Detecting Image Spam Using Local Invariant Features and Pyramid Match Kernel

Haiqiang Zuo Weiming Hu Ou Wu Yunfei Chen Guan Luo

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China

{ hqzuo, wmhu, wuou, yfchen, gluo }@nlpr.ia.ac.cn

ABSTRACT

Image spam is a new obfuscating method which spammers invented to more effectively bypass conventional text based spam filters. In this paper, we extract local invariant features of images and run a one-class SVM classifier which uses the pyramid match kernel as the kernel function to detect image spam. Experimental results demonstrate that our algorithm is effective for fighting image spam.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval –*Clustering, Information filtering.*

General Terms: Algorithms, Security.

Keywords

Image spam, pyramid match kernel, local invariant features.

1. INTRODUCTION

Email spam, also known as unsolicited bulk e-mail or junk e-mail has become a scourge for us who just want to peacefully receive and send email. Many spam-thwarting programs have been developed that inspect words, phrases, mailing histories, IP addresses, and other aspects of an email. Some of which, like the CRM114 [1] has achieved a terminal accuracy of better than 99.99% for filtering text-based spam. Just as the classic battle of virus and antivirus, spammers explore new technologies in an effort to keep one step ahead of spam filters. Spammers' latest obfuscating method involves image spam, in which the main payload of the spam message is carried as an embedded image. Usually, the body of image spam contains no text or only bogus text, and the conventional text based spam filters therefore failed to detect and block it. Most of spams that break through the authors' personal anti-spam defences are image spams. Meanwhile, image spam can be more fascinating and convincing than text alone. Image spam is reported accounting for roughly 40 percent of all spam traffic now, and is still on the rise.

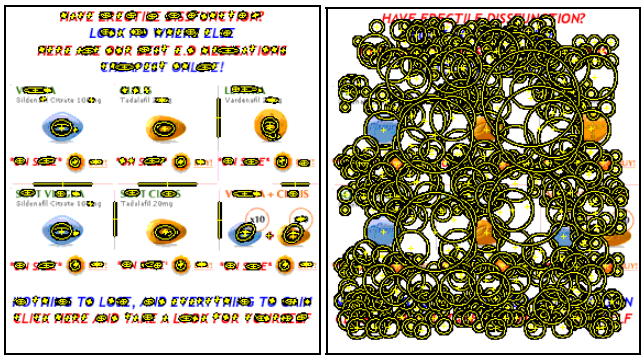
In recent years, many academic researchers and software security companies have turned their attention to investigating more constructive technologies to filter image spam. Approaches mainly differ in the set of features used to represent the image spam. Dredze et al. [2] established a fast image spam detection system which used simple image features like file format, file size, image metadata, average color and so on. Mehta et al. [3] exploited visual features, such as color, texture and shape to train a SVM classifier combining near duplicate detection and gained promising results. Our previous work [4] involved extracting global Fourier-Mellin invariant features to fight image spam.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

Figure 1. Extracting spam image from email.

2. LOCAL INVARIANT FEATURES

To avoid near duplicate detection, spammers often change the layout of the advertisement and produce a series of variations. However, in these variations there will always be some regions remain the same. For example, the picture of a Viagra pill will always appear in a Viagra advertisement image spam. Instead of extracting global features, in this paper, we explore local invariant features. Local invariant features are able to find correspondences in spite of large changes in viewing conditions, occlusions, and image clutter.



(a)MSER (b)SURF

Figure 2. Local invariant features.

A large number of local invariant features have been proposed in the literature. Among them, the MSER[5] and SURF [6] detector showed good accuracy and stability on our image spam detection application. Figure 1 shows the original image extracted from a

spam, and figure 2 demonstrates the local invariant features detected with the MSER and SURF detectors.

3. PYRAMID MATCH KERNEL

Grauman and Darrell [7][8] proposed a vocabulary-guided (VG) pyramid match kernel for computing an approximate bipartite partial matching between two unordered, variable-sized sets of feature vectors. Rather than carve the feature space into uniformly-shaped partitions, the authors let the structure of the feature space determine the partitions. To accomplish this, they performed hierarchical *k*-means clustering on a sample of representative feature vectors drawn from the feature space and built a pyramid tree.

Given the bin structure of the VG pyramid, a point set X is mapped to its pyramid: $\Psi(X) = [H_0(X), \ldots, H_{L-1}(X)]$, where $H_i(X)$ is a k^i-dimensional histogram associated with level i in the pyramid. L, k are the number of levels in the tree and the branching factor respectively.

