

High Speed Image Part Recognition (IPR)

Patrik Ostrihon, COMDOM Software

Reza Rajabiun, COMDOM Software and York University, Toronto

Abstract

As anti-spam filters have improved in their capacity to process text based messages, spammers have learned to convey their advertisements using a number of different “envelopes”. These envelopes include documents, pdf files and graphical formats. Although it is relatively easy to construct antispam filters that read and process content embedded in some of these envelopes, image spam has challenged the analytical capacity of academic and industry researchers. This paper reviews different technologies employed to process image spam, and introduces a new approach to processing content embedded in complex and randomized pictures. High speed Image Part Recognition (IPR) improves on existing ad hoc rules and fingerprint systems because it allows for analysis of image spam content, and can be integrated with self learning Bayesian (statistical) filters.

Corresponding author: Reza Rajabiun <reza.rajabiun@comdomsoft.com>

Copyright © 2007 COMDOM® Software



Introduction

As anti-spam filters have improved in their capacity to process text based messages, spammers have learned to “envelope” their communications in a number of different formats. These envelopes include documents, pdf files and graphical layouts that embed the advertisements that spammers wish to send. Although it is relatively easy to construct filters that read and process content embedded in some of these envelopes, image spam has challenged the analytical capacity of academic and industry researchers. This paper reviews recent methods developed to analyze image spam, and introduces a more robust and economical alternative.

The most pressing problem raised by image spam is the large computational power necessary to process incoming content using traditional Optimal Character Recognition (OCR) techniques. Available OCR methods, first developed to read terrestrial mail envelopes by postal systems after WWII, intended to read writing by senders who wanted their messages to reach their destination. Hence, the recognition processors were designed to expect some degree of convergence between the patterns of writing they observe. In the case of image spam, the situation is radically different, as senders try to hide the content of their messages from the fingerprint or Bayesian filters that are now in place in most telecommunication companies and large organizations that provide access to the end users of the internet. In general, image spam today is smart, in the sense that robots commonly use CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) to randomize the envelope containing their advertisements.¹

Because of the processing costs of image spam, many network administrators have simply limited the ability of their end users to receive messages containing images. This simple solution has the disadvantage that it limits the usefulness of email as a communication device for business and personal use. Over the past year, a number of academic researchers and software security companies have turned their attention to finding more constructive methods to identify and filter image spam, without degrading the quality of the network. This paper reviews two of the recent efforts at developing rule based classifiers and fingerprinting/checksum methods to identify image spam, and introduces a high speed alternative to these mechanisms. COMDOM® Image Part Recognition (IPR) is specially designed to identify the content within low resolution and ever changing images, and hence can be seamlessly integrated with self learning Bayesian content filters.²

¹ <http://www.captcha.net/>

² Which are generally more accurate than ad hoc challenge response and fingerprinting systems employed for spam control. See Antispam Technology Impact Assessment Report: 2007. www.comdomsoft.com.

Although image spam is only one part of the ongoing battles between developers of filtering and spamming software, this class of messages have been recently employed in relatively sophisticated internet schemes, for instance in “pump and dump” scams of corporate securities. It is consequently important to integrate efficient and robust image spam scanning and filtering cores in content filtering systems that are able to learn about the patterns of undesirable content in real time. The next section reviews the approach by Dredze et al. (2007), which relies on ad hoc rules like size and color of a picture to classify it as spam/ham. Then we look at the advances in the fuzzy fingerprinting approach that relies on the mass characterization of spam offered by Wang et al. (2007), and introduce IPR as a more robust alternative to existing approaches to the problem.

