the other approaches in this field, e.g., spam clustering according to common subjects and/or common IP addresses, the clustering of spam images is only one key piece of a complex jigsaw puzzle. Those techniques, when put together properly, can reveal the common source of spam and thus contribute to the capturing and impeding of those elusive cyber criminals. Another contribution is that the proposed ranked clustering algorithm provides a multimodal framework which performs clustering according to multiple features without using a predefined threshold value. The experimental results also show the effectiveness of the proposed clustering algorithm.

5. REFERENCES

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

[2] Carreras, X. and Mrquez, L. 2001. Boosting Trees for Anti-Spam Email Filtering. Inn Proceedings of the RANLP-01.

[3] Drucker, H., Wu, D., and Vapnik, V. N. 1999. Support Vector Machines for Spam Categorization. IEEE Trans. on Neural Networks, vol. 10, no. 5.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Wu, C.-T., Cheng, K.-T., Zhu, Q., and Wu, Y.-L. 2005. Using Visual Features for Anti-Spam Filtering. In Proceedings of ICIP.

[9] 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.

[10] 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.

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

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

[13] Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): pp. 91–110.

[14] Rosenberg , A., and Hirschberg , J. 2007. V-Measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), Prague, Czech Republic, pp. 410–420.