RECAPTCHA: HUMAN-BASED CHARACTER RECOGNITION VIA WEB SECURITY MEASURES

Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, Manuel Blum

Presented by Bhaskar Pilania

Key Aspects

- To discuss how human computing can be used to digitize the human knowledge

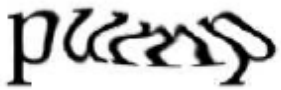
- To explore how can we differentiate between bots and human beings

Introduction to CAPTCHA

- It stands for - Completely Automated Public Turing test to tell Computers and Humans Apart

- It is a grade test which
  - Most humans can pass
  - Current computer program can’t pass

- Users are asked to identify characters in a distorted image



Life without CAPTCHA

- Bots can submit thousands of online forms and suppress human opinion
  - Example – Best graduate school poll hosted on slapdash.com
- Computer programs can create thousands of email accounts for spamming
- Search engine bots can intrude into your website
- Dictionary attacks by bots to crack password systems

CAPTCHA Beyond Security

- CAPTCHAs are representation of certain hard AI problems
- In case the CAPTCHA is not broken by any bot, it ensures security
- If a bot is able to break the CAPTCHA, the AI problem is solved.
- Also solution to such problems are used in Robust Image-Based Steganography.

CAPTCHAs are win-win situation!

Type 1 : MATCHA

- MATCHA instance is described as M = (I; T ; τ)
  - I is distribution of images containing words, T being distribution of their Transformation
- For randomness, verifier flips a unbiased coin and
  - If head comes then, it picks k← I and sets (i,j) = (k,k)
  - For tails, it sets j←I and i← U([I] – {j})
- Verifier sends the prover (i,t(j)) and sets timer for τ; if τ expires then prover rejects
- Prover responds to query by res ε {0; 1]. Verifier accepts/rejects based on res.
- This is performed many times to ensure security as prover can be a program that can break MATCHA by probability .5 simply by sending res = 1.

Type 2 : PIX

- PIX instance is described as $X = (I; T ; L; \lambda; \tau)$
  - I is a distribution over images containing a single word and λ maps an image to the label L contained in it.
- PIX verifier draws $i \leftarrow I$ and $t \leftarrow T$; sends to prover the message (t(i),L) and sets the timer for τ
- P responds with a label $l \in L$
- V accepts if $l = \lambda(i)$ and its timer has not expired, and rejects otherwise
- Various instantiations of pix are in use at major internet portals, like Yahoo and Hotmail.
- Other less conventional ones, like Animal-PIX, presents the prover with a distorted picture of a common animal and asks it to choose between twenty different possibilities.

Why reCAPTCHA?

- According to estimates, humans around the world type more than 100 million CAPTCHAs everyday
- Spending a few seconds typing the distorted characters. In aggregate, this amounts to hundreds of thousands of human hours per day
- This leads to waste of human invaluable mental effort – in doing what computers cannot do
- reCAPTCHA puts this invaluable effort to serve humanity by helping digitize the books

Setting The Background

- The pages are photographically scanned and the resulting bitmap images are transformed into text files by optical character recognition (OCR) software
- In older prints with faded ink and yellowed pages, OCR cannot recognize about 20% of the words
- By contrast, more than 99% of times humans are very accurate at transcribing such prints.
  - Human transcribers are extremely expensive

Overview of reCAPTCHA

- reCAPTCHA displays words taken from scanned texts which could not be recognized by OCR
- The solutions entered by humans are used to improve the digitization process
- To meet primary goal of CAPTCHA(differentiating between humans and computers), users are provided with two words
- the one for which the answer is not known and a second “control” word for which the answer is known.
- If users correctly type the control word, the system assumes they are human and gains confidence that they also typed the other word correctly