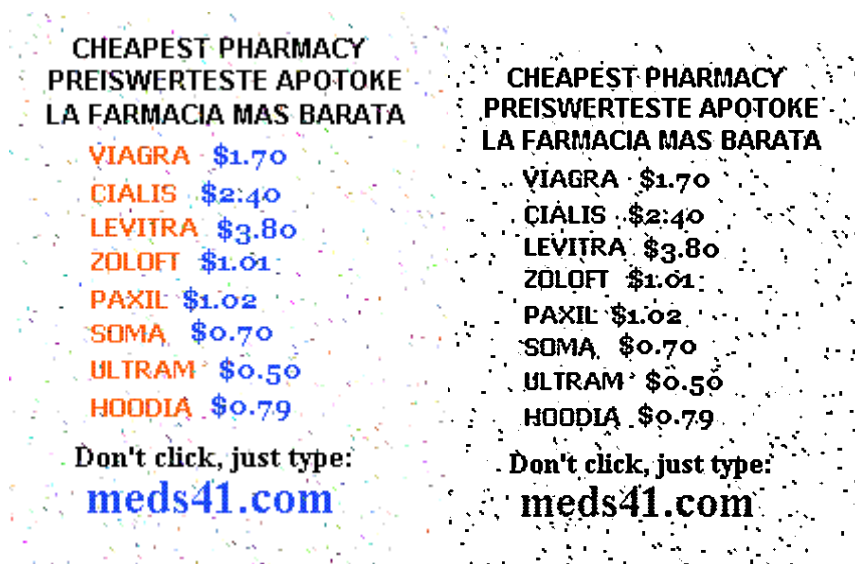
The paper documents why IPR: a) does not suffer from the well known shortcomings of challenge response and signature based systems, notably their ease of manipulation by spammers, and b) imposes much lower computational costs in terms of hardware than OCR. IPR decomposes an image into its constituent parts in order to read the characters and shapes used to construct spam messages. In combination with a high capacity Bayesian classifier, IPR offers a promising approach to fast and robust processing of image spam typically used in sophisticated advertising campaigns, for instance those in the corporate securities industry.

From OCR to ad hoc classifiers

As noted earlier, classical Optimal Character Recognition (OCR) were first developed with the specific aim of automating the processing of mail after World War II, and improved overtime. OCR methods incorporate a range of machine learning techniques that help them adjust to diverse patterns of hand writing, and classify the content after filtering the noise around the letters. For this reason this sets of methods are somewhat tolerant to the type of noise that spammers like to incorporate into messages enveloped in pictures, and hence useful as a point of departure for antispam filters. Moreover, with OCR, once the content of an image has been read, then the patterns of words can be processed by a high capacity Bayesian content filter which classifies the document based on subjective end user preferences.

Despite these two advantages, OCR techniques have not been widely employed to process image spam using commercial or open source filtering and scanning software available today. This may be because OCR is very slow, hence placing a heavy burden on the network infrastructure. Given that OCR methods were designed for people that want their mail envelopes to reach their destinations, spammers are easily able to avoid them. With divergent noise patterns in contemporary forms of image spam, the OCR spam recognition rates are

reduced relative to its other industrial applications. Overall, classical OCR requires high resolution images and is usually suitable for processing sub-base of 600dpi and more. Recognition of images with 75dpi and less (majority of spam images) causes problems to classical OCRs which then results in quite high error rates in recognition. The following example documents simple image spam able to avoid OCRs.



Fuzzy OCR methods have aspired to compensate for the high error rates of classical OCRs by toleration of error rates while making decisions about the images in messages. For example, when making a decision about the words 'vaga' or 'vtaera', a fuzzy OCR system could classify the content as 'viagra'. Despite the improved learning capacity, fuzzy OCRs do not significantly mitigate the computational burden of the classical OCRs, which after all only had to read simple addresses on the back of mostly white envelopes. These limitations highlight why some network administrators employ ad hoc rules for classification of spam, including simply by excluding all images from the network.

Dredze et al. (2007) review recent developments in the area of fuzzy OCRs and similarly argue that despite the advances, this approach remain too slow as a practical solution to complex image spam. In response to this technological gap they develop a general purpose system that can be employed with a variety of antispam packages. In order to deal with the speed problem, they instead devise a classification system that does not rely on reading the content within an image, but instead simply looks at the features of the images. These features include:

- File format
- File size
- Image size
- Average color

- Prevalent color coverage

They claim that with the addition of a simple algorithm that takes advantage of mutual information that computes a score for the joint information of the classes of features, the method achieves a 99% accuracy rate in detecting image spam compiled from their personal corpus of emails. This high level accuracy appears consistent with the static testing environment, but does not seem credible in the real world in which spammers, and even their robots, can change these features in a randomized manner. Although arguably not very practical because of the ease of bypassing this type of approach to processing image spam, their analysis highlights the importance of speed in the next generation of technologies for fighting complex forms of image spam. Indeed the authors concede that the efficacy of this approach would be limited over times as spammers learn to circumvent the system. They however do not address the possibility that such an ad hoc feature selection algorithm can expand in a manner that blocks legitimate messages, leading to false positive, as well as spam detection accuracy problem.

