

reCAPTCHA

reCAPTCHA is a CAPTCHA-like system designed to establish that a computer user is human (normally in order to protect websites from bots) and, at the same time, assist in the digitization of books. reCAPTCHA was originally developed by Luis von Ahn, David Abraham, Manuel Blum, Michael Crawford, Ben Maurer, Colin McMillen, and Edison Tan at Carnegie Mellon University's main Pittsburgh campus.^[1] It was acquired by Google in September 2009.^[2]

reCAPTCHA has completely digitized the archives of *The New York Times* and books from Google Books, as of 2011.^[3] The archive can be searched from the *New York Times* Article Archive, where more than 13 million articles in total have been archived, dating from 1851 to the present day.^[4] Through mass collaboration, reCAPTCHA was helping to digitize books that are too illegible to be scanned by computers, as well as translate books to different languages, as of 2015.^[5]

The system has been reported as displaying over 100 million CAPTCHAs every day,^[6] on sites such as Facebook, TicketMaster, Twitter, 4chan, CNN.com, StumbleUpon,^[7] Craigslist (since June 2008),^[8] and the U.S. National Telecommunications and Information Administration's digital TV converter box coupon program website (as part of the US DTV transition).^[9]

reCAPTCHA's slogan was "Stop Spam, Read Books."^[10] until the introduction of a new version of the reCAPTCHA plugin in 2014 which is now changed to "Tough on Bots, Easy on Humans."^[11]; the slogan has now disappeared from the website^[12] and from the classic version of the reCAPTCHA plugin. A new system featuring image verification was also introduced. In this system, users are asked to just click on a checkbox (the system will verify whether the user is a human or not, for example, with some clues such as already-known cookies or mouse movements within the ReCAPTCHA frame) or, if it fails, select one or more images from a selection of nine images.^[13]

reCAPTCHA



reCAPTCHA

Original author(s)	Luis von Ahn Ben Maurer Colin McMillen Harshad Bhujbal Manuel Blum
Developer(s)	Google Inc.
Initial release	May 27, 2007
Type	Classic version: CAPTCHA New version: checkbox
Website	www.google.com/recaptcha (http://www.google.com/recaptcha)

Contents

Origin

Operation

No CAPTCHA reCAPTCHA

Implementation

Criticism

Security

Derivative projects

References

External links

Origin

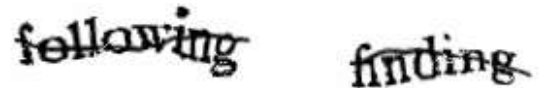
Distributed Proofreaders was the first project to volunteer its time to decipher scanned text that could not be read by OCR. It works with Project Gutenberg to digitize public domain material and uses methods quite different from reCAPTCHA.

The reCAPTCHA program originated with Guatemalan computer scientist Luis von Ahn,^[14] and was aided by a MacArthur Fellowship. An early CAPTCHA developer, he realized "he had unwittingly created a system that was frittering away, in ten-second increments, millions of hours of a most precious resource: human brain cycles".^{[15][16]}

Operation

Scanned text is subjected to analysis by two different optical character recognition programs – one of them, as mentioned the project developer Ben Maurer, is ABBYY FineReader.^[18] Their respective outputs are then aligned with each other by standard string-matching algorithms and compared both to each other and to an English dictionary. Any word that is deciphered differently by both OCR programs or that is not in the English dictionary is marked as "suspicious" and converted into a CAPTCHA. The suspicious word is displayed, out of context, sometimes along with a control word already known.

If the human types the control word correctly, then the response to the questionable word is accepted as probably valid. If enough users were to correctly type the control word, but incorrectly type the second word which OCR had failed to recognize, then the digital version of documents could end up containing the incorrect word. The identification performed by each OCR program is given a value of 0.5 points, and each interpretation by a human is given a full point. Once a given identification hits 2.5 points, the word is considered valid. Those words that are consistently given a single identity by human judges are later recycled as control words.^[19] If the first three guesses match each other but do not match either of the OCRs, they are considered a correct answer, and the word becomes a control word.^[20] When six users reject a word before any correct spelling is chosen, the word is discarded as unreadable.^[20]



An example of how a reCAPTCHA challenge looked in 2007,^[17] containing the words "following finding". The waviness and horizontal stroke were added to increase the difficulty of breaking the CAPTCHA with a computer program.

