

# A Fast Image-Spam Filtering System using Support Vector Machine

Zin Mar Win, Nyein Aye

University of Computer Studies, Mandalay

zinmarwinn19@gmail.com, nyeinaye@gmail.com

## Abstract

The explosion of Image spam emails has prompted the development of numerous spam filtering techniques. This paper proposes an efficient image spam filtering system using three methods. The first method, File properties, analyses high level features in order to reduce computation cost. The second approach uses Hue, Saturation, Intensity (HSI) color model of histogram and the third method uses Hough line Detection. These three methods filter the image spam by analyzing both images including text and image. The images are collected from three different datasets that are Priceton, Image Spam Hunter and Spam Archieve Datasets. Support Vector Machine (SVM) classifies the input image is spam image or normal image. The experimental result shows the accuracy of different methods on different datasets and evaluates computation time. Among the three methods, Hough line can detect the input image within the minimum processing time required.

## 1. Introduction

Image spam is a kind of spam in emails where the message text of spam is presented in an image file. Image spam is an e-mail solicitation that uses graphical images of text to avoid text-based spam filters. These text-based spam filters cannot detect the image. They can detect only text.

Anti-spam filters label an e-mail (with image attached) as spam if they find suspicious text embedded in that image. There are different types of spam: Advertisement, Lottery winning notification, Adult content, Health and Pharmacy, Online Degrees, Bank and Finance, Reactionary and Pornography.



Figure 1. Normal images

FORGET ABOUT SEXUAL PROBLEMS!  
Want to have sexual strength again?  
Enjoy every second with your girlfriend!  
Penis problems treatment.  
Best way to cure ED.  
Need sex? Use effective meds!

<b>VIAGRA</b> *ON SALE*	<b>CIALIS</b> *ON SALE*	<b>LEVITRA</b> *ON SALE*
<b>SOFT VIAGRA</b> *ON SALE*	<b>SOFT CIALIS</b> *ON SALE*	<b>VIAGRA+CIALIS</b> (10) + (10) *ON SALE*

Cialis Soft Tabs as low as \$5.78  
Viagra Professional as low as \$4.07  
Viagra Soft Tabs as low as \$4.1  
Cialis as low as \$5.67  
Valium as low as \$2.39

**HAVE ERECTILE DISSFUNCTION?  
LOOK NO WHERE ELSE!  
HERE ARE OUR BEST E.D MEDICATIONS  
CHEAPEST ONLINE!**

**NOTHING TO LOSE, EVERYTHING TO GAIN!  
CLICK HERE AND CHECK IT OUT**

Figure 2. Spam images

In some image-based spam, the entire spam message is carried as an embedded jpeg or gif image with minimum amount of text. While for others, the image that appears as one are actually a set of images arranged side by side to give the impression that it is just one image. When the image is fully downloaded, the viewer

then gets to see the actual content of the image part of the message.

In the past, spam filtering required the manual construction of pattern matching rule sets. Contemporary spam filtering program indulgence spam detection as a text classification problem utilizing machine-learning algorithms such as neural networks and naïve Bayesian classifiers to learn spam characteristics. OCR-based modules can be used against image spam, to tolerate the analysis of the semantic content embedded into images. The performance of the OCR-based technique in detecting image spam is explained in [4]. The main limitation of this OCR-based spam classification technique is that it requires more processing time.

## 2. Literature Review

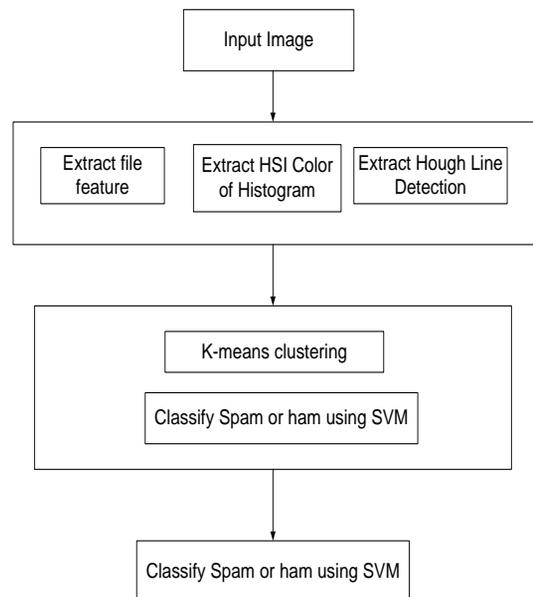
[11] propose File Properties and Histogram (FH) algorithm. FH is a fast method because it does not extract text and analyze the content of email. Incoming email firstly passes traditional anti-spam filters. If the image is suspected to spam, file properties are utilized to detect the image. The second stage, Histogram filtering evaluator checks whether the input image is spam or not. They test the input image on different datasets and calculate the accuracy of spam and non-spam. The results show that the grey histogram filter was able to achieve more than 80% detection rate for all datasets. The color histogram filter was able to achieve more than 86% detection rate because color histogram contains more information than grey histogram.