Fingerprinting

Although there is considerable agreement in industry and the academic community that Bayesian content filtering systems offer the best accuracy rates in the classification of text based messages, in the early 2000s many large operators adopted a second broad class of antispam technologies. As in the case of classical OCR in processing of image email messages, this shift to fingerprinting was partly motivated by the computational requirements of first generation of Bayesian filters. More recent Bayesian filters like COMDOM® Antispam are however at least 5 times faster than leading fingerprint system available to ISPs and large organizations.³

A fingerprint or checksum-based filters exploited the fact that spam messages are sent in bulk. These antispam technology functions essentially by stripping all context that may vary across messages, reduce what remains to a checksum or a fingerprint that defines that particular message within the population of all possible messages. Then, to allow the message to pass through to the end user, the system must compare the checksum with those collected in a centralized database of fingerprints. Hence, there is generally no analysis of the content in the messages.

There are different methods for constructing such a centralized database in real time. Some commercial software produces for example employ feedback buttons on their email client

³ See www.comdomsoft.com for more information on recent advances in the processing speed of Bayesian content filters.

which can be selected to nominate a particular message as spam. Within the system design, this nomination increases the probability of that similar messages are classified as spam within a centralized database. More advanced checksum filters used by ISPs today employ *fuzzy fingerprinting techniques*, which are able to identify outbreaks of new randomization and automation techniques developed by spammers that bypass simple fingerprint systems.

The advantage of this type of filtering is that it lets ordinary users help identify spam, thus vastly increasing the pool of spam fighters. Challenge response and blacklisting systems of the late 1990s primarily had relied on the discretion of administrators in setting up ad hoc rules for the identification of spam. By the early 2000s, spammers had shown their ability to bypass these rule based systems, generally by using techniques able to produce large volumes of spam messages that imply the same content to the human eye, but look different to challenge response and basic fingerprint systems.

The disadvantage of the checksum or fingerprint systems is related precisely to its reliance on ad hoc rules for the identification of text, or enveloped image spam as suggested by Dredze et al. (2007). Spammers can insert unique invisible gibberish—known as hashbusters—into the middle of each of their messages. This makes each message unique, and hence assigned a different checksum in the centralized database that relies on similarities among messages as its core principle for classification. As a result, the centralized databases grow, as does the need for communication between the end users and the expanding central repository of the fingerprints.

Further result of the battle between spammers and checksum systems since the early 2000s has been development of two particularly undesirable technologies by spammers. In the longer term, *Smart spam* and *BGP spectrum agility* are likely to continue to evade systems that rely on similarities among messages and origin authentication, hence polluting email, and increasingly mobile, messaging systems:

- A. Smart spam: Since checksum systems function by identifying spam through monitoring the quantities of similar messages sent in real time, spammers have learned to deploy spam sending robots that are able to send messages that automatically change the shape of individual spam emails. As directed by semi-autonomous robots, the mass mailings that contain the so called hashbusters look different to the fingerprint systems currently in place, but convey the same message to the eyes of the end users. The capacity of spammers to easily send messages that appear unique hence leads to very low spam detection rate in basic fingerprint systems, and has adopted the adoption of fuzzy fingerprinting methods. Image spam is one particular form of smart spam.

- B. In addition to mass mailings that use artificial intelligence techniques (AI) to alter their shape to avoid fingerprint systems, as detailed by Ramachandran and Feamster (2006), much of the total volume of spam is produced through one shot BGP spectrum agility techniques (They find that 11 of the top 20 spam originating ISPs are primarily based in the United States). This means that large numbers of well hidden robots placed on the network of large operators allow spammers to easily cloak their origins, and hence remain undetectable by modified fingerprint systems that aim to increase their detection rates by also blacklisting communication from known spam producing ISPs. (Moreover, they found that around 36% of all spam originates from robots in place in 20 large network providers within which spammers hide their robots and engage in one shot attacks from multiple senders, hence avoiding detection.)

