

# Image Analysis for Efficient Categorization of Image-based Spam E-mail

Hrishikesh B. Aradhye, Gregory K. Myers, James A. Herson  
SRI International, Menlo Park, CA 94025, USA  
hrishikesh.aradhye@sri.com

## Abstract

To circumvent prevalent text-based anti-spam filters, spammers have begun embedding the advertisement text in images. Analogously, proprietary information (such as source code) may be communicated as screenshots to defeat text-based monitoring of outbound e-mail. The proposed method separates spam images from other common categories of e-mail images based on extracted overlay text and color features. No expensive OCR processing is necessary. Our method works robustly in spite of complex backgrounds, compression artifacts, and a wide variety of formats and fonts of overlaid spam text. It is also demonstrated successfully to detect screenshots in outbound e-mail.

## 1. Introduction

As the usage of electronic mail (e-mail) and cellular text messages continues to increase, so does the volume of unsolicited commercial communications (or “spam”) being sent to e-mail and text message users. The volume of spam has long been viewed as a damning threat to the utility of e-mail and text messaging as effective communication media, prompting many proposed solutions to combat the reception of spam. Among these solutions are systems that accept communications only from pre-approved senders and/or formats, or filters that search the text of incoming communications for keywords generally indicative of spam.

Unfortunately, the senders of spam are finding ways to circumvent such systems. For example, one way in which senders have attempted to thwart key-word-based text search systems is to overlay text of the intended message on images linked to HTML formatted e-mail (Fig. 1), so that its content remains perceptible to the viewer and at the same time is shielded from the text-based spam filters. Traditional anti-spam techniques, which typically ignore embedded imagery or perform limited comparisons based on a hash of still-image data, are thus ineffective to combat this approach. Spam e-mails containing imagery account for an estimated 25% of all spam sent today, and this number is expected to increase unless a viable solution is found to counter such communications. Therefore, an automated image-spam detection and categorization method is highly desirable.

The proposed work presents a fast and efficient method for categorization of spam e-mails containing imagery (or links to images) for the purposes of filtering or categorizing the communication. As used herein, the term “spam” refers to any unsolicited electronic communications, including advertisements and communications designed for “phishing” (e.g., designed to elicit personal information by posing as a legitimate institution such as a bank or internet service provider), among others. The proposed methods can also be used for monitoring of outbound e-mails by corporations to detect communications including proprietary or company confidential material (such as screenshots of source code).



(a) Complex backgrounds



(b) Oblique angles



(c) A variety of colors, sizes, fonts, and styles

Figure 1. Examples of spam images and the prevalent complexity of overlaid text



**Figure 2. Illustrative categories of non-spam images**

The method presented herein exploits the text-intensive nature of the imagery involved in the application scenarios discussed above, when compared with the other prevalent categories of images included in e-mail communications (Fig. 2). The basic premise of our method is that the *extent* of text on spam imagery, the text's *visual appearance*, and its *content* all serve as the leading indicators of a spam image, in spite of the complexity and wide variety of such text (Fig. 1). To this end, we use our previously published work toward the detection, rectification, and recognition of text embedded in digital photographs (such as street signs and name plates) as well as that overlaid graphically (such as headline news and captions in broadcast news footage). The following sections lay out the groundwork of our approach, its technical content, and results.

## 2. Related work

### 2.1. Image categorization

The task at hand may be considered as a special case of the generic problem of image categorization, which has several important applications such as information extraction, web mining, web page summarization, and mobile access—and as such has been of great research interest. Only a few relevant prior publications are discussed here. Many of the

generic image categorization methods are based on color histograms and frequency domain analyses, which provide capability to distinguish between images of two distinct classes. A representative application is indoor vs. outdoor classification [1]. Related work by Frankel et al. [2] to separate natural vs. synthetic images and the work by Wang et al. [3] to separate graph vs. photograph images are more relevant to the task at hand. Wang et al. used high frequency wavelet coefficients of image sub-blocks. Frankel et al. proposed several color-based features such as the degree of color saturation and the number of dominant colors. The extensions to this work by Hu and Bagga [4] used additional frequency domain features based on DCT coefficients to achieve the natural vs. synthetic separation. Gavilan et al. [5] use color quantization followed by successive classification of resulting regions using neural networks to distinguish among natural, artificial, portrait, or text images. However, spam images often contain *all* of these natural and synthetic elements (Fig. 1). Therefore, the methods discussed above are not sufficient for the present purpose.

