Image spam filtering using textual and visual information

Giorgio Fumera Ignazio Pillai Fabio Roli
Battista Biggio
Dept. of Electrical and Electronic Eng., Univ. of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
e-mail: {fumera, pillai, roli, biggio}@diee.unica.it

Abstract

In this paper we focus on the so-called *image spam*, which consists in embedding the spam message into images attached to e-mails to circumvent statistical techniques based on the analysis of body text of e-mails (like the "bayesian filters"), and in applying content obscuring techniques to such images to make them unreadable by standard OCR systems without compromising human readability. We argue that a prominent role against image spam will be played by computer vision techniques, in particular visual pattern recognition and image processing techniques. We then discuss two possible approaches to defeat image spam: exploiting the high-level textual information embedded into images by combining OCR and text categorization techniques, and exploiting the low-level image information to detect content obscuring techniques applied to spam images. We also report some results of an experimental investigation on a large data set of spam e-mails, aimed at evaluating the effectiveness of combining standard OCR and text categorization techniques, and preliminary results on the use of low-level features to detect image defects (like broken characters or noise components interfering with characters in a binarized image) which are typical consequences of content obscuring techniques that spammers are using.

1 Introduction

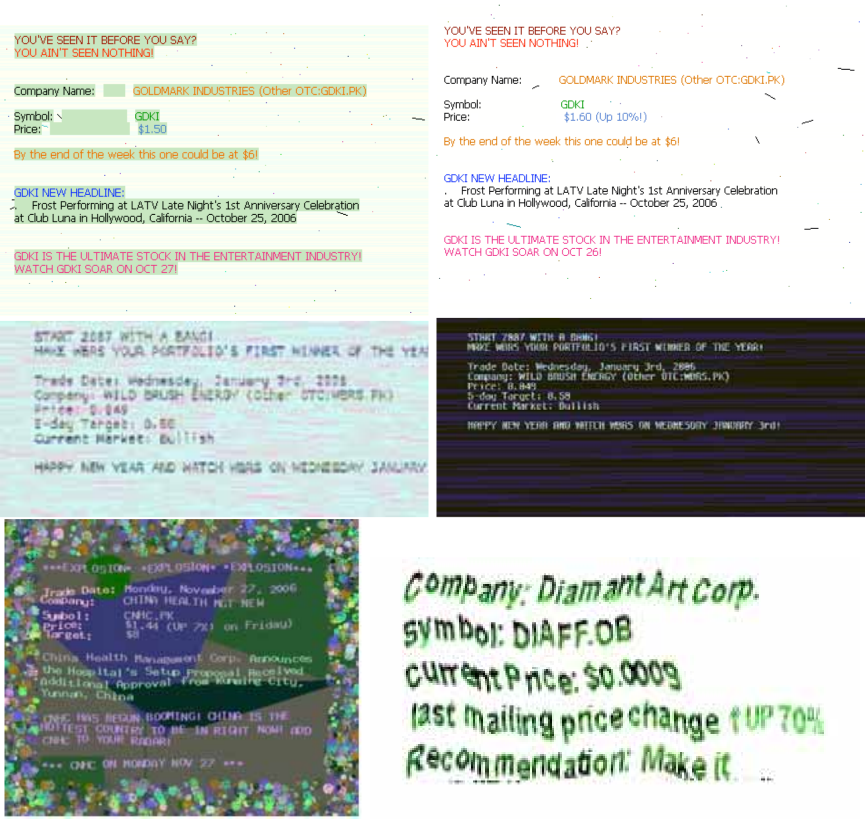
In the past ten years the problem of spam filtering has been addressed by the machine learning community as a text categorization task whose goal is to discriminate between spam and legitimate e-mails on the basis of the text in the e-mails' body, which turned out to ne a more effective approach with respect to the one based on hand-coded rules, like keyword detection. Several text categorization methods have been proposed so far, mainly based on the Naive Bayes (known as "Bayesian filter" in this application field) and Support Vector Machine classification techniques (see for instance [13, 5, 1, 7]). Some of these techniques have also been adopted by commercial and open-source spam filters (for instance, the well-known open-source SpamAssassin[1] filter implements a Naive Bayes text classifier). The first countermeasures taken by spammers consisted in adding bogus text to their e-mails, usually taken from books or news articles, to compromise the effectiveness of statistical technniques. However, a new kind of trick introduced some years ago has rapidly spread during the past year and is now adopted in a large fraction of spam e-mails: it consists in embedding the spam message into attached images to circumvent all spam detection techniques based on the analysis of body text (see the examples in figure 1). This kind of spam is known as *image-based spam* (shortly, *image spam*). Small changes are usually introduced to the same base image (see figure 1, top) to easily circumvent simple techniques based on the analysis of the digital signature of attached images used in some spam filters. The very last evolution, which the authors observed for the first time in their personal e-mails just in November 2006, consists in applying content obscuring techniques with the aim of making the image text unreadable by standard OCR without compromising human readability, as in the example of figure 1 (middle). A remarkable fact is that spammers could exploit to their advantage the content obscuring techniques used to create CAPTCHAs (as suggested by the examples in figure 1, bottom), which were recently introduced just to defend against robot spamming.

It is clear that image spam can make all filtering techniques based on the analysis of the body text of e-mails ineffective. From a high-level perspective, it could be said that spammers are using a new "medium" to convey their message. Accordingly, spam filters will have to exploit also the information coming from images to keep their effectiveness high. To this aim, we argue that in the near future a prominent role will be played by computer vision techniques, in particular visual pattern recognition and image processing techniques. Although pattern recognition techniques against image spam have been implemented by some spam filtering software (for instance, two plug-ins which perform keyword detection on text extracted by OCR have been developed for SpamAssassin[2]), to our knowledge the problem of exploiting computer vision techniques for spam filtering has not yet been addressed in the scientific literature.

The aim of this paper is to discuss two possible approaches for the development of well-grounded image spam detection techniques, under the viewpoint of spam filters made up of different modules each capable to detect a specific characteristic of spam e-mails, whose outputs have then to be properly combined to get a reliable decision. The first approach concerns the possibility of recognizing image spam exploiting the *high-level* textual information embedded into images, through the use of OCR techniques. This raises several issues, among which OCR computational complexity, how to process (possibly noisy) text extracted by OCR, and how to cope with content obscuring techniques. Building on a first investigation of this approach given by the authors in [6], in section 2 we discuss the above issues and report some experimental results

[1] `http://spamassassin.apache.org/`

[2] `http://wiki.apache.org/spamassassin/BayesInSpamAssassin`



YOU'VE SEEN IT BEFORE YOU SAY?
YOU AIN'T SEEN NOTHING!