In [8], they propose a high performance image spam filtering system using three layers. The processing time of the first layer is very fast because it analyzes only header using Naïve Bayesian classifier. The second and third layers analyze the high-level and low-level features of image respectively using SVM. The accuracy rate of the system is about 94%.

Soranamageswari and Meena present a novel approach towards image spam classification. In their research, features are extracted using gradient histogram and normalized for efficient spam classification. The experiments are conducted for different training/testing rule for the Back Propagation Neural Network (BPNN). According to experiments, 90/10 training and testing yield better performance than the other pair of training and testing sets [9].

## 3. The System Architecture

The system contains two parts. The first part is feature extraction and the second part is classification. Feature extraction is used by three methods (File properties, HSI model of Histogram, Hough Line Detection). The second part performs Support Vector Machine (SVM) classification. Before SVM classification, k-means clustering is applied for faster classification.



**Figure 3. System architecture**

### 3.1. Feature Extraction

Transforming the input data into the set of features is called feature extraction. Feature extraction involves reducing the amount of records required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy [7].

#### 3.1.1. File Properties

Image spam e-mails will mostly contain images in jpeg or gif file types. The basic features (Table 1) that can be derived from an image at an extremely low computational cost are the width and the height denoted in the header of the image file, the image file type and the file size. In this study, we focus on all file formats that are commonly seen in emails, which are the Graphics Interchange Format (GIF), and the Joint Photographic Experts Group (JPEG) format, Bitmap (BMP) and Portable Network Graphic (PNG).

In the case of GIF files there will be the presence of virtual frames, which may be either larger or smaller than the actual image width. And this issue can be detected by decoding the image data. The problem imposed in the case of corrupted images is that the lines near to the bottom of the image will not decode properly. Any further decoding of the image data from that point of corruption will be decisive.

Feature  $f_6$  captures the amount of compression achieved by calculating the ratio of pixels in an image to the actual file size. The size of spam images in emails is smaller than 5 KB.

**Table 1. File properties**

Features	Description
$f_1$	Image width denoted in header
$f_2$	Image height denoted in header
$f_3$	Aspect Ratio $f_1/f_2$
$f_4$	File Size
$f_5$	File Area
$f_6$	Compression $f_5/f_4$
$f_7$	File Type (gif, jpeg, png, bmp)

#### 3.1.2. HSI Color Model of Histogram

In statistics, a histogram is a graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson [10]. A histogram is a representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1. The categories are usually specified as consecutive, non-overlapping intervals of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the same size [5]. The rectangles of a histogram are drawn so that they touch each other to indicate that the original variable is continuous [3].

Histograms are used to plot the density of data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the  $x$ -axis is all 1,

then a histogram is identical to a relative frequency plot.

This research extracts HSI color space from the histogram. The HSI color space is very important and attractive color model for image processing applications. This is due to representing colors similarly how the human eye senses colors. The HSI color model represents every color with three components: hue (H), saturation (S), intensity (I). The Hue component describes the color itself in the form of an angle between  $[0,360]$  degrees. 0 degree means red, 120 means green, 240 means blue and 60 degree is yellow, 300 degree is magenta. The Saturation component signals how much the color is polluted with white. The range of the S component is  $[0, 1]$ . The Intensity range is between  $[0, 1]$  and 0 means black, 1 means white [6]. This paper evaluates the mean value of HSI color space.

