The art of breaking and designing captchas

Elie Bursztein
Stanford University

Session ID: HT2-402
Session Classification: Intermediate
Join the Conversation

Full name: workshop  ok
Your full name will appear on your public profile

Username: eliedemo  ok
Your public profile: http://twitter.com/eliedemo

Password: *********  Good

Are you human?

Before we create your account, we need to make sure you're not a computer.

Type the words above

Finish

Create my account

I want the inside scoop—please send me email updates!
World Most-Popular Captchas

- Zkw4 [Megaupload]
- BMVHKY [Reddit]
- 944531 [eBay]
- Zd6bf [CNN]
- RAE3 [Baidu]
- 3nc9z [Authorize]
- 3.2 parks [ReCaptcha]
- 2ccex [NIH]
- 2plcd [Digg]
- guxg47 [Skyrock]
- trustother [Wikipedia]
- aiprocott [Slashdot]
- pmymkuv [Google]
- pmymkuv [Blizzard]
Captcha Design Goal

**Diagram:**
- **X-axis:** Hard for human
- **Y-axis:** Hard for computer
- **'sweet spot'** region
- **AI?**
- **Human**

*Elie Bursztein (@elie)*
*http://elie.im*
How to **break** and **design** CAPTCHA\textregistereds
Based on the analysis of 21 of the most popular schemes
Outline

- How to break text captcha
- How to break audio captcha
- How to make captchas easier for human
- What’s next?
Evaluation metrics

- Learnability
- Accuracy
- Solving time
How to Break Text-Captchas
Think Lego
Pre-processing: captcha binarization
Pre-processing: background removal
How to break a captcha: example
Pre-processing: Line clustering algorithm
Segmentation: cluster separation
Recognition: 3173
Breaker 5 Stages Pipeline

Preprocessing

Segmentation

Post-segmentation

Recognition

Post-recognition

Slashdot captcha

Elie Bursztein (@elie)
http://elie.im
From the **image** to the **matrix representation**
From the **matrix representation** to the **vector representation**
From the vector representation to the segment value (classification)
Breaker efficiency

Solver accuracy = Coverage * Precision^length

Coverage: Segmentation rate
Precision: Recognition rate
Anti-recognition techniques

Blurring

Distortion

Rotation

Fonts

Charsets

Elie Bursztein (@elie)
http://elie.im

RSACONFERENCE2012
SVM learning rate

% success vs Trainning set size for different conditions:
- 09
- AZ09
- azAZ09
- Distortion
- 3 fonts
- 5 fonts
- Angles
KNN learning rate

% success

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

10 20 50 100 200 500

Trainning set size

09  AZ09  azAZ09  Distortion  3 fonts  5 fonts  Angles
Anti-recognition taxonomy

Background Confusion

Lines

Collapsing
Breaking World of Warcraft

Elie Bursztein (@elie)
http://elie.im
Breaking Captcha.net

Elie Bursztein (@elie)
http://elie.im
Breaking Wikipedia
Breaking Digg
Breaking Slashdot

Elie Bursztein (@elie)
http://elie.im
Breaking eBay

Elie Bursztein (@elie)
http://elie.im
Failing to break eBay
Breaking Baidu

Elie Bursztein (@elie)
http://elie.im
<table>
<thead>
<tr>
<th>Service</th>
<th>Segmentation rate</th>
<th>Solving rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorize</td>
<td>84%</td>
<td>66%</td>
</tr>
<tr>
<td>Baidu</td>
<td>98%</td>
<td>5%</td>
</tr>
<tr>
<td>Blizzard</td>
<td>75%</td>
<td>70%</td>
</tr>
<tr>
<td>Captcha.net</td>
<td>96%</td>
<td>73%</td>
</tr>
<tr>
<td>CNN</td>
<td>50%</td>
<td>16%</td>
</tr>
<tr>
<td>Digg</td>
<td>86%</td>
<td>20%</td>
</tr>
<tr>
<td>eBay</td>
<td>95%</td>
<td>43%</td>
</tr>
<tr>
<td>Google</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>MegaUpload</td>
<td>n/a</td>
<td>93%</td>
</tr>
<tr>
<td>NIH</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td>Recapcha</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Reddit</td>
<td>71%</td>
<td>42%</td>
</tr>
<tr>
<td>Skyrock</td>
<td>30%</td>
<td>2%</td>
</tr>
<tr>
<td>Slashdot</td>
<td>52%</td>
<td>35%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>57%</td>
<td>25%</td>
</tr>
</tbody>
</table>
Learning rate for real schemes

Elie Bursztein (@elie)
http://elie.im
Building a breaker guidelines

- Immediate visual feedback
- Visual debugging
- Algorithm independence
- Exposing algorithm parameters
Decaptcha main interface
Apply design principles

- Core design principles
  - Randomize length
  - Randomize character size
  - Wave the captcha
- Use anti-recognition as a means of strengthening captcha security
- Don’t use a complex charset
  - Bad for human (see our research on this)
  - Useless for security
- Use collapsing or lines
Designing Better Captchas
Think Lego again

- Decompose in features
- Analyze
  - feature in isolation
  - features interaction
Real vs Generated

- \( s_b6qn \)
- \( E7U4 \)
- \( 865301 \)
- \( \text{spereque} \)
- \( \text{earns relief} \)
- \( e5sd6AV \)
## Real vs Generated