Company Name: GOLDMARK INDUSTRIES (Other OTC:GDKI.PK)

Symbol: GDKI
Price: $1.50

By the end of the week this one could be at $6!

GDKI NEW HEADLINE:
Frost Performing at LATV Late Night's 1st Anniversary Celebration
at Club Luna in Hollywood, California -- October 25, 2006

GDKI IS THE ULTIMATE STOCK IN THE ENTERTAINMENT INDUSTRY!
WATCH GDKI SOAR ON OCT 27!

YOU'VE SEEN IT BEFORE YOU SAY?
YOU AIN'T SEEN NOTHING!

Company Name: GOLDMARK INDUSTRIES (Other OTC:GDKI.PK)

Symbol: GDKI
Price: $1.60 (Up 10%!)

By the end of the week this one could be at $6!

GDKI NEW HEADLINE:
Frost Performing at LATV Late Night's 1st Anniversary Celebration
at Club Luna in Hollywood, California -- October 25, 2006

GDKI IS THE ULTIMATE STOCK IN THE ENTERTAINMENT INDUSTRY!
WATCH GDKI SOAR ON OCT 26!

START 2007 WITH A BANG!
MAKE WBRS YOUR PORTFOLIO'S FIRST WINNER OF THE YEA

Trade Date: Wednesday, January 3rd, 2006
Company: WILD BRUSH ENERGY (Other OTC:WBRS.PK)
Price: 0.049
5-day Target: 0.58
Current Market: Bullish

HAPPY NEW YEAR AND WATCH WBRS ON WEDNESDAY JANUARY

START 2007 WITH A BANG!
MAKE WBRS YOUR PORTFOLIO'S FIRST WINNER OF THE YEAR!

Trade Date: Wednesday, January 3rd, 2006
Company: WILD BRUSH ENERGY (Other OTC:WBRS.PK)
Price: 0.049
5-day Target: 0.58
Current Market: Bullish

HAPPY NEW YEAR AND WATCH WBRS ON WEDNESDAY JANUARY 3rd!

EXPLOSION* *EXPLOSION* *EXPLOSION

Trade Date: Monday, November 27, 2006
Company: CHINA HEALTH MGT NEW
Symbol: CNHC.PK
Price: $1.44 (UP 7%) on Friday)
Target: $8

China Health Management Corp. Announces
the Hospital's Setup Proposal Received
Additional Approval from Kunming City,
Yunnan, China

CNHC HAS BEGUN BOOMING! CHINA IS THE
HOTTEST COUNTRY TO BE IN RIGHT NOW! ADD
CNHC TO YOUR RADAR!

*** CNHC ON MONDAY NOV 27 ***

Company: Diamant Art Corp.
symbol: DIAFF.OB
current Price: $0.0009
last mailing price change ↑UP 70%
Recommendation: Make it

Figure 1: Examples of images attached to spam e-mails, taken from the authors' personal mail boxes. Two nearly identical images with small changes to defeat signature-based techniques (top). Images with content obscuring techniques against OCR (middle). Content obscuring techniques similar to the ones used in CAPTCHAs (bottom).

aimed at evaluating the effectiveness of combining standard OCR and text categorization techniques. We then propose in section 3 a different and somewhat complementary approach, which consists in using low-level image features with the specific goal of detecting the use of content obscuring techniques, which are likely to make OCR ineffective. Using a conceptual metaphor, we could say that, differently from previous works, this approach is aimed to detect the "noise" (the adversarial clutter contained in the image) instead of the "signal" (the spam text message). We describe a possible implementation of this approach aimed at detecting image defects like broken characters or the presence of noise component which interfere with characters in a binarized image (which are typical consequences of content obscuring techniques that spammers are using), and report some preliminary experimental results.

2 Analysis of text information embedded into images

In this section we discuss the main issues related to the use of OCR techniques against image spam. We then give an overview of an approach recently proposed, based on combining OCR and text categorization techiques, and present some experimental results aimed at evaluating its potential effectiveness.

2.1 Proposed approach

Using OCR techniques to extract text embedded into attached images is an obvious attempt, and has already been implemented in some commercial or open source filter. However it is not so obvious how such text should be processed to the purposes of an anti-spam filter, and there are no studies on this topic in the scientific liteature with the exception of [6]. In particular, the use of OCR techniques in anti-spam filters raises two immediate issues: is the computational complexity of OCR affordable to spam filters? to what extent can a OCR system be effective, if content obscuring techniques are used by spammers? We nevertheless believe that the feasibility of using OCR techniques in spam filters can not be ruled out a priori without a thorough investigation of their actual effectiveness. With regard to the computational complexity, it could be limited by implementing a hierarchical architecture in which the OCR is used only if other modules are not able to reliably determine if the e-mail is spam or legitimate. With regard to content obscuring techniques, it is likely that they can not be exploited in every kind of spam: for instance, we believe that content obscuring can not be excessive in phishing, in which e-mails should look as if they come from reputable senders, and thus should be as "clean" as possible. Accordingly, OCR techniques could be effective at least for certain kinds of image spam, and could be exploited into a module of spam filter whose output has to be combined with the outputs of other modules to reach a final and more reliable decision. Moreover, other modules based on low-level image processing techniques can be used for cases in which standard OCR techniques can be expected to be not effective: this will be discussed in section 3.

Consider now the problem of how to analyze text extracted by OCR from attached images. Straightforward approaches like the ones based on keyword detection have already been implemented in spam filters (as the SpamAssassin plug-in mentioned in section 1). We consider here a different approach investigated first by the authors in [6], based on analyzing text extracted from images using the same text categorization techniques applied to text in e-mail's body. This approach can be justified by the fact that text embedded into spam images plays the same role as text in the body of e-mails without images, namely to convey the spam message. These two kind of text can then be viewed simply as a different "coding" of the message carried by an e-mail. Accordingly, given that text categorization techniques proved to be effective for body text, it is worth investigating whether they can be exploited also for image text. This approach requires to insert text extracted from images, as well as the one in the e-mail's body, in the design and operating cycle of a text classifier. For a detailed discussion on several possible implementations of this approach and on related issues we refer the readers to [6].[3] In the following we give some results of an experimental investigation aimed at evaluating the potential effectiveness of this approach.