The arms race between the developers of the checksum software and the writer of the spam-generating software since the early 2000s has contributed to the increase in the total volume of spam. According to Organization for Economic Cooperation and Development as of 2005 80% of all messages were spam, a figure that has arguably grown by today⁴ One particular manifestation of the arms since 2005 has been the increased use of methods for enveloping spam inside ever changing documents, pdf files, and images. Messages embedded in these envelopes not only autonomously change shape, but also significantly increase the computational costs associated with processing the spam flows. The result has been more than a simple increase in the volume of spam as noted by OECD, but also a radical expansion of the volume of bits that need to be processed by large network owners and operators.

From an economic perspective, the implication is that spammers have sufficient technical capacity to produce a positive response rate in environments that rely on the most advanced of fuzzy fingerprint systems in place today. Since the spammers do not incorporate the costs on hardware, software, and administration by the recipients, the false negative problem in these systems endogenously increases the volume of spam present on the network.

From a software engineering perspective, the ineffectual battle between fingerprint/checksum systems has motivated the development of heavy bundles that incorporate many different types of filters. The combination of multiple layers of filters, many of which are relatively ad hoc, slows down processing relative to a pure fuzzy fingerprint system that, at least in theory, was first in the early 2000s to respond to the needs of ISPs and large organizations for fast and accurate spam filters.

⁴ <http://www.oecd-antispam.org/>

Bundling of different filters has also lead to an interesting feature of real fuzzy fingerprint systems. In theory, the pure checksum method should not lead to any false positives. This is because each fingerprint can be produced to be unique at the limit. However, as spammers have learned to avoid the fingerprint checks, commercial bundles have resorted to using ad hoc challenge response solutions. For instance, some university and corporate administrators of networks simply limit the capacity of end user to receive images because of the bandwidth and processing costs. This means that even legitimate personal and business communications presented in such a format cannot be received.

Wang et al. (2007) extend the traditional fingerprinting approach to the filtering of smart text spam to those messages enveloped in graphical formats. They observe that image spam commonly employs similar pictures that only differ how they apply specific classes of templates and randomization methods by spammers. Common techniques for the construction of complex image spam include:

- Waves
- Animations
- Deformities
- Rotations

They also list at least 17 different randomization techniques used to add noise to the basic templates used to construct the spam messages. The length of this list highlights the enhanced capacity of spammers to generate smart spam that aims to bypass fingerprinting or ad hoc rule based systems. The authors claim that their model generates low false positive rates, consistent with the notion that in theory fingerprint systems should never reject messages that are not sent in mass. They further claim that such a system can reach over 96% accuracy rates, again in accordance with the standard spam detection rates for fuzzy fingerprinting of text under static test conditions.

Also improving the processing speed relative to fuzzy OCRs, as was the objective for method offered by Dredze et al (2007), the Wang et al. (2007) approach exhibits much lower spam detection rates, even under static tests described by the authors. Under more realistic conditions, it is relatively easy to see that spammers can easily bypass the rules relating to the features of pictures, and employ new randomization techniques (besides the 17 highlighted by Wang et al. (2007)). Low detection rates expected from the two proposals stem fundamentally from their lack of attention to the content of the advertisements, and primary focus on the shape of the envelope.

Image Part Recognition and Bayesian classification

Statistical filtering of spam was first proposed in by Sahami et al (1998). A statistical filter is a kind of document classification system, and a number of machine learning researchers have turned their attention to problems in implementing such an approach to mitigating the costs of spam. Statistical filtering was popularized by Paul Graham's influential 2002 article A Plan for Spam, which proposed the use of naive Bayes classifiers to predict whether messages are spam or not – based on collections of spam and non-spam (ham) email submitted by users.⁵ Statistical filtering, once set up, requires no maintenance per se: instead, users mark messages as spam or ham, and the filtering software learns from these judgments.

Thus, a statistical filter does not reflect the software author's or administrator's biases as to content, but accounts for user preferences. For instance, a biochemist who researches variations of pills sold as “*Viagra*” may teach the filter not to mark legitimate professional correspondence as spam. Similarly, price sensitive buyers of discounted sexual enhancers can signal the type of content they want to receive. In the longer term, the use of Bayesian filters forces senders to become more targeted in the mass emails they produce to grab attention. Loder et al. (2006) provide a useful economic model for the analysis of different approaches to combating spam, relative to a perfect Bayesian filter. COMDOM® Antispam provides the first highly optimized application of the ideal Bayesian filter designed for the requirements of ISPs and large organizations.

