10-2022

An Accessible Web CAPTCHA Design for Visually Impaired Users

Manso Alqarni

Fangyi Yu

Rupendra Raavi

Mahadeo Sukhai

5. An Accessible Web CAPTCHA Design for Visually Impaired Users

Mansour Alqarni,
Ontario Tech University
Mansour.alqarni@ontariotechu.net

Fangyi Yu,
Ontario Tech University
Fangyi.Yu@ontariotechu.net

Rupendra Raavi,
Ontario Tech University
Rupendra.Raavi@ontariotechu.ca

Mahadeo Sukhai,
ARIA Team
mahadeo.sukhai@cnib.ca

Abstract

In the realm of computing, CAPTCHAs are used to determine if a user engaging with a system is a person or a bot. The most common CAPTCHAs are visual in nature, requiring users to recognize images comprising distorted characters or objects. For people with visual impairments, audio CAPTCHAs are accessible alternatives to standard visual CAPTCHAs. Users are required to enter or say the words in an audio-clip when using Audio CAPTCHAs. However, this approach is time-consuming and vulnerable to machine learning algorithms, since automated speech recognition (ASR) systems could eventually understand the content of audio with the improvement of the technique. While adding background noise may deceive ASR systems temporarily, it may cause people to have difficulties deciphering the information, thus reducing usability. To address this, we designed a more secure and accessible web CAPTCHA based on the capabilities of people with visually impairments, obviating the need for sight via the use of audio and movement, while also using object detection techniques to enhance the accessibility of the CAPTCHA.

Keywords:
Accessibility, CAPTCHAs, Computer Vision, Human-computer Interaction, Impairment.

1. Introduction

CAPTCHAs (Completely Automated Public Turing Tests to Distinguish Computers from Humans) are often used online to distinguish between human users and non-human bots (Von Ahn et al., 2004). Most CAPTCHAs do this by requiring users to perform visual-processing tasks that are easy for humans but complex for bots (Kaur & Behal, 2014). These visual-processing tasks, on the other hand, are unavailable to the world's 285 million persons with visual impairments (PVIs), 39 million of whom are completely blind and 246 million of whom have impaired vision. Rather than that, PVIs depend on audio CAPTCHAs, which use auditory processing tasks to distinguish people from bots.

Audio CAPTCHAs are substantially less useful than their visual counterparts in their present condition (Sasse et al., 2001). While visual CAPTCHAs take an average of 9.8 seconds to solve and have a 93% success rate, audio CAPTCHAs take an average of 51 seconds to complete and have a 50% success rate (Bursztein et al., 2010). This performance and accuracy disparity exists because current audio CAPTCHAs are modeled after their visual equivalents rather than using audio-specific designs (Bigham et al., 2009). As such, present audio CAPTCHAs impose unreasonably high demands on the users who rely on them in terms of attention and memory capacity. This implies that visual CAPTCHAs are not a

comparable challenge to audio CAPTCHAs; audio CAPTCHAs are more troublesome for PVIs than visual designs are for fully sighted individuals (Holman et al., 2007).

Interference in audio is one of the most significant challenges people have while attempting to solve audio CAPTCHAs. Many PVIs, for example, depend on screen readers to assist themselves with navigating user interfaces. When these users begin entering the letters they hear in a CAPTCHA task, their screen reader system reads each typed letter aloud while they listen for the next character in the task. As a result of the auditory conflict between the written and spoken letters, needless user aggravation and mistakes are created.

2. CAPTCHA Solutions for PVIs

2.1 Crowdsourcing and Friend-sourcing

A lot of PVIs rely on distant support services or seek direct assistance from friends and family (Wu & Lada, 2014). While both strategies involve collaboration, crowdsourcing and friend-sourcing can assist PVIs in overcoming the accessibility issues inherent in today's online CAPTCHA challenges. Prior work (Bigham et al., 2010) has examined the possibility of linking PVIs with remote support from sighted assistants, for instance, answer inquiries about the situation or seek verbal advice when utilizing inaccessible interfaces. These applications, however, are restricted to descriptive instruction, limiting their effectiveness in solving task-based CAPTCHAs.

