ReMAPTCHA: A Map-based Anti-Spam Method that Helps to Correct OpenStreetMap

Stefan KELLER

University of Applied Sciences, Rapperswil / Switzerland · sfkeller@hsr.ch

This contribution was double-blind reviewed as extended abstract.

Abstract

This is a report about a prototype software called ReMAPTCHA, which is a map-based anti-spam method. In fact, it is a self-contained variant of a reCAPTCHA. It can also serve to correct topological errors ("almost connections") in OpenStreetMap.

1 Introduction

A CAPTCHA (AHN et al. 2003, AHN et al. 2004) is a software that prevents automated bots from registering themselves in online systems, where they aim to leave unwanted web links or initiate, or collect unsolicited mails. It's an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart". A CAPTCHA generates images with distorted text and asks user to type in the original one. Since computers can't read the distorted text while humans can (CHELLAPILLA et al. 2005), bots cannot intrude to sites protected by CAPTCHAs.

Fig. 1: Typical CAPTCHA image with distorted letters (here: "HSR")

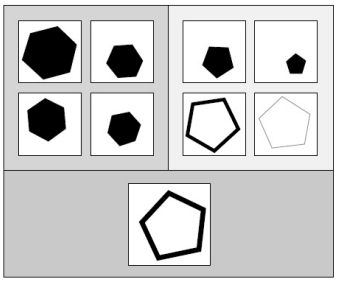

Fig. 2: A BONGO image. To which side does the block on the bottom belong? (BONGARD 1970)

There are alternatives to CAPTCHAs, like "honeypots", and CAPTCHA variants, e.g. based on image recognition (BONGARD 1970), sound, math puzzles ("What is 7 minus 3 times 2?") or trivia questions ("What tastes better, a toad or a popsicle?"). But the CAPTCHAs like those explained above are seen as the superior solution currently available

Vogler, R., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2014): GI_Forum 2014. Geospatial Innovation for Society.

to prevent spam attacks from a security and usability point of view (PENNINGER & FEDERRATH 2012).

reCAPTCHAs (AHN et al. 2008) are an evolutionary refinement of CAPTCHAs. ReCAPTCHAs utilize the mental effort exerted by users so that it's not wasted. Each reCAPTCHA has two parts, one contains the "control word" which is already known by the system and the other contains the "unknown word" which is captured from text book and needs to be digitized. The user is asked to read both words in random order. The user's input will be verified against the "control word" so the user needs to type in this word correctly in order to pass the reCAPTCHA test. Furthermore, if the "control word" is entered correctly, the system assumes the user's answer for the "unknown word" is reliable. With reCAPTCHA, users not only solve CAPTCHA words, but also help to digitize book texts. The reCAPTCHA was patented to Google and is being offered as a free web service.

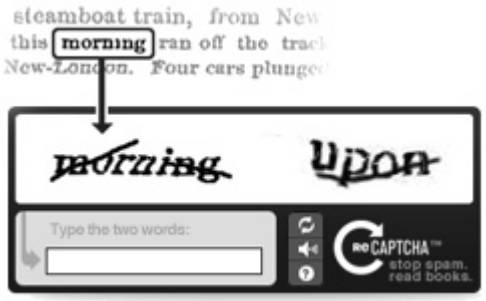

Fig. 3:
A sample of reCAPTCHA in which 'upon' is "control word" and 'morning' is "unknown word" (AHN et al. 2008)

In this contribution, a variant of such a reCAPTCHA is being presented called ReMAPTCHA – which is a combination of the words "map" and "reCAPTCHA". Instead of digitizing book texts, the ReMAPTCHA helps to correct errors in the OpenStreetMap (OSM) database.

OSM is a wiki-like project for the collaborative creation of a world map. Since OSM allows users to edit the map at any time, there are inevitable errors in the database. A typical type of error is a missing connection between roads (ways), which could be a topological error. While two roads may look like they are connected in the rendered map, there may still be a gap between them in the database. In this case, navigation systems using OSM data will give insufficient results. As of end of 2013 there are about 400.000 potential error cases in OSM where two roads are closer than 5 meters.