Given two point sets' pyramid encodings $\Psi(X)$ and $\Psi(Y)$, the pyramid match kernel is defined as:

$$K_\Delta(\Psi(X),\Psi(Y)) = \sum_{i=0}^{L-1}\sum_{j=1}^{k^i}\omega_{ij}\left[\min(n_{ij}(X),n_{ij}(Y)) - \sum_{h=1}^{k}\min(c_h(n_{ij}(X)),c_h(n_{ij}(Y)))\right]$$

Where $n_{ij}(\bullet)$ denote the number of histogram counts corresponding to the j^{th} bin entry of histogram $H_i(\bullet)$, and $c_h(n_{ij}(\bullet))$ refer to the number of histogram counts for the h^{th} child bin of that entry, $1 \le h \le k$. The weights ω_{ij} are the sum of stored maximal to-center distances from either input set.

4. EXPERIMENTS

Our legitimate image email (also called non-spam or ham) corpus is made up 2839 images, which include 832 our homegrown ham images and Dredze et al.'s valid 2007 personal ham images. Dredze et al. have removed all images considered private and the released ham data are smaller than they used in references [2]. Our image spam corpus have 3885 images in all, 2173 images taken from SpamArchive corpus (removing duplicates), and 1712 images taken from our personal spam.

The MSER and SURF detectors are used to find interest points in each image, and SIFT[9], shape contexts (SC) [10], SURF descriptors are used to compose the feature sets. A one-class SVM classifier which uses the pyramid match kernel as the kernel function is trained using the spam corpus alone. Choosing a one-class classifier in our experiments in that the spam dataset is easily accessible, but a representative set of legitimate emails is usually difficult to collect, due to privacy concerns. The legitimate emails provided in our experiments are only used as part of the test collection and evaluate the classifier's accuracy.

We assessed our method by using 10-fold cross-validation. The spam corpus was randomly divided into 10 folds, and one fold was left together with the non-spam corpus as the test set and the other folds were used for training. The experiment was repeated 10 times, and the classification result was calculated by averaging over 10 runs.

Figure 3 demonstrates our experimental results. It can be seen that when using the SURF detector and descriptor, the algorithm reaches the best performance. The highest accuracy we achieved is over 98.5% which is comparable with most of the text-based spam filtering algorithm, and that is a promising result.

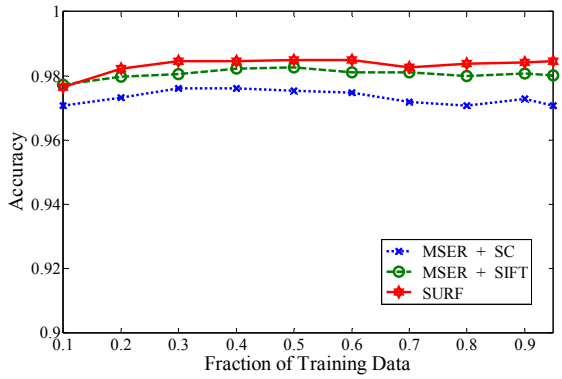


Figure 3. Experimental results.

The number of average local invariant features detected by SURF detector in each image is 447, however the number of average local invariant features detected by MSER detector in each image is 153. That means using MSER detector only needs a smaller time and space consumption. With this in mind, using the MSER detector and SIFT descriptor is also a good choice.

5. CONCLUSIONS

We have extracted local invariant features of images and run a one-class SVM classifier which uses the pyramid match kernel as the kernel function to detect image spam. Promising results have been obtained in our experiments.

6. ACKNOWLEDGMENTS

This work is partly supported by NSFC (Grant No. 60825204 and 60672040) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

7. REFERENCES

[1] CRM114 - the Controllable Regex Mutilator. http://crm114.sourceforge.net/

[2] M. Dredze, R. Gevaryahu, A. Elias-Bachrach. Learning Fast Classifiers for Image Spam. CEAS, 2007.

[3] B. Mehta, S. Nangia, M. Gupta. Detecting image spam using visual features and near duplicate detection. WWW, 2008.

[4] H.Q. Zuo, X. Li, O. Wu, W.M. Hu , G. Luo. Image spam filtering using Fourier-Mellin invariant features. ICASSP, 2009.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. BMVC, 2002.

[6] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. ECCV, 2006.

[7] K. Grauman and T. Darrell. Approximate Correspondences in High Dimensions. NIPS, 2007.

[8] Lee, John J. LIBPMK: A Pyramid Match Toolkit http://hdl.handle.net/1721.1/41070

[9] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. IJCV, 2004.

[10] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. PAMI, 2002.