2.2 Experimental results

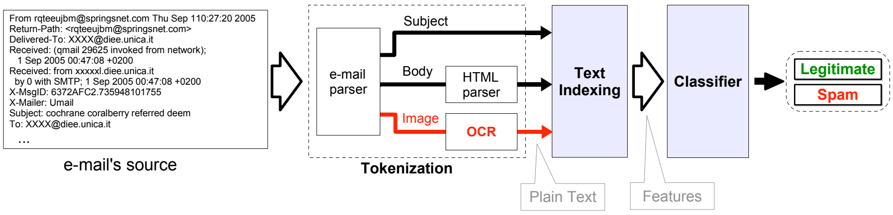
Among the possible implementations of the above approach, in [6] the authors analyzed the simplest one, consisting in designing a single text classifier using the well known bag-of-words approach, with a vocabulary of terms extracted only from the body of training e-mails (even for e-mails with attached images), and in training such classifier on feature vector representation of training e-mails constructed using again only the body text. The text extracted from attached images (if any) by OCR is used only at the classification phase of testing e-mails, as described below. The scheme of the processing steps is shown in figure 2.

The experimental results we present here refer to the 'submit' data set of the publicly available SpamArchive[4] corpus of spam e-mails, available in January 2006, made up of 142, 897 spam e-mails, of which about 10% contained attached images. Legitimate e-mails were taken from the Enron[5] data set [10], which was the only *large* and publicly available corpus of legitimate e-mails to our knowledge. About 100, 000 e-mails were selected, to obtain a ratio between spam and legitimate of about 3:2, according to recent surveys on e-mail traffic. A limit of the Enron data set for the purposes of our experiments is that it does not contain e-mails with attached images. However we preferred not to construct "artificial" legitimate e-mails by adding non-spam images. To make several runs of the experiments, the two corpora of e-mails were subdivided in

[3] This paper is available on-line at `http://jmlr.csail.mit.edu/papers/v7/`

[4] `www.spamarchive.org`

[5] `http://www.cs.cmu.edu/~enron/`



```
From rqteeujbm@springsnet.com Thu Sep 110:27:20 2005
Return-Path: <rqteeujbm@springsnet.com>
Delivered-To: XXXX@diee.unica.it
Received: (qmail 29625 invoked from network);
  1 Sep 2005 00:47:08 +0200
Received: from xxxxxl.diee.unica.it
 by 0 with SMTP; 1 Sep 2005 00:47:08 +0200
X-MsgID: 6372AFC2.735948101755
X-Mailer: Umail
Subject: cochrane coralberry referred deem
To: XXXX@diee.unica.it
  ...
```

Figure 2: Scheme of the proposed approach to implement a spam filter module based on both the text in the e-mail's body and the text embedded into attached images. The traditional document processing steps (tokenization for plain text extraction, indexing for constructing the feature vector representation of e-mails, and classification) are extended by including in the tokenization phase the extraction of text from attached images using OCR techniques. In the implementations of this approach investigated in this paper, text extracted from attached images is used only at the operation phase.

ten subsets according to chronological order. Each subset was further subdivided in chronological order into training, validation and test sets, respectively containing 45%, 15% and 40% of the e-mails.

At the design phase of the text classifier, the terms in the e-mails' body were extracted after removing all HTML tags (if any). Punctuation and stop words were then removed and stemming was carried out, using the software SMART.[6] Several kinds of features related to the bag-of-words approach for document representation were evaluated. Below we report results reated to the well known tf-idf features. Several sizes of the feature set were also considered: to this aim, feature selection was performed with the classifier-independent Information Gain criterion. At classification phase, text from attached images was extracted using the demo version of the commercial software ABBYY FineReader 7.0 Professional,[7] without performing the preliminary training of this software, and using default parameter settings.

A SVM was used as text classifier, given its state-of-the-art performance on text categorisation tasks. The SVM-light[8] software [9] was used for SVM training. A linear kernel was used, which is a typical choice in text categorisation works. Classification performance was evaluated in terms of the receiver operating characteristic (ROC) curve. To fix a trade-off between the false positive and false negative misclassification rates, denoted in the following respectively as FP and FN, the decision threshold on the output produced by the SVM classifier was computed on validation e-mails (not used in the training phase) by minimizing the FN rate for a given maximum allowed value of the FP rate.

In [6] two different methods were tested for constructing the feature vector representation of testing e-mails with attached images: merging body text with

[6] ftp://ftp.cs.cornell.edu/pub/smart/
[7] http://www.abbyy.com/
[8] http://svmlight.joachims.org/

image text (denoted below with 'B+I', which stands for 'Body text + Image text'), and using only the image text ('I'). In subsequent experiments, we also classified both feature vectors B and I, and took the maximum of the corresponding SVM outputs to classify the a-mail ('max {B,I}').[9] The two latter methods were suggested by the fact that often, though not always, the body of spam e-mails with attached images contains only bogus text. The above methods were compared with the baseline results obtained using only body text ('B').

We point out that using the methods B+I, I and max {B,I} only spam e-mails with attached images could be classified in a different way with respect to the baseline method B, all the other experimental conditions being equal. We therefore focus in the following on the test set misclassification rate attained on these e-mails only. In table 1 we report the test set misclassification rate on spam e-mails with attached images, averaged over the ten runs of the experiments and obtained with different numbers of tf-idf feaures and different ROC operating points (namely FP values). Note that the considered FP values are too high for a real spam filter. However in our experiments only a single classifier was used, while a higher overall effectiveness can be obtained using also the other modules of spam filters.

Firstly, it can be seen that even when text extracted from attached images was not used (see table entries denoted with B), many e-mails containing attached images (from 52% to 95%, depending on the number of features and on the required FP rate) were correctly classified. This means that in this data set, e-mails with text embedded into images often contained also body text which allowed to identify them as spam. However, higher misclassification rates (up to 48%) were attained when lower values values of the FP rate were required. Nevertheless, when text extracted from attached images was used at classification phase as described above (see table entries denoted with B+I, I and max {B,I}), a lower misclassification rate was always attained. In particular, for the lowest required FP rate value of 0.01, the misclassification rate was reduced up to one third. It is worth noting that, being equal the number of features and the required value of the FP rate, the highest improvement was always attained by separately classifying the feature vectors corersponding to the body text and to image text, and then taking the maximum of the classifier outputs (corresponding to table entries denoted with max {B,I}). This means that, at least for this data set, the spam message could be often recognised either using only the body text, or using only the text embedded into images. Accordingly, separately computing the corresponding scores and taking their maximum turned out to be a better solution than both using only the text embedded into images (I), and mixing the two kinds of text to construct a single feature vector (B+I).

