Multimed Tools Appl (2018) 77:12293–12329
DOI 10.1007/s11042-017-4883-7

Exploring the influence of CAPTCHA types to the users response time by statistical analysis

Darko Brodić[1] · Alessia Amelio[2] · Radmila Janković[1]

Received: 28 October 2016 / Revised: 23 May 2017 / Accepted: 29 May 2017 /
Published online: 7 June 2017

Abstract CAPTCHA stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart. It is a test program that solves a given task for preventing the attacks made by automatic programs. If the response to CAPTCHA is correct, then the program classifies the user as a human. This paper introduces a new analysis of the impact of different CAPTCHAs to the Internet user's response time. It overcomes the limitations of the previous approaches in the state-of-the-art. In this sense, different types of CAPTCHAs are presented and described. Furthermore, an experiment is conducted, which is based on two populations of Internet users for text and image-based CAPTCHA types, differentiated by demographic features, such as age, gender, education level and Internet experience. Each user is required to solve the different types of CAPTCHA, and the response time to solve the CAPTCHAs is registered. The obtained results are statistically processed by Mann-Whitney *U* and Pearson's correlation coefficient tests. They analyze 7 different hypotheses which evaluate the response time in dependence of gender, age, education level and Internet experience, for the different CAPTCHA types. It represents an invaluable study in the literature to predict the best use of a given CAPTCHA for specific types of Internet users.

Keywords CAPTCHA · Web · Response time · Usability · Statistical analysis · Internet user

✉ Darko Brodić
dbrodic@tfbor.bg.ac.rs

Alessia Amelio
aamelio@dimes.unical.it

[1] Technical Faculty in Bor, University of Belgrade, Vojske Jugoslavije 12, 19210 Bor, Serbia

[2] DIMES, University of Calabria, Via P. Bucci Cube 44, 87036 Rende (CS), Italy

1 Introduction

1.1 CAPTCHA basics

CAPTCHA, which represents a puzzle program to be solved, is closely related to three main elements: (i) the Turing test, (ii) the Human-Computer Interaction (HCI), and (iii) the Human Interactive Proofs (HIP).

The Turing test represents a pioneer method which determines the capability of the machine to think like a human [46]. In the test, we have three participants. The first one is the examiner which is a human. It asks the questions and judges the answers. Other two participants are the human and the machine, whose task is to answer the questions asked by the examiner. The test has a specific time duration. After that time, the examiner should decide if the answerer is a human or a machine. If the examiner cannot differentiate the answer obtained by both the answerers, then the machine has an Artificial Intelligence (AI) similar to the human. In that case, the machine has "passed" the Turing test, because it answers "just like a human". Essentially, the machine is considered to have AI, if it mimics the human answers to the test. Still, the test has some limitations. They occur in the case when the questions are formulated in Yes/No manner. In such case, the Turing test has no significance. However, if the answers to the questions are expanded by many possibilities, then the Turing test can be of great importance. Still, if the questions are connected to the knowledge based on information sources, such as Google search engine, then the machine, i.e. the computer can outperform the human in the Turing test. At the end, we can conclude that the Turing test did not actually test computers' intelligence only. Instead, it explores whether a computer behaves like a human. Hence, it gives a cross-section of the human and intelligent behavior.

Human-Computer Interaction (HCI) determines the interaction in the communication between humans and computers. According to the cognitive psychology, the humans have a specific way of processing data. A similar case is related to computers, too. Hence, their interaction can be effective if they are accustomed one to another. In this way, a HCI system consists of the following elements: (i) a human commonly referred as a user, and (ii) a computer. Furthermore, a HCI model is divided into five different levels: (i) task level, (ii) semantic level, (iii) syntactic level, (iv) interactive level, and (v) a level of physical devices [40]. Then, the obtained information is processed by: (i) reasoning, (ii) problem solving, (iii) skill acquisition, and (iv) error. These elements have important implications to the HCI interactive design. However, to be effective the HCI should be user centric.

Human Interactive Proofs (HIPs) are used to make a differentiation between the human users and the computer robot programs [8, 21]. They require a specific type of interaction by a user which is difficult to emulate by the robot programs. Bot is an abbreviation of software robot that runs automated tasks over the Internet. HIP is designed to enable the following [12]:

- To differentiate the humans from the computers,
- To differentiate a category of the humans,
- To differentiate a specific human from the category of humans.

In this way, HIP incorporates the test program which is subjected to the human and the computer. As a result, only a specific category of humans can positively solve the test. At the end, the test results should be validated by the computer [15]. From all aforementioned, it is clear that HIP and the Turing test are deeply associated. The Turing test has a task to confirm if the computer has an appropriate AI. On the contrary, if there exists a question to which the computer cannot give the right answer, but humans can, then HIP can be used.

Taking into account all aforementioned, the CAPTCHA is a program created to differentiate humans from bots during the logging to a website [47]. Accordingly, the bot is a program, which tries to emulate the human users. It includes elements of AI as well as ability of automated reasoning. In that sense, CAPTCHA is referred to the Turing test as well as to HIP, but also depends on HCI. However, some differences exist. Unlike the Turing test, the CAPTCHA is controlled by an integrated examiner. Hence, if we rephrase the Turing test, we obtain a CAPTCHA. Accordingly, CAPTCHA is a test program that is used for solving a given type of task, which is more suitable for humans than for bots. If the answer to the CAPTCHA is correct, then the program classifies the user as a human.

The aim of the CAPTCHA is to stop the attacks made by bots. Todays research about CAPTCHA is focusing on the development of test programs which could be easily solved by people and represent a heavy problem to bots. The CAPTCHA's main tasks are the following [7]:

- Prevention of spam on forums and e-mails,
- Prevention of opening a large number of orders on websites that offer free services, such as Gmail, Yahoo, etc.,
- Protection of the user accounts from the attacks that extract the user's passwords,
- Validation of the online surveys by answering the questionnaire that determines the differences between humans and bots,
- Protection of online pools.

In order to incorporate a high level of security, the CAPTCHA has to meet the following requirements:

- The solution to the CAPTCHA must not be conditional. It means that it shouldn't depend on the user's language and/or age. This leads to the conclusion that it should be intuitive as much as possible,
- The solving of the CAPTCHA should be easy for the humans and hard for the bots in order to differentiate humans from bots. Also, it should be completed by humans in no longer than 30 seconds [43], with a success rate of at least 90% [8],
- The creating of the CAPTCHA must not disturb the user privacy. It further means that it has not to be linked to the user.

