Breaking CAPTCHAs

Xiaoyang Wen

COMP 857 Machine Learning
UNC Chapel Hill
What’s CAPTCHA?

- Completely Automated Public Turing Test to Tell Computers and Humans Apart
- Visual CAPTCHA: Small images requiring you to tell their contents

FYI: CAPTCHA images are called Human Interactive Proofs (HIPs) images
• Why do we need them?
• Why do we need them?
• Why do we need them?

• distinguish human and programs (or bots)
• Why do we need them?
  • distinguish human and programs (or bots)
  • preventing comment spam in blogs/forums
• Why do we need them?

• distinguish human and programs (or bots)
  • preventing comment spam in blogs/forums
  • protecting website/email account registrations
• Why do we need them?
  • distinguish human and programs (or bots)
    • preventing comment spam in blogs/forums
    • protecting website/email account registrations
    • protecting email addresses from scrapers
Why do we need them?

- distinguish human and programs (or bots)
  - preventing comment spam in blogs/forums
  - protecting website/email account registrations
  - protecting email addresses from scrapers
  - preventing dictionary attacks
• Why do we need them?
  • distinguish human and programs (or bots)
    • preventing comment spam in blogs/forums
    • protecting website/email account registrations
    • protecting email addresses from scrapers
    • preventing dictionary attacks
    • online polls
Why do we need them?

- distinguish human and programs (or bots)
  - preventing comment spam in blogs/forums
  - protecting website/email account registrations
  - protecting email addresses from scrapers
  - preventing dictionary attacks
  - online polls
  - email spams prevention
• Why do we need them?
  • distinguish human and programs (or bots)
    • preventing comment spam in blogs/forums
    • protecting website/email account registrations
    • protecting email addresses from scrapers
    • preventing dictionary attacks
    • online polls
    • email spams prevention
    • .....
• Why do we need them?
  • distinguish human and programs (or bots)
    • preventing comment spam in blogs/forums
    • protecting website/email account registrations
    • protecting email addresses from scrapers
    • preventing dictionary attacks
    • online polls
    • email spams prevention
    • ..... 
  • A easy and lightweight method to achieve its goal
• Roughly three types
• Roughly three types

Text-based
Roughly three types

- Text-based
- Real-Image-based
• Roughly three types

Text-based

2*3=?
The capital of England?

Real-Image-based
• Current usage status on the internet

  • Widely used (in email systems and internet forums/message boards)

  • For example, in 2007, 65 among top 100 internet forums in China are using CAPTCHAs during account registration or before posting

  • According to reCAPTCHA.com, guess how many CAPTCHAs are solved around the world every day

  • One, and the only one in most cases, major technique to do the job (telling ‘bots’ apart from human users)
• Current usage status on the internet

  • Widely used (in email systems and internet forums/message boards)

  • For example, in 2007, 65 among top 100 internet forums in China are using CAPTCHAs during account registration or before posting

  • According to reCAPTCHA.com, guess how many CAPTCHAs are solved around the world every day

    200 million

  • One, and the only one in most cases, major technique to do the job (telling ‘bots’ apart from human users)
• Well...are they really able to do its job?
Well...are they really able to do its job?
• Well…are they really able to do its job?

• Probably not, a majority of them can be cracked much more easily than expected
• Well...are they really able to do its job?

• Probably not, a majority of them can be cracked much more easily than expected

• Ironically, they bother real human users
• Well...are they really able to do its job?

• Probably not, a majority of them can be cracked much more easily than expected

• Ironically, they bother real human users

• 200 million CAPTCHAs are solved by humans around the world each day
Well...are they really able to do its job?

- Probably not, a majority of them can be cracked much more easily than expected
- Ironically, they bother real human users
- 200 million CAPTCHAs are solved by humans around the world each day
  - if 10 seconds for each, about 150,000 working hours per day
• Who are interested in CAPTCHAs?
• Who are interested in CAPTCHAs?
  • ‘Good’ guys
Who are interested in CAPTCHAs?