Let us finally comment on the influence of OCR noise on the accuracy of the text classifier. This issue was addressed in preliminary experiments carried out using a corpus of spam e-mails collected by the authors (the same experimental

[9] When training the SVM classifier spam e-mails were labelled with $+1$, and legitimate ones with -1. The corresponding decision function is to label a testing e-mail as spam, if the corresponding SVM output is higher than a fixed threshold value.

number of		maximum allowed FP rate (validation set)		
features	text	0.05	0.03	0.01
1000	B	0.143 (0.064)	0.222 (0.090)	0.479 (0.157)
	B+I	0.086 (0.071)	0.186 (0.128)	0.363 (0.148)
	I	0.110 (0.121)	0.186 (0.145)	0.371 (0.119)
	max {B,I}	0.035 (0.031)	0.071 (0.052)	0.227 (0.101)
5000	B	0.083 (0.040)	0.136 (0.056)	0.361 (0.140)
	B+I	0.035 (0.033)	0.073 (0.042)	0.278 (0.141)
	I	0.062 (0.065)	0.103 (0.072)	0.269 (0.149)
	max {B,I}	0.017 (0.016)	0.031 (0.023)	0.132 (0.066)
10000	B	0.062 (0.035)	0.110 (0.049)	0.346 (0.139)
	B+I	0.025 (0.026)	0.067 (0.049)	0.262 (0.136)
	I	0.047 (0.053)	0.089 (0.090)	0.244 (0.149)
	max {B,I}	0.011 (0.012)	0.024 (0.021)	0.117 (0.071)
20000	B	0.047 (0.029)	0.074 (0.032)	0.300 (0.121)
	B+I	0.029 (0.027)	0.059 (0.058)	0.249 (0.137)
	I	0.054 (0.066)	0.085 (0.088)	0.229 (0.139)
	max {B,I}	0.011 (0.012)	0.019 (0.018)	0.091 (0.051)

Table 1: Fraction of misclassified testing spam e-mails among the ones containing attached images, for three different values of the maximum allowed FP rate (computed on the validation set), and of four different numbers of features. Reported values are averaged across the ten runs of the experiments. Standard deviation is reported between parentheses. The column "text" refer to the way in which text extracted from images was used to construct the feature vector representation of tesing e-mails: B: only the text in the body field was used. B+I: the text in the body field and the text extracted from attached images were mixed before building the feature vector. I: only the text extracted from attached images was used. max {B,I}: both feature vectors B and I were computed, and classified, and the maximum of the corresponding scores was used to classify the e-mail.

setting above was used). In these experiments we compared the misclassification rates on e-mails with attaches images attained by extracting the embedded text automatically by OCR, and manually (thus without the possible errors introduced by OCR). We found that the resulting misclassification rates were nearly identical, which means that in this application OCR noise should not degrade significantly the performance of a text classifier.

To sum up, the above results provided evidence that combining OCR and text categorization techniques, even standard ones, can be effective in recognizing spam images, at least when content obscuring techniques are not used (which also in the future is likely to be the case of some kinds of image spam like phishing). This suggests that it is worth to further investigate the issue of develping modules of spam filters focused on image spam based on OCR techniques, possibility optimized to this task for reducing the OCR computational complexity.

3 Detecting content obscuring techniques using low-level image features

Some authors proposed techniques for recognizing spam images based on detecting the presence of text, and on characterizing text areas with low level features like their size relative to the image [2, 14] and their colour distribution [2]. A classifier was then trained on such features to discriminate spam images from legitimate ones. The rationale of these approaches is that images which contain text are likely to be spam. Some vendors have already included in their filters image processing modules based on this kind of low level features.

Here we suggest a different approach which is based on "actively" looking for a specific characteristics of spam images, rather than trying to discriminate spam images from legitimate ones based on "generic" low level features extracted from text areas as in the approaches mentioned above. We argue that a useful approach could be to analyze images with the aim to detect whether content obscuring techniques were used. The rationale of this approach is that images which are obscured in a way aimed to make OCR difficult are likely to be spam. This approach can also be viewed as complementary to the one discussed in section 2, which was based on *reading* the embedded text using OCR tools: we showed that OCR combined with text categorization techniques can be effective for recognizing spam images in which no content obscuring techniques are used. The approach we are considering can thus help in recognizing images on which OCR is likely to be ineffective. In the following we will discuss a possible implementation of this approach, which is currently being investigated by our research group.

In principle, different kinds of content obscuring techniques against OCR can be used, like the ones in the examples of figure 1 (middle and bottom). Here we focus on techniques whose effect is to make OCR ineffective by resulting in a low quality binarized image (note that binarization is the first step of OCR systems). The two spam images in figure 1 (middle) are an example of such kind of techniques. The aim of binarization in OCR is to remove from an image any background and non-text component. Text detection and character segmentation critically depend on a successful outcome of this step. As a result of an improper image binarization, characters can be broken up into smaller pieces, or can be merged together. Non-text objects can be kept in the foreground as well and interfere with characters. Accordingly, when an image is found to contain text (note that the mere *presence* of text can be detected even on a complex background, for instance using techniques likes the ones surveyed in [8]) a possible way to detect content obscuring techniques which result in a low quality binarized image is to analyze the binarized image to detect the presence and the extent of the above defects. The outcome of this analysis can then be used to measure the likelihood that the image is spam.

The above problem is similar to the one of how to measure the *quality* of a binarized image in terms of the degree of difficulty it can pose to an OCR. This problem was addressed in some works in the OCR literature; for instance, in [4] a method was proposed for predicting OCR performance based on simple features associated with degraded (broken or merged) characters. Another interesting kind of measure was suggested to us by the "BaffleText" CAPTCHA proposed in [3]. BaffleText uses random masking to degrade text images, resulting in image defects similar to the ones we are interested in. In [3] the *complexity* of an image for a human reader (not for an OCR) was evaluated using *perimetric complexity*, a measure used in the psychophysics of reading literature (see for instance [12]). Perimetric complexity is defined as the squared length of the boundary between black and white pixels (the "perimeter") in the whole image, divided by the black area, P^2/A. Both the measures used in [4] and perimetric complexity can not be directly applied to measure the degree and extent of character breaking and merging for text of arbitrary length and possibly different character sizes. However we found that perimetric complexity, with some changes, could be usefully exploited to this purpose. In the following we briefly explain how we are investigating the use of perimetric compelxity, and report some preliminary experimental results.

Note first that the perimetric complexity of a *single* object is scale-invariant. We found that its value for the clean image of a single character lies approximately in the range $[25, 200]$. If an image contains only clean text, most of the connected components in the binarized image will correspond to single characters, and will be characterized by P^2/A values in the above range. If the binarized image is degraded, on average broken characters and connected components originated from background noise like dots, clumps and small line segments (see the second, third, and fourth examples in figure 1), will be characterized by a lower P^2/A value and a lower area. Instead, characters interfering with large noise components will exhibit on average larger P^2/A and area values. One way to highlight this is to plot the histogram of the area of each connected component of the binarized image, relative to the whole area of black pixels, versus its perimetric complexity.