### 2.2. Overlay text detection

Text overlaid on images and videos is intended to convey specific information to the viewer, and as such it is desirable to reliably locate and recognize such text for many different domain applications. A recent survey shows that 50% of all web images contain text [6]. However, detection of overlaid text in such video and still imagery is a difficult problem because of the involved complexity of form, content, and low resolution of the text in compressed images such as JPEGs, and the unconstrained nature of the background (Fig. 1). Therefore, the problem of overlay text detection and extraction has attracted recent research interest in the image processing and multimedia analysis community. Previous work on the extraction of text regions overlaid on images can be roughly classified into component-based, edge-based, and texture-based methods.

Component-based methods [7] model text characters as monochrome regions that are extracted using segmentation or color clustering. Heuristic constraints on size and height-to-width ratios are then applied to refine detected text components. Edge-based methods [8] exploit the frequent occurrence of vertical edges in roman text, which are detected and connected using a smoothing filter and linking. Texture-based methods [9] propose algorithms to characterize the texture of a block of text (using, for example, wavelets or spatial variance), followed by a classification step.

Our text detection and location process is in part component-based as well as edge-based, and has previously been discussed in prior publications by the authors. A discussion of the relative merits and disadvantages of our approach to text detection is beyond the scope of this paper, since any other text detection method could alternatively be used for the present purpose without significantly affecting the proposed feature extraction and decision-making steps.

### 3. Approach

The proposed spam image categorization approach detects vertically oriented edge transitions and connected components of similar intensity in a gray-scale image, and links those that are compatible in size and relative position to form lines of text. It works in three main steps. The text regions in the image, if any, are first extracted. The subsequent step defines a small number of reliable spam-indicative features from the image, using in part the extracted text regions. Subsequently, support vector learning is applied to make a spam-non-spam decision for each image. An aggregation step may be required to make a single categorization for an e-mail containing many sub-images and/or attachments. The following sections describe each of these steps in detail.

#### 3.1. Text region extraction

Our method of text detection assumes that the characters in a line of text are of approximately the same intensity, and that there is sufficient resolution to separate individual characters as components. The gray-scale image is first thresholded at  $N$  thresholds that uniformly divide the intensity range [0-255] (we used  $N=8$ ). Connectivity analysis is performed on each of the  $N$  thresholded images. Based on the extent of vertical overlap between neighboring blobs, collinear, character-like blobs (as determined by their sizes relative to their spacing) are linked together at each of the  $N$  levels to form candidate text lines. In any of the thresholded images, characters within the same word may appear merged if the background varies or the character intensity is on the borderline of a threshold. Therefore, before linking, each blob may be replaced by multiple smaller blobs, based on the position overlap of blobs at an adjacent threshold level. Candidate lines of text at larger rotation angles often appear as broken pieces of text, since the linking is based on vertical overlap. Therefore, extensions of candidate text lines are investigated for potential merges with pieces of text along the same lines.



Figure 3. Examples of spam text detection results

The blob-linking process is performed twice to detect text lines of both polarities. The distances between adjacent blobs are typically less than those between vertical edges. This allows us to impose tighter tolerances on the linking to prevent the false linking of non-text to the ends of text regions. Also, this text detection method was designed to detect both in-scene text and superimposed graphical text, and is therefore capable of handling plain backgrounds (such as the backgrounds for street signs) as well as cluttered backgrounds (such as the background imagery of advertisements and superimposed text). The detection parameters were tuned to result in few false positives, so that the end-user content-based indexing and query system is provided with only those regions that are highly likely to be true text lines. Illustrative performance of our method can be seen in Figure 3.

#### 3.2. Spam-indicative feature extraction

Once the text regions in an image have been extracted, we proceed with characterizing the image in terms of a few simple features that are most indicative of spam images. The features are briefly described below, all of which are real numbers in the range [0,1].