The original reCAPTCHA method was designed to show the questionable words separately, as out-of-context correction, rather than in use, such as within a phrase of five words from the original document.^[21] Also, the control word might mislead context for the second word, such as a request of "/metal/ /fife/" being entered as "metal file" due to the logical connection of filing with a metal tool being considered more common than the musical instrument "fife".

In 2012, reCAPTCHA began using photographs of house numbers taken from Google's Street View project, in addition to scanned words.^[22]

In 2014, reCAPTCHA implemented another system in which users are asked to select one or more images from a selection of nine images.^[13]

In 2017, reCAPTCHA was improved to require no interaction for most users.^[23]

No CAPTCHA reCAPTCHA

In 2013, reCAPTCHA began implementing behavioral analysis of the browser's interactions with the CAPTCHA to predict whether the user was a human or a bot before displaying the captcha, and presenting a "considerably more difficult" captcha in cases where it had reason to think the user might be a bot. By end of 2014 this mechanism started

to be rolled out to most of the public Google services.^[24] Because NoCAPTCHA relies on the use of Google cookies that are at least a few weeks old, reCAPTCHA has become nearly impossible to complete for people who frequently clear their cookies.

In 2017, Google improved this mechanism, calling it an "invisible reCAPTCHA". According to former Google "click fraud czar" Shuman Ghosemajumder, this capability "creates a new sort of challenge that very advanced bots can still get around, but introduces a lot less friction to the legitimate human."^[25]

Implementation

The reCAPTCHA tests are displayed from the central site of the reCAPTCHA project, which supplies the words to be deciphered. This is done through a JavaScript API with the server making a callback to reCAPTCHA after the request has been submitted. The reCAPTCHA project provides libraries for various programming languages and applications to make this process easier. reCAPTCHA is a gratis service (that is, the CAPTCHA images are provided to websites free of charge, in return for assistance with the decipherment),^[26] but the reCAPTCHA software itself is not open source.^[27]

Also, reCAPTCHA offers plugins for several web-application platforms, like ASP.NET, Ruby, or PHP, to ease the implementation of the service.^[28]

Criticism

Some have criticized Google for using reCAPTCHA as a source of unpaid labor.^[29] They say Google is unfairly using people around the world to help it transcribe books, addresses, and newspapers without any compensation.

The use of reCAPTCHA has been labelled "a serious barrier to internet use" for people with sight problems or disabilities such as dyslexia by BBC journalist Stephanie Hegarty.^[30]

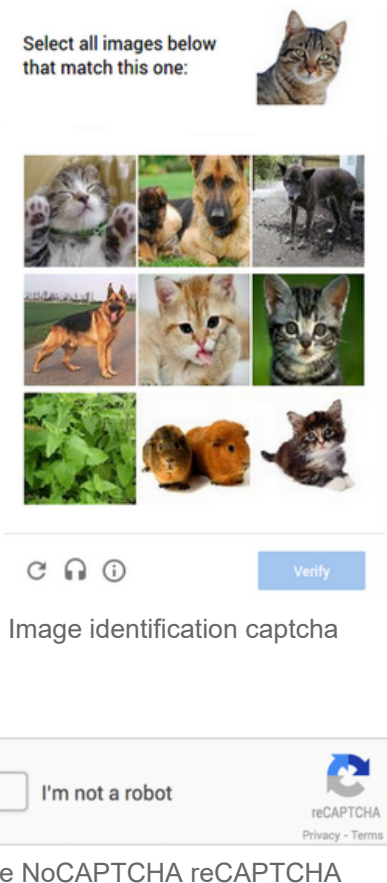
reCAPTCHA is also a barrier to Internet use in areas of the world where there is heavy Internet censorship and the underlying enabling sites are blocked.

Software engineer Andrew Munsell, in his article "Captchas Are Becoming Ridiculous" states "A couple of years ago, I don't remember being truly baffled by a captcha. In fact, reCAPTCHA was one of the better systems I'd seen. It wasn't difficult to solve, and it seemed to work when I used it on my own websites." ^[31] Munsell goes on to state, after encountering a series of unintelligible images that despite refreshing "Again, and again, and again. The captchas were not only difficult for a computer to read, but impossible for a human." Munsell then provided numerous examples.