1.2 CAPTCHA types

The first CAPTCHA was designed by Broder's team in 1997 for Altavista, to prevent automatic adding URL to a database of a web browser [28]. The development of different types of CAPTCHA has been widespread in recent years. According to its elements, the CAPTCHA can be divided into the following categories [17]:

- Text-based CAPTCHA,
- Image-based CAPTCHA,
- Audio-based CAPTCHA,
- Video-based CAPTCHA,
- other CAPTCHA types.

Text-based CAPTCHA is the most widespread CAPTCHA type. It asks the user to decrypt a text which is usually distorted in different ways. Unfortunately, this type of CAPTCHA can be successfully attacked by bots due to the existence of good decoders. Figure 1 shows an example of text-based CAPTCHA.

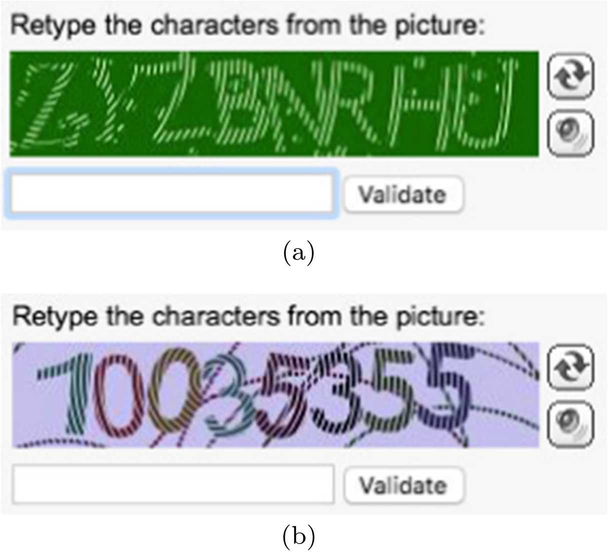

Retype the characters from the picture:
Validate
(a)
Retype the characters from the picture:
Validate
(b)

Fig. 1 An example of text-based CAPTCHA: (**a**) with text only, and (**b**) with numbers only

Image-based CAPTCHA is usually considered as the most advanced and safest type of CAPTCHA. It requires the users to find out and to point to a desired image from an image list. Because it is based on image details, it represents for a bot an extremely difficult task to be solved. Figure 2 shows an example of image-based CAPTCHA.

The FaceDCAPTCHA has been introduced as an extension to the image-based CAPTCHA [17]. It is a CAPTCHA that incorporates elements of face detection. It is one of the newest CAPTCHA types that includes a high level of security. It exploits a research about the human brain, which is very effective in the process of natural face segmentation in spite of used complex backgrounds. Figure 3 shows an example of FaceDCAPTCHA.

Audio-based CAPTCHA is usually a text-based CAPTCHA which includes an addition that is especially designed for the people with disabilities (typically blind users). This addition represents an element whose purpose is an audio reproduction of characters that the user has to input. Unfortunately, this type of CAPTCHA is typically attacked in approximately 70% of cases, due to rapid development of speech analysis and recognition algorithms. Figure 4 shows an example of audio-based CAPTCHA.

Video-based CAPTCHA is an extension of text-based CAPTCHA in the video format. It represents a video which includes a passing text with a specific different color compared to the video. The user's task is to recognize the passing text and type it in the given text box within CAPTCHA. Still, the modern Optical Character Recognition (OCR) programs challenge this task in many cases. Figure 5 shows an example of video-based CAPTCHA.

Other types of CAPTCHA represent those CAPTCHAs that cannot be part of the previous categorization. Figure 6 shows such types of CAPTCHA.

Basically, the CAPTCHA is taking advantage of human capacity in the printed or handwritten text read, using speech, image and facial recognition. The articles about CAPTCHA have researched mainly the safety and security standpoint ignoring the difficulties of users to solve the task. Although the CAPTCHA protects user accounts and passwords, its solution is often a stationary obstacle not only for bots, but also for humans.

