Image Spam Filtering Using Visual Information

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

Abstract

We address the problem of recognizing the so-called image spam, which consists in embedding the spam message into attached images to defeat techniques based on the analysis of e-mails' body text, and in using content obscuring techniques to defeat OCR tools. We propose an approach to recognize image spam based on detecting the presence of content obscuring techniques, and describe a possible implementation based on two low-level image features aimed at detecting obscuring techniques whose consequence is to compromise the OCR effectiveness resulting in character breaking or merging, or in the presence of noise interfering with characters in the binarized image. A preliminary experimental investigation of this approach is reported on a personal data set of spam images.

1 Introduction

Spam filtering consists in discriminating spam e-mails from legitimate ones, and its adversarial environment makes it a very challenging task. Several techniques are currently used in commercial and open-source spam filters, and are being investigated by computer science researchers. Some of them are aimed at recognizing spam on the basis of the text in the e-mail's body. To this aim, the use of text categorization techniques has been investigated in the machine learning and pattern recognition fields in the past ten years [11, 5, 7, 1]. To defeat such techniques, spammers have first reacted using tricks like misspelling words or adding bogus text to their e-mails. Recently however they introduced a different kind of trick, which has rapidly spread during the past year, and can make all techniques based on the analysis of body text ineffective: it consists in embedding the spam message into attached images and is known as *image spam*. Moreover, content obscuring techniques are being increasingly used to defeat the attempts of using OCR tools (see Fig. 7). It is also worth noting that spammers could exploit to their advantage techniques used to create CAPTCHAs (see Fig. 7, bottom), which were introduced just to defend against robot spamming. To effectively tackle this new kind of spam, specific computer vision and image analysis techniques are clearly required. To our knowledge, this topic has not yet been addressed in the pattern recognition field. A first contribution was given by the authors in [6], where the use of state-of-the-art text categorization techniques on text extracted by OCR tools from images attached to e-mails was thoroughly investigated. This analysis showed that OCR tools can be effective against image spam, for cases in which no content obscuring techniques are used by spammers. This is a relevant result since it is likely that such techniques can not be exploited in every kind of spam: for instance, content obscuring can not be excessive in *phishing* e-mails, which should look as if they come from reputable senders, and thus should be as "clean" as possible.

In this paper we focus instead on the detection of spam images in which content obscuring techniques are used to the extent that standard OCR tools are likely to be ineffective. Differently from techniques recently proposed against image spam, based on discriminating spam images from legitimate ones through the use of generic low-level image features related to the presence of embedded text [2, 12], we propose in Sect. 2 an approach whose rationale is to detect specifically the use of content obscuring techniques applied to embedded text, and to assess the likelihood of the image being spam through measures of the extent of image clutter. This can be viewed as a complementary approach to the one based on OCR tools investigated in [6], since it aims at detecting the "noise" (the adversarial clutter contained in the image) instead of the "signal" (the spam text message). We also point out that our approach can be exploited in the context of 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 (for instance, this is the case of the well-known open-source SpamAssassin filter,

http://spamassassin.apache.org/).

A possible implementation of our approach, which we are currently investigating, is described in Sect. 3. This implementation is aimed at detecting content obscuring techniques whose consequence is to compromise the OCR effectiveness through character breaking or merging, or through the presence of noise components (like small dots) which interfere with characters. To this aim, two measures of the degree of image obscuring are proposed. In Sect. 4 a preliminary experimental analysis of our approach is reported, to assess the capability of the proposed features to detect the considered content obscuring techniques. We finally discuss some open issues and future research directions in Sect. 5.

2 Proposed approach

Recently some authors proposed techniques for recognizing spam images based on detecting the presence of embedded text, and on characterizing text areas with low level features like their size relative to the image [2, 12] or 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. We point out however that the discriminant capability of the above features is not likely to be satisfactory, since they are related on relatively generic characteristics of images, and since it is very difficult to collect representative samples of legitimate images for classifier training.

