A Discriminative Classifier Learning Approach to Image Modeling and Spam Image Identification

[1]Byungki Byun, [1]Chin-Hui Lee, [2]Steve Webb, and [2]Calton Pu

[1]School of Electrical & Computer Engr.
Georgia Institute of Technology
Atlanta, GA 30332-0250, USA
+1-404-894-2901
{yorke3, chl} @ ece.gatech.edu

[2]College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280, USA
+1-404-894-3152
{webb, calton} @ cc.gatech.edu

ABSTRACT

We propose a discriminative classifier learning approach to image modeling for spam image identification. We analyze a large number of images extracted from the SpamArchive spam corpora and identify four key spam image properties: color moment, color heterogeneity, conspicuousness, and self-similarity. These properties emerge from a large variety of spam images and are more robust than simply using visual content to model images. We apply multi-class characterization to model images sent with emails. A maximal figure-of-merit (MFoM) learning algorithm is then proposed to design classifiers for spam image identification. Experimental results on about 240 spam and legitimate images show that multi-class characterization is more suitable than single-class characterization for spam image identification. Our proposed framework classifies 81.5% of spam images correctly and misclassifies only 5.6% of legitimate images. We also demonstrate the generalization capabilites of our proposed framework on the TREC 2005 email corpus. Multi-class characterization again outperforms single-class characterization for the TREC 2005 email corpus. Our results show that the technique operates robustly, even when the images in the testing set are very different from the training images.

1. INTRODUCTION

Text-based learning filters have grown in sophistication and effectiveness in filtering email spam [3, 5, 17]. In response, spammers have adopted a number of countermeasures to circumvent these text-based filters. Currently, one of the most popular spam construction techniques involves embedding text messages into images and sending either pure image-based spam or a combination of images and text (typically legitimate-looking text with legitimate content). This strategy, usually called "image-spam," has been successful in bypassing text-based spam filters, posing a new challenge for spam researchers [24].

Attempts to use optical character recognition (OCR) techniques to convert spam images back to text for processing by text-based filters have been foiled [15]. An effective response by spammers is the application of CAPTCHA[1] (**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part) techniques, which are designed to preserve readability by humans but capable of effectively confusing the OCR algorithms [7, 20]. In this paper, we explore a different image analysis approach that is capable of discerning more physical properties. Our hypothesis is that these distinctive properties can help image filters separate spam images from legitimate images.

Several approaches that extract spam images' properties have been proposed. In [24], the existence of text-regions, the number of banners and graphic images, and the location of images were extracted as spam images' distinctive properties. In [1], the existence of text-regions, saturation of color, and the heterogeneity of color were identified. In both papers, good results were reported for detecting spam images.

Despite the success of these previous studies, the most indicative properties of spam images still need to be found to enhance model discrimination performance and robustness against image variation. In the two approaches discussed above, the existence of text-regions was assumed to be the most indicative property. In this study, we analyze a number of spam images and define four specific properties: color moments, color heterogeneity, conspicuousness, and self-similarity. We use several image and signal processing techniques, such as extracting color information, clustering, and evaluating the outputs of log-Gabor filter banks, to perform feature extraction. We also exploit multi-class characterization of spam and legitimate images to account for the diversity in images that covers a wide range of visual content. In [1, 24], spam images are modeled within a single class. However, as mentioned above, images that are transmitted in emails are difficult to model. Due to these difficulties, it is desirable to adopt multi-class characterization techniques that characterize spam and legitimate images with multiple models. These models are defined as sub-classes, and in our experimental framework, spam and legitimate images are further divided into three sub-classes, respectively.

Since we use more than one model to describe spam and legitimate images, we need to define a decision rule. We propose using three decision rules: selecting the maximum score, taking an arithmetic average, and taking a geometric average. Then, an MFoM-based discriminative learning algorithm is proposed to identify spam images. With the decision rule that selects maximum scores, experiments show that our framework works better than an approach exploiting single-class characterization. Specifically, our framework yields a spam image identification

CEAS 2007 – Fourth Conference on Email and Anti-Spam, August 2-3, 2007, Mountain View, California USA

[1] http://www.captcha.net/

rate of 81.5% and a false positive rate of only 5.6%. We also test our framework on a different dataset to prove that our framework has strong generalization capabilities. Concretely, we used our framework to evaluate a test set of images that were extracted from the TREC 2005 email corpus. Although these images are quite distinct from the images used in the training set, our framework was able to maintain a high level of performance. These positive results are consistent with our hypothesis, and they also encourage further investigation into the spam images as an effective component technology for separating complex spam emails (e.g., those containing text and multimedia information) from legitimate emails.

The remainder of this paper is organized as follows. In Section 2, we define multi-class characterization in detail. In section 3, we describe distinctive properties for feature extraction. Section 4 proposes decision rules for multi-class characterization, and Section 5 explains our MFoM-based learning algorithm. With Section 6, we conclude this paper.

2. MULTI-CLASS CHARACTERIZATION

In most image modeling techniques ranging from content-based to concept-based, a single class model is learned for a single image class. For example, in image categorization problems, a single classifier is constructed to model each individual image class [4, 12, 21]. Similarly, in a concept model implementation problem, each concept will be modeled by a single classifier that characterizes the concept [2, 8]. In many other cases, this single class characterization scheme has been adopted as well.