Security

The main purpose of a CAPTCHA system is to prevent automated access to a system by computer programs or "bots". On 14 December 2009, Jonathan Wilkins released a paper describing weaknesses in reCAPTCHA that allowed a solve rate of 18%.^{[33][34][35]}

On 1 August 2010, Chad Houck gave a presentation to the DEF CON 18 Hacking Conference detailing a method to reverse the distortion added to images which allowed a computer program to determine a valid response 10% of the time.^{[36][37]} The reCAPTCHA system was modified on 21 July 2010, before Houck was to speak on his method. Houck



modified his method to what he described as an "easier" CAPTCHA to determine a valid response 31.8% of the time. Houck also mentioned security defenses in the system, including a high security lock out if an invalid response is given 32 times in a row.^[38]



An example of how reCAPTCHA challenges were presented in 2010,^[32] containing the words "and chisels"

On 26 May 2012, Adam, C-P and Jeffball of DC949 gave a presentation at the LayerOne hacker conference detailing how they were able to achieve an automated solution with an accuracy rate of 99.1%.^[39] Their tactic was to use techniques from machine learning, a subfield of artificial intelligence, to analyse the audio version of reCAPTCHA which is available for the visually impaired. Google released a new version of reCAPTCHA just hours before their talk, making major changes to both the audio and visual versions of their service. In this release, the audio version was increased in length from 8 seconds to 30 seconds, and is much more difficult to understand, both for humans as well as bots. In response to this update and the following one, the members of DC949 released two more versions of Stiltwalker which beat reCAPTCHA with an accuracy of 60.95% and 59.4% respectively. After each successive break, Google updated reCAPTCHA within a few days. According to DC949, they often reverted to features that had been previously hacked.

On 27 June 2012, Claudia Cruz, Fernando Uceda, and Leobardo Reyes (a group of students from Mexico) published a paper showing a system running on reCAPTCHA images with an accuracy of 82%.^[40] The authors have not said if their system can solve recent reCAPTCHA images, although they claim their work to be intelligent OCR and robust to some, if not all changes in the image database.

In an August 2012 presentation given at BsidesLV 2012, DC949 called the latest version "unfathomably impossible for humans" – they were not able to solve them manually either.^[39] The web accessibility organization WebAIM reported in May 2012, "Over 90% of respondents [screen reader users] find CAPTCHA to be very or somewhat difficult."^[41]

reCAPTCHA frequently modifies its system, requiring spammers to frequently update their methods of decoding, which may frustrate potential abusers.

Only words that both OCR programs failed to recognize are used as control words. Thus, any program that can recognize these words with nonnegligible probability would represent an improvement over state of the art OCR programs.^[20]

Derivative projects

reCAPTCHA had also created project Mailhide, which protects email addresses on web pages from being harvested by spammers.^[42] By default, the email address is converted into a format that does not allow a crawler to see the full email address; for example, "mailme@example.com" would be converted to "mai...@example.com". The visitor would then click on the "..." and solve the CAPTCHA in order to obtain the full email address. One can also edit the pop-up code so that none of the address is visible.


References

1. "reCAPTCHA: About Us" (<https://web.archive.org/web/20100611210259/http://recaptcha.net/aboutus.html>). Archived from [the original](http://recaptcha.net/aboutus.html) (<http://recaptcha.net/aboutus.html>) on 2010-06-11. Retrieved 2018-08-14.
2. "Teaching computers to read: Google acquires reCAPTCHA" (<http://googleblog.blogspot.com/2009/09/teaching-computers-to-read-google.html>). Google. Retrieved 2009-09-16.
3. "Deciphering Old Texts, One Woozy, Curvy Word at a Time" (<https://www.nytimes.com/2011/03/29/science/29recaptcha.html>). The New York Times. March 28, 2011. Retrieved November 20, 2017.
4. "New York Times Article Archive" (<https://www.nytimes.com/ref/membercenter/nytarchive.html>). *The New York Times*. 2007-09-25. ISSN 0362-4331 (<https://www.worldcat.org/issn/0362-4331>). Retrieved 2017-11-21.