The approach we propose in this work is based instead on looking for a specific characteristic of spam images with embedded text, namely the presence of content obscuring techniques. The rationale of this approach is that images with embedded text which are obscured in a way aimed to make OCR ineffective are likely to be spam. In principle, different kinds of content obscuring techniques can be used against OCR: adding random background noise (like small dots) which interfere with text, using a non-uniform background, using pixels of different colours for each character, distorting text lines or single characters etc. To this aim, methods developed for building CAPTCHAs can be used as well. It is thus not likely that a single method can be able to detect all possible kinds of content obscuring techniques. On the other hand, specialized methods to detect a precise obscuring technique (for instance, the presence of random dots) are likely to exhibit a very limited generalization capability and to be easily defeated by spammers, as happens for text analysis techniques based on keyword search. We therefore propose a midway approach based on the following consideration: since the goal of content obscuring techniques is to make OCR algorithms ineffective, they could be detected by looking at their effects on the image processing steps carried out by OCR algorithms. The first processing step is image binarization. In a low-quality binarized image characters could 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. Many content obscuring techniques, even among the ones mentioned above, result indeed in such kind of defects in the binarized image. Accordingly, when an image has embedded text (whose simple presence can be detected even in complex images using techniques like the ones surveyed in [8]), a possible way to detect such kind of obscuring techniques is to analyze the binarized image to detect the presence of the above defects and to measure their extent. The outcome of this analysis can be used as the output of a spam filtering module which has subsequently to be properly combined with the outputs of other modules (aimed at detecting other characteristics of spam e-mails) to reach a final decision about the "spamminess" of the input e-mail, as explained in Sect. 1. In the next section we propose two possible low-level image features aimed at measuring the extent of character breaking and merging, and of text interference with noise components.

3 Low-level features to detect content obscuring techniques

The problem of detecting the presence and measuring the extent of image defects which can compromise the effectiveness of OCR tools, and in particular the ones mentioned in the previous section, has some analogies with the problem of measuring image *quality* addressed 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. This method is not directly applicable to our task since it is based on strict assumptions which do not hold in spam images, for instance equally sized 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 [10]). 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. Perimetric complexity measured on the whole image is not suitable to our task either, for instance because it strongly depends on text length. We nevertheless found that it can be usefully exploited to our purposes if it is applied to *individ-*

ual components of a binarized image, as described below.

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 is approximately in the range $[15, 150]$. 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 characters are broken, the resulting components will have on average a lower area than clean characters (the area is defined as the number of black pixel), and we observed that some of them (but not all) have a lower P^2/A than 15. Small components originated from background noise like dots, clumps and small line segments (see Fig. 7, second image from top), will be characterized on average by a lower P^2/A value than 15 and a lower area than clean characters. Instead, merged characters and large noise components will exhibit on average larger P^2/A and area values. These considerations suggested us two measures of the degree of image degradation, one related to broken characters and small noise components, the other to merged characters and large noise components. The starting point is a binarized (b/w) image, on which all connected components are first identified (in this work, 8-connectivity is used to this aim), and the P^2/A value is computed for each of them.

To measure the extent of character breaking or the presence of small noise components interfering with characters, we first subdivide the image into a grid of $p \times q$ equally sized blocks $\mathrm{b}_{ij}, i = 1, \ldots, p; j = 1, \ldots, q$ ($p = q = 10$ was used in this work). For each block b_{ij} we consider the components whose centre of mass belongs to the block, and compute the number c_{ij} of components with P^2/A in the range $[15, 150]$ (these should correspond to characters), and the number n_{ij} of components with P^2/A lower than 15 (these should correspond to small noise components or to broken characters). Then, considering only the blocks with $c_{ij} > 0$, i.e. containing at least one component with a character-like complexity (we denote their number with b_0), we compute for each of them the fraction of noise components $f_{ij} = n_{ij}/(n_{ij} + c_{ij}) \in [0, 1)$. High values of f_{ij} mean that in the considered block there are relatively few character-like components and many noise-like components (or components resulting from broken characters): such a block is thus likely to contain highly degraded text. Low f_{ij} values should instead denote the presence of clean text. We finally compute the average fraction of noise components over all blocks, $f_{\text{noise}} = \frac{1}{b_0} \sum_{\mathrm{b}_{ij}: c_{ij} > 0} \frac{n_{ij}}{n_{ij}+c_{ij}} \in [0, 1]$. If $c_{ij} = 0$ for all blocks, this means that all image components have a P^2/A value lower than 15 and are thus "noisy": f_{noise} is defined accordingly as 1.

The second measure we devised is aimed to detect the presence of merged characters and large noise components: its rationale is that such objects have on average both a larger P^2/A and a larger area than clean characters, as explained above. Their presence could then be detected by computing the average P^2/A value over all image components, each weighted with its relative area, defined as the ratio between the number of its black pixels and the number A_{tot} of black pixels in the whole image: $\overline{P^2/A} = \frac{1}{N} \sum_{k=1}^{N} \frac{P_k^2}{A_k} \times \frac{A_k}{A_{\text{tot}}}$, where P_k and A_k denote the perimeter and area of the k-th component, and N is the number of components in the whole image. Note that for an image of $P \times Q$ pixels the P^2/A value of a single component ranges between 4 (corresponding to a single black pixel) to $8PQ$ (corresponding to a "chessboard" with squares of 1×1 pixels, which is the 8-neighborhood connected component with the highest perimetric complexity).