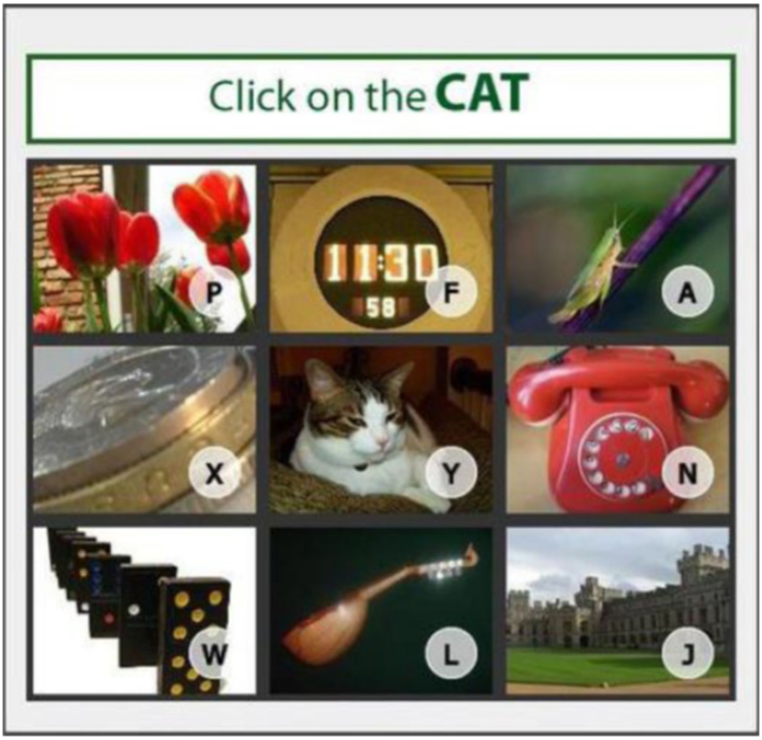


Fig. 2 An example of image-based CAPTCHA

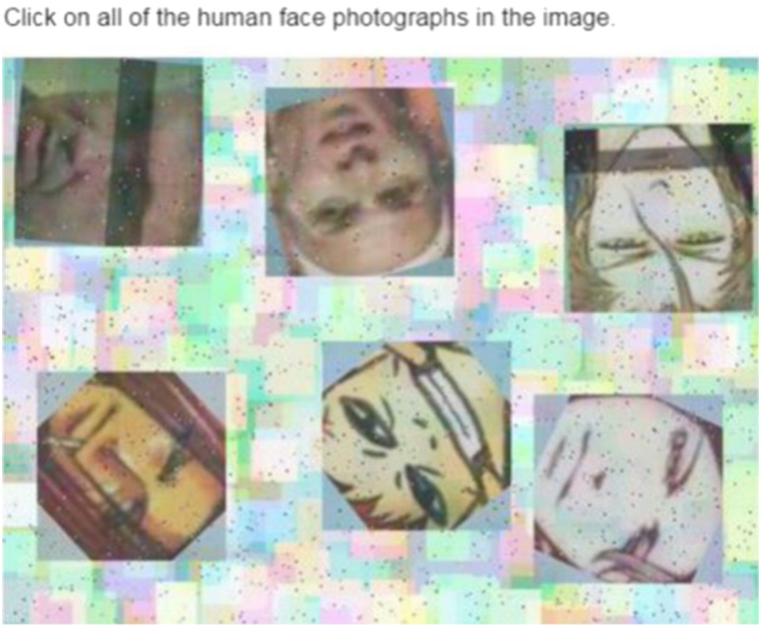


Fig. 3 An example of FaceDCAPTCHA

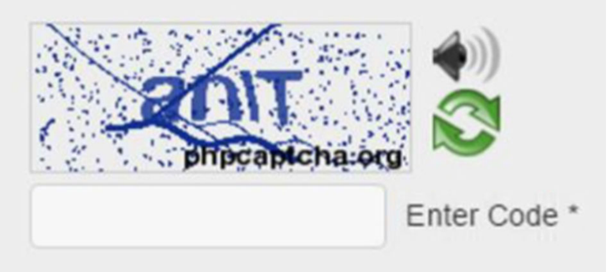


Fig. 4 An example of audio-based CAPTCHA

The paper is organized as follows. Section 2 describes the related works. Section 3 introduces the experiment. Section 4 defines the statistical setting used for the experiment. Section 5 presents the experimental results and the comparison results, and discusses them. Finally, Section 6 draws conclusions and outlines future work directions.

2 Related works

Statistical analysis has the aim to collect, explore, and present the data in order to discover trends and patterns. Analysis is started with a statistical model or a statistical population to be studied. In recent times, different statistical analysis tools have been proposed in multiple domains.

In the domain of pattern recognition, an atomic generative probabilistic model for complex activity recognition has been introduced, preserving the consistency of its atomic actions and their temporal dependencies [33]. In [30] the activity recognition has been performed on sensor data representing recorded human activities. The proposed model

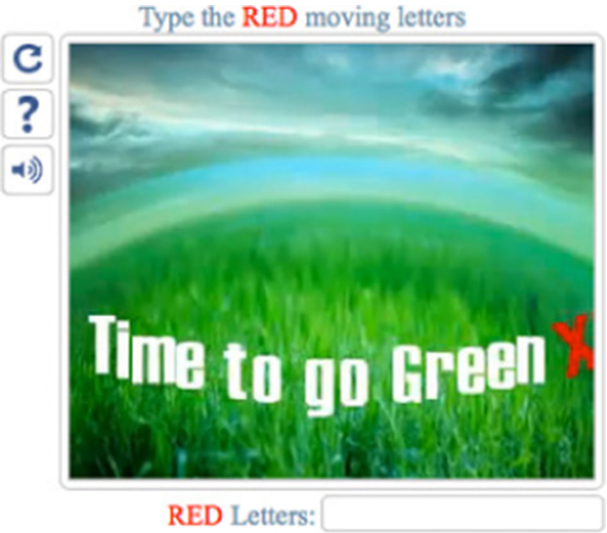
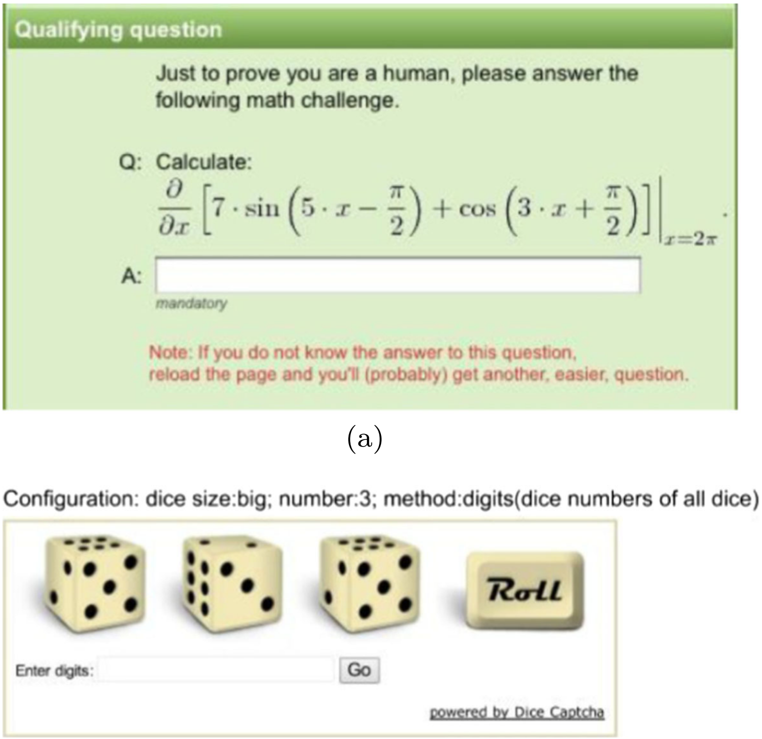


Fig. 5 An example of video-based CAPTCHA



Qualifying question

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

Q: Calculate:

$$\frac{\partial}{\partial x}\left[7\cdot\sin\left(5\cdot x-\frac{\pi}{2}\right)+\cos\left(3\cdot x+\frac{\pi}{2}\right)\right]\Bigg|_{x=2\pi}.$$

A:

mandatory

Note: If you do not know the answer to this question, reload the page and you'll (probably) get another, easier, question.

(a)

Configuration: dice size:big; number:3; method:digits(dice numbers of all dice)

Enter digits:

(b)

Fig. 6 Examples of other types of CAPTCHA: (**a**) QRBGS CAPTCHA [18], (**b**) Dice CAPTCHA [11]

consists of two parts, including a temporal pattern mining algorithm for tracking the temporal dependence between the actions, and an adaptive multi-task learning algorithm for capturing the relationships between the activities represented as tasks, and the characteristics which are specific or shared between the tasks. The obtained temporal patterns are used to characterize the activities in order to automatically recognize them [32]. Also, a multi-source learning framework for career path modeling and prediction has been proposed [31]. It integrates multiple social network sources of multiple user views by penalizing the differences among the sources and the abrupt changes when the user's tasks are associated with neighboring time steps. Ref. [9] has proposed a probabilistic framework for human motion tracking by fusing the low-level and high-level approaches for overcoming their limitations. The two approaches proceed in parallel and each time the system merges their contributions by a probabilistic method which complements the advantages of both. In order to select the trackers based on the motion types, the algorithm for trackers sampling is based on low and high-dimensional trackers [29]. The two methods which are used respectively for low-dimensional and high-dimensional trackers sampling are combined to adaptively sampling the trackers based on the motion type. In [37] an unsupervised method for physical activity recognition from accelerometer data of the smartphones has been introduced. The features are extracted from the raw data deriving from the accelerometer, on which the MCODE classification method is employed for the recognition of the physical activity.