‘Good’ guys

- Email system admins, online forum admins, security researchers.....
• Who are interested in CAPTCHAs?
  • ‘Good’ guys
    • Email system admins, online forum admins, security researchers.....
  • ‘Bad’ guys
• Who are interested in CAPTCHAs?

  • ‘Good’ guys
    • Email system admins, online forum admins, security researchers.....

  • ‘Bad’ guys
    • Spammers, attackers who want to do something(collecting accounts, posting forum threads, etc) automatically and extensively, security researchers.....
Who are interested in CAPTCHAs?

- ‘Good’ guys
  - Email system admins, online forum admins, security researchers.....

- ‘Bad’ guys
  - Spammers, attackers who want to do something(collecting accounts, posting forum threads, etc) automatically and extensively, security researchers.....
  - Sometimes, ‘bad’ guys are very rich: $500,000 offer to break reCAPTCHA
• Who are interested in CAPTCHAs?

  • ‘Good’ guys
    • Email system admins, online forum admins, security researchers.....

  • ‘Bad’ guys
    • Spammers, attackers who want to do something(collecting accounts, posting forum threads, etc) automatically and extensively, security researchers.....
    • Sometimes, ‘bad’ guys are very rich: $500,000 offer to break reCAPTCHA

• An interesting arms race
• Who are interested in CAPTCHAs?

• ‘Good’ guys
  • Email system admins, online forum admins, security researchers.....

• ‘Bad’ guys
  • Spammers, attackers who want to do something(collecting accounts, posting forum threads, etc) automatically and extensively, security researchers.....
  • Sometimes, ‘bad’ guys are very rich: $500, 000 offer to break reCAPTCHA

• An interesting arms race
  • If ‘good’ guys take a lead: having a practical solution protecting lots of websites
• Who are interested in CAPTCHAs?

• ‘Good’ guys
  • Email system admins, online forum admins, security researchers.....

• ‘Bad’ guys
  • Spammers, attackers who want to do something(collecting accounts, posting forum threads, etc) automatically and extensively, security researchers.....
  • Sometimes, ‘bad’ guys are very rich: $500,000 offer to break reCAPTCHA

• An interesting arms race
  • If ‘good’ guys take a lead: having a practical solution protecting lots of websites
  • If ‘bad’ guys take a lead: solving a hard AI problem = an pretty good advance in AI research
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race

Simple rules
• How CAPTCHAs evolve in the arms race

• How breaking skills evolve in the arms race

Simple rules

More complex rules with training data
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race

Simple rules

More complex rules with training data

Machine Learning join the war
• How CAPTCHAs evolve in the arms race
• How breaking skills evolve in the arms race

Simple rules

More complex rules with training data

Machine Learning join the war
• Coded rules

• Rules with some machine learning skills

• Mostly by machine learning techniques
• Coded rules
  • Jeff Yan, et al. *Breaking visual CAPTCHAs with Naive Pattern Recognition Algorithms.* ACSAC 2007
  • One of my former projects (Cracking as many as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills

• Mostly by machine learning techniques
• Coded rules
  
  
  • One of my former projects (Cracking as many as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills
  

• Mostly by machine learning techniques
• Coded rules
  • Jeff Yan, et al. *Breaking visual CAPTCHAs with Naive Pattern Recognition Algorithms.*
    ACSAC 2007
  • One of my former projects (Cracking as many as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills
    Advances in Neural Information Processing Systems, 2004

• Mostly by machine learning techniques
    CVPR 2003
  • Philippe Golle. *Machine learning attacks against the Asirra CAPTCHA*
• Coded rules

  • Jeff Yan, et al. *Breaking visual CAPTCHAs with Naive Pattern Recognition Algorithms.*
    ACSAC 2007

  • One of my former projects (Cracking as much as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills

    Advances in Neural Information Processing Systems, 2004

• Mostly by machine learning techniques

    CVPR 2003

  • Philippe Golle. *Machine learning attacks against the Asirra CAPTCHA*
• High level thoughts on how to crack CAPTCHAs
• High level thoughts on how to crack CAPTCHAs
• High level thoughts on how to crack CAPTCHAs
  • CAPTCHAs are
    • with noises/clutter
    • characters distorted
    • several characters (normally alphanumerical)
High level thoughts on how to crack CAPTCHAs