Spammers have attempted to fight statistical filtering by inserting many random but valid “noise” words or sentences into their messages in their attempts to fight such filters by increasing the likelihood and a message will be classified as neutral. Attempts to hide the noise words include setting them in tiny font or the same color as the background. However, these noise countermeasures seem to have been largely ineffective according to Graham-Cumming (2006). In this context, a Bayesian filter can learn to converge to perfection, as long as it is fast enough in relating the preferences of end users, and smart enough in responding to changing content of communications that must be classified.

Although there has been substantial agreement on the long term benefits of using Bayesian methods to fight spam in the academic community, arguably the most important problem has been constructing the proposed software. The adoption of fingerprint or checksum methods in the early to mid 2000s was partly motivated by lower systems requirements in centralized checksum systems, relative to early commercial or open source Bayesian filters. COMDOM® Antispam was designed to minimize systems resources, including hardware, software, and administrative oversight, hence addressing the gap in the market for robust, efficient, and self

⁵ <http://www.paulgraham.com/antispam.html>

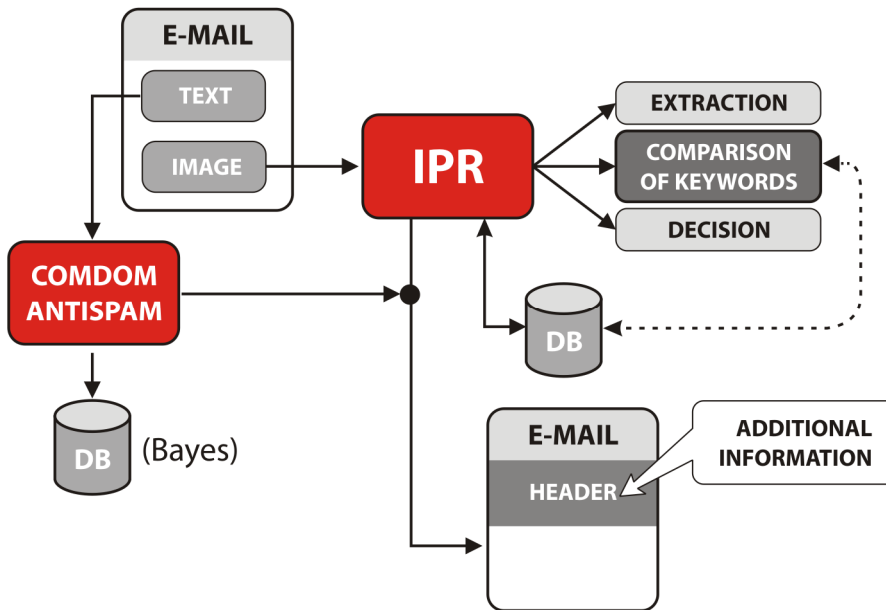
learning anti-spam needed in ISPs and large organizations. High speed Image Part Recognition® has been developed to complement the radical improvement in performance offered by COMDOM® Antispam in the treatment of text based spam. However, the technology may be integrated with other Bayesian filters since it is designed to recognize characters that make up the content of the advertisements enveloped in images.

Consequently, COMDOM Image Part Recognition retains the advantages of OCRs in terms of their ability to characterize the content of the advertisement in a manner that can be checked against the local or global database in a Bayesian system. However, IPR is specially developed to account for the incentives of spammers to obfuscate the letters that convey the message to the human eye. Specifically, image spam commonly employs low-resolution images (<75dpi), arguably because spammers know the limitations of existing fuzzy OCRs to read such messages accurately. Moreover, IPR autonomously can learn about the patterns of noise incorporated by spammers in their advertisement, and hence does not rely on a predetermined set of rules and administrative intervention when integrated with COMDOM® Antispam.

Instead of focusing on the features of the envelope, or reducing pictures to a fuzzy checksum, IPR relies on the Achilles Hell of spammers, the content of their advertisements. The recognition system in IPR segments incoming pictures into a number of constituent parts, and based on “learned” parts from possible images, reconstructs complex letters in image spam constructed using CAPTCHA methods. By deconstructing an image into its constituent parts and checking these parts against patterns of letter parts, the system is designed to be resistant to noise and other deformations of images used in image spam templates. (such as “affine transformations”, ie. rotation, scaling, shearing or shifting).