Four examples are reported in figures 3-6. In the top of each figure we show the original spam image (taken from the authors' personal mail boxes): in the first one no content obscuring technique was used; in the other ones three different kinds of content obscuring techniques are used, but only in the second and third image such techniques result in the kind of image defects we are considering. In the second and fourth row of each figure we report the same image binarized respectively by the open source `gocr`[10] OCR tool, used in the SpamAssassin plug-in mentioned in section 1, and by the ABBYY commercial OCR tool mentioned in section 2.2. Finally, below each binarized image we report the corresponding histogram of the relative area of each component versus its complexity (for simplicity, the complexity values are discretized into a finite number of intervals).

The text in the spam image in figure 3 is not obscured, and its histograms show that most components have P^2/A values in the range $[25, 200]$. The "spamminess" of this image could be detected through the text extracted by OCR, as shown in section 2. The text in the image in figure 4 is instead obscured, and many small components of low complexity are generated by the binarization step: some correspond to broken characters, others to background line segments. Such components result in the presence of several non-zero histogram values for P^2/A values lower than 25. Instead, in the image in figure 5 the overlapping line segments generate components with high complexity, as well as large blocks of black pixels with low complexity but high area. These noise components are revealed by the non-zero histogram values outside the interval $[25, 200]$. This suggests that the presence of content obscuring techniques which result in character breaking or in the presence of noie components interfering with characters, like in the images of figures 4 and 5, could be detected for instance by comparing the histogram of the binarized image with a reference or template histogram which could look like the ones in figure 3. This comparison should finally output a score (in the example, it could be a measure of histogram similarity) indicating the likelihood that this kind of content obscuring techniques has been used in the considered image. Consider finally the spam image in figure 6, where a CAPTCHA-like obscuring technique is used. This technique does not result in the considered kind of image defects. In this case the histograms are indeed quite similar to the ones of figure 3. Clearly, a different approach has to be used to detect the CAPTCHA-like technique of images like this one. To this aim, works both in the OCR literature and on CAPTCHAs could give useful suggestions, as in the case of the perimetric complexity measure exploited above.

Let us now give a preliminary evaluation of our method on samples of non-spam images, focusing on photographs which are a typical kind of images attached to legitimate e-mails. To this aim, consider the three images shown in figures 7-9 (top-left), which were downloaded from the Internet. The first one contains relatively uniform regions, while the other two are characterized by the presence of high frequency textures. We already pointed out that the approach described in this section makes sense only if applied to text areas. For the sake of completeness, let us consider what happens when it is applied to images which

[10] `http://jocr.sourceforge.net/`

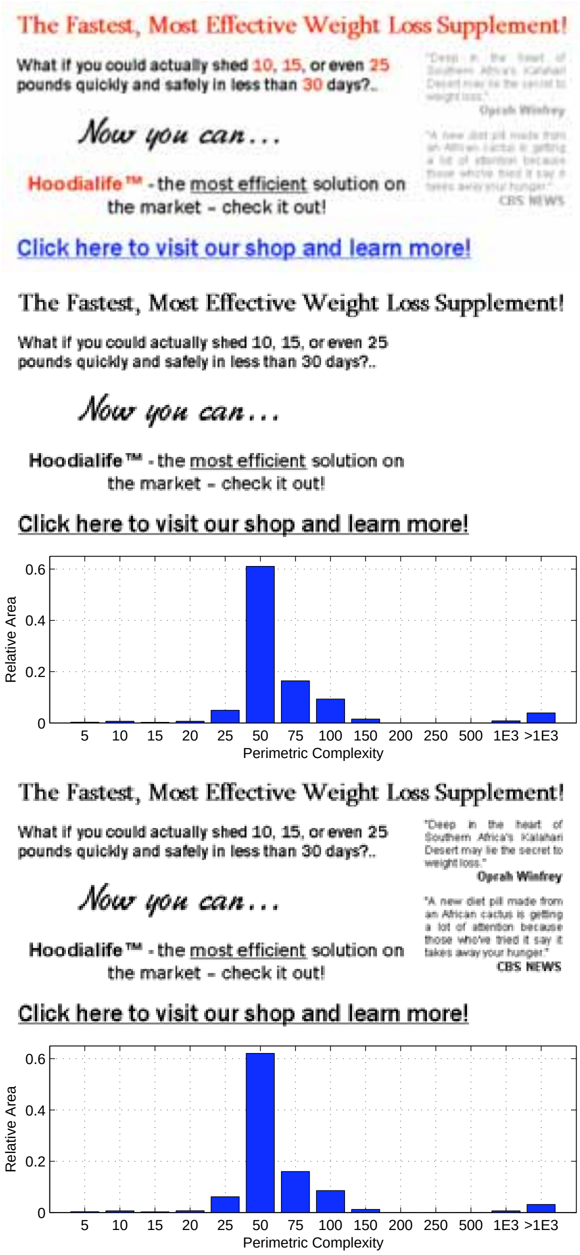
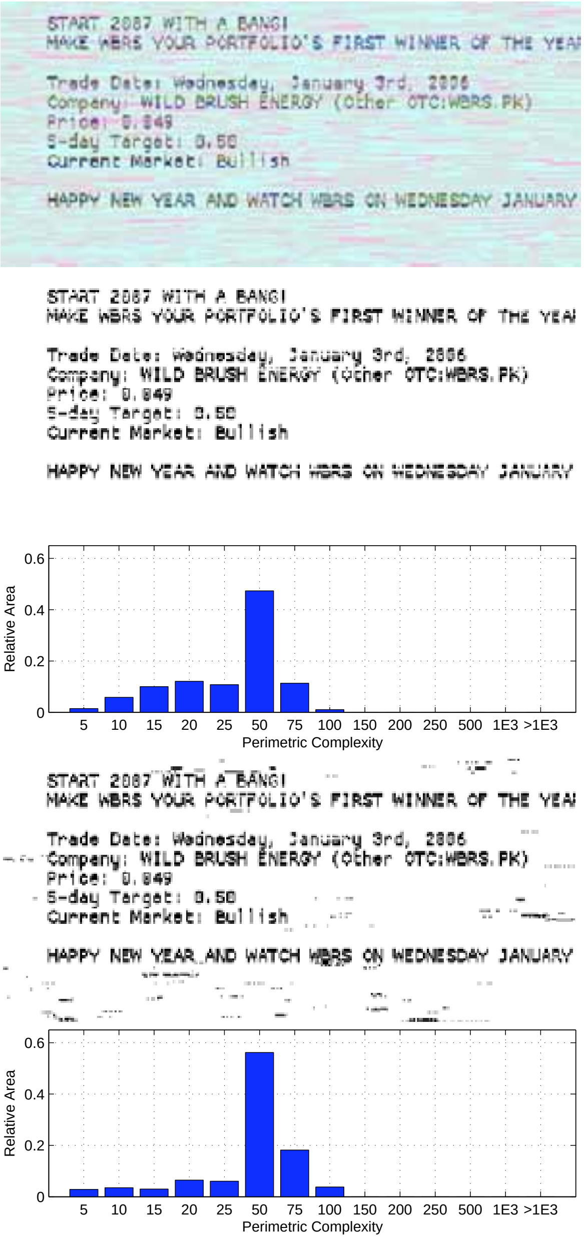


Figure 3: A spam image without content obscuring techniques. From top: the original image, the image binarized by `gocr`, the corresponding histogram of the relative area of each connected component vs its perimetric complexity (see text), the image binarized by the ABBYY OCR, and the corresponding histogram.