- CAPTCHAs are
  - with noises/clutter
  - characters distorted
  - several characters (normally alphanumerical)

So a general process might be..
- pre-processes like de-noise
- character segmentation
- try to recognize one by one
• De-noise
  • De-noise in colored image
    • replace a pixel with an ‘average’ one around it

  • Turn an image to a black white one

  Colored image → Grey-scale image → Binary image
  A good choice of grey threshold value could clean up background

• De-noise in B&W image
  • find the ‘lonely’ pixel(s)
• Character segment

• Histogram of vertical pixel counts

• Dealing with some segment difficulties
  • threshold value choice
  • more segments and less segments
• Train to get a sample pool

• Manually label some sample CAPTCHAs as training data

• Feature extraction
  • a binary string

• Sample pool building
  • each CAPTCHA design has its own pool
• Recognizing characters
  • Going through the same processing steps
  • Feature extraction
  • Find K Nearest Neighbors in the sample pool
    • use simple Hamming distance
    • which appears the most is the final result
  • 3 8 3 3 3 3 B 3 B E then 3 is the final result
• a simple illustration of the whole process
Overall experiment result

- collected 62 types (all) of CAPTCHAs from the top 100 forums
- 51 types results on the left figure
- the other 11 types: less than 30% success rate (in reality, less than 30% is not that bad for automatic programs)
• Other segmentation methods
- **Other segmentation methods**
  - **Snake segmentation**
    - Set some moving rules for the ‘snake’
      - moving down as much as possible until reaching a foreground pixel
      - when it can only move L and U, moves L for one pixel, and move down as much as possible...
    - .....
• Other character features we can use
• Other character features we can use
  • Pixel Counts for characters
• Other character features we can use
  • Pixel Counts for characters

<table>
<thead>
<tr>
<th>Letter</th>
<th>Pixel Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>183</td>
</tr>
<tr>
<td>B</td>
<td>217</td>
</tr>
<tr>
<td>C</td>
<td>159</td>
</tr>
<tr>
<td>D</td>
<td>192</td>
</tr>
<tr>
<td>E</td>
<td>163</td>
</tr>
<tr>
<td>F</td>
<td>133</td>
</tr>
<tr>
<td>G</td>
<td>190</td>
</tr>
<tr>
<td>H</td>
<td>186</td>
</tr>
<tr>
<td>I</td>
<td>121</td>
</tr>
<tr>
<td>J</td>
<td><strong>111</strong></td>
</tr>
<tr>
<td>K</td>
<td><strong>178</strong></td>
</tr>
<tr>
<td>L</td>
<td><strong>111</strong></td>
</tr>
<tr>
<td>M</td>
<td>233</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Letter</th>
<th>Pixel Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>239</td>
</tr>
<tr>
<td>O</td>
<td><strong>178</strong></td>
</tr>
<tr>
<td>P</td>
<td><strong>162</strong></td>
</tr>
<tr>
<td>Q</td>
<td>229</td>
</tr>
<tr>
<td>R</td>
<td>208</td>
</tr>
<tr>
<td>S</td>
<td>194</td>
</tr>
<tr>
<td>T</td>
<td>175</td>
</tr>
<tr>
<td>U</td>
<td>164</td>
</tr>
<tr>
<td>V</td>
<td><strong>162</strong></td>
</tr>
<tr>
<td>W</td>
<td>234</td>
</tr>
<tr>
<td>X</td>
<td>181</td>
</tr>
<tr>
<td>Y</td>
<td>153</td>
</tr>
<tr>
<td>Z</td>
<td>193</td>
</tr>
</tbody>
</table>

Table 2. A letter–pixel count lookup table for letters A-Z. (Note: ‘J’ and ‘L’ have the same pixel count. So are ‘K’ and ‘O’, and ‘P’ and ‘V’.)
• Other character features we can use (cont.)

• Some characters have similar or the same pixel counts, like J and L or P and V

• Try to set up some rules to explore their shape differences
• Other character features we can use (cont.)