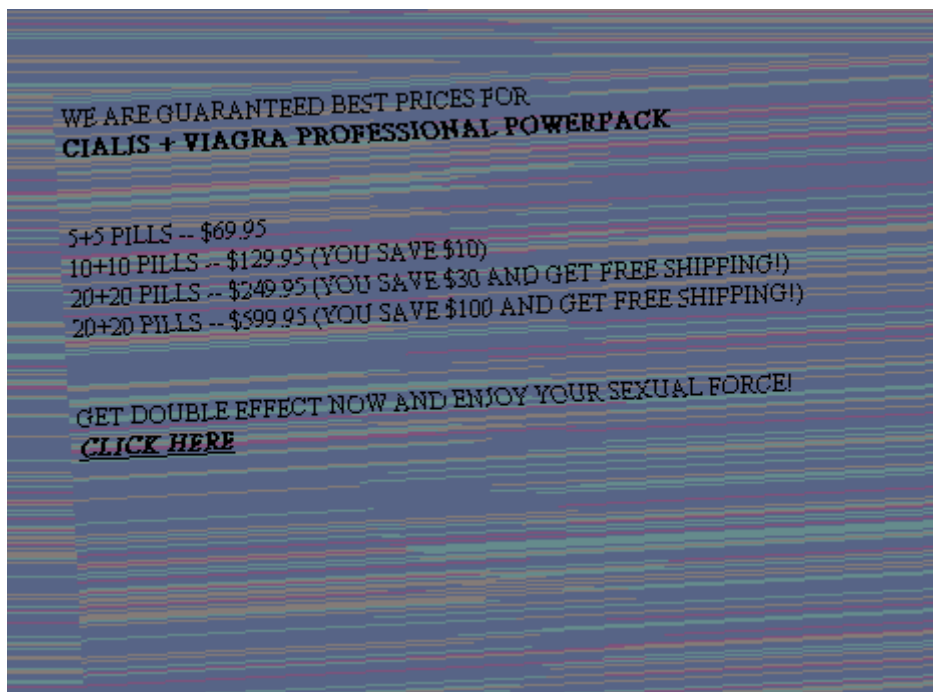
This design enables very fast recognition and classification of the image, despite low resolution and other types of noise. Battles between CAPTCHA algorithm writers and robots that aim to bypass these visual security barriers to computers have produced a wide variety of techniques for the randomization and obfuscation of text, now used by spamming robots. In the case of image spam processing, advances in Artificial Intelligence (AI) resulting from the battles show that computers can easily be trained to remove background clutter that envelopes texts. Moreover, these advances have meant that spam robots are easily able to send messages in pictures that are so highly distorted, so that they are unreadable even by the human eye. Since it does not make sense to send advertisements that cannot be read by human recipients, spammers using CAPTCHA are forced to finely balance computer and human capabilities in designing their messages. The implication is a finite set of image parts that are readable by humans, which are scanned, classified, and reconstructed as letters in IPR.

IPR is an application of the general class of probabilistic and inferential methods in computer vision designed for handling noise and speed in processing spam. (Forsyth and Ponce, 2003, Part IV for a review). As in other approaches to image analysis within this class, IPR constructs a database of pattern templates using classifiers, compares pattern templates, and makes the decision of a picture is spam based on variations in differences. The next figure presents the basic architecture of IPR and its role within a high capacity Bayesian system.



The next figures illustrate the relative robustness of this approach that looks at the differences in the properties of letters for high speed transformation of complex image content (our internal tests show that IPR is 10 to 50 times faster than commercial OCRs). This content can then be further processed by high capacity Bayesian filter that learns about the patterns of desirable and undesirable content based on end user or administrative preferences as illustrated above. The first set represents the output from standard filtering techniques that produce the input to be scanned and processed by IPR. The second depict the wide range of variations on letters that IPR can rapidly process, and the last figure highlights how patterns of objects other than letters that can also be analyzed using this approach to filtering image spam.