In the domain of information retrieval, Ref. [34] has introduced a video retrieval method using spatio-temporal features for matching to a given query. It is refined by a learning method for capturing child-adult interactive behaviors according to a collection of home videos. Ref. [36] has presented a spatio-temporal, multi-task and multi-view learning framework for predicting the water quality by integrating heterogeneous urban data from multiple sources. In [35] the water quality has been forecasted by using quality and hydraulic data deriving from monitor stations, pipe networks, meteorology and point of interests. After identifying the factors influencing the water quality, a multi-task and multi-view learning framework has been presented, integrating the data from multiple domains into a unified learning approach. In [41] a new model for automatically predicting the political ideology of users from social media posts has been introduced. The approach has the aim to identify politically oriented groups of users, as well as moderate or neutral ones, and to predict the political ideology of unknown users.

The related works on CAPTCHA often employ statistical approaches in different aspects. They can be divided according to their properties. Hence, they are split into the following property areas [1]: (i) Security, (ii) Practicality, and (iii) Usability.

Usually, the main concerns of creating and using CAPTCHA are in the domain of its security. Hence, the majority of related works have addressed that problem. It represents a central problem of the CAPTCHA, but it is not the only one that has a great importance. Many researchers have proposed to improve the CAPTCHAs in terms of their security. The text-based CAPTCHA has been very vulnerable to the bots' attacks. Hence, many techniques have been proposed to improve it. Ref. [44] has proposed a few techniques, which have been introduced using the handwritten text in the CAPTCHA. They are: (i) doubling of text, (ii) different orientation, (iii) mirroring of each word in the text, and (iv) overlapping text with some curve text lines. Also, a model of scattering has been proposed to improve the text-based CAPTCHA [2]. The latest development of the text-based CAPTCHA has proposed the use of a CAPTCHA technique that is based on chaotic logistic map and projective S-box [22]. An improvement of the image-based CAPTCHA security has been proposed in [23, 24, 27]. Ref. [23] has suggested the extraction of image fragments as well as the change of the image orientation. Ref. [27] has proposed a complex, but efficient method called ARTiFACIAL for image-based facial CAPTCHA, which transforms the facial elements in the image. Also, an interesting approach to create an image-based CAPTCHA has been introduced in [24]. It uses the so-called ageCAPTCHA which extracts images from a public image database. The image is then cropped to a specific size in order to obtain the rectangle that contains the face of indeterminate age.

Practicality is very important from the programmer's point of view. Unfortunately, this issue has no connection with the users of the CAPTCHA. However, some of the proposed solutions provide a detailed way how to create the CAPTCHA [22, 23, 27].

Still, the usability represents one of the important problems related to the using of the CAPTCHA. Hence, it especially concerns the users of the CAPTCHA. Unfortunately, this problem is rarely observed and tested. Ref. [40] has proposed a user friendly image-based CAPTCHA scheme based on the human appearance characteristics. Still, it received a human success rate of 62% and 83%, which is not satisfactory. Ref. [38] has conducted experiments on a small population of twenty Internet users to obtain their response to different types of CAPTCHA. The authors obtained statistically significant differences between CAPTCHAs according to all dependent variables, but not in task's completion time. The usability of the text-based CAPTCHA has been researched in ref. [26]. Although it explores different age groups, the population is too small, i.e. only 24 samples. Also, this study uses only one demographic factor (user's age) for the analysis of the CAPTCHA's solution time. Finally, only the text-based CAPTCHA is under consideration. A wider population of 107 participants has been used for testing their ability to solve the text and image-based CAPTCHAs [3]. However, this study has the limitation of using only university students with an age between 17 to 26 years. A further study has introduced a new type of CAPTCHA called AgeCAPTCHA [24]. It incorporates publicly available images. After that, it extracts faces of a certain age by means of face detection and age estimation algorithms. At the end, the faces are cropped, which reduces the attacks. The CAPTCHA is tested on 267 participants. Although this CAPTCHA brings a higher security level, it is characterized by longer solution time compared to the text-based CAPTCHA and Microsoft's Asirra CAPTCHA. One of the most recent studies has used two different experiments [4]. The first experiment includes 131 participants, which are explored according to their cognitive preferences between verbal and image ones. The second experiment includes 125 participants, which examine the users' speed of successfully solving the text-based and image-based CAPTCHAs. Although the obtained results in this study concerning user's individual differences are quite interesting, they cannot be applied to a wider population because the tested population sample is only a student population. Accordingly, the result of this study lacks a level of generality. Finally, an advanced statistical analysis has been introduced in [5, 6], using the association rules for evaluating the dependence of the response time to text and image-based CAPTCHA from the co-occurrence of some demographic factors of the users. The main limitation of this method is that discretization of the response time and the other numerical features is required, which might compromise the efficacy of the evaluation.

To summarize, the main limitations of the current approaches for analysis of the CAPTCHA's usability in the state-of-the-art are the following:

- The small sample population, i.e. the low number of users involved in the experiment (i.e. the statistical significance of the tested population),
- The limited number of demographic factors which are tested to evaluate the solution time to the CAPTCHA (e.g. the age),
- The lack of generality of the analysis (e.g. only university students of age between 17 and 26 years),
- The reduced number of considered CAPTCHA types (e.g. mainly text-based CAPTCHA).
- the discretization of the numerical variables involved in the analysis.

In this paper, we present a new study of the CAPTCHA's usability problem. It represents the complexity of the CAPTCHA's tasks from the user's viewpoint, i.e. user-centric view of the problem. The proposed study successfully overcomes the limitations of the previous approaches in the state-of-the-art:

- It uses a larger population of at least 100 Internet users,
- It proposes a higher number of demographic factors to be tested, which are: (i) age, (ii) gender, (iii) education level, and (iv) Internet experience level,
- It presents a more general analysis, including a variegated population of students, employees, clerks, teachers and engineers of age between 18 and 52 years,
- It explores 9 CAPTCHA types, including different types of text and image-based CAPTCHA,
- It does not require the feature discretization.