However, in spam image identification, the above scheme may not be suitable because the variety of images that are transmitted in emails is too large to capture in a single model. A spam image might contain text messages with a uniform background; it might contain a complex background without any text messages, or it might contain any number of other variations. The diversity of legitimate images is also gigantic. Since multi-class characterization uses multiple models to characterize a single class, we believe it should be applied to help cover the wide variety of images found in emails. Thus, in our proposed framework, a spam image class and a legitimate image class are described with several models, which are called *sub-classes*.

In [1], they defined four categories ("photos," "baby," "graphics," and "screenshot") to represent legitimate images. However, multi-class characterization was not applied in their study. Each category was regarded as a separate image class – not a sub-class. Thus, instead of unified spam image identification results (i.e., spam vs. legitimate), four separate classification results (i.e., spam vs. each legitimate category) were reported.

Multi-class characterization is extremely advantageous in spam image categorization because it improves both effectiveness and robustness. With a single-class characterization method, more sophisticated class models (e.g., support vector machines) must be used to reflect the complexity of the problem. However, using a multi-class characterization method, simple classifiers can effectively describe complicated image classes. Additionally, robustness can be achieved through multi-class characterization. Robustness is a very important issue in spam image identification because the images found in email can consist of virtually anything. Multi-class characterization is intended to cover as many variations as possible by introducing sub-classes, and as a result, it provides more robustness than single-class characterization.

To implement multi-class characterization, two issues must be addressed: (1) grouping images into several sub-classes and (2) designing decision rules and classifiers. In the following section, we apply multi-class characterization to spam and legitimate images. We explain how to design decision rules and classifiers in Sections 4 and 5.

2.1 Multi-Class Characterization of Spam Images

For multi-class characterization, we prepared spam images from SpamArchive and grouped them into two groups: synthetic and non-synthetic images. Synthetic images are images made with any artificial techniques, whereas non-synthetic images are images with no artificial modifications. Typical non-synthetic images are sexual or female images, and most synthetic images are advertising pictures.

Synthetic images can be partitioned into two regions: message regions and background regions. However, some of the images do not contain message regions. Additionally, background regions are extremely diverse. For example, both uniform and complex backgrounds are represented. Therefore, synthetic images are characterized further based on two criteria: complexity of background and existence of text messages. Non-synthetic images can be characterized by their content. In particular, non-synthetic images can be grouped into sexual and non-sexual images. Based on these observations, 5 sub-classes were generated. Table 1 summarizes the multi-class characterization results for spam images.

Table 1. Multi-class Characterization of Spam Images

	Synthetic	**Text**	**Content**
Text_Simple	Yes	Yes	N/A
Text_Complex	Yes	Yes	N/A
No_Text	Yes	No	N/A
Non_Synthetic_Sexual	No	No	Sexual
Non_Synthetic_Others	No	No	Others

We can characterize spam images in many different ways. In particular, in our experiments, the bottom three sub-classes are combined and an "Other" sub-class is defined. Figure 1 shows examples for every sub-class except Non_Synthetic_Sexual because that sub-class contains pornographic content.

2.2 Multi-Class Characterization of Legitimate Images

Unfortunately, a standardized legitimate email image data set has not been made publicly available. Moreover, it is difficult to obtain enough legitimate email image data due to several reasons including copyright issues. The Enron Corpus (the largest publicly available legitimate email corpus) does not contain any attachments, and we were only able to extract 288 legitimate images from the TREC 2005 email corpus. Thus, we utilized images across the 20 different classes from Corel CDs, instead. As mentioned earlier, the goal of multi-class characterization is to

cover as many variations as possible. Therefore, we also added more images from Google Image Search[2] by searching for four keywords: "sports," "baby," "maps and directions," and "cartoons," to emulate more realistic situations. The keywords are chosen based on our observations that images belonging to such keywords are most frequently seen in legitimate emails.

The characterization of legitimate images is quite simple because metadata is available. In particular, predefined class names are the metadata for CorelCDs, and keywords are used for Google Image Search. All images from CorelCDs with "sports" and "baby" images from Google Image Search are integrated into one sub-class: "photos." The rest of the images are characterized based on their metadata: "maps and directions" and "cartoons." Figure 2 shows examples of each legitimate image's sub-class.

Figure 1. Spam image examples; (a) Text_Simple, (b) Text_Complex, (c) No_Text, (d) Non_Synthetic

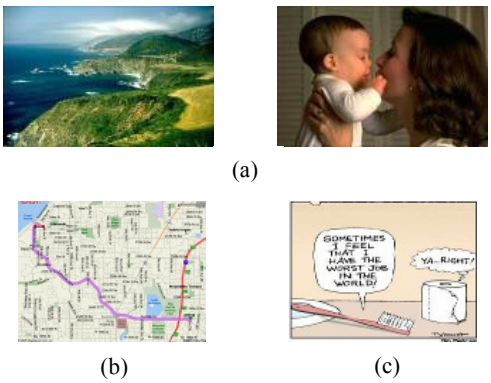

Figure 2. Legitimate image examples; (a) photos, (b) maps and directions, (c) cartoons

[2] http://images.google.com/

3. DISTINCTIVE PROPERTIES OF SPAM IMAGES

Extracting distinctive properties from spam images is a crucial part of a spam image identification task. Well-extracted features provide discrimination and robustness. In this section, we identify four key properties of spam images: color moments, color heterogeneity, conspicuousness, and self-similarity. We describe the characteristics of these properties and explain feature extraction procedures.

