High Speed Image Part Recognition (IPR)

Prepared by:

Partik Ostrihon, COMDOM Software
Reza Rajabiun, COMDOM Software and York University, Toronto

High Speed Image Part Recognition

Overview

1. Problem: Messages enveloped in images, pdf..... Image spam: Average (<20%) Spikes (80%), sophisticated (pump and dump)
2. Implication: Large processing power, OCR for high resolution (600dpi or more)
3. ------> IPR for low resolution (75dpi and less/10-50x faster than commercial OCR)
4. Bad ideas: Limiting end user access to images, reputation based approaches/blacklists (BGP Spectrum Agility)
5. New ideas/old ideas

COM DOM ANTISPAM

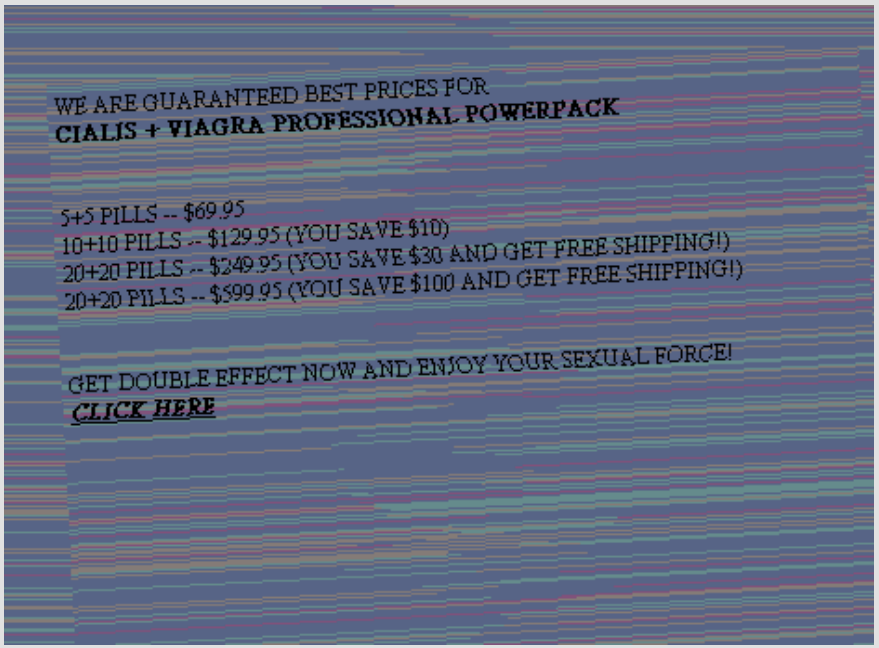
High Speed Image Part Recognition

Presentation

1. From OCR to ad hoc classifiers
2. Fingerprinting
3. Image Part Recognition and Bayesian classification
4. Examples
5. Discussion

1. Introduction

Complex envelopes



1. Introduction

Analytical problem:

a) Spam requires a positive response rate

b) To pass through spam filters: obfuscate/randomize

c) Principle: Cannot obfuscate more than allowed by the limits of human visual and analytic perceptions to get a positive response rate.

Implication: Infinite set of randomized envelopes, but a closed set of content patterns for advertisements (size depending on alphabet, education of population.....)

2. From OCR to ad hoc classifiers

Response to OCR Techniques

Spammers solution: Add CAPTCHA-like techniques making embedded text analysis difficult.

- changing fonts, shapes
- adding noise (dots, lines, random shapes ...)
- colouring
- image size variations, rotation, slanting ...

Implication: Larger more complex spam-low resolution

2. From OCR to ad hoc classifiers

Ad hoc classifiers

Dredze et al. (2007): “fast learning classifiers”

Highlights: importance of throughput to antispam/TCO, but

Learning about features of envelopes, not their content

Features (9 total):

- File format (does it match metadata)
- File size
- Image size
- Average color
- Color saturation-(Aradhye et al (2005).
  - Claim: 90-99% accuracy under static tests conditions. Throughput increase from seconds per image to 3-4 milliseconds

3. Fingerprinting

Checksum-based filters

A fingerprint or checksum-based filters exploited the fact that spam messages are sent in bulk.

Functions essentially by striping 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.

Adoption in early 2000s motivated by slow nature of first generation Bayesian (statistical) content filters.

In text based filtering: Low accuracy, but fast (low false positives in theory, but not in commercial fingerprinting bundles that also add other layers like sender reputation)

High Speed Image Part Recognition

3. Fingerprinting

“Near Duplicate Detection”

Wang et al. (2007): Also because of efficiency considerations.