• Some characters have similar or the same pixel counts, like J and L or P and V

• Try to set up some rules to explore their shape differences
• Other character features we can use (cont.)

• Some characters have similar or the same pixel counts, like J and L or P and V

• Try to set up some rules to explore their shape differences

Two cuts of the middle line for P
One cut for V
• Coded rules
  • Jeff Yan, et al. Breaking visual CAPTCHAs with Naive Pattern Recognition Algorithms.
    ACSAC 2007
  • One of my former projects (Cracking as much as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills
    Advances in Neural Information Processing Systems, 2004

• Mostly by machine learning techniques
    CVPR 2003
  • Philippe Golle. Machine learning attacks against the Asirra CAPTCHA
• This paper’s main goal is not to break any one HIP in particular with the highest possible success rate, but to learn about the common strengths and weaknesses of HIPs on the market

• Generic method is to write a custom algorithm to locate the characters, and then use machine learning for recognition

• Once the segmentation problem is solved, solving HIP becomes a pure recognition problem which can trivially be solved using simple machine learning skills

• A Convolution Neural Network is then trained and used as the character recognizer
• HIP of Mailblocks
  • Select the red channel, binarize and erode it, extract the largest connected components (CCs), and breakup CCs that are too large into two or three CCs

![Image of segmentation and recognition results]

• Then train a Neural Network

• 88.8% for segmentation, 95.9% for recognition, 66.2% for overall success rate
• HIP of Register

• The image was smoothed, binarized, and the largest 5 connected components were identified.

• Then train a Neural Network

• 95.4% for segmentation, 87.1% for recognition, 47.8% for overall success rate
• HIP of Yahoo version 2

• Remove 6 pixel border, up-sample, dilate first then erode, select large CCs with sizes close to HIP char sizes.

• Then train a Neural Network

• 58.4% for segmentation, 95.2% for recognition, 45.7% for overall success rate
• HIP of Google/Gmail

• Covert to grayscale, up-sample, threshold and separate connected components

a) cousli

b) womair

• Then train a Neural Network

• 10.2% for segmentation, 89.3% for recognition, 4.89% for overall success rate
• Lessons learned from breaking HIPs
• Lessons learned from breaking HIPs

• Segmentation could be much more computationally expensive, compared to recognition after that
Lessons learned from breaking HIPs

- Segmentation could be much more computationally expensive, compared to recognition after that.
- Therefore, a good HIP designer should focus on providing a harder segmentation problem.
Using Machine Learning to Break Visual Human Interaction Proofs (HIPs)

- **Lessons learned from breaking HIPs**
  - Segmentation could be much more computationally expensive, compared to recognition after that.
  - Therefore, a good HIP designer should focus on providing a harder segmentation problem.

- A possible automatic segmentor design:
  - Label characters based on their correct position and train a recognizer.
  - Apply the recognizer at all locations in the HIP.
  - Collect all candidate characters identified with high confidence by the recognizer.
• A possible automatic segmentor design
  • Label characters based on their correct position and train a recognizer
  • Apply the recognizer at all locations in the HIP
  • Collect all candidate characters identified with high confidence by the recognizer
• A possible automatic segmentor design
  • Label characters based on their correct position and train a recognizer
  • Apply the recognizer at all locations in the HIP
  • Collect all candidate characters identified with high confidence by the recognizer

Works like a sliding window detector
A possible automatic segmentor design

- Label characters based on their correct position and train a recognizer
- Apply the recognizer at all locations in the HIP
- Collect all candidate characters identified with high confidence by the recognizer

Works like a sliding window detector

a threshold of 0.5
• Coded rules
  - One of my former projects (Cracking as much as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills

• Mostly by machine learning techniques
• High level goal of this paper is to explore object recognition in clutter

• Their specific targets are EZ-Gimpy and Gimpy CAPTCHAs (then used by Yahoo)

• Their breaking algorithm is based on shape context matching along with the help from dictionary models

• Two algorithms are introduced
• In this paper, represent and compare objects as shapes
• Shapes are coded by ‘shape contexts’
  • *shape context* is a set of highly discriminative descriptors, introduced in a paper which measures similarity between shapes and exploit it for object recognition
• In this paper, represent and compare objects as shapes

• Shapes are coded by ‘shape contexts’

• *shape context* is a set of highly discriminative descriptors, introduced in a paper which measures similarity between shapes and exploit it for object recognition

• Intuition: Shape is a good indicator of (deformable) the same objects
In this paper, represent and compare objects as shapes.