Figure 4 illustrates a typical error case where the two roads are actually connected but there is a gap between the two in the OSM database. A user looking at the image using common human sense will help to correct the error by answering a connection-related question generated by the ReMAPTCHA.

2 Method

A ReMAPTCHA consists of two "control words" and a part of a satellite image showing the situation. The two "Control words" are distorted and label the roads in the ReMAPTCHA (see Figure 4 and 5).

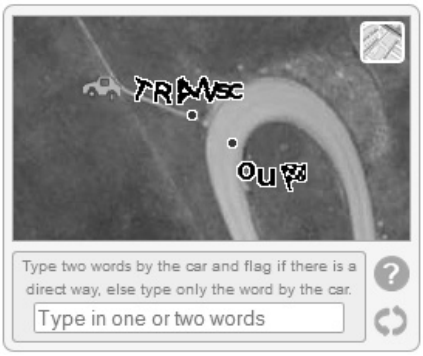

Fig. 4: A case in which 2 roads are connected

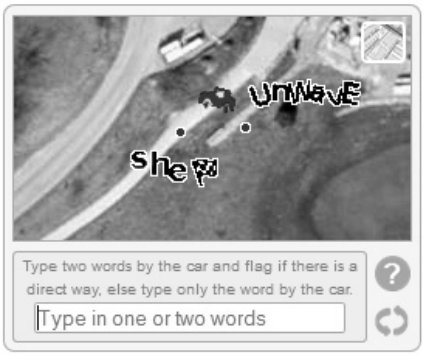

Fig. 5: A case in which 2 roads are not connected

Figure 4 shows the prototype of the ReMAPTCHA project. The user is asked a security question like this: "*Type two words by the car and flag if there is a direct way, else type only the word by the car.*". In the case of figure 4, the correct answer would be "**TRANSC ou**" because the two roads look like connected. So the answer contains the two labels on the direct route from the car to the flag. Another acceptable answer would be just "**TRANSC**" (without "ou"). With this answer, the user would state that the two roads are unconnected. This fact is the "unknown part" of the ReMAPTCHA. Whatever the user answers, his answer must contain at least one "control word" in order to pass the challenge. Just in case, there is a button on the upper right showing the OSM map as an additional source of information.

In the other case, as shown in figure 5, the roads are not connected in reality, so the correct answer would be "**unwave**" (without "**she**").

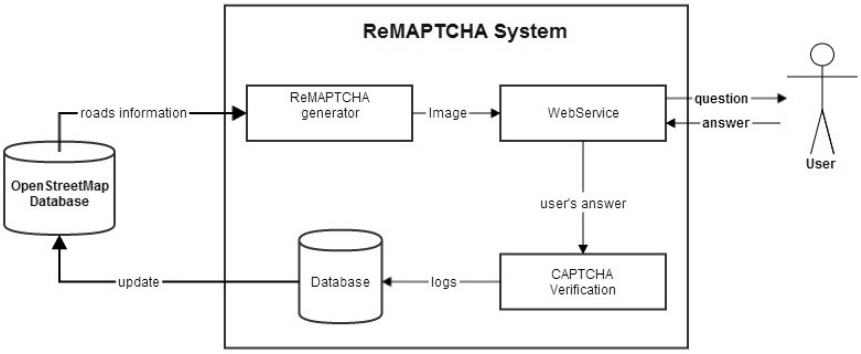

Fig. 6: Data flow of the ReMAPTCHA

The user's input will be then verified by checking if it correctly matches the "control words". In figure 3, the system will verify both "control words" even the second one "**ou**".

If they match, the system will recognize the user as human and grant him privilege. This is the CAPTCHA part of our algorithm.

Besides solving the "control words" in the image, the user also answers the question about connection between two roads. For example, by typing the second word a user confirms the connection between the two roads. This is the ReCAPTCHA part of our algorithm.

3 Result and Discussion

A first prototype of the ReMAPTCHA algorithm was realized in our lab and is in testing and evaluation phase. The software is written in Python; the data is managed in the open source database PostgreSQL. The finalized algorithm will be published, so that the security of the system can be assessed by independent researchers.