3.1 Color Moments

The first distinctive property of spam images is color moments. Color is one of the most widely used visual features in image retrieval problems, and it is relatively robust to background complication and invariant to image size or orientation [16].

In spam images, several notable color characteristics can be found such as discontinuous distributions, high intensity, dominant peaks, etc. Figure 3 illustrates such color characteristics, comparing color distributions of a spam image (the upper images), with those of a legitimate image (the lower images). In the upper-right image, it is easily noticeable that the distribution is not continuous and that dominant peaks exist all over the channels, while the lower-right image does not have these characteristics.

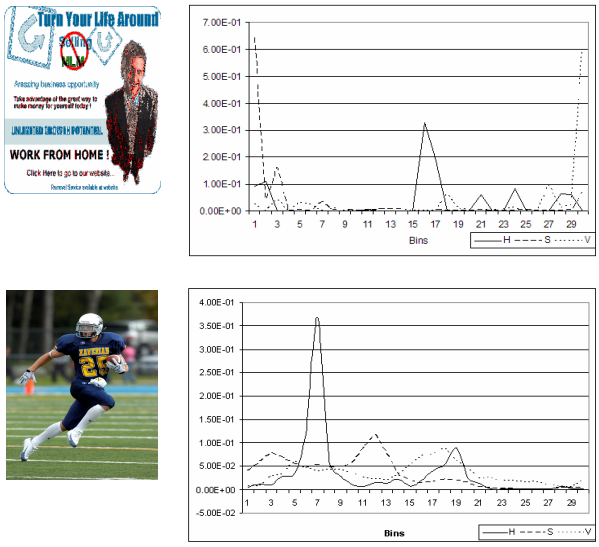

Figure 3. Color distributions of a spam image and a legitimate image in the HSV color space; focus on discontinuity and dominant peaks in the upper-right image.

The simplest way to extract these color characteristics is to use color histograms [19]. However, since most of the information is embedded in low-order moments, color moments can be used instead [18]. In our framework, the first and second central moments are computed. First, all images are transformed to the HSV color space. HSV is a color representation method that is similar to the way humans perceive color [10, 11]. Then, in the HSV color space, the first and second central moments are computed for every channel.

3.2 Color Heterogeneity

Legitimate images typically convey a much larger number of colors than spam images. Even in a person's face, large variations can be observed, which account for shades or illuminations in the face. In contrast with legitimate images, color in spam images usually stays constant. The background is generally filled with the same colors, and a single sentence usually consists of the same fonts and colors.

In [1], this property was defined as a color heterogeneity feature. They first identified text regions and non-text regions, quantizing each region with at most N colors through a minimum variance quantization algorithm. Then, RMS errors between the original images and the quantized images were calculated. We compute the RMS errors between the original images and the quantized images as well, but we do not make any separation between the text-region and non-text regions because we assume that color heterogeneity is a global feature.

Figure 4 shows four legitimate images and four spam images: (1)-(4) are legitimate, and (5)-(8) are spam. (1)-(3) are in the "photo" sub-class, and (4) is in "cartoon" sub-class. (5) and (6) are in "Text_Complex," and (7) and (8) are in "Text_Simple."

Figure 4. Legitimate and spam image examples; (1)-(4) legitimate images, (5)-(8) spam images

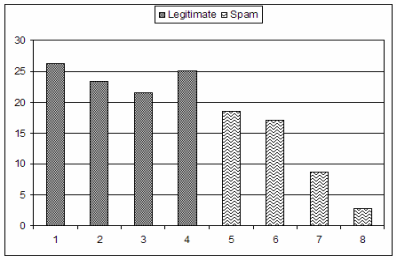

Figure 5. Color heterogeneities of images in Figure 4; N is set to 8. X-axis corresponds to the image numbers in Figure 4. Y-axis represents the computed color heterogeneities.

Figure 5 shows the above images' resulting color heterogeneity feature values. As the figure shows, all of the four legitimate images have larger values than the four spam images. Additionally, the images in the "Text_Complex" sub-class ((5) and (6)) have larger values than images in the "Text_Simple" sub-class ((7) and (8)). This observation illustrates the need for multi-class characterization. Those two sub-classes exhibit considerably different color heterogeneity values so single-class characterization would blur the discriminative characteristics of spam images.

Figure 6 plots the distributions of RMS errors of spam images and legitimate images. This figure clearly shows that most of the spam images have fewer RMS errors than most of the legitimate images.