Shapes are coded by ‘shape contexts’

- *shape context* is a set of highly discriminative descriptors, introduced in a paper which measures similarity between shapes and exploit it for object recognition.

- Intuition: Shape is a good indicator of (deformable) the same objects.

Fig. 1. Examples of two handwritten digits. In terms of pixel-to-pixel comparisons, these two images are quite different, but to the human observer, the shapes appear to be similar.
Algorithm A is used to break EZ-Gimpy. It consists of 3 steps:

- Perform a series of quick tests to hypothesize locations of letters in the image.
- Extract strings of these hypothesized letters that form candidate words.
- Choose the most likely word(s) by evaluating a matching score for each of these words.
Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA

- **Step 1 Finding Letter-Location Hypotheses**
  - This step prunes a large space of letter locations (26 letters that could occur anywhere in the image) down to a small number of \(<\text{letter, location}>\) tuples.
  - First, it gets a set of \(<\text{letter, location, score}>\) tuples, and apply a threshold over it to get a final set of \(<\text{letter, location}>\) tuples: 5 to 30 tuples, normally.

![Diagram](image)
**Step 2 Extracting Candidate Words**

- Having a set of `<letter, location>`, construct a directed acyclic graph (DAG) in which there is a node for each letter hypothesis $l_i$, and an edge $e(l_i, l_j)$ between two nodes if $l_j$ can be used as the letter succeeding $l_i$ in a word.

- There will be lots of paths. Computer all the tri-grams in the DAG. We prune a dictionary word unless all of its tri-grams are present.

- For the remaining, we check that each of them is actually found as a path through the DAG, since the n-2 tri-grams might not actually occur in sequence. 1 to 10 words left then, normally

---

(a) profit

(b) profit

(c) profit

(d) profit
• **Step 3 Choosing the Most Likely Word**
  
  • Having a very small set of final candidate words, we score each word by averaging score for matching each of its letters.
  
  • The way to score each letter is based on deformable matching cost, introduced by Belongie et al. before, of individual letter.

![Diagram](image-url)
• Step 3 Choosing the Most Likely Word

• Having a very small set of final candidate words, we score each word by averaging score for matching each of its letters.

• The way to score each letter is based on deformable matching cost, introduced by Belongie et al. before, of individual letter.

Overall success rate: 83%
• Algorithm B  *Used to break Gimpy*

• It’s a **holistic** algorithm which attempts to find the entire word at once, instead of looking for letters.

• Reason: in Gimpy, it’s highly clutter, many of the parts can be occluded or highly ambiguous. Nearly impossible to determine which characters is present without a ‘see’ of the whole word.

Figure 6: Examples of letter sized patches from Gimpy CAPTCHAs. It is very difficult to read these isolated letters without the long range context in which they appear. We must instead read whole words at once.
• Algorithm B

• Gimpy contains 10 words overlaid in 5 pairs, separately. So we can examine each pair, independently.

• The most salient parts of the words are often the beginning or ending few letters.
  • Given the opening and ending bigrams, the number of remaining possible candidate words is very small.
  • Therefore, we can determine opening/closing bigrams first, we will be left with a very small set of candidate words.
• Algorithm B (cont.)

• In 411 Gimpy dictionary words, there are 128 distinct opening and 112 closing bigrams.

• Similar to Step 1/2 of Algorithm A, we can prune the dictionary to get a short list.

• Given a word guess, remove it, and recognize the second word based on remaining pixels.

• Score them similarly to Algorithm A.

• Top 3 scores are chosen
• Coded rules


  • One of my former projects (Cracking as much as possible CAPTCHAs using in China), 2007

• Rules with some machine learning skills


• Mostly by machine learning techniques


• What is Asirra CAPTCHA?