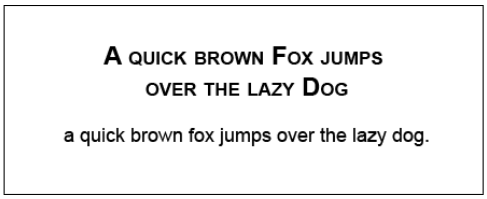
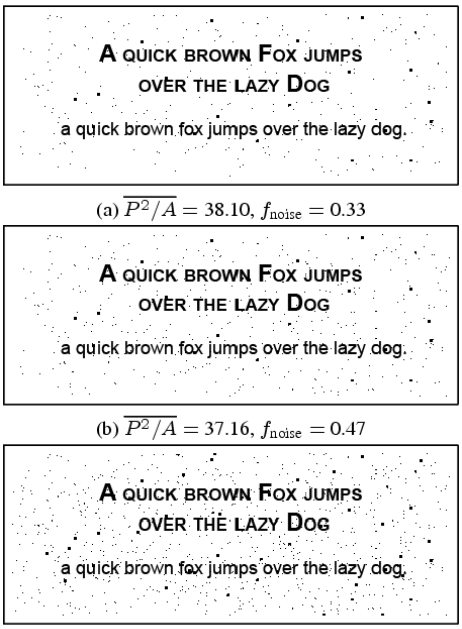


Figure 1: A clean artificial image: $\overline{P^2/A} = 38.85$, $f_{\text{noise}} = 0.013$

In the next section we present some preliminary experiments to demonstrate how the measures defined above could be exploited to detect the presence and measure the extent of image degradation due to broken or merged characters and to the presence of noise components.

4 Experimental results

We give first a demonstration of the capability of the two features defined in the previous section to measure the extent of the considered kinds of image degradation on the artificial image shown in Fig. 1, in which all the 26 characters of the English alphabet are present, both uppercase and lowercase. We considered four kinds of degradation: small random noise components of different size and density (Fig. 2), characters broken by a grid made up of 1-pixel wide white lines with different spacing (Fig. 3), reduced character spacing resulting in merged characters (Fig. 4) and characters merged with a grid made up of 1-pixel wide black lines with different spacing (Fig. 5). Note that obscuring techniques similar to the second one above were observed in real spam images recently received by the authors (see Fig. 7, third image from top), while the last technique was proposed in the EZ-Gimpy visual CAPTCHA which was used by Yahoo (see [9]). In Figs. 2-5 three different levels of image degradation are shown for each kind of degradation, together with the corresponding values of f_{noise} and $\overline{P^2/A}$. From Figs. 2 and 3, in which the obscuring techniques result in broken

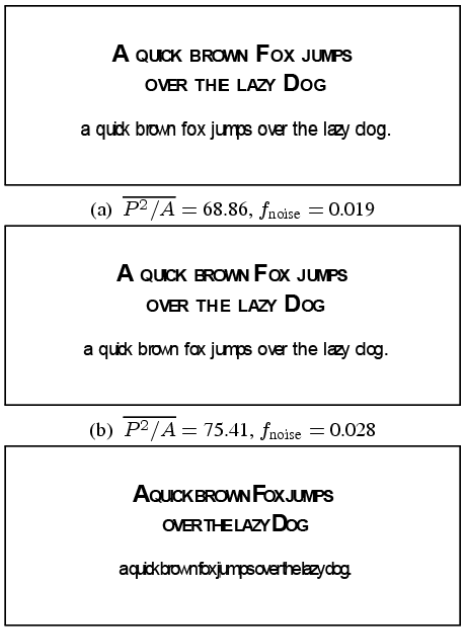


(a) $\overline{P^2/A} = 38.10$, $f_{\text{noise}} = 0.33$

(b) $\overline{P^2/A} = 37.16$, $f_{\text{noise}} = 0.47$

(c) $\overline{P^2/A} = 34.86$, $f_{\text{noise}} = 0.63$

Figure 2: Small noise components added to the clean image.