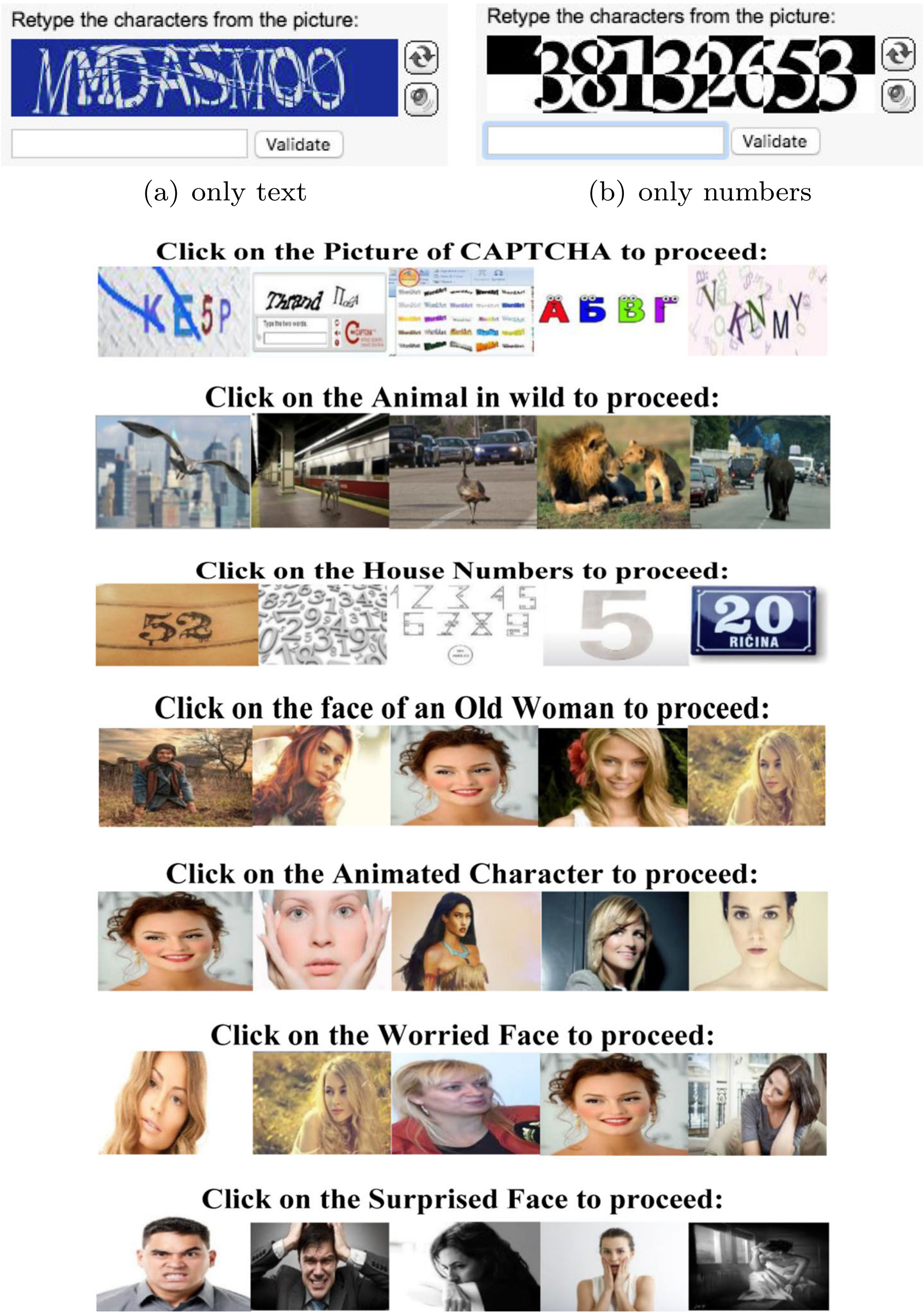
All aforementioned will enable the ability to conduct the experiment of CAPTCHA whose primary focus is the exploration of user's ability to resolve certain types of CAPTCHA in a given preset time (typically 30 or 45 seconds). The aim of our research is to identify the following: (i) the user's response time to solve different types of CAPTCHA, and (ii) the suitability of different types of CAPTCHA to a particular group of Internet users. Accordingly, we create 7 hypotheses, which should be (dis)proved using statistical tools. At the end, the presented results are used to suggest the "good" CAPTCHA for certain types of Internet users. As a final result we differentiate the CAPTCHA's suitability to a certain group of Internet users that operate with laptop computers. Hence, it will be a cornerstone for further creation of combined CAPTCHA's elements and types which are almost impossible to be solved in a predefined time by bots, and possible to be solved by humans. Such research will contribute to perceive the elements of usability, especially for the programmers of CAPTCHA, which will further improve their elements of practicality. Thus, the obtained results can be invaluable to the scientific and professional public including the users of laptop computers and the designers and programmers of CAPTCHA. At the end, it will lead to the creation of combined types of CAPTCHA which are suitable for humans unlike the bots.

3 Experiment

The experiment of the CAPTCHA includes two different experiments with text-based and image-based CAPTCHAs. The first experiment, based on 2 different text-based CAPTCHAs, is tested on a community of 102 Internet users. Furthermore, the second experiment which includes 7 different image-based CAPTCHAs is tested on a population of 100 Internet users. All Internet users are volunteer students, employees, clerks, teachers and engineers, who signed an online consent form before starting the experiment. By this form, they gave their consent to anonymously provide and use their data only for research and study purposes. Each of them is required to solve different types of CAPTCHA on the same laptop computer, and the response time to find the solution to the CAPTCHAs is registered. All users are divided according to four demographic factors: (i) gender (male and female), (ii) age (below and above 35 years old, hence from 18 to 35 years and from 36 to 52 years), (iii) education level (higher and secondary), and (iv) years of Internet use (represented in a number of years). The tests are conducted on 9 CAPTCHA types, including two text-based CAPTCHAs with: (i) text only, and (ii) numbers only, and seven image-based CAPTCHAs including: (i) animals in the wild, (ii) home numbers, (iii) face of an old woman, (iv) animated face, (v) worried face, (vi) surprised face, and (vii) the picture of the CAPTCHA. Figure 7 shows the different types of CAPTCHA used in the experiment. Then, the obtained results are statistically processed in order to present the advantages and disadvantages of specific types of CAPTCHA.