Identify three specific fingerprinting filters:

a) Color histogram filter

b) Haar Wavelet filter (low resolution info about original image)

c) Orientation histogram feature

Generating checksums controlled against centralized database

Claim:

1) High detection rate in aggregating these signatures and low false positive rates (should be zero?)

2) with a speed of app. 90 ms for the bundle

High Speed Image Part Recognition

4. Image Part Recognition and Bayesian classification

IPR

Application of Probabilistic and inferential methods in computer vision (Forsyth and Poncer, 2003).

Achilles Hell of spammers: Their message

Problem with older content filters for industrial use:

Too slow, leading to adoption of ad hoc rules about features, fingerprinting (smart spam), and reputation bundles (BGP Spectrum Agility).

But more accurate in dynamic terms

New Bayesian filters: 30 times faster than older ones, 5x faster than centralized fingerprinting.

IPR: Retains advantage of OCRs in allowing for integration with second generation content filters that are both more accurate and faster.

COM DOM ANTISPAM

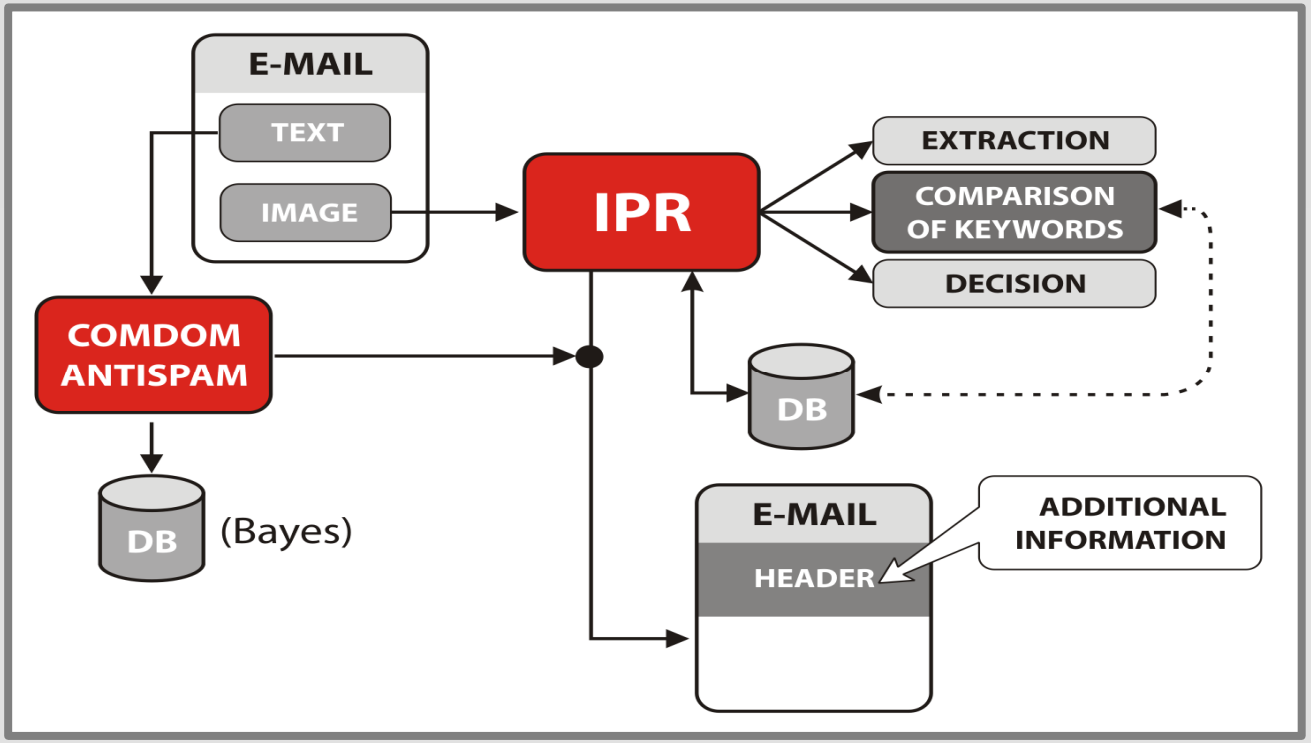
4. Image Part Recognition and Bayesian classification

Probabilistic approach

a) building pattern templates using classifiers
b) compare template difference
c) decision rule: variation of differences