1. **Extent of text feature:** The extent of text in the image is defined as the fraction of the total area of the image that falls within the extracted text

regions. Text may be inherently present in natural scene images in the form of road signs and logos (Fig. 2(a)–(b)). Synthetic images (such as graphics and maps, Fig. 2(c)–(d)) may include text as well. However, the extent of text features as defined above is intuitively expected to be higher for spam.

2. **Color saturation features:** As defined by Frankel et al. [2], color saturation is quantified as the fraction of the total number of pixels in the image for which the difference  $\max(R,G,B) - \min(R,G,B)$  is greater than some threshold  $T$  (set to 50 by Frankel et al. [2]; Hu and Bagga [4]; and in this work). We evaluate this fraction for both text and non-text parts of the image separately, leading to two color saturation features. When compared with images of natural scenes, we expect the spam images to be generally more saturated due to the presence of synthetic graphics. However, when compared with generic computer-generated graphics images, we expect the spam images to be less saturated due to the presence of natural elements.
3. **Color heterogeneity features:** First, the original color image is scaled by the maximum possible intensity such that the intensities in the RGB channels are within the range  $[0,1]$ ; it is then converted to an indexed image using minimum variance quantization such that the number of colors in the indexed image is at most  $k$ . The RMS errors between the original image and the indexed image are then calculated individually for the text and non-text parts of the image, which form our two color heterogeneity features. We chose  $k=10$  for non-text portions and  $k=8$  for text portions. Since the graphical portions of spam images comprise far fewer colors than natural scene images, we expect their heterogeneity to be lower. However, other relevant categories of images, such as screenshots (Fig. 2(d)), often consist of even fewer colors, leading to much lower heterogeneity for their non-text regions.

### 3.3. Support vector learning

Given the above 5 features extracted from several training spam and non-spam images, we used Support Vector Machines (SVMs) to learn to differentiate among the classes of interest. The advantages of SVMs include ease of parameter selection and configuration, and global optimization. We used the  $SVM^{light}$  [10] tool to design our classifiers, each of which separates spam images from non-spam images of one of four specific categories of interest: natural photographs (indoor and outdoor), baby pictures, graphics/maps, and

screenshots. We thus used the one-vs-one scheme of representing a multi-class classification problem as a set of binary decisions. A polynomial kernel with degree 2 gave the best results. As a side experiment, we also trained analogous classifiers to separate screenshot images from natural photographs, baby pictures, and graphics to assist outbound mail filtering of proprietary information.

### 3.4. Evidence aggregation

Spam images are often broken into several sub-images which are arranged as a single composite image using HTML tables. Alternatively, the spam e-mail may contain several separate images. Therefore, it may be necessary to aggregate the image-based evidence to form a single spam–non-spam categorization for a given e-mail. We first compute a composite feature vector as an area-based weighted average of the feature vectors of individual sub-images:

$$f_{agg} = \sum_{i=1}^M A_i f_i / \sum_{i=1}^M A_i, \text{ where } f_{agg} \text{ is the}$$

aggregated feature vector,  $f_i$  is the feature vector of the  $i^{\text{th}}$  sub-image,  $A_i$  is the area of the  $i^{\text{th}}$  sub-image, and  $M$  is the total number of sub-images. The aggregated feature vector is then subjected to the SVM-based decision-making process.

## 4. Experimental results

We monitored image-based spam e-mail incoming at a server. As noted before, a given e-mail may contain multiple images and/or sub-images. Not all sub-images per e-mail contain any relevant message or graphics; some exist only for formatting or other purposes. We arranged the images in two distinct datasets. Images in the first dataset (SPAM-1) recorded which of the constituent sub-images belonged to the same e-mail. It contained 497 images/sub-images collected from 130 unique e-mails. The other dataset (SPAM-2) did not specifically record the sub-image information and contained 1245 images collected from several hundred e-mails. The non-spam images in our dataset were collected by querying the Google-images search engine with the search words “photo”, “baby”, “graphics”, and “screenshot”, resulting in 711, 186, 306, and 283 images, respectively. Note that the search results were not perfect, and had to be manually verified for relevance with the query word. Also, the four non-spam categories are admittedly rather

arbitrary. We hope that a large corpus of real spam and non-spam emails is made available in the future to facilitate such experiments.