Taking into account the examined population of data, we establish 7 hypotheses which will be (dis)proved using statistical tools. These hypotheses have been carefully formulated for investigating the main dependences between demographic features and response time to CAPTCHA [5, 6]. They are the following:

1. The group of Internet users with higher education level will have a faster response time in solving the text-based CAPTCHA (Hypothesis 1),
2. The response time of the image-based CAPTCHA is the longest one for Internet users with secondary education level (Hypothesis 2),
3. A faster response time in solving the CAPTCHA is presumable for image-based compared to text-based CAPTCHA in the case of male users (Hypothesis 3),


Retype the characters from the picture:
MMDASMOO
Validate
Retype the characters from the picture:
38132653
Validate
Click on the Picture of CAPTCHA to proceed:
KE5P
АБВГ
Click on the Animal in wild to proceed:
Click on the House Numbers to proceed:
52
5
20
RIČINA
Click on the face of an Old Woman to proceed:
Click on the Animated Character to proceed:
Click on the Worried Face to proceed:
Click on the Surprised Face to proceed:

(a) only text

(b) only numbers

(c) image

Fig. 7 The different types of: (**a**-**b**) text, and (**c**) image-based CAPTCHAs used in the experiment

4. There exists a statistically significant difference between gender groups in solving the text-based CAPTCHA with only numbers and the image-based CAPTCHA with picture of the CAPTCHA (Hypothesis 4) [5],
5. There exists a statistically significant difference between age groups in solving the CAPTCHA (Hypothesis 5),
6. Subjects who use the Internet less time will have a longer response time in solving the CAPTCHA (Hypothesis 6),
7. There exists a statistically significant difference between gender groups in the average response time to solve the CAPTCHA (Hypothesis 7),

which have to be proved or disproved, i.e. confirmed or rejected. The explanation of this process is given below.

4 Evaluation measures

Hypotheses are used to test the validity of a claim that is made on a data population. Any statistical test is closely related to previously established hypotheses that should be proved or disproved by such a tool. Accordingly, any researcher tries to define any statistical hypothesis linked with the research data. We can see a hypothesis as a clear assumption about the tested data and its characteristics. Using statistical tools, the initial hypothesis is proved or disproved. The initial research hypothesis claims something about the research data. It is called hypothesis H_1. Usually, the researcher believes that this hypothesis is valid according to some common knowledge. In the contrast, the null-hypothesis, i.e. H_0 is a hypothesis which denies the claims given in the research hypothesis H_1. We can say that this approach related to the hypotheses H_0 and H_1 is simplified, but quite clear one. Thinking more deeply, the null-hypothesis typically refers to the common view of something, while the research hypothesis is the claim about the cause of a phenomenon that researcher believes is true. Taking into account all aforementioned, the statistical test proves/disproves either the research hypothesis H_1 or the null-hypothesis H_0 or vice versa.

4.1 Mann-Whitney *U* test

We use the Mann-Whitney U test to prove or disprove our hypotheses. It is a non-parametric test used to compare differences between two independent groups N_1 and N_2. The dependent variable is usually ordinal or continuous. Still, some pre-assumptions should be met in order to obtain a valid result [39]:

- Assumption 1: the dependent variable (output) should be ordinal or continuous,
- Assumption 2: the independent variable (input) should be composed of two categorical, independent groups,
- Assumption 3: observations from both groups N_1 and N_2 should be independent; this means that no correlation exists between observations in the same group or between groups,
- Assumption 4: the independent variables should not be normally distributed.

Basically, the Mann-Whitney U test analyses the differences in the ranked positions of scores in different groups by computing the mean and total ranks. This test relies on the scores which are ranked from the lowest to the highest ones. Accordingly, the group with the lowest mean rank is the group with the greatest number of lower scores within it. Also, the group with the highest mean rank has a greater number of high scores within it.

The null-hypothesis H_0 of this test is that the two groups have the same distribution of scores. On the contrary, the research hypothesis H_1 is that the distributions of the two groups are different in some aspect (e.g. center, spread and/or shape).

The most important value of this test is the significance p-value. It is a function of the results of the observed sample relative to a statistical model, which measures how extreme the observation is. Basically, it evaluates how well the research data support the arguments that the claim in the null-hypothesis H_0 is true. The p-value should receive a value between 0 and 1. The smaller is the value that p receives, the larger is the statistical significance. Basically, it evaluates if the research hypothesis H_1 under consideration can or cannot adequately explain the observation. The rejection or confirmation of H_1 depends on the arbitrarily pre-defined threshold value α, which is referred to as the level of significance.

The significance p-value according to its values can be interpreted as follows:

- $p < \alpha$ shows a strong evidence against the null-hypothesis H_0. As a consequence, H_0 is disproved (H_1 is proved).
- $p > \alpha$ shows a weak evidence against the null-hypothesis H_0. As a consequence, H_0 is proved (H_1 is disproved).
- p around the boundary area of α is considered to be marginal, which means that it can go either way. Hence, the conclusions can be drawn by its own.

It follows the statistical measure U which is calculated using the following equation [14]:

$$U = n_1 n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1, \tag{1}$$

where U is the result of the Mann-Whitney U test, n_1 is the size of the first independent group, n_2 is the size of the second independent group, and R_1 represents the sum of ranks of the first group. The evaluation is based on the obtained U value. If this value is higher than the critical U value, then the two groups will have the same distribution of scores (H_0 proved and H_1 disproved), otherwise if U value is lower than the critical one, then the two distributions will be different (H_0 disproved and H_1 proved). Critical U values are tabulated for small groups of size ≤ 20. In the case of larger groups of size > 20, the U value approaches a normal distribution by a Z-test as follows:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}, \tag{2}$$

where U is the value obtained by the test. In this case, the evaluation is based on the Z value. If the absolute value of Z is < 1.96, then the two groups will have the same distribution of scores (H_0 proved and H_1 disproved), otherwise if the absolute value of Z is > 1.96, then the two distributions will be dissimilar in some way (H_0 disproved and H_1 proved) [42, 45].

In order to (dis)prove our hypothesis, firstly we evaluate if there is a statistically significant difference between the groups w.r.t. (with respect to) the dependent variable. In particular, we observe the asymptotic 2-tailed p-value (*Asymp. Sig. 2-tailed*) which is computed by using an approximation of the p-value's true distribution. It represents a better choice than the exact p-value for groups of larger size [13]. If this p-value $> \alpha$, our hypothesis is disproved. Otherwise, if this p-value $< \alpha$, we consider the Z value (because in our case the sample size of the groups is always > 20). If the absolute Z value < 1.96, then the two groups will have a similar proportion of high and low values of the dependent variable.

Hence, we will disprove our hypothesis. Otherwise, if the absolute Z value > 1.96, we will consider the mean rank values showing the difference level between the groups (in fact, the distributions of the two groups should not have the same shape). In particular, the group corresponding to the highest mean rank will be the group with higher values of the dependent variable. On the contrary, the group with the lowest mean rank will correspond to the group with lower values of the dependent variable. It will provide the correct information for (dis)proving our hypothesis. It is worth noting that the p and Z values can also be evaluated at the same time, because it will determine the same result. Figure 8 shows the flow diagram of the algorithm ($\alpha = 0.05$).