4. Image Part Recognition and Bayesian classification

Organization



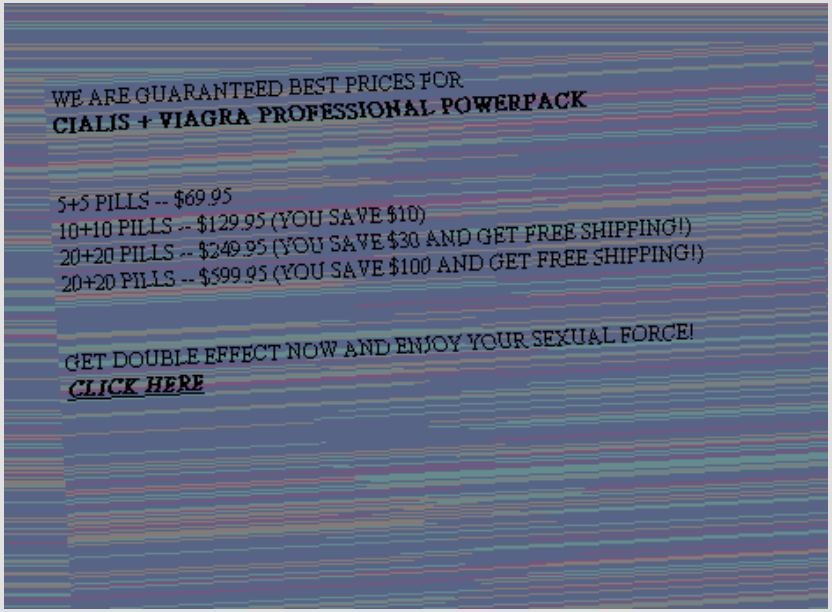
5. Examples

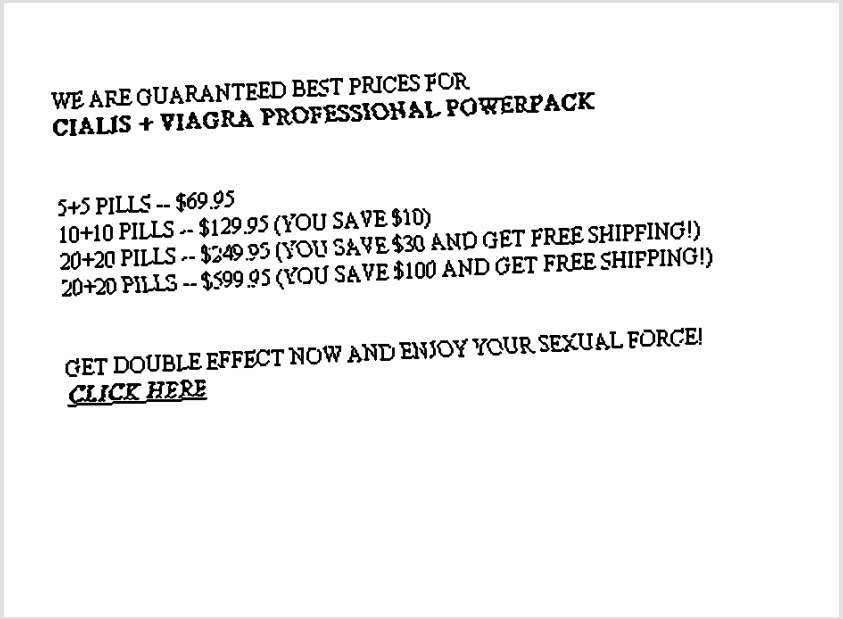
Filtering A(input)-OCR limits

High Speed Image Part Recognition

5. Examples

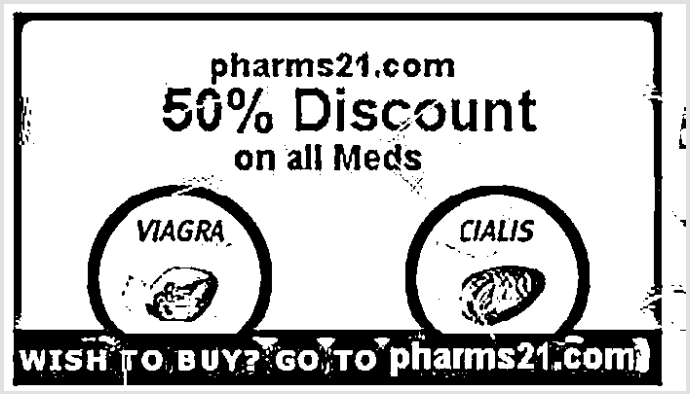
Filtering (inputA)