5. "Massive-scale online collaboration" (http://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration.html). *www.ted.com*. Retrieved 2015-10-24.
6. "reCAPTCHA FAQ" (<http://www.google.com/recaptcha/faq>). Google. Retrieved 2011-06-12.
7. Rubens, Paul (2007-10-02). "Spam weapon helps preserve books" (<http://news.bbc.co.uk/2/hi/technology/7023627.stm>). BBC.
8. "Fight Spam, Digitize Books" (<http://blog.craigslist.org/2008/06/fight-spam-digitize-books/>). Craigslist Blog. June 2008.
9. "TV Converter Box Program" (<https://web.archive.org/web/20091104004349/https://www.dtv2009.gov/>). *dtv2009.gov*. Archived from the original (<https://www.dtv2009.gov/>) on 2009-11-04.
10. "reCAPTCHA: Stop Spam, Read Books" (<http://www.google.com/recaptcha>). Google. Retrieved 2013-07-10.
11. "reCAPTCHA: Easy on Humans, Hard on Bots" (<https://www.google.com/recaptcha/intro/index.html>). *www.google.com*. Retrieved 2018-02-01.
12. "recaptcha" (<http://www.google.com/recaptcha/intro/index.html>). Google. Retrieved 2015-01-03.
13. Greenberg, Andy (December 3, 2014). "Google Can Now Tell You're Not a Robot with Just One Click" (<https://www.wired.com/2014/12/google-one-click-recaptcha/>). *Wired*. Retrieved October 1, 2015.
14. "'Full Interview: Luis von Ahn on Duolingo', Spark, November 2011" (<http://www.cbc.ca/spark/2011/11/full-interview-luis-von-ahn-on-duolingo/>). Canadian Broadcasting Corporation. 2011-11-30. Retrieved 2013-07-10.
15. Hutchinson, Alex (March 2009). "Human Resources: The job you didn't even know you had". *The Walrus*. pp. 15–16.
16. Hutchinson, Alex. "Human Resources: The job you didn't even know you had" (<http://thewalrus.ca/human-resources/>). *The Walrus*. Retrieved 7 December 2015.
17. Contributor. "reCAPTCHA: Using Captchas To Digitize Books – TechCrunch" (<https://techcrunch.com/2007/09/16/recaptcha-using-captchas-to-digitize-books/>). *techcrunch.com*.
18. "What is the best OCR software on the market?" (<https://www.quora.com/What-is-the-best-OCR-software-on-the-market/answer/Ben-Maurer>). Retrieved 2016-03-21.
19. Timmer, John (2008-08-14). "CAPTCHAs work? for digitizing old, damaged texts, manuscripts" (<https://arstechnica.com/news/ars/post/20080814-captchas-workfor-digitizing-old-damaged-texts-manuscripts.html>). *Ars Technica*. Retrieved 2008-12-09.
20. Luis; Maurer, Ben; McMillen, Colin; Abraham, David; Blum, Manuel (2008). "reCAPTCHA: Human-Based Character Recognition via Web Security Measures" (PDF). *Science*. **321** (5895): 1465–1468. doi:10.1126/science.1160379 (<https://doi.org/10.1126/science.1160379>). PMID 18703711 (<https://www.ncbi.nlm.nih.gov/pubmed/18703711>).
21. "'questionable validity of results if words are presented out of context', Google Groups, August 29, 2008" (http://groups.google.com/group/recaptcha/browse_thread/thread/c53efad7ac89fd2). Google. Retrieved 2013-07-10.
22. March 29th, 2012 (2012-03-29). "Google Now Using ReCAPTCHA To Decode Street View Addresses" (<https://techcrunch.com/2012/03/29/google-now-using-recaptcha-to-decode-street-view-addresses/>). TechCrunch. Retrieved 2013-07-10.
23. Certification, Digital (2017-03-14). "Digital Certification: The Digital Rating For Websites" (<https://digital-certification.com/blog/google-improves-their-captcha-with-no-user-interaction-required/>). *Digital Certification | Blog*. Retrieved 2017-03-14.
24. "Are you a robot? Introducing 'No CAPTCHA reCAPTCHA'" (<http://googleonlinesecurity.blogspot.com/2014/12/are-you-robot-introducing-no-captcha.html>). Google. 2014-12-03. Retrieved 2015-04-14.
25. "Google just made the internet a tiny bit less annoying" (<http://www.popsci.com/google-invisible-recaptcha#page-3>). *Popular Science*. 2017-03-10. Retrieved 2017-04-05.
26. "FAQ" (<https://archive.is/20120716142737/http://recaptcha.net/faq.html>). reCAPTCHA.net. Archived from the original (<http://recaptcha.net/faq.html>) on 2012-07-16.
27. "reCAPTCHA: Stop Spam, Read Books" (<http://www.google.com/recaptcha>). Google. Retrieved 2014-01-14.
28. "Developer's Guide – reCAPTCHA — Google Developers" (<https://developers.google.com/recaptcha/intro?csw=1>). Google. Retrieved 2014-01-14.
29. "Massachusetts woman's lawsuit accuses Google of using free labor to transcribe books, newspapers" (<http://www.bizjournals.com/boston/blog/techflash/2015/01/massachusetts-womans-lawsuit-accuses-google-of.html>). *Boston Business Journal*.