### 3.1.3. Hough Line Detection

The Hough Line Transform is used to detect straight lines. To apply the transform, an edge detection is pre-processed. There are many edge detectors: Sobel, Prewitt, Roberts, Laplacian of a Gaussian (LoG), Zero Crossings and Canny. Among them, the system uses canny edge detector because it detects true weak edges [2]. The algorithm of canny edge detector is as follow:

1. The image is smoothed using a Gaussian filter with a specified standard deviation,  $\sigma$  to reduce noise.
2. The local gradient,  $g(x,y)=[G_x^2+G_y^2]^{1/2}$ , and edge direction,  $\alpha(x,y) = \tan^{-1}(G_y/G_x)$ , are computed at each point.
3. The edge points determined in 2 give rise to ridges in the gradient magnitude image. The ridge pixels are then thresholded using two thresholds,  $T_1$  and  $T_2$ , with  $T_1 < T_2$ . Ridge pixels greater than  $T_2$  are said to be “strong” edge pixels. Ridge pixels with values between  $T_1$  and  $T_2$  are said to be “weak” edge pixels.

4. Finally, the algorithm performs edge linking by incorporating the weak pixels that are 8-connected to the strong pixels.

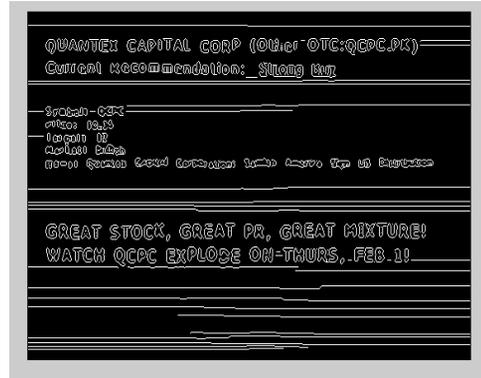


Figure 4. Canny edge detector

After edge detection, Hough Line Detection finds the number of lines in image. Firstly, it convert Cartesian space (Image space) to parameter space (Hough space) and initialize to zero. Hough Line detection algorithm is explained in Figure 5.

1. Initialize  $H[\rho, \theta] = 0$
2. for each edge point  $I[x,y]$  in the image  
for  $\theta = 0$  to 180  
 $\rho = x \cos \theta + y \sin \theta$   
 $H[\rho, \theta] += 1$
3. Find the value(s) of  $(\rho, \theta)$  where  $H[\rho, \theta]$  is maximum.
4. The detected line in the image is given by  
 $\rho = x \cos \theta + y \sin \theta$

Figure 5. Hough transform algorithm

### 3.2. Classification

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Classification is

considered an instance of supervised learning. i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance [1].

### 3.2.1. K-means Clustering

Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing. Data clustering has many engineering applications including the identification of part families for cellular manufacturing.

The K-means algorithm is a popular data clustering algorithm. To use K-means, it requires the number of clusters in the data to be pre-specified. Finding the appropriate number of clusters for a given data set is generally a trial-and-error process which made more difficult by the subjective nature of deciding what constitutes ‘correct’ clustering [12].

This research explores a method based on information obtained during the K-means clustering operation itself to select the number of clusters, K. The method employs an objective evaluation measure to suggest suitable values for K, thus avoiding the need for trial and error.

### 3.2.2. SVM Classification

SVM is a useful technique for data classification. SVM classifies data by finding the best hyperplane that separates all data points of one class from another class. The best hyperplane for SVM is the one with the largest margin between the two classes.

Training set:  $(\mathbf{x}_i, y_i), i=1,2,\dots,N;$   
 $y_i \in \{+1, -1\}$

Hyperplane:  $\mathbf{w}\mathbf{x} + b = 0$

where  $\mathbf{x} = (x_1, x_2, \dots, x_d), \mathbf{w} = (w_1, w_2, \dots, w_d),$

$\mathbf{w}\mathbf{x} = (w_1x_1 + w_2x_2 + \dots + w_dx_d)$

if  $\mathbf{w}\mathbf{x} + b > 0, y_i = 1$

if  $\mathbf{w}\mathbf{x} + b < 0, y_i = -1$

## 4. Experimental Results

The system measures the performance in terms of Accuracy. The Accuracy represents the ratio between the number of correctly identified spam (TP) and ham (TN) to the total number of images.

$$TP = \frac{\text{No. of Spam images classified as Spam}}{\text{No. of total Spam images}} \quad (1)$$

$$FP = \frac{\text{No. of Normal images classified as Spam}}{\text{No. of total Normal images}} \quad (2)$$

$$\text{Accuracy}(A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

The accuracy is calculated on an Intel core I3 machine and it is developed by using MATLAB. The classification can be done using Support Vector Machine. K-means clustering can speed up the processing time by reducing the training data. Figure 6 show the accuracy on each dataset. SA means Spam Archieve Dataset and ISH means Image Spam Hunter Dataset. Figure 7 show the accuracy on all datasets. Figure 8 compares the processing time in milliseconds.