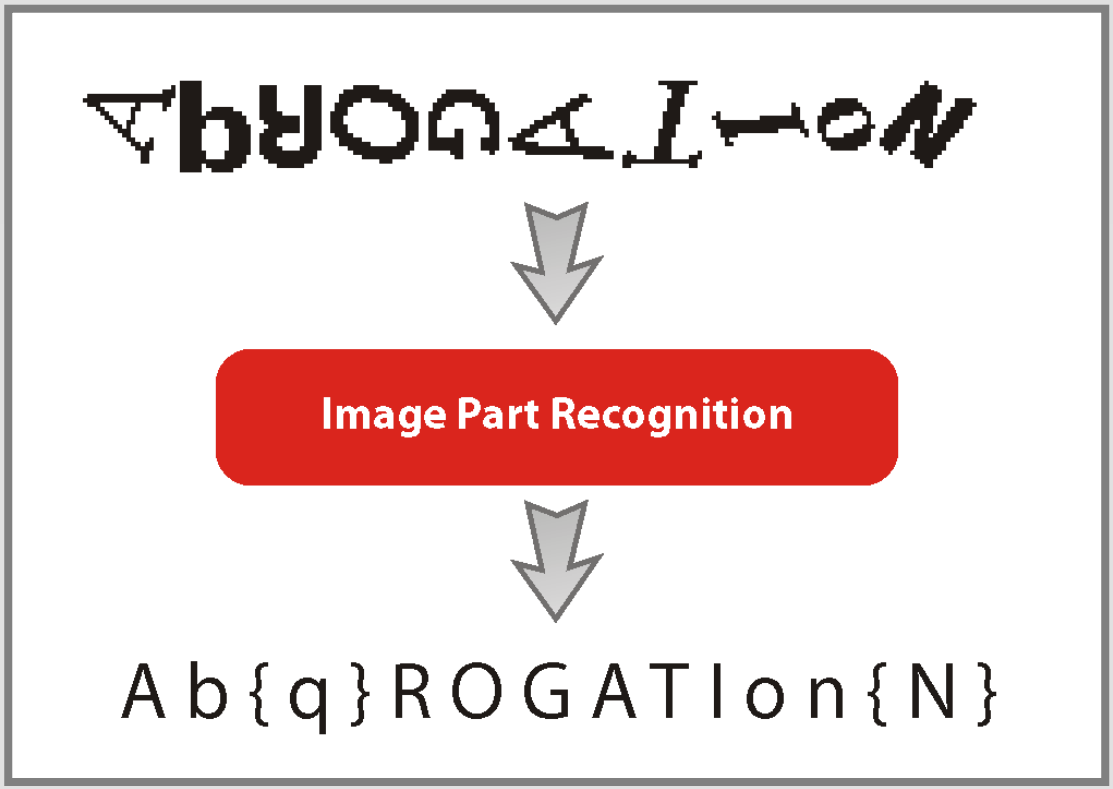
5. Examples

Filtering C



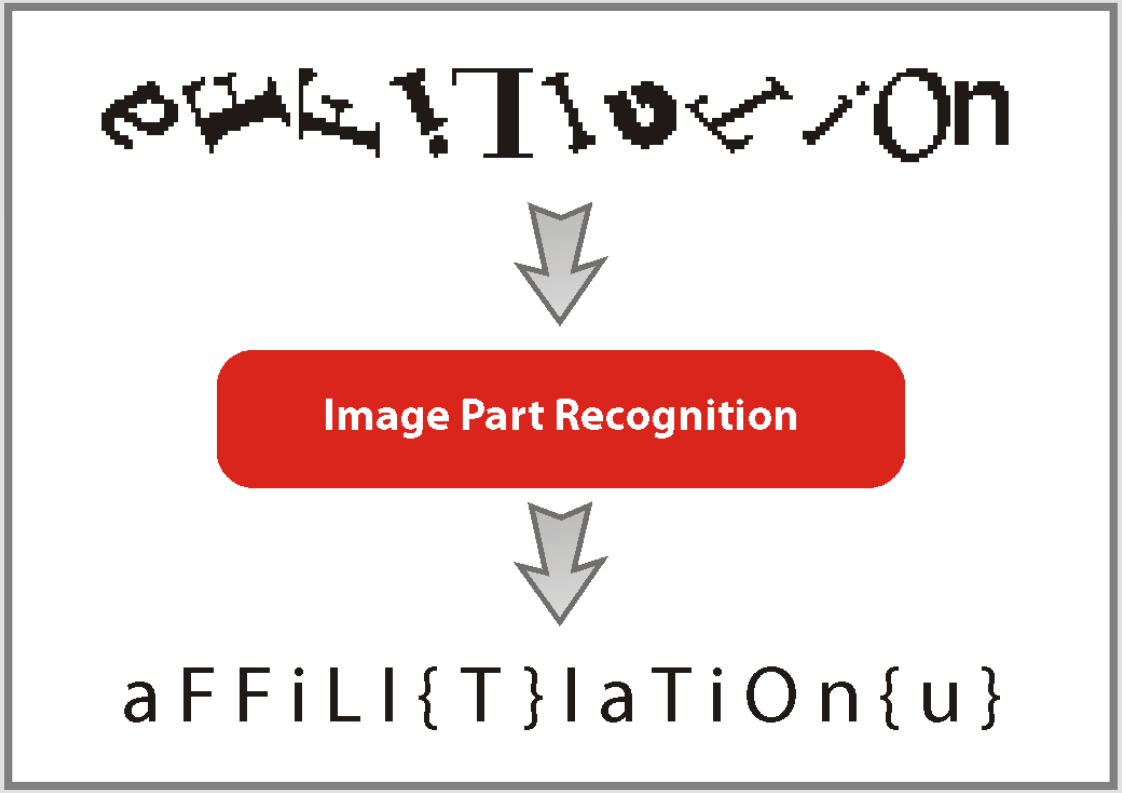
5. Examples

Figure 1