30. "BBC News – The evolution of those annoying online security tests" (<https://www.bbc.com/news/magazine-18367017>). *bbc.com*. Retrieved 2014-09-22.
31. "Captchas Are Becoming Ridiculous | Andrew Munsell" (<https://www.andrewmunsell.com/blog/captchas-are-becoming-ridiculous>). *andrewmunsell.com*. Retrieved 2014-09-22.
32. Firewall, The. "Those Scrambled Word Tests For Stopping Spambots Are Tough For Humans Too" (<https://www.forbes.com/sites/firewall/2010/06/18/those-scrambled-word-tests-for-stopping-spambots-are-tough-for-humans-too/>). *forbes.com*.
33. "Strong CAPTCHA Guidelines" (<http://bitland.net/captcha.pdf>) (PDF).
34. "Google's reCAPTCHA busted by new attack" (https://www.theregister.co.uk/2009/12/14/google_recaptcha_busted/).
35. "Google's reCAPTCHA dented" (<http://www.h-online.com/security/news/item/Google-s-reCAPTCHA-dented-888859.html>).
36. "Def Con 18 Speakers" (<http://www.defcon.org/html/defcon-18/dc-18-speakers.html#Houck>). *defcon.org*.
37. "Decoding reCAPTCHA Paper" (<https://web.archive.org/web/20100819053439/http://n3on.org/projects/reCAPTCHA/docs/reCAPTCHA.docx>). Chad Houck. Archived from the original (<http://n3on.org/projects/reCAPTCHA/docs/reCAPTCHA.docx>) on 2010-08-19.
38. "Decoding reCAPTCHA Power Point" (<https://web.archive.org/web/20101024210642/http://n3on.org/projects/reCAPTCHA/docs/reCAPTCHA.pptx>). Chad Houck. Archived from the original (<http://n3on.org/projects/reCAPTCHA/docs/reCAPTCHA.pptx>) on 2010-10-24.
39. "Project Stiltwalker" (<http://www.dc949.org/projects/stiltwalker/>).
40. Claudia Cruz-Perez; Oleg Starostenko; Fernando Uceda-Ponga; Vicente Alarcon-Aquino; Leobardo Reyes-Cabrera (30 June 2012). "Breaking reCAPTCHAs with Unpredictable Collapse: Heuristic Character Segmentation and Recognition". In Carrasco-Ochoa, Jesús Ariel; Martínez-Trinidad, José Francisco; Olvera López, José Arturo; Boyer, Kim L. *Pattern Recognition* (https://dx.doi.org/10.1007/978-3-642-31149-9_16). Lecture Notes in Computer Science. **7329**. México. pp. 155–165. doi:10.1007/978-3-642-31149-9_16 (https://doi.org/10.1007%2F978-3-642-31149-9_16). ISBN 978-3-642-31148-2.
41. "Screen Reader User Survey #4 Results" (<http://webaim.org/projects/screenreadersurvey4/#captcha/>).
42. "Mailhide: Free Spam Protection" (<http://www.google.com/recaptcha/mailhide/>). Google.

External links

- Official website (<https://www.google.com/recaptcha>) 
- Repository (<https://github.com/google/recaptcha>)
- ReCAPTCHA: The job you didn't even know you had (<http://www.walrusmagazine.com/articles/2009.03-technology-human-resources-recaptcha-alex-hutchinson/>) Two-page article in *The Walrus* magazine
- Luis; Maurer, Benjamin; McMillen, Colin; Abraham, David; Blum, Manuel (2008). "reCAPTCHA: Human-Based Character Recognition via Web Security Measures". *Science*. **321** (5895): 1465–1468. doi:10.1126/science.1160379 (<https://doi.org/10.1126%2Fscience.1160379>). PMID 18703711 (<https://www.ncbi.nlm.nih.gov/pubmed/18703711>).
- Massive-scale online collaboration (https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration), a TED talk by Luis von Ahn

Retrieved from "<https://en.wikipedia.org/w/index.php?title=ReCAPTCHA&oldid=863294508>"

This page was last edited on 9 October 2018, at 21:45 (UTC).

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.