4.2 Pearson's correlation coefficient test

If the Man-Whitney U test cannot be used for some reason (e.g. the evaluation of the hypothesis requires the violation of at least one assumption of the test), the Pearson's correlation coefficient test is employed.

The null hypothesis H_0 of this test is that input (independent variable) and output (dependent variable) not correlated. On the contrary, the research hypothesis H_1 is that input and output are correlated.

The result of this test is considered as statistically significant only if the p-value receives values lower than the significance level α [10]. The p-value is defined as the probability of obtaining a result equal to or more extreme than what was actually observed, when the null-hypothesis H_0 is true [19]. A low p-value below the significance threshold α suggests that the sample provides enough evidence for rejecting the null-hypothesis H_0 for the entire population.

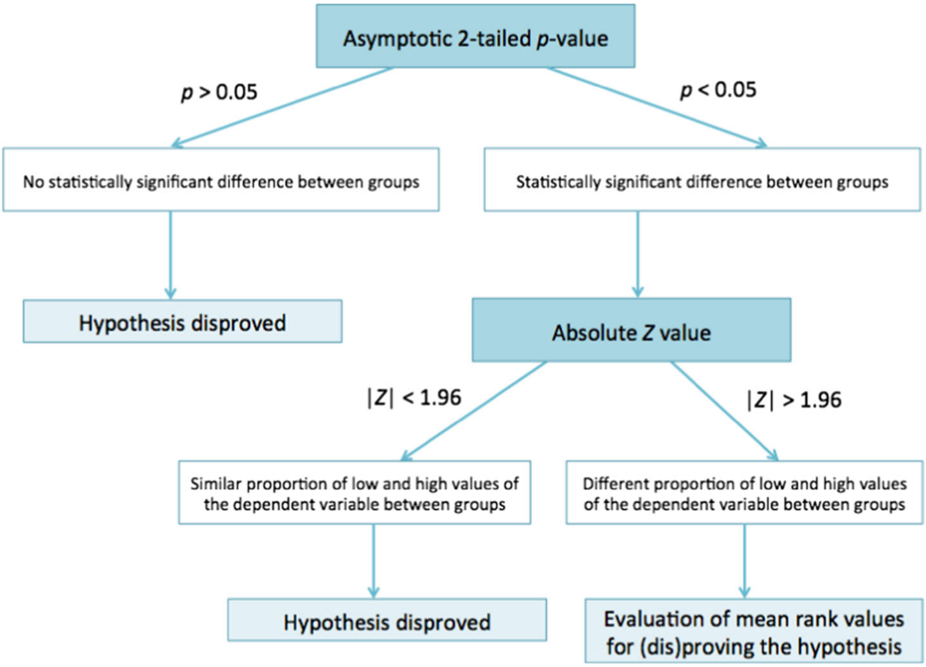

Asymptotic 2-tailed p-value
p > 0.05
p < 0.05
No statistically significant difference between groups
Statistically significant difference between groups
Hypothesis disproved
Absolute Z value
|Z| < 1.96
|Z| > 1.96
Similar proportion of low and high values of the dependent variable between groups
Different proportion of low and high values of the dependent variable between groups
Hypothesis disproved
Evaluation of mean rank values for (dis)proving the hypothesis

Fig. 8 Flow diagram of the algorithm using the Man-Whitney U test

It follows the Pearson's correlation coefficient r which quantifies the correlation level between output and input variables. Hence, it measures the linear dependence (correlation) between an output and an input variable. It is calculated as:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}. \tag{3}$$

The Pearson's correlation coefficient r expresses the strength and direction of a linear relationship between two variables x given as input and y given as output, where x_i is the value of the variable x for the instance i, $\overline{x}$ is the mean of the x values, y_i is the value of the variable y for the instance i, and $\overline{y}$ is the mean of the y values. It can take values between -1 and +1. If the value is positive, then the correlation is direct or positive. On the contrary, if the value is negative, then the relationship is inverse or negative. The larger is the value of the Pearson's coefficient, the stronger is the linear relationship between the variables. A coefficient of zero indicates no linear relationship at all, so if one variable changes, the other doesn't change at all. Typically, it can be interpreted according to its value as: (i) a perfect negative linear relationship, when r = -1, (ii) a strong negative linear relationship, when r = -0.70, (iii) a moderate negative linear relationship, when r = -0.50, (iv) a weak negative linear relationship, when r = -0.30, (v) no linear relationship, when r = 0, (vi) a weak positive linear relationship, when r = +0.30, (vii) a moderate positive linear relationship, when r = +0.50, (viii) a strong positive linear relationship, when r = +0.70, (ix) a perfect positive linear relationship, when r = +1.

In order to (dis)prove our hypothesis, the Pearson's correlation coefficient test is evaluated by measuring: (i) the exact p-value (*Sign. 2-tailed*), and (ii) the r coefficient value. Firstly, we evaluate if there is a statistically significant difference between the groups (i.e. exact p-value $< \alpha$). If the p-value $> \alpha$, our hypothesis is disproved. Otherwise, if the p-value $< \alpha$, r is evaluated between the input (independent variable) and the output (dependent variable) for quantifying the strength or weakness of their correlation according to this value. Hence, our hypothesis will be (dis)proved according to the r value. Figure 9 shows an overview of the algorithm ($\alpha = 0.05$).

5 Results and discussion

The statistical results have been generated by employing SPSS Statistics program version 20, on a laptop computer with Dual-Core CPU at 1.7 GHz, 8 GB RAM, and Windows 10 operating system, for the analysis of each hypothesis [20]. In particular, the Mann-Whitney U test and the Pearson's correlation coefficient test with $\alpha = 5\%$ of significance level have been used for the evaluation. It indicates a chance of only 5% of observing our samples, considering that the null (opposite) hypothesis H_0 is true. Hence, the results corresponding to observed p-values below 0.05 are considered as statistically significant.