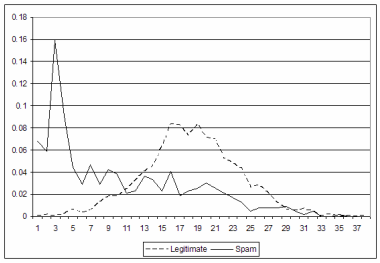

Figure 6. Distributions of color heterogeneity feature; N is set to 8. X-axis corresponds to heterogeneity values. Y-axis represents probabilities.

3.3 Conspicuousness

Spammers want spam messages to be easily noticeable to receivers so that desired actions can be generated (e.g., reading the message, clicking a link, etc.). Thus, it is natural for spam images to use highly contrasted colors. We define this property as conspicuousness, which means "obvious to the eye." In practice, we can identify many spam images that use highly saturated colors with contrasting white or black or white contrasting black. In Figure 4, image (7) uses a highly saturated yellow background color. In Figure 4, image (8) uses pure blue and red in text messages, which is contrasted with a white-like background.

This property will become manifest looking from the SV plane, which is a subspace of an HSV color space. If an image is highly conspicuous, high density is expected to be seen at white, black, and saturated colors. In the SV plane, such colors corresponded to three points: (0,1), (0,0), and (1,1), respectively. Based on this idea, translating the conspicuousness of images into a feature vector is done as follows.

First, we represent an image in the SV plane. Then, using a k-means algorithm, we learn M centroids of the image's SV distributions. With the M centroids, we compute average distances between the centroids and the above three points. In particular, we search for the closest point for each centroid and add the distance up. Finally, we average the total sum, and the average is the extracted conspicuousness.

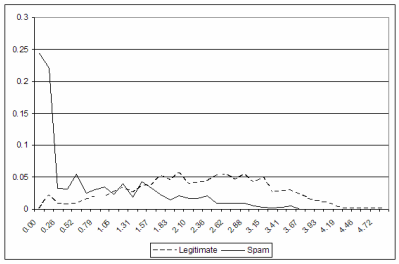

Figure 7. Distributions of conspicuousness feature; M is set to 8. X-axis represents the values of conspicuousness feature and y-axis represents probabilities.

Figure 7 shows distributions of the extracted conspicuousness feature for spam and legitimate images. This figure shows that for about half of the spam images, the extracted feature values are less than 0.13, whereas the feature values are greater than 0.13 for over 97% of the legitimate images.

3.4 Self-Similarity

In spam images, some characteristic patterns can be found from text messages or the background. These patterns produce another distinctive property of spam images called self-similarity. The reason why this feature is called self-similarity is because a similarity between macroblocks of images is measured. The term *macroblock* usually refers to a square block of MxM pixels, where M is small. The uniform background of spam images would be the simplest example that high self-similarity is measured. In spam images, the background is much more uniform than legitimate images. If the uniform background is segmented into macroblocks, only one type of macroblock will be seen, and in turn, high self-similarity will be measured. Text messages in spam images also present high self-similarity. In text-embedded spam images, the fonts and sizes that are used are typically different from sentence to sentence. However, in a single sentence, font and size are constant. Therefore, similar macroblocks will be generated, and in turn, high self-similarity would be expected. Based on our experiments, text messages in different fonts and sizes create similar macroblocks as well.

To extract self-similarity, we need to learn representative patterns from data first. We expect that noisy variations in the data would be removed while emphasizing self-similarity through this learning process. First, we segment a whole image into several macro blocks. The size of a macroblock is globally set to 32x32. When the size of an image is smaller than the size of our macroblock, the size is set to the minimum of the two. Similarity between macroblocks is computed using a log-Gabor filter bank [6, 13, 14]. Unlike a typical Gabor filter bank, log-Gabor filters can design arbitrarily large bandwidth while keeping a DC component at zero [13]. We apply a log-Gabor filter bank to each macroblock and compute a mean and a variance of each filter bank's output. We then exploit an unsupervised clustering algorithm such as a k-means clustering algorithm. Centroids of the resulting clusters become the representative patterns.

Now, images are indexed with these patterns. First, the patterns are numbered with 1 to N, where N is the number of centroids, and each of the macroblocks is indexed with the closest pattern. We then count how often each pattern has appeared in an image. This will generate N dimensional vectors. We finally normalize these vectors as the sum of each element equals to 1. These become the extracted self-similarity vectors.

4. MULTI-CLASS DECISION RULES

In a typical spam image identification problem, for which single-class characterization is used, the decision rule is fairly simple: choose a class that maximizes certain scores. Let X be a test vector, and let C_+ and C_- represent a spam image class and a legitimate image class, respectively. Moreover, let Λ_+ and Λ_- be the corresponding parameter sets for C_+ and C_-. Then, the decision rule can be described in Eq (1)

$$\begin{aligned} &\text{Choose } C_+ \text{ if } g(X,\Lambda_+) > g(X,\Lambda_-), \\ &\text{Choose } C_- \text{ otherwise,} \end{aligned} \tag{1}$$

where $g(X,\Lambda_+)$ and $g(X,\Lambda_-)$ are class score functions for positive and negative classes, respectively.

