

**IPAM Research in Industrial Projects for Students (RIPS)
Proposal Summer 2009**

Symantec Project Proposal: An Image Spam Problem

Industrial Partner: Symantec Research Labs
Industrial Advisers: Dr. Marc Dacier & Darren Shou {marc_dacier,
darren_shou}@symantec.com
Faculty Mentor: Dr. Arthur Szlam

1 Background

Spam is usually defined as junk or unsolicited email sent by a third party. While it is certainly an annoyance to users and administrators, spam is also a serious security concern as it can be used to deliver trojans, viruses, and phishing attempts. It could also cause a loss of service or degradation in the performance of network resources and email gateways. Image spam is an obfuscating method in which the text of the message is stored as a GIF or JPEG image and displayed in the email. This prevents text-based spam filters from detecting and blocking spam messages. A typical spam image has several variants in which the image is altered without changing the relevant visual content.

Between July 1 and December 31, 2007, spam made up 71 percent of all email traffic monitored at the gateway, a 16 percent increase over the last six months of 2006, when 61 percent of email was classified as spam. In the second half of 2007, the daily average percentage of image spam was seven percent. This is down from a daily average of 27 percent during the first six months of 2007 (ISTR 2008).

2 The Problem

The abstract version of the image spam problem is to identify images that contain horizontal/vertical lines of text. Whereas standard OCR algorithms are effective but not very fast, we would like to discover or develop an effective and efficient solution to this problem. We just need to know there is text, but we don't need to know what the text actually is. Note that we don't want to determine good images vs. spam images.

Spammers intentionally distort and create 'noise' in these images to make detection more difficult, for example, blurring of text outlines, construction of the image from multiple image layers assembled within an HTML e-mail, or use of animated image formats can also be very effective. Current anti-spam systems can fail to detect image spam because the images lack uniformity in terms of size, style and the arrangement of symbols.

3 Candidate Starting Points & Directions

X. Chen, J. Yang, J. Zhang, and A. Waibel. "Automatic detection and recognition of signs

from natural scenes.” IEEE transactions on image processing, 13(1):87–99, 2004.

H. Goto and H. Aso, “Extracting curved lines using local linearity of the text line”, IJDAR, vol.2 pp. 111- 118, 1999.

F. Hones and J. Litcher, “Layout extraction of mixed mode documents”, Machine vision and applications, vol.7, pp.237-246, 1994.

P.K.Loo and C.L.Tan, “Word and sentence extraction using irregular pyramid”, Proc. DAS, pp. 307-318, 2002.

J. Ohya, A. Shio, and S. Akamatsu. “Recognizing characters in scene images.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(2):214–220, 1994.

H. Wang. “Automatic character location and segmentation in color scene images.” Proceedings of 11th International Conference on Image Analysis and Processing, pages 2–7, 2001.

4 Research Tasks and Deliverables

- Task 1: Explore the design space of standard OCR algorithms
 - Task 2: Manipulate them to identify images that contain horizontal/vertical lines of text
 - Task 3: Empirically compare them
 - Task 4: Prototype the most promising approach
-
- Deliverable 1: A technical report detailing approaches taken, experimental results and conclusions
 - Deliverable 2: Presentation to Symantec Research Labs and separately to IPAM RIPS at Project's day
 - Deliverable 3: The prototype source/software

5 References

Internet Security Threat Report: September 2008. Symantec Corporation