The results of our experiments with five-fold cross-validation are summarized in Tables 1 and 2. For the SPAM-1 dataset in Table 1, the evidence aggregation step was carried out as well. The results indicate that it is indeed possible to filter spam and sensitive outbound e-mail using the methods proposed in this work. The false-positives vs. false-negatives trade-off can be further adjusted using the cost adjustment parameters available in *SVM<sup>light</sup>*.

## 5. Conclusion

A method for efficient categorization of two types of image-based e-mails of tremendous practical interest is presented. Our method is largely based on the extraction of text regions in the images of interest and subsequent feature analysis and support vector classification. Incoming spam e-mail was separated from four common classes of non-spam incoming e-mails with image attachments without the use of expensive OCR processing. Analogously, to prevent image-based communication of proprietary material, the proposed method was successfully applied for the detection of screenshots in outbound e-mail. The success of our method for these tasks is encouraging and, in our opinion, merits further investigation.

## References

[1] M. Szummer, and R.W. Picard, "Indoor-Outdoor Image Classification," In *Proc. IEEE Intl. Workshop on Content-*

*Based Access of Image and Video Databases*, 1998, pp. 42–51.

[2] C. Frankel, M. Swain, and V. Athitsos, "Webseer: An Image Search Engine for the World Wide Web," Univ. of Chicago Technical Report TR96-14, 1996.

[3] J.Z. Wang, J. Li, and G. Wiederhold, "SIMPLicity: Semantics-sensitive Integrated Matching for Picture Libraries," *IEEE T PAMI*, 23(9), 2001, pp. 947–963.

[4] J. Hu, and A. Bagga, "Categorizing Images in Web Documents," *IEEE Multimedia*, 11(1), 2004, pp. 22–30.

[5] D. Gavilan, H. Takahashi, and M. Nakajima, "Image Categorization Using Color Blobs in a Mobile Environment," *Computer Graphics Forum (EG 2003)*, 22(3), 2003, pp. 427–432.

[6] T. Kanungo, and C. Lee, "What Fraction of Images on the Web Contain Text?," In *Proc. of the First Intl. Workshop on Web Document Analysis*, 2001, pp. 43–46.

[7] R. Lienhart, and W. Effelsberg, "Automatic Text Segmentation and Text Recognition in Video Indexing," *ACM/Springer Multimedia Systems*, Vol. 8, 2000, pp. 69–81.

[8] T. Sato, T. Kanade, E.K. Hughes, and M.A. Smith, "Video OCR for Digital News Archives," in *IEEE Workshop on Content Based Access of Image and Video Databases*, 1998.

[9] H. Li, D. Doermann, and O. Kia, "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing*, 9(1), 2000, pp. 147–156.

[10] T. Joachims, "Making Large Scale SVM Learning Practical," In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*, MIT Press, 1999.

**Table 1. Experimental Results: Detection of Spam in Incoming E-mail**

True Class	SPAM-2	Photo	SPAM-2	Baby	SPAM-2	Graphics	SPAM-2	Screensho t
<b>Classification</b>								
<b>Spam</b>	<b>87%</b>	12%	<b>82%</b>	18%	<b>81%</b>	27%	<b>81%</b>	12%
<b>Non-Spam</b>	13%	<b>88%</b>	18%	<b>82%</b>	19%	<b>73%</b>	19%	<b>88%</b>
True Class	SPAM-1	Photo	SPAM-1	Baby	SPAM-1	Graphics	SPAM-1	Screensho t
<b>Classification</b>								
<b>Spam</b>	<b>73%</b>	1%	<b>81%</b>	4%	<b>80%</b>	14%	<b>71%</b>	7%
<b>Non-Spam</b>	27%	<b>99%</b>	19%	<b>96%</b>	20%	<b>86%</b>	29%	<b>93%</b>

**Table 2. Experimental Results: Detection of Screenshots in Outgoing E-mail**

True Class	Screenshot	Photo	Screenshot	Baby	Screenshot	Graphics
<b>Classification</b>						
<b>Spam</b>	90%	7%	89%	18%	81%	27%
<b>Non-Spam</b>	10%	93%	11%	82%	19%	73%