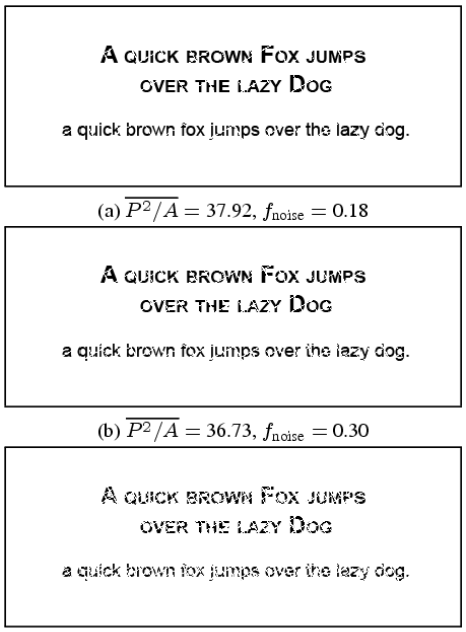


(a) $\overline{P^2/A} = 68.86$, $f_{\text{noise}} = 0.019$

(b) $\overline{P^2/A} = 75.41$, $f_{\text{noise}} = 0.028$

(c) $\overline{P^2/A} = 106.28$, $f_{\text{noise}} = 0$

Figure 4: Merging characters by reducing their spacing.

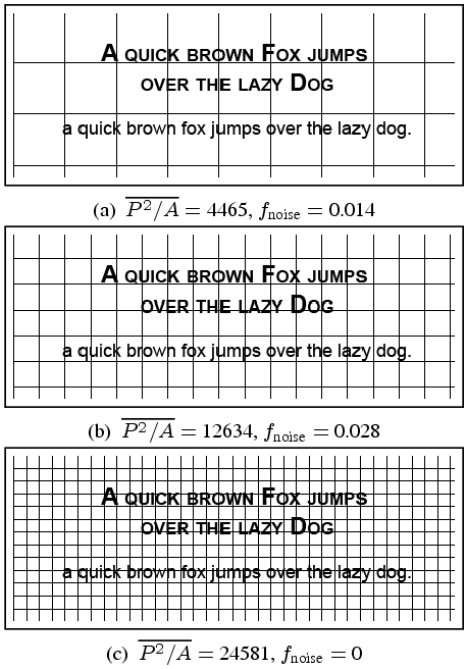


(a) $\overline{P^2/A} = 37.92$, $f_{\text{noise}} = 0.18$

(b) $\overline{P^2/A} = 36.73$, $f_{\text{noise}} = 0.30$

(c) $\overline{P^2/A} = 29.35$, $f_{\text{noise}} = 0.53$

Figure 3: Characters broken by a grid of white lines.



(a) $\overline{P^2/A} = 4465$, $f_{\text{noise}} = 0.014$

(b) $\overline{P^2/A} = 12634$, $f_{\text{noise}} = 0.028$

(c) $\overline{P^2/A} = 24581$, $f_{\text{noise}} = 0$

Figure 5: Merging characters with a grid of black lines.

characters or small noise components, it can be seen that the f_{noise} measure increases approaching 1 as the degradation level increases, starting from a value near zero for the clean image of Fig. 1. This agrees to the desired behaviour. The value of $\overline{P^2/A}$ is in the range $[15, 150]$ of character values for the clean image. Merged characters due to reduced spacing (Fig. 4) leads to increasing $\overline{P^2/A}$, although they remain in the range $[15, 150]$. Instead, merging characters with a grid (Fig. 5) leads to much larger values of $\overline{P^2/A}$ far from the range $[15, 150]$, which increase for increasing degradation level.

We now report some preliminary experimental results on a data set of 186 real spam images collected at the personal authors' mailboxes (since no publicly available data sets of spam images is available yet, to our knowledge). These images are available at `http://ce.diee.unica.it/spam-images.zip`. Content obscuring techniques clearly aimed at defeating OCR tools were applied by spammers on 96 of these images (see the examples in Fig. 7), while the remaining 90 images were either clean, or contained a limited amount of random noise probably aimed at defeating detection techniques based on image digital signatures, which is however negligible to OCR. Note that some of the obscuring techniques against OCR do not result in the kinds of image degradation considered in Sect. 3 (as in Fig. 7, bottom). Binarization of these images was carried out using the demo version of the commercial software ABBYY FineReader 7.0 Professional (`http://www.abbyy.com/`), using default parameter settings. We found that on 29 out of 96 noisy images the outcome of the binarization was a good quality image, despite the use of content obscuring techniques. Fig. 6 shows the values of $\overline{P^2/A}$ and f_{noise} for all 186 images in a two-dimensional plot. The main observation which can be drawn from this plot is that clean binarized images form a relatively compact cluster, while degraded binarized images are spread across a larger region of the $\overline{P^2/A}$–f_{noise} space. There is some overlapping between the two clusters, but we observed that images in the overlapping region correspond to intermediate levels of degradation (as in the examples of Figs. 2-5). This is a good indication that the above measures could be exploited to obtain an overall measure of the extent of image degradation. For instance, they could be used as features for a classification algorithm like a one-class classifier, whose continuous valued output (properly scaled) could be the output of a spam filtering module aimed at detecting the considered kinds of image degradation.