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Real Bypass</th>
<th>Real MTK</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorize</td>
<td>95</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>Baidu</td>
<td>90</td>
<td>93</td>
<td>90</td>
</tr>
<tr>
<td>Blizzard</td>
<td>89</td>
<td>95</td>
<td>91</td>
</tr>
<tr>
<td>Ebay</td>
<td>93</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>Microsoft</td>
<td>88</td>
<td>80</td>
<td>91</td>
</tr>
<tr>
<td>Google</td>
<td>88</td>
<td>86</td>
<td>94</td>
</tr>
<tr>
<td>Recaptcha</td>
<td>72</td>
<td>75</td>
<td>93</td>
</tr>
<tr>
<td>Yahoo</td>
<td>89</td>
<td>87</td>
<td>89</td>
</tr>
</tbody>
</table>
Evaluation system
Experiment details

<table>
<thead>
<tr>
<th>Round</th>
<th>Task</th>
<th>N possible</th>
<th>N sampled</th>
<th>N tests per sample</th>
<th>Total tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Baseline (&quot;Control&quot;)</td>
<td>1</td>
<td>1</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>Real world captchas</td>
<td>8</td>
<td>8</td>
<td>1000</td>
<td>8000</td>
</tr>
<tr>
<td>3</td>
<td>Features in isolation</td>
<td>496</td>
<td>496</td>
<td>200</td>
<td>99200</td>
</tr>
<tr>
<td>4</td>
<td>2 feature interactions</td>
<td>60950</td>
<td>60950</td>
<td>5</td>
<td>304750</td>
</tr>
<tr>
<td>5</td>
<td>3 feature interactions</td>
<td>1303224</td>
<td>25000</td>
<td>10</td>
<td>250000</td>
</tr>
<tr>
<td>6</td>
<td>4 feature interactions</td>
<td>113951684</td>
<td>25000</td>
<td>10</td>
<td>250000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>912150</td>
</tr>
</tbody>
</table>
Some of the features tested

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blurring</td>
<td>3tr2bb</td>
<td>ch6h</td>
</tr>
<tr>
<td>Text color</td>
<td>tx3soh</td>
<td>b7l</td>
</tr>
<tr>
<td>Font</td>
<td>0AGP22</td>
<td>fFBDu</td>
</tr>
<tr>
<td>Background color</td>
<td>0zt99n</td>
<td>4umtr35</td>
</tr>
<tr>
<td>Collapsing</td>
<td>02qeh</td>
<td>Jf8b</td>
</tr>
<tr>
<td>Tilting</td>
<td>O a x y n</td>
<td>4il</td>
</tr>
<tr>
<td>Waving</td>
<td>a Zhang</td>
<td>gjx</td>
</tr>
<tr>
<td>Distortion</td>
<td>0izjw6</td>
<td>jn6s</td>
</tr>
<tr>
<td>Line angle</td>
<td>ul8mgx</td>
<td>wybr</td>
</tr>
<tr>
<td>Line shape</td>
<td>pe3prq</td>
<td>k2r576</td>
</tr>
<tr>
<td>Line position</td>
<td>0abqng</td>
<td>ljXg</td>
</tr>
<tr>
<td>Line size</td>
<td>5h3txk</td>
<td>CMK</td>
</tr>
<tr>
<td>Line coverage</td>
<td>llc</td>
<td>Q56e</td>
</tr>
</tbody>
</table>
Angle of rotation

![Graph showing solving time and accuracy against rotation angle]

Elie Bursztein (@elie)
http://elie.im
Collapsing

![Graph showing solving time and accuracy vs character gap width]
Character size

![Graph showing the relationship between character size, solving time, and accuracy.](image-url)
Resolution invariant

Accuracy

captcha length (number of characters)

<= 1024
> 1024
all captchas

Elie Bursztein (@elie)
http://elie.im
2D interactions

Elie Bursztein (@elie)
http://elie.im
Length vs Angle interaction
Perception Does Not Match Number
How to Break Audio-Captcha
Audio Captchas

The not-so-fine print
For added security, please enter the verification code hidden in the image.

Elie Bursztein (@elie)
http://elie.im
Creating Audio Captcha

Voices

Captcha Maker

Noises

Elie Bursztein (@elie)
http://elie.im
Noise intensity (RMS/SNR)

Authorize

Dig

Microsoft

Elie Bursztein (@elie)
http://elie.im
Sound representation

WAV → DFT → Cep → TFR → TDC → TCR
Solving an audio captcha

Elie Bursztein (@elie)
http://elie.im
Dealing with random noise

- Statistical learning
- Supervised learning
- RLS (Regularized least square) classifier
### Results

<table>
<thead>
<tr>
<th>Service</th>
<th>Length</th>
<th>Coverage</th>
<th>Digit</th>
<th>Captcha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorize</td>
<td>5</td>
<td>100</td>
<td>97</td>
<td>89.2%</td>
</tr>
<tr>
<td>Digg</td>
<td>5</td>
<td>100</td>
<td>76</td>
<td>41.4%</td>
</tr>
<tr>
<td>eBay</td>
<td>6</td>
<td>85.6</td>
<td>92.5</td>
<td>82.9%</td>
</tr>
<tr>
<td>Microsoft</td>
<td>10</td>
<td>80.6</td>
<td>89.6</td>
<td>48.9%</td>
</tr>
<tr>
<td>Recaptcha</td>
<td>8</td>
<td>99.9</td>
<td>40.5</td>
<td>1.5%</td>
</tr>
<tr>
<td>Yahoo</td>
<td>7</td>
<td>99.1</td>
<td>74.7</td>
<td>45.4%</td>
</tr>
</tbody>
</table>
Recaptcha semantic noise
Confusion matrices
How many captchas do you need?
Apply

- Within 3 months
  - Make sure you have a strong captcha scheme
  - Ensure that your site is accessible

- Within 6 months
  - Log your captchas failure rate and monitor them
  - Have a backup captcha scheme in case your scheme is broken
Thank you

Questions ?

Follow-me !

Twitter: @elie

Captcha research: http://elie.im/captcha