There are other trade-offs to crowdsourcing or soliciting assistance from friends. Crowdsourcing may raise privacy and security concerns for remote control — where a stranger takes over the PVI's system to answer CAPTCHAs on their behalf — friend-sourcing is more in line with PVIs' established workflows: PVIs typically seek support from friends and family members to address physical world accessibility problems (Abdrabo et al., 2016). Nevertheless, friend-sourcing might also create privacy concerns, since users may not want their continuing behaviors exposed to their friends. Additionally, friend-sourcing might help PVIs save money by eliminating the need for professional crowd workers. However, friend-sourcing is slower and less dependable compared with crowdsourcing according to a study on people with Alzheimer disease (Bateman et al., 2017), and PVIs may prefer to avoid flooding their social networks with repetitive requests for assistance (Rzeszotarski & Morris, 2014).

Zhang et al. (Zhang et al., 2021) developed and implemented a novel framework, termed WebAlly, that enables PVIs to obtain immediate assistance from friends or trained crowd workers at the point of need. Their solution, which was implemented as a Google browser extension, consists of mainly two parts: (1) A one-way request from the requester and (2) a synchronous cooperation procedure between the helper and the requester. Their framework, however, was restricted to only one specific type of CAPTCHA challenge (the Google reCAPTCHA) and cannot generalize to other visual-based CAPTCHA challenges (such as dragging puzzles; differentiating and typing distorted letters). Additionally, their approach necessitates interaction between PVIs and their supporters, which may impair PVIs' perceived independence when doing everyday duties.

2.2 Audio CAPTCHAs

In comparison to crowdsourcing and friend-sourcing, audio CAPTCHAs are more pervasive and commonly adopted alternatives to PVIs. There are two categories of audio CAPTCHAs, according to prior research: rule-based and content-based. Rule-based tasks require users to interpret what they hear. For instance, “count the number of times you hear the digit eight”. This type of tasks can alleviate the strain on short-term memory, as they need simply the recall of a running total (Bock et al., 2017). Whereas for the content-based tasks, users need to translate the voice in an audio recording to text. For example, “type the letters in the audio recording in the textbox to pass the CAPTCHA”.

Sauer et al. investigated the impact of content-based designs that closely approximate the latest design standard for audio and visual CAPTCHAs. They played eight digits in distorted voices and asked people to input them sequentially in their experiment. As a result, they found that this strategy dramatically increased the cognitive burden from PVIs, which require them to remember the CAPTCHA sequence or utilize external tools to swiftly identify the objects they heard. The content-based designs are proved to be not usable based on the experiment result of a 46 percent CAPTCHA passing success rate and lengthy average time required to complete tasks (Sauer et al., 2008).

Fanelle et al. (Fanelle et al., 2020) developed CAPTCHAs with a reduced short-term cognitive burden, requiring users to recall only one or two entities at a time. They achieved this using rule-based design and the elimination of acoustic interference. They created four new CAPTCHAs: Math (e. g., answer the sum of two digits), Pauses (e. g., record each letter played in the recording), Character (e.g., count how many times the word “3” is spoken and type the result), and Categories (e.g., count the number of sounds made by trains). They recruited PVIs globally for a comprehensive user study to assess the usability of their CAPTCHA designs, as well as participants' perceptions and opinions. Additionally, they evaluated the CAPTCHAs' security using cutting-edge natural language processing technologies (Solanki et al., 2017). Consequently, Pauses is the least secure design (67 percent of problems were resolved by machines), followed by Match (22 percent were solved by machines). Furthermore, participants regarded the Character design as the most usable and satisfying, while the Categories design was the quickest and most accessible.

3. Methodology

In contrast to the previously discussed CAPTCHA types, we suggest a web CAPTCHA design that incorporates object detection techniques, requiring PVIs to simply obey the audio instruction to turn their heads in order to pass the CAPTCHAs.