One issue of CAPTCHAs is that humans are disrupted when forced to solve it. The ReMAPTCHA contains far less characters per word, as compared, for example, to the prevalent Google reCAPTCHA.

As shown by (CHELLAPILLA & SIMARD 2005, THAYANANTHAN et al. 2003) it is possible to implement programs that can break text image CAPTCHA with a success rate of more than 90%. In contrast to a conventional CAPTCHA – in which text are read horizontally from left-to-right order – the "control words" in our ReMAPTCHA algorithm are placed among two-dimensions with arbitrary base line orientation. This makes the "control words" harder to be deciphered leading to an increased security (Bursztein et al. 2010).

Future work includes the following:

The same ReMAPTCHA image is served multiple times (about 5) to different users (perhaps within a certain time window). All users' judgments about connections will be logged into database. Only if all responses are 100% consistent (whether the connection is confirmed or denied) the correction will be considered as reliable, and the OSM database will be automatically updated.

A user (or bot) who fails to solve the "control words" will be given another chance with another ReMAPTCHA. However, the system will switch to "safe mode" after three unsuccessful tries, showing only challenges where the correct answer is known by the system.

Developing a distortion algorithm by combining multiple types of distortion (e.g. using faded text, adding noise, applying artificial transformation).

Increasing the reliability of users' judgments by showing potential errors around the user's location based on "ip address geocoding". The underlying assumption here is that the user is more familiar with their surroundings.

Increasing the reliability of users' judgments by providing satellite or orthophoto images, thereby introducing more "ground truth" (although availability of such base map material is limited).

In conclusion, the ReMAPTCHA algorithm not only prevents web applications from spam attacks in a more secure way, but also utilizes the user's effort to correct the OSM database. In a next step, ReMAPTCHA will be put online for user testing. After having passed all acceptance tests it will be offered to the public, like to registration pages of OpenStreetMap websites (wikis, local main site, etc.). With approximately 1.500 new users registered per day (according to OSM statistics page, January 2014), we hope that the ReMAPTCHA service can help to correct many errors, therefore serving to enhance the capturing process and data quality of OSM.

References

AHN, L., MAURER, B., MCMILLEN, C., ABRAHAM, D. & BLUM, M. (2008), reCAPTCHA: Human-Based Character Recognition via Web. Security Measures, 321, 1465-1468. Published Sept. 2008 at http://www.cs.cmu.edu/~biglou/reCAPTCHA_Science.pdf (accessed 31. Jan. 2014).

AHN, L., BLUM, M. & LANGFORD, J. (2004), Telling Human And Computer Apart Automatically. Communication of ACM, 47, 57-60.

AHN, L., BLUM, M., HOPPER, N. & LANGFORD, J. (2003), Advances in Cryptology. BIHAM, E. (Ed.), Lecture Notes in Computer Science, 2656, 294-311.

BONGARD, M. M. (1970), Pattern Recognition. Rochelle Park, NJ, Spartan Books.

BURSZTEIN, E., BETHARD, S., MITCHELL, J C., JURAFSKY, D. & FABRY, C. (2010), How Good Are Humans At Solving Captchas? A Large Scale Evaluation. In: Security and Privacy.

CHELLAPILLA, K. & SIMARD, P. Y. (2005), Advances in Neural Information Processing Systems 17. SAUL, L. K., WEISS, Y. & BOTTOU, L. (Eds), MIT Press, Cambridge, MA, 265-272.

CHELLAPILLA K., LARSON, K., SIMARD, P. & CZERWINSKI, M. (2005), Designing Human Friendly Human Interaction Proofs (HIPs). In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 711-720. Association for Computing Machinery, New York.

PENNINGER S. & FEDERRATH. H. (2012), Sicherheit und Usability von CAPTCHAs. digma: Zeitschrift für Datenrecht und Informationssicherheit, 2/2012, 84-86. http://epub.uni-regensburg.de/23565/ (accessed 31. Jan. 2014).

THAYANANTHAN A., STENGER, B., TORR, P. H. S. & CIPOLLA, R. (2003), Shape Context and Chamfer Matching in Cluttered Scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 127-133. IEEE Computer Society, Los Alamitos, CA.