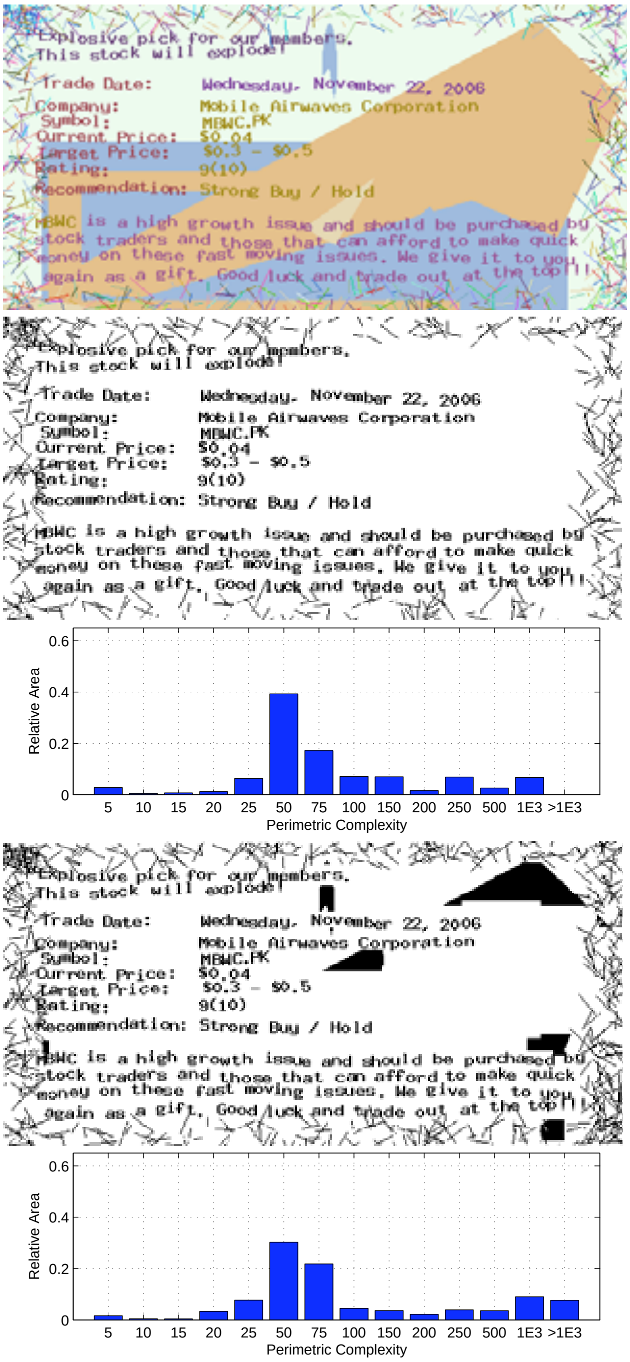


START 2007 WITH A BANG!
MAKE WBRS YOUR PORTFOLIO'S FIRST WINNER OF THE YEA

Trade Date: Wednesday, January 3rd, 2006
Company: WILD BRUSH ENERGY (Other OTC:WBRS.PK)
Price: 0.049
5-day Target: 0.50
Current Market: Bullish

HAPPY NEW YEAR AND WATCH WBRS ON WEDNESDAY JANUARY

Figure 4: A spam image with a content obscuring technique resulting in character breaking and small background clutter (see the caption of figure 3).



Explosive pick for our members.
This stock will explode!

Trade Date: Wednesday, November 22, 2006

Company: Mobile Airwaves Corporation
Symbol: MBWC.PK
Current Price: $0.04
Target Price: $0.3 - $0.5
Rating: 9(10)
Recommendation: Strong Buy / Hold

MBWC is a high growth issue and should be purchased by stock traders and those that can afford to make quick money on these fast moving issues. We give it to you again as a gift. Good luck and trade out at the top!!!

Figure 5: A spam image with a content obscuring technique resulting in large background clutter (see the caption of figure 3).

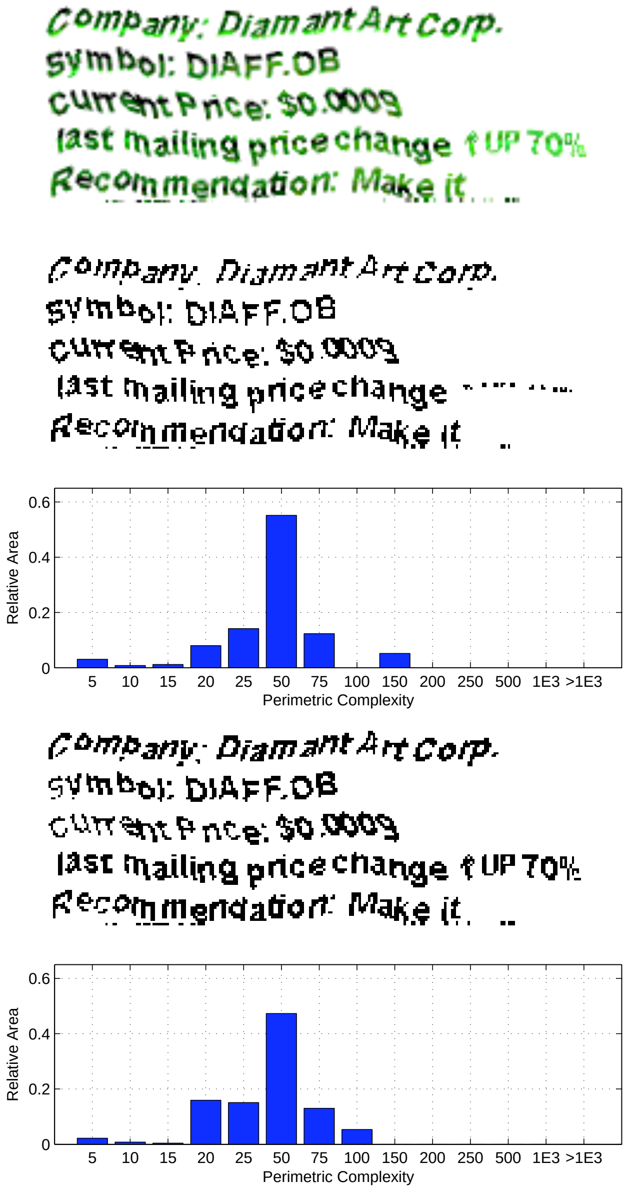


Figure 6: A spam image with a typical CAPTCHA-like content obscuring technique resulting in character distortion, without significant character breaking or merging, or background clutter (see the caption of figure 3).