• Developed by MSR, and is published in ACM CCS 2007. *Estimated breaking success rate: 0.2%*

• This ‘attack’ paper is published in ACM CCS 2008, only one year later. *Breaking success rate: 10.3%*
• What is Asirra CAPTCHA?

• Developed by MSR, and is published in ACM CCS 2007. *Estimated breaking success rate: 0.2%*

• This ‘attack’ paper is published in ACM CCS 2008, only one year later. *Breaking success rate: 10.3%*
What is Asirra CAPTCHA?

- Developed by MSR, and is published in ACM CCS 2007. *Estimated breaking success rate: 0.2%*
- This ‘attack’ paper is published in ACM CCS 2008, only one year later. *Breaking success rate: 10.3%*
Machine Learning Attacks Against the Asirra CAPTCHA
This paper introduced a classifier which is 82.7% accurate in telling apart the images of cats and dogs used in Asirra.
• This paper introduced a classifier which is 82.7% accurate in telling apart the images of cats and dogs used in Asirra.

• Therefore, this classifier allows us to solve a 12-image Asirra challenge with probability 10.3%.
• This paper introduced a classifier which is 82.7% accurate in telling apart the images of cats and dogs used in Asirra.

• Therefore, this classifier allows us to solve a 12-image Asirra challenge with probability 10.3%.

• This classifier is a combination of two SVMs (with Radial Basis Kernel) trained on color and texture features of Asira’s cat and dog images.
• **Color Features**

  • Using HSV model of color, rather than RGB
    
    • closer to human perception of color
    
    • Subdivide the Hue channel into $C_h$ bands of equal width; the Saturation channel into $C_s$ bands of equal width; the Value channel into $C_v$ bands of equal width

  • Model a color feature vector $F(N, C_h, C_s, C_v)$

  • a boolean vector length $N^2 C_h C_s C_v$
• **Color Features**

  • Using HSV model of color, rather than RGB
    
    • closer to human perception of color
    
    • Subdivide the Hue channel into \( C_h \) bands of equal width; the Saturation channel into \( C_s \) bands of equal width; the Value channel into \( C_v \) bands of equal width

  • Model a color feature vector \( F(N, C_h, C_s, C_v) \)

  • a boolean vector length \( N^2 C_h C_s C_v \)

<table>
<thead>
<tr>
<th>Color features</th>
<th># images</th>
<th>Classifier accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td># features</td>
<td></td>
<td>mean</td>
</tr>
<tr>
<td>Total</td>
<td>5,000</td>
<td>76.3</td>
</tr>
<tr>
<td>Training set</td>
<td>4,000</td>
<td>76.3</td>
</tr>
<tr>
<td>15,760</td>
<td></td>
<td>76.3</td>
</tr>
<tr>
<td>15,760</td>
<td></td>
<td>77.1</td>
</tr>
</tbody>
</table>

Table 2: Accuracy of SVM classifiers trained on a combination of color features.
• Texture Features

• From training images, we first build a set of sub-images called, *texture tiles*.

• Texture tiles are a set of small sub-images which are added into this set iteratively until a given size is reached.

  • Distance between two sub-image: the average Euclidian distance between the pixels of the two in RGB space

  • Given a sub-image, if there already exists one tile whose distance to it is below a threshold, discard the sub-image; Otherwise, add it in.
Texture Features

- From training images, we first build a set of sub-images called, *texture tiles*.

- Texture tiles are a set of small sub-images which are added into this set iteratively until a given size is reached.
- Distance between two sub-image: the average Euclidian distance between the pixels of the two in RGB space
- Given a sub-image, if there already exists one tile whose distance to it is below a threshold, discard the sub-image; Otherwise, add it in.

It's like a uniform collection of sub-images over all images.
• Texture Features (cont.)
  • The feature vector associated with an image is the vector of distances between the image $A$ and each of the $K$ texture tiles in $T$.
  • Therefore, the length of the feature vector is $K$
• **Texture Features (cont.)**

• The feature vector associated with an image is the vector of distances between the image A and each of the $K$ texture tiles in $T$.