Exemplifying reCAPTCHA

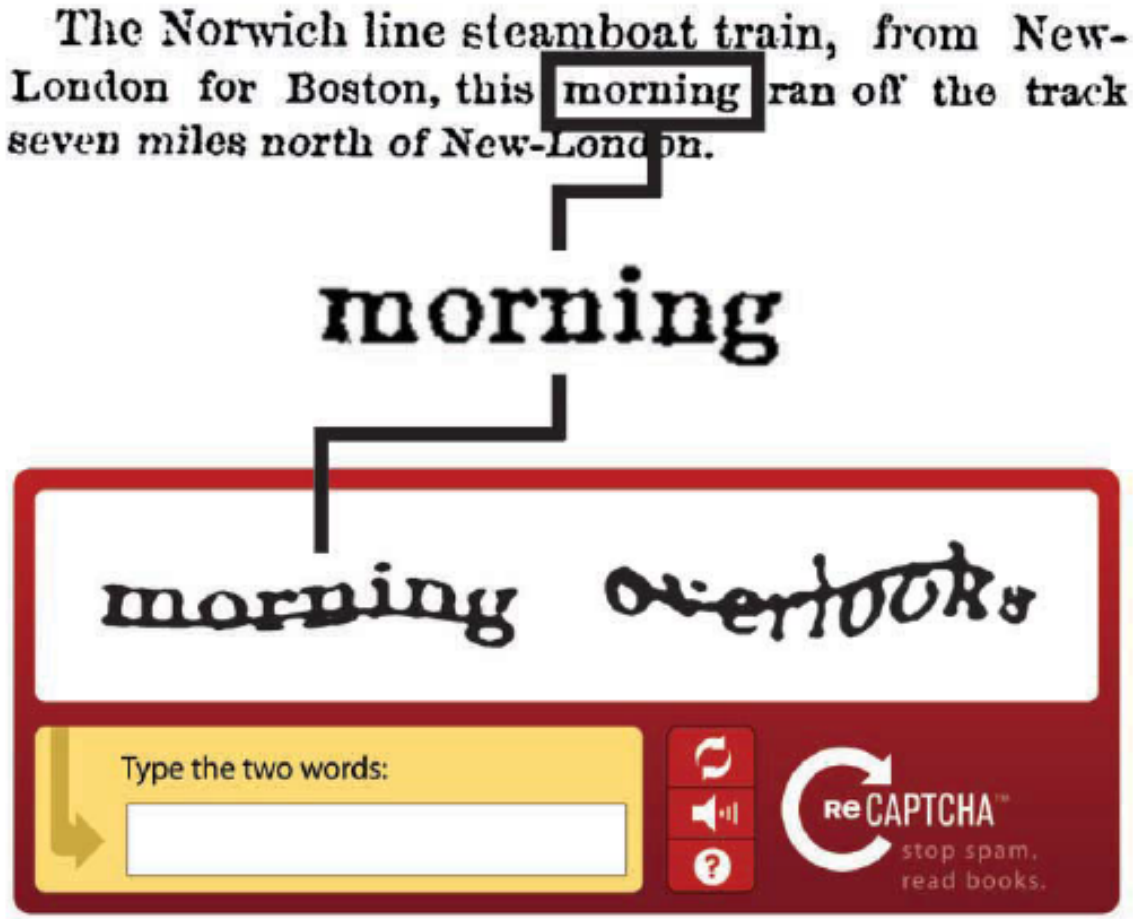


Fig. 1. The reCAPTCHA system displays words from scanned texts to humans on the World Wide Web. In this example, the word "morning" was unrecognizable by OCR. reCAPTCHA isolated the word, distorted it using random transformations including adding a line through it, and then presented it as a challenge to a user. Because the original word ("morning") was not recognized by OCR, another word for which the answer was known ("overlooks") was also presented to determine if the user entered the correct answer.

Working of reCAPTCHA

- Two different OCR programs analyze the image; and deciphered output is compared to each other and an english dictionary.
- Any discrepancy is marked as suspicious, placed in a image then distorted to be used as CAPTCHA
- A vocabulary of 100,000 control words is used to avoid random guesses by bots
- Only words that both OCR programs failed to recognize are used as control words.
- Any program that can recognize these words with non negligible probability would represent an improvement over state of the art OCR programs.

Working of reCAPTCHA

- To account for human error in the digitization process, reCAPTCHA sends every suspicious word to multiple users, with a different random distortion.
- Mechanism of votes is used before putting confidence in digitization.
  - Each human guess counts to 1 vote; each OCR guess being 0.5
  - A guess must obtain at least 2.5 votes before it is chosen as the correct spelling of the word for the digitization process.
- If the first three human guesses match each other, but differ from both of the OCRs' guesses, then (and only then) the word becomes a control word in other challenges. Why?
- When six users reject a word before any correct spelling is chosen, the word is discarded as unreadable

Experiments and Results

- A random sample of 50 old scanned were chosen and manually transcribed by two professionals to compare accuracy of reCAPTCHA and OCR.
- Each word counted as a “hit” if the algorithm deciphered the entire word correctly or a “miss” if any of the letters were wrong.
- The error rate was defined as the number of misses divided by the total number of words.
- The results of one OCR program were run through the same process for comparison.
- reCAPTCHA system achieved an accuracy of 99.1% at word level while OCR was accurate for only 83.5%
- The percentage of words on which both OCR systems made a mistake was 7.3%

Analysis of Results

- reCAPTCHA can achieve an accuracy comparable to the "gold standard" accuracy
- This may be primarily because words presented by reCAPTCHA are shown individually, in isolation from the original context.
- Also, as reCAPTCHA uses a combination of OCR and multiple humans, which in some cases turns out to be more resilient to accidental typographical mistakes

Some More Analysis

- reCAPTCHA is more secured than conventional CAPTCHA that generate their own randomly distorted characters.
  - CAPTCHAs come from a limited distribution of possible transformations which machine learning algorithms, after some training, can recognize the distorted characters.
  - While reCAPTCHAs have distortion due to natural fading, scanning noise and distorting.
- A small very small fraction of control words could be correctly deciphered by computer programs. These bots are directly used to as enhancements to OCR's

Factors Affecting Success

- Accuracy of speakers where English is not native language was lower than 'hits' of native speakers

- Success rate is proportional to the length of the control word:
  - Four character words have a success rate of 93.7% and Seven character words, 96.7%
  - This can be explained by longer words providing more context for the users

Achievements and Conclusion

- Till date 1.2 billion CAPTCHAs have been solved, amounting to over 440 million suspicious words correctly deciphered.
- Over 17,600 books have been transcribed till date
  - Assuming 100,000 words per book(400 pages, 250 words per page)
- Conventionally it would have taken workforce of more than 1500 people deciphering words working 40 hours per week
  - assuming an average rate of 60 words per minute
- "Wasted" human processing power is harnessed to solve problems that computers cannot yet solve
- It makes valuable contributions to solve hard AI problems by enhancing OCR

Perspectives and Future Work

- reCAPTCHAs can be used at
  - mail client of sender to prevent spamming
  - At social websites to prevent crawling

- Success of reCAPTCHA project gave birth to duolingo.com
  - Free language education for the world
  - Simultaneously transforming web in various different languages

References

- [1] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford, in CAPTCHA: Using Hard AI Problems For Security

- [2] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, Manuel Blum, in reCAPTCHA: Human-Based Character Recognition via Web Security Measures