We conclude by pointing out some cases in which the proposed measures failed to detect image degradation. Some images were obscured using different background colours, which resulted in large noise components interfering with characters (see the example in Fig. 7, fourth image from top). However this was not detected by the $\overline{P^2/A}$ measure, since the noise components turned out to exhibit a P^2/A value in the same range as clean characters. Moreover, some clean images had a textured frame around the text but not interfering with it. However, after binarization the frame generated noise components exhibiting a wide range of P^2/A values: some of them exhibiting a character-like P^2/A value were erroneously considered as characters surrounded either by small noise components (by the f_{noise} measure), or by large noise components (by the f_{noise} measure). A possible improvement to the proposed measures to avoid these problems could be to take into account also the size of the bounding box of components with a character-like P^2/A values.

Figure 7: Examples of real spam images (details) from the data set used for the experiments, with different kinds of content obscuring techniques. From top: multi-coloured text interfering with non-uniform background; small random noise interfering with text; characters broken by lines of the same colour as the background; non-uniform background; a CAPTCHA-like technique.

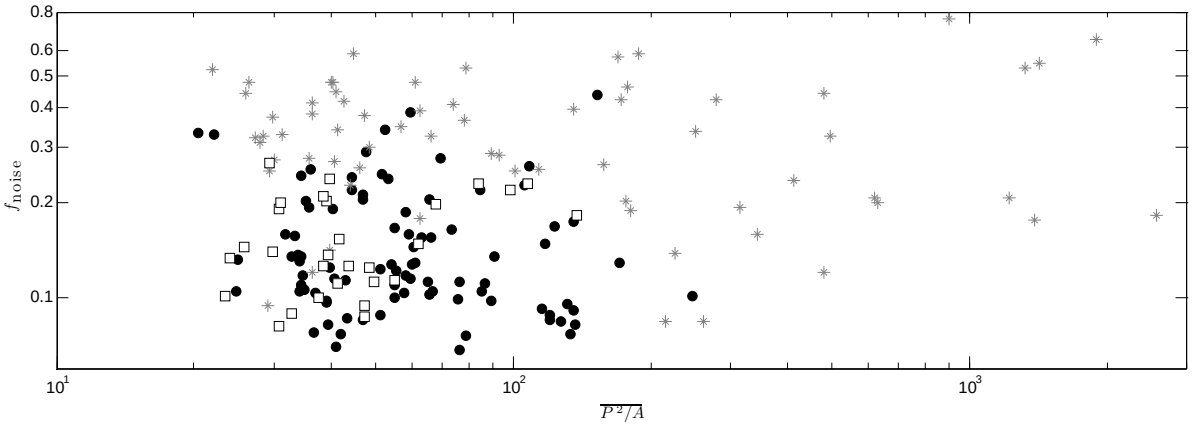


Figure 6: Plot of f_{noise} vs $\overline{P^2/A}$ for the 186 spam images (log-log scale). •: clean images; □: obscured images resulting in clean binarized images; *: obscured images resulting in degraded binarized images.

5 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 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 proposed a possible approach against image spam based on recognizing the use of content obscuring techniques (the "noise", namely the adversarial clutter contained in the image) which can make OCR ineffective, through the detection of their effects on the processing steps of OCR algorithms. We then proposed two measures of image degradation focused on obscuring techniques which result in character breaking or merging and in background noise interfering with characters. Preliminary experimental results showed that the proposed approach is an interesting research direction against image spam. Works in the OCR literature, in particular related to image quality measures, and recent works on CAPTCHAs could give further 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 *Proc. ACM Int. Conf. on Research and Developments in Information Retrieval*, pages 160–167, 2000.

[2] H. Aradhye, G. Myers, and J. A. Herson. Image analysis for efficient categorization of image-based spam e-mail. In *Proc. 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 *Proc. Int. Conf. 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, 2002. http://paulgraham.com/spam.html.

[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] G. Mori and J. Malik. Recognizing objects in adversarial clutter: breaking a visual captcha. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, volume I, pages 134–141, 2003.

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

[11] 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.

[12] C.-T. Wu, K.-T. Cheng, Q. Zhu, and Y.-L. Wu. Using visual features for anti-spam filtering. In *Proc. IEEE Int. Conf. on Image Processing*, volume III, pages 501–504, 2005.