• Therefore, the length of the feature vector is $K$

<table>
<thead>
<tr>
<th>Texture features</th>
<th># tiles</th>
<th>$\delta$</th>
<th># Images</th>
<th>Classifier accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature set</td>
<td></td>
<td></td>
<td>Total</td>
<td>Training set</td>
</tr>
<tr>
<td>$G_1$</td>
<td>1,000</td>
<td>40.0</td>
<td>5,000</td>
<td>4,000</td>
</tr>
<tr>
<td>$G_2$</td>
<td>5,000</td>
<td>40.0</td>
<td>5,000</td>
<td>4,000</td>
</tr>
<tr>
<td>$G_2$</td>
<td>5,000</td>
<td>40.0</td>
<td>10,000</td>
<td>8,000</td>
</tr>
</tbody>
</table>

Table 3: Accuracy of SVM classifiers trained on texture features extracted from Asirra images. The texture features are described in section 2.2. The accuracy of the classifier is the fraction of cat and dog images classified correctly in the test set.
• A Combination of two SVMs
  • Output of Color SVM is weighted by 1/3
  • Output of Texture SVM is weighted by 2/3
• A Combination of two SVMs

• Output of Color SVM is weighted by $1/3$

• Output of Texture SVM is weighted by $2/3$

<table>
<thead>
<tr>
<th>Features</th>
<th># Images</th>
<th>Classifier accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>$(F_1 \cup F_2 \cup F_3) + G_2$</td>
<td>5,000</td>
<td>4,000</td>
</tr>
<tr>
<td>$(F_1 \cup F_2 \cup F_3) + G_2$</td>
<td>10,000</td>
<td>8,000</td>
</tr>
</tbody>
</table>

Table 4: Accuracy of the combined outputs of the color classifier of section 2.1 and the texture classifier of section 2.2. The color classifier is given half the weight of the texture classifier (see section 2.3).
• A very hot CAPTCHA now: reCAPTCHA (A research project lead by Luis Von Ahn)

30 Million served daily

Google acquired reCAPTCHA last month
• A very hot CAPTCHA now: reCAPTCHA (A research project lead by Luis Von Ahn)

30 Million served daily
Google acquired reCAPTCHA last month

• What’s the idea behind this?
• A very hot CAPTCHA now: reCAPTCHA (A research project lead by Luis Von Ahn)

30 Million served daily

Google acquired reCAPTCHA last month

• What’s the idea behind this?
  • A lots of human labor/time is consumed solving CAPTCHA. Can we use it?
• What’s the idea behind this?
• at least, we can use the time to do something useful
• What’s the idea behind this?
• at least, we can use the time to do something useful

The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge, has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active. The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge, and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.
• What’s the idea behind this?
• at least, we can use the time to do something useful
The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active. The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

The Breckinridge and Lane Democrats, having taken courage at the recent eastern advises, are energetically organizing for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active. The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

The Breckinridge and Lane Democrats, having taken courage at the recent eastern advises, are energetically organizing for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active. The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.
• More about reCAPTCHA

  • You can use it today to protect your email address, just like them:

  • Google acquired reCAPTCHA last month
    • For security purpose
    • Also to help Google Books
  • $500,000 dollar offer to crack it
  • Also my project target
• More about reCAPTCHA
  • You can use it today to protect your email address, just like them:
    Professor Mike and Fabian
  • Google acquired reCAPTCHA last month
    • For security purpose
    • Also to help Google Books
  • $500,000 dollar offer to crack it
  • Also my project target
• Two funny things to share
  • maybe one more CAPTCHA type...
• Two funny things to share
• maybe one more CAPTCHA type...

Qualifying question

Just to prove you are a human, please answer the following math challenge.

Q: Calculate:
\[
\frac{\partial}{\partial x} \left[ 4 \cdot \sin \left( 7 \cdot x - \frac{\pi}{2} \right) \right]_{x=0}
\]

A: [manditory]

Note: If you do not know the answer to this question, reload the page and you’ll get another question.
• Two funny things to share
• maybe one more CAPTCHA type...