Smart Image Spam: Filtering and scanning with IPR



Filtering (Input for IPR)

WE ARE GUARANTEED BEST PRICES FOR
CIALIS + VIAGRA PROFESSIONAL POWERPACK

5+5 PILLS -- \$69.95
10+10 PILLS -- \$129.95 (YOU SAVE \$10)
20+20 PILLS -- \$249.95 (YOU SAVE \$30 AND GET FREE SHIPPING!)
20+20 PILLS -- \$399.95 (YOU SAVE \$100 AND GET FREE SHIPPING!)

GET DOUBLE EFFECT NOW AND ENJOY YOUR SEXUAL FORCE!
[CLICK HERE](#)



Recognition of words:

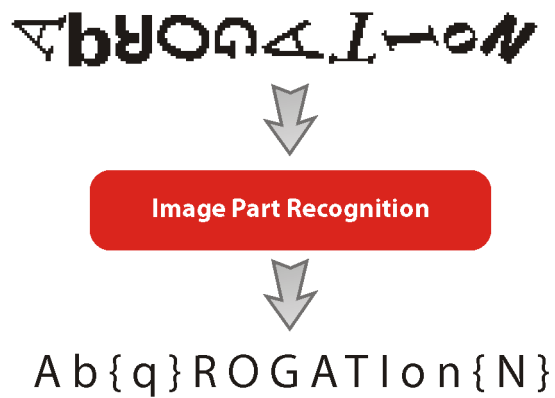


Figure 1



Figure 2

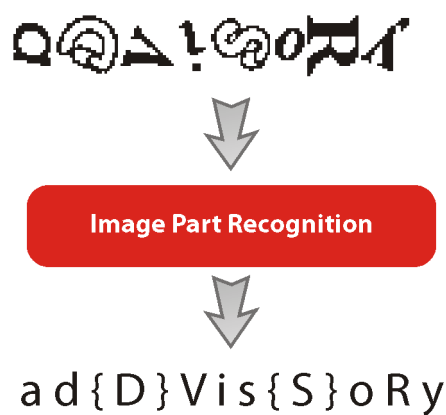


Figure 3







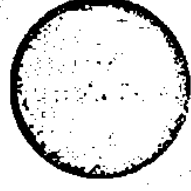

Other patterns:

Today's Bestsellers

 <p>Viagra Our price \$1.79</p>	 <p>Viagra Soft Tabs Our price \$2.05</p>	 <p>Cialis Soft Tabs Our price \$3.93</p>
 <p>Cialis Our price \$2.69</p>	 <p>Phentermine Our price \$5</p>	 <p>Xanax Our price \$2.99</p>
 <p>Valium Our price \$2.48</p>	 <p>Levitra Our price \$3.96</p>	 <p>Soma Our price \$0.67</p>

[ORDER NOW](#)

Today's Bestsellers

 <p>Viagra Our price \$1.79</p>	 <p>Viagra Soft Tabs Our price \$2.05</p>	 <p>Cialis Soft Tabs Our price \$3.93</p>
 <p>Cialis Our price \$2.69</p>	 <p>Phentermine Our price \$5</p>	 <p>Xanax Our price \$2.99</p>
 <p>Valium Our price \$2.48</p>	 <p>Levitra Our price \$3.96</p>	 <p>Soma Our price \$0.67</p>

[ORDER NOW](#)

References

Dredze, M. R. Gevaryahu, and A. Elias-Bachrach. (2007). Learning Fast Classifiers for Image Spam, CEAS 2007, Mountain View, U.S.A.

Forsyth, D. and Ponce, J. Computer Vision: A Modern Approach. Prentice Hall, 2003.

Graham-Cumming, J. Does Bayesian Poisoning Exist. 2006. Virus Bulletin.

Loder, T. M. Van Alstyne, and R. Wash. An Economic Response to Unsolicited Communication. 2006. Advances in Economic Analysis & Policy, 6, 1.

Ramachandran, A. and N. Feamster. Understanding the Network-Level Behavior of Spammers. 2006. SIGCOMM 06, Pisa, Italy.

Sahami, M. S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk email. 1998. AAAI Workshop on Learning for Text Categorization.

Wang, Z. W. Josephson, Q. Lv, M. Charikar, and K. Li. Filtering Image Spam with Near-Duplicate Detection. 2007. CEAS 2007, Mountain View, U.S.A.