Recently, neural networks have returned to the forefront of classification challenges thanks to AlexNet (Krizheysky et al., 2012). A deeper version of the VGG (Simonyan et al., 2014) and the GoogLeNet (Szegedy et al., 2015) suggested an "inception" design to tackle the problem of depth and breadth limitations in deep convolutional neural networks. As time passed by, the residual neural network (ResNet) (Zhang et al., 2016) offered a new residual neural network block design which allowed a weaker and small link to counteract network depth's disappearance of gradients. When DenseNet (Huang et al., 2017) was introduced, it enhanced the accuracy and increased the efficiency of the network, and the computational cost was significantly lowered.

R-CNN (Girshick et al., 2015), Fast R-CNN (Ren et al., 2015), Faster R-CNN (Liu et al., 2016) and other multi-step networks are widely used. Convolution is required for each R-CNN region proposal box, which is a considerable time investment. Fast R-CNN here does a single convolutional proceeding over the complete picture which increases the speed drastically, and the convolution features that are extracted will be directly given as an input to the RPN to obtain feature information from each proposal box, which further enhances the speed and accuracy. Despite this, the two-step network's rate is much lower than that required for real-time detection. YOLO (Shafiee et al., 2017) and SSD, two prominent one-step networks, are at the top of this list of one-step networks. Moreover, object detection algorithms such as YOLO use the intersection over union (IOU) to calculate accuracy which makes themselves more effective than other algorithms.

4. Our Proposal and Implementation

We propose to use the deep learning-based facial recognition model called tiny face detection, built on top of the several ingle-shot shot detector (SSD) units (Liu et al., 2016). Our deep learning model is written in JavaScript, which enables us to deploy it via the web. After developing the model, we distributed it through the web.

The main objective of our model is to detect if there is a human in front of the webcam. Once PVIs click the "identify" button, the webcam will open and determine if a person is present based on the facial recognition model. If a person is recognized, then the CAPTCHA is successfully passed; otherwise, it indicates that the CAPTCHA has failed and prompts the user to try again. Our CAPTCHA design is illustrated in Figure 4.

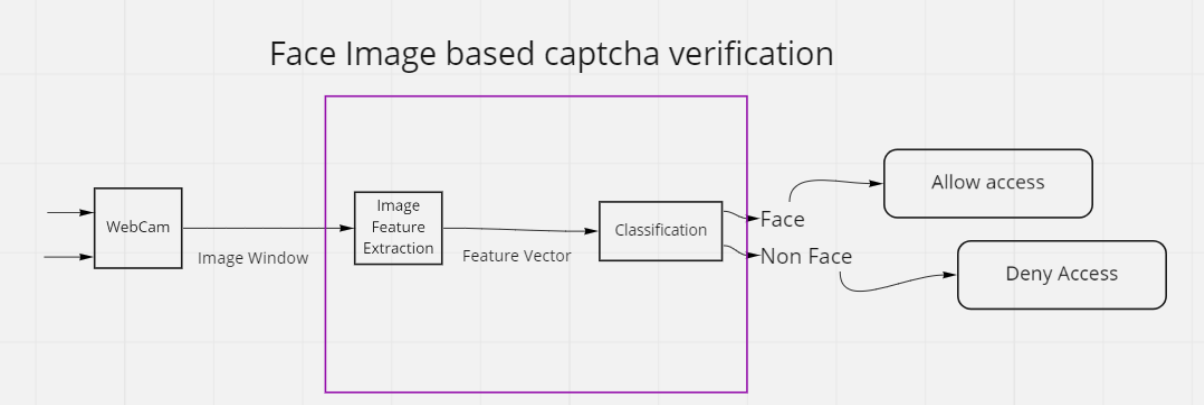

Figure 4: Our CAPTCHA Design (Webcam Model Face Recognition)

In our web CAPTCHA design, a user activates the camera first, and then our deep learning model extracts features from each frame of the webcam's recorded image. After that, the model evaluates whether to grant or prohibit access depending on the extracted features. Figure 5 shows several experiments we ran for testing our CAPTCHA.

5. Evaluation of the Deep Learning Model

In the process of evaluating a model, there are certain metrics on which the model should be evaluated against about what makes a certain method trustworthy. These metrics are more important when it comes to setting face recognition evaluation practice standards in terms of how models are evaluated.