Q: Calculate:
\[ \frac{\partial}{\partial x} \left[ 4 \cdot \sin \left( 7 \cdot x - \frac{\pi}{2} \right) \right]_{x=0}. \]

A: [Enter answer]

Note: If you do not know the answer to this question, reload the page and you’ll get another question.
• Two funny things to share
• maybe one more CAPTCHA type...

Qualifying question

Just to prove you are a human, please answer the following math challenge.

Q: Calculate:
\[
\frac{\partial}{\partial x} \left[ 4 \cdot \sin \left( 7 \cdot x - \frac{\pi}{2} \right) \right]_{x=0}
\]

A: [mandatory]

Note: If you do not know the answer to this question, reload the page and you’ll get another question.

No premium user. Please enter the one that can NOT be created from the unfolded pattern. 29 seconds remain.

A B C D E F
• Two funny things to share
• maybe one more CAPTCHA type...

‘Geek’ type?
• Two funny things to share (cont.)
  • human-being CAPTCHA solvers exist!
• Two funny things to share (cont.)

• human-being CAPTCHA solvers exist!

Some ads from human-labor CAPTCHAs solving companies

* We have 30 pc 90 worker & we have 300 captcha team. Your any captcha project we done quickly. We have high experience captcha worker
* Sir, We have 10 systems with good typing skill workers. We can easily do 25k per day
* I have 40 PCs and 55 Persons working in my office for data entry work. I person can do 800 captcha entry per hour. We can deliver you good quantity with quality
* Hello Sir, I will kindly introduce myself. This is shivakumar. we have a team to type capcthas 24/7 and we can type more than 200k captchas per day
* WE ARE PROFESSIONAL CAPTCHA ENTRY OPEATORS AND WE CAN DO EVEN 25000 ENTRIES PER DAY AS MY COMPANY IS A 25 SEATER FIRM SPEALISED IN DATA ENTRY
  * Our Team is very much interested in your project and we could easily handle more than 50,000 captcha entries per day
  * We having more then 10 teams, we are operating 24/7 data entry works and delivering 700k/day captchas daily
  * I have a team of 7 people, willing to do captchas at $2 per 1000 entries. Please consider my bid. We can definitely provide 50K captchas per day
  * My team is equipped to offer the services. 20 person team, T1 business speed internet with an on hand technical staff. We are able to start right away
  * Dear Sir I am an expert in account creation, will provide you the accounts as per your requirements. I ensure the guaranteed satisfaction always. I charge only $40/1000
  * captcha typing teams for 24/7 ready. Rate $1.25 for 1K, up to 100K captchas per day
  * $1 per 1K of entries, and ready to produce 50 K entries per day. Kindly look forward procedures to provide the chance to us
  * My rate $4.00 per 1k My team can work 24/7. They are jobless now
  * $3 per 1000 image entry, Ready start. 24/7 service like also. We have 30 pc 90 worker & 39 big captcha team so your any target we solve
  * Dear Sir, We have quoted 50 $ per 50000 entries., Kindly look forward procedures to provide a chance to work with you
Discussion
• What is a good CAPTCHA image design?
• What is a good CAPTCHA image design?

<table>
<thead>
<tr>
<th>easy for bots</th>
<th>easy for bots</th>
</tr>
</thead>
<tbody>
<tr>
<td>hard for human</td>
<td>easy for human</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>hard for bots</th>
<th>hard for bots</th>
</tr>
</thead>
<tbody>
<tr>
<td>hard for human</td>
<td>easy for human</td>
</tr>
</tbody>
</table>

*the ideal case but is there any?*
• Under the attack of machine learning techniques, is text-based CAPTCHAs dead?
• The answer is most likely to be YES
• Can Real-Image-Based CAPTCHAs save themselves?
• Under the attack of machine learning techniques, is text-based CAPTCHAs dead?
  • The answer is most likely to be YES
• Can Real-Image-Based CAPTCHAs save themselves?

Please select all the cat photos:

[Images of cats]

Use the slider to rotate the image to its natural, upright position, then press the submit button.