Multi-class characterization broadens the choices of decision rules in spam image identification. Eq. (1) still functions as an overall decision rule; however, since multiple sub-classes are used to describe an image class, class score functions for a spam and legitimate class, $g(X,\Lambda_+)$ and $g(X,\Lambda_-)$ can be computed differently.

The first method is to assign the maximum scores between sub-classes to $g(X,\Lambda_+)$ and $g(X,\Lambda_-)$.

$$\begin{aligned} g(X,\Lambda_+) &= \max_{1\le j\le M}(g(X,\Lambda_{+j})), \\ g(X,\Lambda_-) &= \max_{1\le j\le N}(g(X,\Lambda_{-j})). \end{aligned} \tag{2}$$

Here, we let $g(X,\Lambda_{+j})$ and $g(X,\Lambda_{-j})$ be the sub-class's score functions, where Λ_{+j} and Λ_{-j} are the parameter sets for sub-class C_{+j} and C_{-j}. C_{+j} is the j[th] sub-class of a spam image class C_+, and C_{-j} is the j[th] sub-class of a legitimate image class C_-.

Another way is to compute arithmetic averages of the sub-classes, which is defined as:

$$\begin{aligned} g(X,\Lambda_+) &= \frac{1}{M}\sum_{j=1}^{M} g(X,\Lambda_{+j}), \\ g(X,\Lambda_-) &= \frac{1}{N}\sum_{i=1}^{N} g(X,\Lambda_{-i}), \end{aligned} \tag{3}$$

where M and N are the numbers of sub-classes for spam images and legitimate images, respectively.

The third approach is to compute geometric averages of the sub-classes. As an example, the equation to compute a geometric average for a spam image class is shown in Eq. (4)

$$g(X,\Lambda_+) = \log\left[\frac{1}{M}\sum_{j=1}^{M}\exp\{g(X,\Lambda_{+j})\}^{\eta}\right]^{\frac{1}{\eta}}, \tag{4}$$

where η is a positive constant.

Since different decision rules yield different decision boundaries, the performances of above methods should vary. The first method concerns the most relevant sub-classes in a spam and a legitimate image class. The second and third methods measure average scores. We determine the best decision rule empirically, and in section 6, the performances of above three approaches are measured and compared against each other.

5. MFOM-BASED CLASSIFIER LEARNING

In classifier learning with multi-class characterization, we estimate parameters Λ for classifiers, given training data $\{X,C\}$, with $\Lambda = \{\Lambda_{+j},\Lambda_{-i} \mid 1\le j\le M, 1\le i\le N\}$ and $C = C_+ \cup C_-$. Here, we adopt an MFoM-based learning approach proposed in [9]. In an MFoM-based learning approach,

the classifier parameters Λ are estimated by directly optimizing a certain performance metric of classifiers (e.g., detection errors, F1, or precision).

A typical performance metric used is an average detection error rate (DER). The DER is defined as an arithmetic average of false positive rate and false negative rate:

$$DER = \frac{1}{2N}\left\{\sum_i \frac{FP_i}{T-|C_i|} + \frac{FN_i}{|C_i|}\right\}, \quad (5)$$

where N is the number of classes, T is the total number of training data, $|(\cdot)|$ is a cardinality, and FP_i and FN_i are false positive error and false negative error for the i^{th} class.

However, we cannot optimize this performance metric analytically since FP_i and FN_i are discrete entities. Therefore, in MFoM-based learning, a continuous and differentiable function, called a class loss function $l_i(X,\Lambda)$, is introduced to approximate FP_i and FN_i for class i. Since both FP_i and FN_i are computed from the error counts made from class i, a class loss function should simulate these error counts. In particular a class loss function should be close to zero when X is correctly classified and should be close to one when X is misclassified. Summing over the entire training data set, FP_i and FN_i can be approximated with $l_i(X,\Lambda)$ as:

$$\begin{aligned} FP_i &= \sum_X (1-l_i(X,\Lambda))I(X \notin C_i), \\ FN_i &= \sum_X l_i(X,\Lambda)I(X \in C_i), \end{aligned} \quad (6)$$

where $I(\cdot)$ is an indicator function.

For choosing a class loss function, any functions satisfying the above behavior would work. Typically, a sigmoid function is adopted, defined as:

$$l_i(X,\Lambda) = \frac{1}{1+\exp\{-\alpha(d_i(X,\Lambda)+\beta)\}}, \quad (7)$$

where α and β are all positive constants, serving as parameters that determine the size of the learning window and the offset of decision boundary, and $d_i(X,\Lambda)$ is a function of X and Λ, called a class misclassification function.

Defining a class misclassification function is the most important consideration in an MFoM-based classifier learning approach. $d_i(X,\Lambda)$ should be consistent with the class loss function's behavior. In particular, $d_i(X,\Lambda)$ should be negative for correct decisions and positive for incorrect ones. Using the same notation used in Section 4, we can defined $d_i(X,\Lambda)$ as:

$$d_i(X,\Lambda) = -g(X,\Lambda_i) + g(X,\Lambda_i^-), \quad (8)$$

where Λ_i^- is the parameter set of a competing class for class i, which is Λ_- for C_+ or Λ_+ for C_-, and $g_i(X,\Lambda)$ is constructed by one of the methods discussed in Section 4. It is easy to check that the decision rule in Section 4 is equivalent to the following:

$$\begin{aligned} &\text{Choose } C_i \text{ if } d_i(X,\Lambda) < 0, \\ &\text{Choose } C_i^- \text{ otherwise,} \end{aligned} \quad (9)$$

where C_i^- represents a competing class for class i, which is C_- for C_+ or C_+ for C_-. If a classifier is assumed to be perfect, $l_i(X,\Lambda)$ becomes close to zero when $X \notin C_i$; otherwise, $l_i(X,\Lambda)$ becomes close to one.