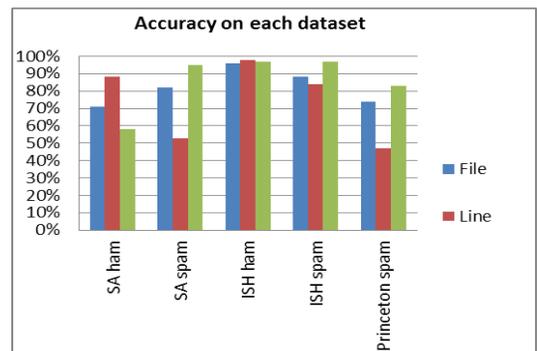


Figure 6. Accuracy on each dataset

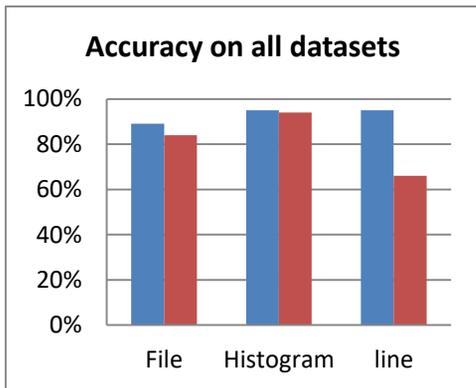


Figure 7. Accuracy on each dataset

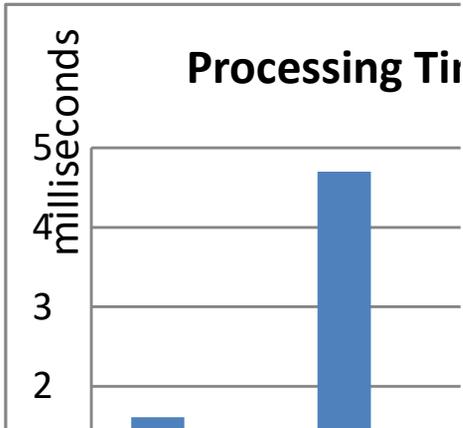


Figure 8. Processing time

## 5. Conclusion

Images are collected from different Image spam datasets: Princeton, Spam Archive and Image Spam Hunter. Princeton Image spam benchmark contains a total of 1071 spam images that are separated into 178 groups. The Spam Archive dataset includes 3253 spam images and 486 ham images. There are 810 ham images and 932 spam images in Image Spam Hunter dataset. According to my experimental results, Hough Line Detection method gives more accuracy than the other two methods in natural images but histogram method produces the best accuracy in spam images.

Compared with time complexity, Hough Line Detection takes the minimum processing time.

## References

- [1] Alpaydin, Ethem (2010). *Introduction to Machine Learning*. MIT Press. p. 9. [ISBN 978-0-262-01243-0](#)
- [2] C.Gonzalez, E.Woods and L. Eddins, "Digital Image Processing using MATLAB", Prentice Hall, p398
- [3] Charles Stangor (2011) "Research Methods For The Behavioral Sciences", Wodsworth, Cengage Learning, ISBN 9780840031976
- [4] G. Fumera, I. Pillai, and F. Roli, " Spam Filtering based on Analysis of Text Information Embedded into Images", *Journal of Machine Learning Research* (special issue on Machine Learning in Computer Security), vo.7, pp 2699-2720, 2006
- [5] Howitt, D. and Cramer, D.(2008) *Statistics in Psychology*. Prentice Hall
- [6] <http://www.blackice.com/colorspaceHSI.htm>
- [7] L. Tzong-Jye, T Wen-Liang and Lee Chia-Lin, "A High Performance Image-Spam Filtering System", *Ninth International Symposium on Distributed Computing and Application to Business, Engineering and Science*, IEEE, 2010, pp 445-449
- [8] M.Soranamageswari, Dr.C. Meena, "A Novel Approach towards Image Spam Classification", *International Journal of Computer Theory and Engineering*, Vol,3, No1, February, 2011, p-84-88
- [9] Pearson, K. "Contributions to the Mathematical Theory of Evolution II. Skew Variation in Homogeneous Material", *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences* 186: p.343-414
- [10] Pe Hu, W Xiangming, W Zheng and L Xinqi "Filtrng Image Spam using File Properties and Color Histogram", *International Conference on MultiMedia and Information Technology*, IEEE, Location, 2008, pp. 276-279.
- [11] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics anProbability*", Berkeley, University of California Press, 1:281-297