do not contain text. In figures 7-9 (first row) we show next to the original image the binarized one (obtained using thw ABBYY OCR tool mentioned above) and the corresponding histogram of relative area vs perimetric complexity. For the image in figure 7 it can be seen that all components have P^2/A values around 60, in the range of clean characters. In the case of figures 8 and 9, most of the area belongs to components with high P^2/A complexity, but there are also some character-like components. We point out that the histograms of these images are not very similar to the ones of images with either clean or noisy text, which suggests that our approach could also discriminate images which contain text from images which do not. However we believe that a more robust approach is to check first the presence of text with specific techniques [8], and then to apply a detector of obscuring techniques only to images which do contain text.

Consider now the same three images above with some embedded text, to simulate legitimate images with embedded text. To this aim we used a sentence containing al 26 characters of the English alphabet (see figures 7-9, middle row, left). In the first two images we placed the text on regions with uniform background (this would be a reasonable choice in a legitimate image), while in the third one the text is placed on a complex background and is clearly more difficult to read for an OCR tool. The corresponding binarized imges and histograms are shown again next to the original images. If the histogram is computed on the whole image, in the case of figure 7 it correctly denotes the presence of clean text. However the histogram for figure 8 erroneously denotes the presence of obscured text, since about half the area of foreground components corresponds to high complexity objects, although they do not interfere with text. Finally, the histogram of figure 9 correctly denotes the presence of obscured text due to high complexity components; however we point out that, as in the case of figure 8, such histogram does not allow to distinguish noise components which interfere with text from noise components which do not.

Consider finally the correct way to apply the method proposed in this section, namely computing the histogram only on areas of the binarized image which contain text. These areas are shown in figures 7-9, bottom row, left, together with the corresponding binarized images and histograms. Now the histograms correctly denote the presence of clean text in figures figures 7 and 8, and the presence of noisy text in figure 9. It is worth noting that the noisy text of figure 9 would increase the likelihood that the corresponding e-mail is spam. However this does not necessarily means that it would be labelled as spam, since the actual label will depend also on the outputs of the other modules of a spam filter. In other words, this problem is similar to the one caused to bayesian filters by the presence of typical spam words like “viagra”, “buy”, etc. in the body of a legitimate e-mail.

To sum up, the above preliminary results show that measures of image quality related to embedded text (like perimetric complexity) could be directly exploited, or suggest the development of new measures, to develop filtering modules against image spam aimed at detecting the use of different kinds of content obscuring techniques.

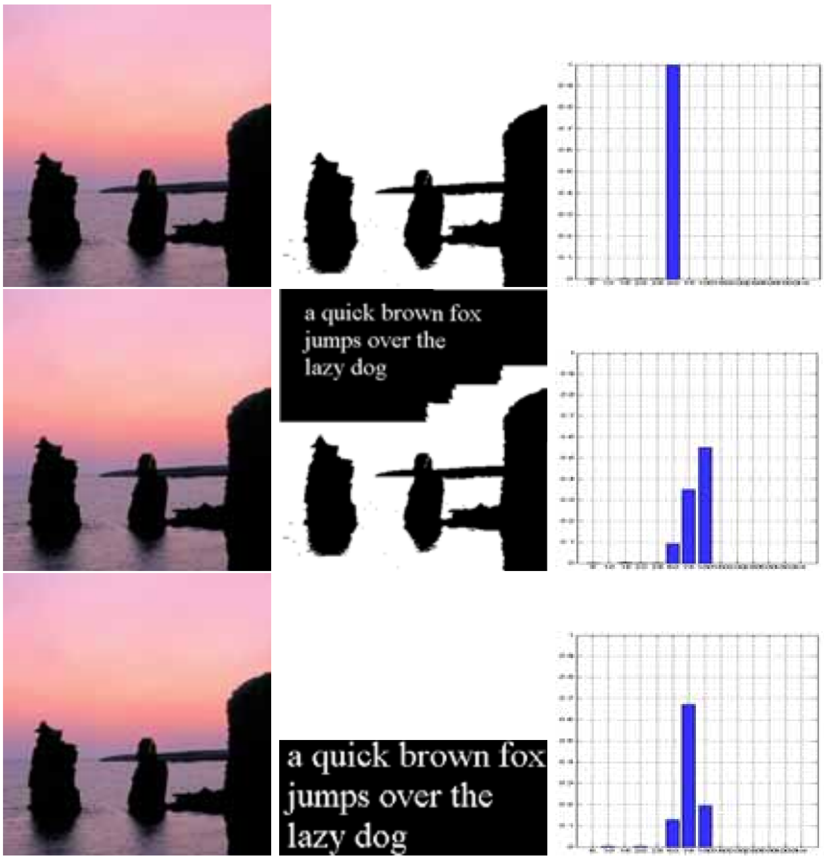


Figure 7: A legitimate image (a photograph): without embedded text (top row), with embedded text (middle row) and the text area only (bottom row), together with the corresponding binarized image and the histogram of component area vs perimetric complexity.

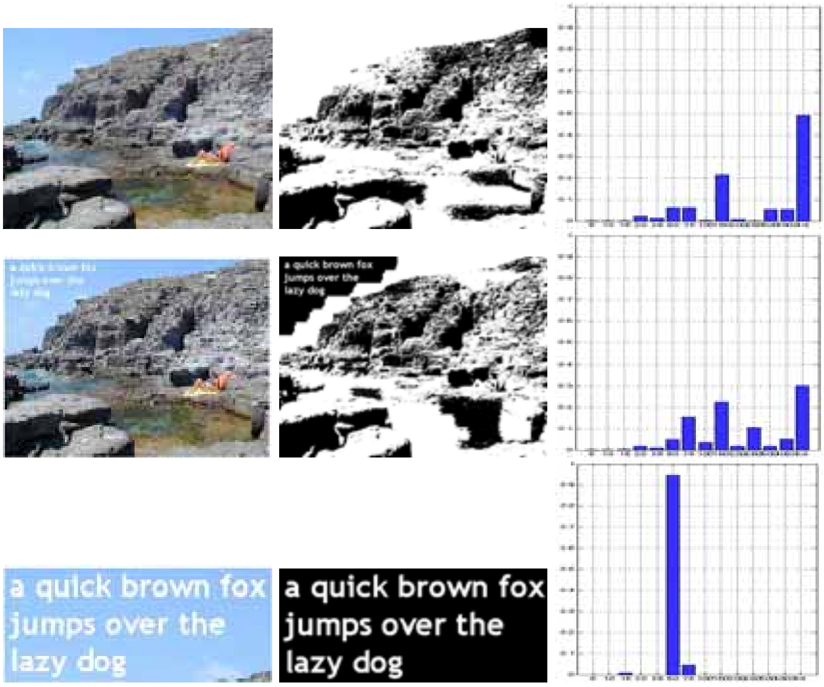


Figure 8: See caption of figure 7.