In sum, a DER is now approximated with a continuous and differentiable function as a result. The approximated function now becomes our objective function. Summing over all classes, the overall objective function $L(T,\Lambda)$ can be written as follows:

$$L(T,\Lambda) = \frac{1}{2N}\left\{\sum_i \sum_X \frac{(1-l_i(X,\Lambda))I(X \notin C_i)}{T-|C_i|} + \frac{(l_i(X,\Lambda))I(X \in C_i)}{|C_i|}\right\}, \quad (10)$$

The similar approaches can be applied for any other performance metrics such as precision, recall, or F1 to the corresponding construct objective functions. In our framework, a DER is the preferred metric. Finally, the maximization of the objective function $L(T,\Lambda)$ can be done by a generalized probabilistic decent algorithm.

One of the important properties of MFoM-based learning approaches is that a relative distance between class i and its competing classes will be maximized when the performance metric is optimized. To see how this distance is maximized, let us look at the class misclassification function $d_i(X,\Lambda)$ for class i. It is easy to see that the absolute value of $d_i(X,\Lambda)$ is the separation between class i and its competing class. Since a bigger separation implies a smaller error, the value of $d_i(X,\Lambda)$ will be increasing as the objective function is maximized. This might imply the robustness of an MFoM-based classifier learning approach.

6. EXPERIMENTAL RESULTS

We extracted 90,170 image files from the SpamArchive spam corpora and used them as our spam images. Unfortunately, most of the extracted files were malware propagation vehicles (i.e., malware embedded in a file with an image file suffix). Many of the files also contained formatting errors. After eliminating these malformed files and repeated images, the data set contained 669 distinct images. We will refer to these images as our SpamArchive image data set. For legitimate images, we obtained 1267 images from Corel CDs and 358 images from Google Image Search. We will refer to these images as the C+G image data set.

In Section 2, spam images are characterized with 5 sub-classes, and legitimate images are characterized with 3 sub-classes. Through several preliminary experiments, we decided to integrate three sub-classes ("Non_Text," "Non_Synthetic_Other," and "Non_Synthetic_Sexual") into one sub-class ("Others"). Then, we were left with a total of 6 sub-classes: 3 for spam images and 3 for legitimate images. The following is a summary of the resulting multi-class characterization used and the numbers of images that belong to each sub-class:

a) The SpamArchive image data set
   Text_Simple : 360 images,
   Text_Complex : 141 images,
   Others : 168 images

b) The C+G image data set
Photos : 1404 images,
Cartoons : 115 images,
Maps : 105 images

Next, we determined the parameters for feature extraction. For the color heterogeneity feature, we define the number of colors used for quantization to be 8. The number of centroids is also set to 8 for the conspicuousness feature. When extracting the self-similarity feature, we set the number of patterns to be equal to 64.

Feature fusion is done by concatenating four features directly, which yields a 72-dimensional feature vector. To decide the best feature fusion method, we carried out classification experiments with two other feature fusion methods, which are an MBT approach [23] and a weighted sum approach [22], and compared the results. Among the three feature fusion methods, concatenating four features outperforms the other two methods consistently.

For an MFoM-based learning approach, a linear discriminant function is adopted for a sub-class' scoring functions $g(X,\Lambda_{+j})$ and $g(X,\Lambda_{+i})$. A linear discriminant function is defined as:

$$g(X,\Lambda) = w^T \cdot X + b, \qquad (11)$$

where w and b are parameters. In our framework, therefore, w_{+j} and b_{+j} for the j[th] sub-class of spam images and w_{-i} and b_{-i} for the i[th] sub-class of legitimate images are going to be estimated. α and β are determined empirically in a way that for initial iteration, the objective function fluctuates and then converges subtly. For different decision rules, different α and β should be used.

6.1 Comparison of Decision Rules

To determine the best decision rule for spam image identification with multi-class characterization, we evaluated the three decision rules specified in Section 4 using the SpamArchive and C+G datasets. First, the class misclassification function $d_i(X,\Lambda)$ is derived from each of the decision rules, and each MFoM-based classifier is trained accordingly. 10% of the images are randomly chosen for testing, and the rest of the images are used in the training stage. We repeat this random selection 10 times and average all of the results. Table 2 gives the results for the first decision rule – selecting the maximum values (D1), the second decision rule – arithmetic average (D2), and the third decision rule – geometric average (D3).