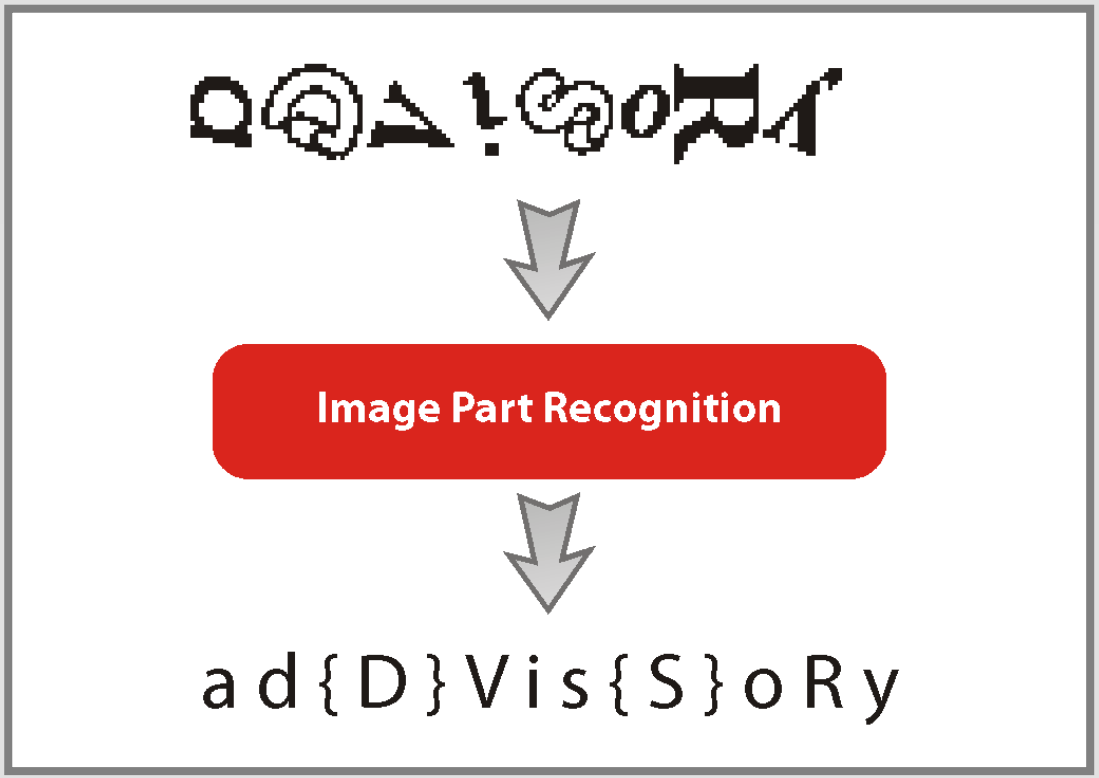
5. Examples

Figure 2



5. Examples

Figure 3



High Speed Image Part Recognition

5. Examples

Other patterns