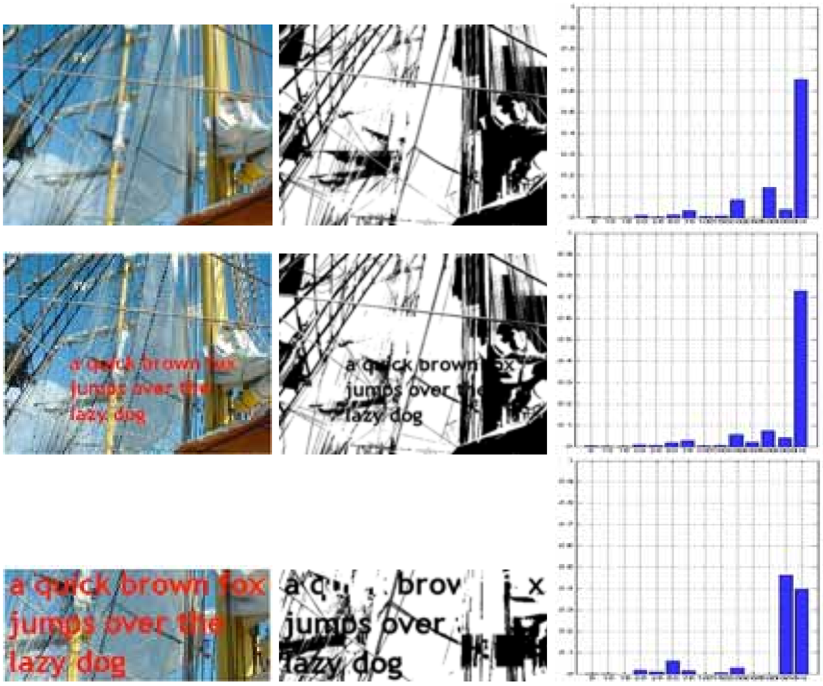


Figure 9: See caption of figure 7.

4 Conclusions

Image spam is becoming one of the main kinds of spam, and the use of content obscuring techniques against OCR is likely to rapidly spread, although we believe that they can not be exploited in some kinds of spam, for instance phishing. This suggests that computer vision and pattern recognition techniques will play a prominent role in the development of the next generation of spam filters. In this paper we considered a spam filter architecture made up of several modules arranged in parallel or hierarchically, each acting as a detector of specific characteristics of spam e-mails, whose outputs have to be properly combined to reach a final reliable decision about the "spamminess" of an input e-mail.

In this context, we discussed two possible approaches against image spam. One is based on exploiting the high-level textual information embedded into images, through the use of OCR techniques and text categorization techniques which proved to be effective for e-mail's body text. We reported experimental results which show that OCR can be effective for images in which no content obscuring techniques are used. The second approach can be viewed as complementary to the first one, since it is aimed at detecting the use of content obscuring techniques (the "noise", namely the adversarial clutter contained in the image) which can make OCR ineffective, instead of extracting the high-level textual information (the "signal"). We proposed a possible implementation of this approach for the specific case of obscuring techniques which result in character breaking or in noise components interfering with characters, which compromise the binarization pre-processing step performed by OCR algorithms.

We believe that works in the OCR literature, concerning in particular image quality measures, and recent works on CAPTCHAs could give useful suggestions for the development of this research direction.

References

[1] A. Androutsopoulos, J. Koutsias, K.V. Cbandrinos and C.D. Spyropoulos. An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd ACM International Conference on Research and Developments in Information Retrieval*, pages 160–167, Athens, Greece, 2000.

[2] H.B. Aradhye, G.K. Myers and James A. Herson. Image Analysis for Efficient Categorization of Image-based Spam E-mail. In: *Proc. 8th Int. Conf. Document Analysis and Recognition*, pages 914–918, 2005.

[3] H. S. Baird and M. Chew. BaffleText: a Human Interactive Proof. In: *Proc. IS&T/SPIE Document Recognition & Retrieval Conf.*, 2003.

[4] L.R. Blando, J. Kanai and T.A. Nartker. Prediction of OCR Accuracy Using Simple Image Features. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 319–322, 1995.

[5] H. Drucker, D. Wu and V.N. Vapnik. Support vector machines for spam categorization. *IEEE Transaction on Neural Networks*, 10(5):1048–1054, 1999.

[6] G. Fumera, I. Pillai and F. Roli. Spam filtering based on the analysis of text information embedded into images. *Journal of Machine Learning Research* (special issue on Machine Learning in Computer Security), 7:2699–2720, 2006.

[7] P. Graham. A plan for spam. `http://paulgraham.com/spam.html`, (2002)

[8] K. Jung, K.I. Kim and A.K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition* 37:977–997, 2004.

[9] T. Joachims. Making large-Scale SVM learning practical. In B. Schölkopf, C. Burges and A. Smola, editors, *Advances in kernel methods - Support vector learning*, pages 41–46, MIT-Press, 1999.

[10] B. Klimt and Y. Yang. The Enron corpus: A new data set for e-mail classification research. In *Proceedings of the European Conference on Machine Learning*, pages 217–226, 2004.

[11] A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *Proceedings of the AAAI Workshop on learning for text categorization*, pages 41–48, 1998.

[12] D.G. Pelli, C.W. Burns, B.Farell and D.C. Moore-Page. Feature detection and letter identification. *Vision Research*, 46:4646–4674, 2006.

[13] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz. A Bayesian approach to filtering junk e-mail. AAAI Technical Report WS-98-05, Madison, Wisconsin, 1998.

[14] C.-T. Wu, K.-T. Cheng, Q. Zhu and Yi-L. Wu. Using visual features for anti-spam filtering In *Proceedings of the IEEE International Conference on Image Processing*, Vol. III, pages 501–504, 2005.