Table 2. Comparison of Decision Rules

	Spam Image Identification Rate(%)	False Positive(%)
D1	**81.5**	5.6
D2	77.3	7.3
D3	80.3	**5.3**

This table shows that the first decision rule (D1) achieves the best for identifying spam images, and the third decision rule (D3) has the lowest false positive rate. The second decision rule (D2) performs worst, compared to the other two rules. For overall performance, the first decision rule appears to be the best for spam image identification with multi-class characterization.

6.2 Multi-Class vs. Single-Class Characterization

In this section, we compare multi-class characterization approaches and single-class characterization approaches. For a single-class characterization approach, the decision rule defined in Eq. (1) is adopted. For the class discriminant functions $g(X,\Lambda_+)$ and $g(X,\Lambda_-)$, a linear discriminant function, specified as in Eq. (11), is used, and an MFoM-based classifier is trained. For a multi-class characterization approach, the results are borrowed from Section 6.1, which were obtained using the first decision rule (D1). Table 3 shows the comparison between single-class characterization and multi-class characterization.

Table 3. Multi-Class vs. Single-Class

	Spam Image Identification Rate(%)	False Positive(%)
Single-Class	79.3	12.3
Multi-Class	**81.5**	**5.6**

This table shows that a multi-class characterization approach works better than a single-class characterization. The improvement in false positive rate is quite significant, and the spam image identification rate improvement is also noticeable. Therefore, as claimed, multi-class characterization is more effective and suitable for spam image identification. Additionally, since low false positive rates are typically more important than high spam identification rates in spam image identification, we can achieve another preferred characteristic through multi-class characterization.

6.3 Comparison with Other Techniques

It is impossible to compare other previously proposed techniques [1, 24] with our proposed framework directly. The main reason concerns differences in the datasets that were used. In [1], legitimate images were retrieved from Google Image Search, and in [24], legitimate images only consisted of images selected from CorelCDs.

However, it is possible to position our proposed framework by analyzing those other techniques' results. Table 4 shows the performance of [1]. In that work, single-class characterization was used. Since only pair-wise performances were reported, we average all of their performances and compute an overall performance. They also used two distinct datasets, SPAM-1 and SPAM-2, for spam images.

Table 4. Spam Image Identification Result – (Aradhey, et al., 2005)

Dataset	Spam Image Identification Rate(%)	False Positive(%)
SPAM-1	76.25	6.5
SPAM-2	82.75	17.25
Average	79.5	11.875

Table 5 gives the results in [24]. In that work, three classification results, depending on SVM parameters, were reported. Among

those results, we take a classification result most widely used in [24]

Table 5. Spam Detection Results – (Wu, et al., 2005)

Spam Image Identification Rate(%)	False Positive(%)
81.4	0.98

Based on the figures in the above two tables, we observe that our proposed framework generally obtains better results and sometimes comparable results. Since we use multiple sources to construct our datasets, whereas other techniques relied on a single source such as either CorelCDs or an image search engine, we believe that our training and test datasets are more realistic and more difficult to model. Therefore, we believe that our proposed framework works as well as (if not better than) previous techniques.

6.4 Generalization Capabilities

In spam image identification, generalization is rather important because the variation of images is immense. We test generalization capabilities of our approaches with completely different datasets. In the training stage, the SpamArchive and C+G image data sets, totaling 2461 images, are used. In the testing stage, images extracted from the TREC 2005 email corpus are used. 1,249 spam image files and 288 legitimate image files were extracted and processed using our normalization processes. Unfortunately, these data sets contained overlapping images (i.e., images found in both legitimate and spam emails), which we were forced to remove. From our observations, most overlapping images were either icons or background templates. 171 spam images and 152 legitimate images were retained, and the results of our analysis are shown in Table 6.

Table 6. Spam Image Identification Results on TREC 2005 Corpus

	Spam Image Identification Rate(%)	False Positive(%)
Single-Class	84.8	25
Multi-Class	**86.6**	**19.1**

In Table 6, we observe that our false positive rate increased by 13.5%. One explanation for this increase is the difference between C+G images and TREC 2005 legitimate images. This result clearly reflects the need for a standardized data set for legitimate email images. On the other hand, we observed improvements in the spam image identification rate, compared with the results in Table 4. This proves the generalization capabilities of our proposed framework for identifying spam images.

We also observe that multi-class characterization achieves better than single-class characterization, even for the TREC 2005 email corpus. False positive rate is improved by 5.9% and spam image identification rate is increased by 1.8% as well. The improvement in false positive rate is more eminent as observed in Section 6.2. If we recall that the legitimate images used for training are not from real emails, it can be claimed that a multi-class characterization approach is much advantageous for spam image identification tasks since it is very difficult to obtain appropriate training data.

7. CONCLUSION AND FUTURE WORK

In this paper, we propose a framework for spam image identification that applies multi-class characterization and MFoM-based learning. We identify four key properties of spam images: color moments, color heterogeneity, conspicuousness, and self-similarity. By combining multi-class characterization and these four properties for feature extraction, experimental results show that our proposed technique is effective. For our SpamArchive spam image data set, we obtain a spam image identification rate of 81.5% and a false positive rate of only 5.6% for the legitimate images. The robustness of the technique is demonstrated by an experiment using the same classifier (without retraining) on the images in the TREC 2005 email corpus. Our approach achieves 86.6% spam image identification with 19.1% false positives. The higher false positive rate is due to the differences between our legitimate corpus used for classifier training and the TREC 2005 data set. However, with a comparison between single-class characterization and multi-class characterization for the TREC 2005 email corpus, we achieve a much better false positive rate when a multi-class characterization approach is used.