To maintain the audit's credibility, the benchmarks must be consistent, both in terms of ethical expectations and standards, and the data itself. It may be difficult to compare one year's achievements with the previous year's if ethical standards and performance criteria are not consistent. Audit scheduling is the only part of the procedure that has yet to be standardized. In the absence of a regular audit period, there is no expectation that standards or expectations will be met consistently, and this is a significant problem. Benchmark features, such as data resolution, can be affected by equipment modifications, such as those made to digital cameras. In our research, photo sizes and resolutions ranged from 32x32 to 3072x2048 or even higher among benchmarks. When the amount of pixels is utilized as the direct input to methods such as deep learning, it becomes more difficult to establish which aspects of reported performance measurements are dependent on these other characteristics.

Figure 5: CAPTCHA Testing

In these regards, we are planning to use several standardized metrics to evaluate our model, including accuracy, precision, and recall. All these measurements take into account four critical values: true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

A true positive is an outcome for which the model forecasts the positive class correctly. A true negative, similarly, is an outcome for which the model properly predicts the negative class. On the other hand, a false positive occurs when the model forecasts the positive class inaccurately, meaning that the sample is indeed negative, whereas a false negative is an outcome in which the model forecasts the negative class incorrectly, where the sample is in the positive class.

Accuracy is the ratio of the correctly labeled subjects to the whole pool of subjects. Precision means the percentage of the predicted results which are positive. On the other hand, recall refers to the percentage of total positive results correctly classified by our model. The formulas are shown as follows:

$$Accuracy = (TP + TN)/(TP + TN + FN + FP)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

6. Limitations

Our web CAPTCHA has several drawbacks such as the inability to discriminate between a photograph and a regular person. Our CAPTCHA recognizes the human face in photographs and allows users to proceed (as shown in Figure 6). This might result in serious security vulnerabilities if attackers utilize photographs to circumvent our CAPTCHAs. To address this problem, we intend to embed audio instructions into the web CAPTCHA and require users to follow the instructions (e.g., head adjustments, head rotation right, left, up and down) to pass our CAPTCHA.

7. Conclusion

In this paper, we propose to develop a web CAPTCHA design using a deep learning face detection model in order to enhance the usability for PVIs. In the future, we plan to build a real-time detection model to recognize human by asking the users to move their head in different directions to alleviate any visual load for PVIs.

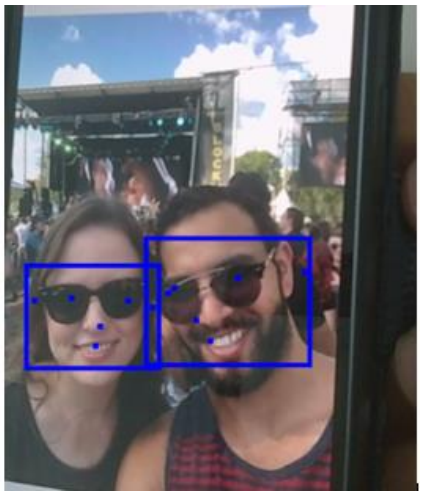

Figure 6: Our model detects the photo as a person and allows passing the CAPTCHA

References:

Abdrabo, D. A., Gaber, T., & Wahied, M. (2016). Assistive Technology Solution for Blind Users Based on Friendsourcing. *In The 1st International Conference on Advanced Intelligent System and Informatics*, November 28-30, 2015, Beni Suef, Egypt (pp. 413-422). Springer, Cham.

Alnfiai, M. (2020). A Novel Design of Audio CAPTCHA for Visually Impaired Users. *International Journal of Communication Networks and Information Security*, 12(2), 168-179.

Bateman, D. R., Brady, E., Wilkerson, D., Yi, E. H., Karanam, Y., & Callahan, C. M. (2017). Comparing crowdsourcing and friendsourcing: a social media-based feasibility study to support Alzheimer disease caregivers. *JMIR research protocols*, 6(4), e6904.

Berton, R., Gaggi, O., Kolasinska, A., Palazzi, C. E., & Quadrio, G. (2020, January). Are CAPTCHAs preventing robotic intrusion or accessibility for impaired users? In 2020 *IEEE 17th Annual Consumer Communications & Networking Conference* (pp. 1-6). IEEE.

Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S. and Yeh, T. (2010, October). Vizwiz: nearly real-time answers to visual questions. *In Proceedings of the 23rd annual ACM symposium on User interface software and technology* (pp. 333-342).

Bock, K., Patel, D., Hughey, G., & Levin, D. (2017). unCAPTCHA: A Low-Resource Defeat of reCAPTCHA's Audio Challenge. *In 11th USENIX Workshop on Offensive Technologies*.

Bigham, J. P., & Cavender, A. C. (2009, April). Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use. *In Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1829-1838).

Bursztein, E., Bethard, S., Fabry, C., Mitchell, J. C., & Jurafsky, D. (2010, May). How good are humans at solving CAPTCHAs? A large-scale evaluation. *In 2010 IEEE symposium on security and privacy* (pp. 399-413). IEEE.

Fanelle, V., Karimi, S., Shah, A., Subramanian, B., & Das, S. (2020). Blind and Human: Exploring More Usable Audio CAPTCHA Designs. *In Sixteenth Symposium on Usable Privacy and Security* (pp. 111-125).

Girshick, R. (2015). Fast r-cnn. *In Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Holman, J., Lazar, J., Feng, J. H., & D'Arcy, J. (2007, October). Developing usable CAPTCHAs for blind users. *In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility* (pp. 245-246).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

Jain, M., Tripathi, R., Bhansali, I., & Kumar, P. (2019, October). Automatic generation and evaluation of usable and secure audio ReCAPTCHA. *In The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 355-366).

Jiang, H., & Learned-Miller, E. (2017, May). Face detection with the faster R-CNN. *In 2017 12th IEEE international conference on automatic face & gesture recognition* (FG 2017) (pp. 650-657). IEEE.

Jiang, N., & Dogan, H. (2015, July). A gesture-based CAPTCHA design supporting mobile devices. *In Proceedings of the 2015 British HCI Conference* (pp. 202-207).

Kaur, K., & Behal, S. (2014). CAPTCHA and its techniques: a review. *International Journal of Computer Science and Information Technologies*, 5(5), 6341-6344.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. *In European conference on computer vision (pp. 21-37)*. Springer, Cham.

Meutzner, H., Gupta, S., & Kolossa, D. (2015, April). Constructing secure audio CAPTCHAs by exploiting differences between humans and machines. *In Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2335-2338).

Noorjahan, M. (2019). A biometric based approach for using CAPTCHA-to enhance accessibility for the visually impaired. Disability and rehabilitation. *Assistive Technology*, 15(2), 153-156.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rzeszotarski, J. M., & Morris, M. R. (2014, April). Estimating the social costs of friendsourcing. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2735-2744).

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3), 122-131.

Sauer, G., Hochheiser, H., Feng, J., & Lazar, J. (2008, July). Towards a universally usable CAPTCHA. *In Proceedings of the 4th Symposium on Usable Privacy and Security* (Vol. 6, p. 1).

Shafiee, M. J., Chywl, B., Li, F., & Wong, A. (2017). Fast YOLO: A fast you only look once system for real-time embedded object detection in video. *arXiv preprint arXiv:1709. 05943.*

Shirali-Shahreza, S., Penn, G., Balakrishnan, R., & Ganjali, Y. (2013, April). Seesay and hearsay CAPTCHA for mobile interaction. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2147-2156).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Solanki, S., Krishnan, G., Sampath, V., & Polakis, J. (2017, November). In (cyber) space bots can hear you speak: Breaking audio CAPTCHAs using ots speech recognition. *In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 69-80).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Von Ahn, L., Blum, M., & Langford, J. (2004). Telling humans and computers apart automatically. *Communications of the ACM*, 47(2), 56-60.

Wu, S., & Adamic, L. A. (2014, April). Visually impaired users on an online social network. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3133-3142).

Zhang, Z., Zhang, Z., Yuan, H., Barbosa, N. M., Das, S., & Wang, Y. (2021). WebAlly: Making Visual Task-based CAPTCHAs Transferable for People with Visual Impairments. *In Seventeenth Symposium on Usable Privacy and Security* (pp. 281-298).