Our results are consistent with the hypothesis that spam images are fundamentally different from normal images, and therefore, we can apply image analysis techniques to distinguish them. This is an interesting and encouraging result since early attempts to apply OCR algorithms to convert images to text have been defeated by spammers applying simple CAPTCHA techniques.

Our approach is a component technology that can be combined with text-based filters to improve the recall and precision of filtering text-and-multimedia spam messages. Beyond email, our techniques are also applicable in other spam application areas such as web spam, blog spam, and instant messaging spam, where non-text (multimedia) data has been growing in volume and importance.

We are currently working on refinements for our multi-class MFoM-based characterization of spam images. For example, additional properties and a more refined class categorization may improve the precision and recall of the method. On the application side, incorporating our approach with text-based learning filters for email spam filtering is a natural next step. Other spam media such as web spam are also interesting but relatively unexplored areas. Another area that will improve our research and the performance of our approach is the availability of large-scale data sets for both spam images and legitimate images.

8. ACKNOWLEDGMENTS

This work has been supported by grants from the Air Force Office of Scientific Research. The authors would like to thank Filippo Vella who initially built up our image processing systems. We also thank anonymous reviewers for their insightful feedback and comments.

9. REFERENCES

[1] Aradhey, H. B., et al., Image Analysis for Efficient Categorization of Image-based Spam E-mail, *Proc. of ICDAR*, 2005.

[2] Carbonetto, P., Freitas, N., and Barnard, K., A Statistical Model for General Contextual Object Recognition, *Proc. of ECCV,* 2004.

[3] Carreras, X. and Mrquez, L., Boosting Trees for Anti-Spam Email Filtering, *Proc. of RANLP-01*, 2001.

[4] Chen, Y. and Wang, J., Image Categorization by Learning and Reasoning with Regions, *The Journal of Machine Learning Research*, vol. 5, pp. 913-939, 2004.

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

[6] Field, D. J., Relations between the statistics of natural images and the response properties of cortical cells, *Journal of Optical Society of America*, vol.4, No. 12., 1987

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

[8] Gao, S., Wang, D. –H., and Lee, C. –H., Automatic Image Annotation Through Multi-Topic Text Categorization, *Proc. of ICASSP*, vol. 2, pp. 377-380., 2006.

[9] Gao, S., Wu, W., Lee, C. –H., and Chua T. –S., A MFoM Learning Approach to Robust Multiclass Multi-Label Text Categorization, *Proc. of ICML*., 2004

[10] Gonzalez, R. C. and Woods, R. E., Digital Image Processing 2ed, *Prentice Hall Press*, pp. 295., 2002.

[11] Gonzalez, R. C., Woods, R. E., and Eddins, S. L., Digital Image Processing using Matlab, *Prentice Hall Press*., 2004.

[12] Hu, J. and Bagga, A., Categorizing Images in Web Documents, *IEEE Multimedia*, vol. 11, no. 1, pp. 22-30., 2004.

[13] Kovesi, P. D(1999), "Image Feature from Phase Congruency," *Videre: Journal of Computer Vision Research*, The MIT Press, vol. 1, no. 3., 1999.

[14] Kovesi, P. D., MATLAB and Octave Functions for Computer Vision and Image Processing, *http://www.csse.uwa.edu.au/~pk/research/matlabfns*, 2000.

[15] Leavitt, N. (2007), Vendors Fight Spam's Sudden Rise, *IEEE Computer*, vol. 40, no. 3, pp 16-19., 2007.

[16] Rui, Y., Huang, T. S., and Chang, S. –F. (1999), Image Retrieval: Current Techniques, Promising Directions and Open Issues, *Journal of Visual Communications and Image Representation*, vol. 10, no. 4, pp. 39-62., 1999.

[17] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E., A Bayesian Approach to Filtering Junk Email, *AAAI Workshop on Learning for Text Categorization*., 1998.

[18] Stricker, M. and Orengo, M., Similarity of Color Images, *Proc. of SPIE,* pp 381-392., 1995.

[19] Swain, M. and Ballard, D., Color Indexing, *Int. Journal of Computer Vision*, vol.1, pp 11-32., 1991.

[20] Symantec Corp., The State of Spam, *A Monthly Report – January*, 2007.

[21] Szummer, M. and Picard, R., Indoor-Outdoor Image Classification, *IEEE Workshop on CAIVD*, pp. 42., 1998.

[22] Vella, F., et al., Information Fusion Techniques for Automatic Image Annotation, *Proc. of VISAPP*., 2007.

[23] Wang, D. -H., et al., Discriminative Fusion Approach for Automatic Image Annotation, *Proc. of MMSP*., 2005.

[24] Wu, C. -T., Cheng, K. -T., Zhu, Q., and Wu, Y. -L., Using Visual Features For Anti-Spam Filtering, *Proc. of ICIP*., 2005.