The Mann-Whitney U test is based on the four assumptions specified in Section 4.1 [39]. In all our hypotheses, the dependent variable is the response time to solve the text and/or image-based CAPTCHAs. Because it is a continuous variable, the first assumption is respected. Furthermore, all our hypotheses, except the sixth one, have an independent variable composed of two categorical groups. In particular, the first and second hypotheses require the education level as an independent variable, composed of two groups of users respectively with higher and secondary education. The fourth and seventh hypotheses require the gender as independent variable, dividing the Internet users into male and

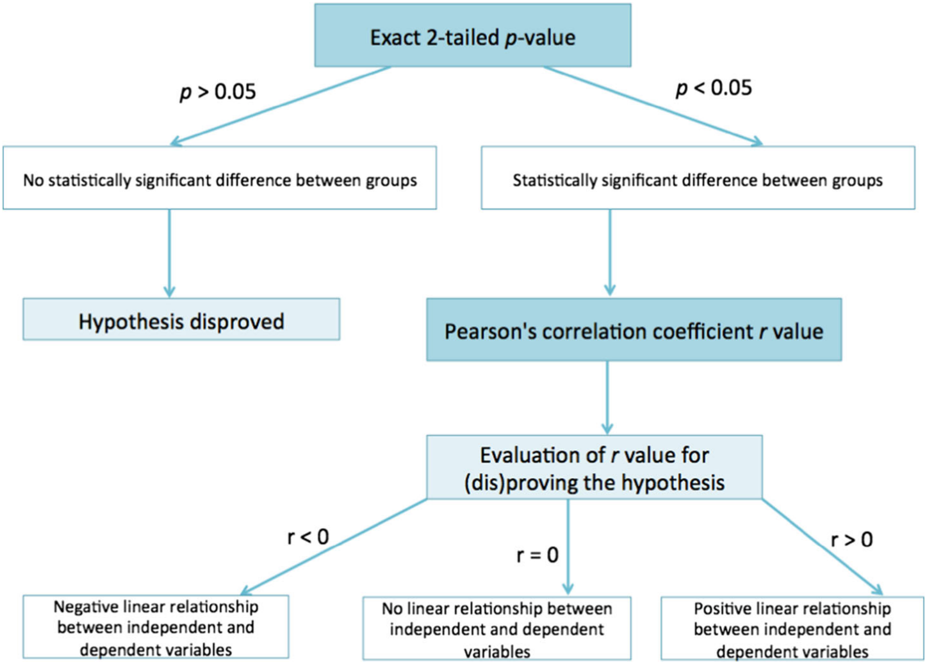

Exact 2-tailed p-value
p > 0.05
p < 0.05
No statistically significant difference between groups
Statistically significant difference between groups
Hypothesis disproved
Pearson's correlation coefficient r value
Evaluation of r value for (dis)proving the hypothesis
r < 0
r = 0
r > 0
Negative linear relationship between independent and dependent variables
No linear relationship between independent and dependent variables
Positive linear relationship between independent and dependent variables

Fig. 9 Overview of the algorithm using the Pearson's correlation coefficient test

female. Finally, the fifth one considers the age as independent variable, dividing the users into below and above 35 years old. Because the sixth hypothesis violates the second assumption (Internet experience is the independent variable, determining up to 9 groups of users of different experience), the Mann-Whitney *U* test is substituted by the Pearson's correlation coefficient test. Again, the third assumption is respected, too: Internet users are different for each group, and there is no participation of a user in more than one group. It is valid for the two groups of Internet users in both the text and image-based CAPTCHA: (i) male and female, (ii) higher and secondary education, and (iii) below and above 35 years. It is also valid for users with a different Internet experience level. Finally, the fourth assumption has been verified for each independent variable required by each hypothesis: (i) gender, (ii) age, (iii) education level and (iv) Internet experience. We observed that any of the used independent variables is normally distributed w.r.t. the response time to solve the text and image-based CAPTCHAs. Furthermore, the distributions of the two groups of the independent variables have been separately analyzed. It confirmed that the two distributions have a different shape. Consequently, in the Mann-Whitney *U* test, the comparison of the mean ranks will be performed.

Table 1 shows the results obtained by the Mann-Whitney *U* test for the first hypothesis. It is based on the assumption that Internet users with higher education level will have a faster response time to solve the text-based CAPTCHA. Still, the independent variable is the education level, composed of the two groups of users with higher and secondary education. The dependent variable is the response time to solve the text-based CAPTCHA with only text and numbers.

The users with higher education level are 52, while the users with secondary education level are 50, for both the text and numbers CAPTCHA. Firstly, we notice that there is

Table 1 Hypothesis 1: Results of the Mann-Whitney *U* test for the text-based CAPTCHA

CAPTCHA type	Education	*n*	Mean Rank	Z	Asymp. Sig. (2-tailed)
Text	Higher	52	43.320	−2.848	**0.004**
	Secondary	50	60.010		
Numbers	Higher	52	41.350	−3.534	**0.000**
	Secondary	50	62.060		

Statistically significant results are marked in bold. The size of each group is represented by *n*

a statistically significant difference (p-value < 0.05) in the values of the response time between the groups. Also in both cases the absolute value of Z is higher than 1.96, which implies that the distribution of the values of the response time is different between the groups. In particular, we observe that the group with the highest mean rank is the secondary education one. In fact, its mean rank has an increment of around 17 w.r.t. higher education for the text (from 43.320 to 60.010), and of around 21 w.r.t. higher education for the numbers (from 41.350 to 62.060). Hence, we can conclude that the response time to solve the text-based CAPTCHA with only text and numbers is statistically higher for the group of users with a secondary education rather than for the group of users with a higher education. In this way, the first hypothesis is proved.

Table 2 shows the results of the Mann-Whitney *U* test for the second hypothesis. It assumes that the response time to solve the image-based types of CAPTCHA is longer for the group of Internet users with a secondary education level. In this case, the image-based types of CAPTCHA are under consideration. Hence, the independent variable is the education level, considering the two groups of Internet users with higher and secondary

Table 2 Hypothesis 2: Results of the Mann-Whitney *U* test for the image-based CAPTCHA

CAPTCHA type	Education	*n*	Mean Rank	Z	Asymp. Sig. (2-tailed)
Animals in the wild	Higher	52	48.520	−711.000	0.477
	Secondary	48	52.650		
Home numbers	Higher	52	39.920	−3.795	**0.000**
	Secondary	48	61.960		
Face of an old woman	Higher	52	45.050	−1.956	0.050
	Secondary	48	56.410		
Animated face	Higher	52	50.360	2618.500	0.959
	Secondary	48	50.660		
Worried face	Higher	52	58.130	−2.736	**0.006**
	Secondary	48	42.240		
Surprised face	Higher	52	53.680	−1.142	0.254
	Secondary	48	47.050		
Picture of the CAPTCHA	Higher	52	35.080	−5.533	**0.000**
	Secondary	48	67.210		

Statistically significant results are marked in bold. The size of each group is represented by *n* education, who solved the image-based types of CAPTCHA. Accordingly, the dependent variable is the response time to solve the image-based types of CAPTCHA.

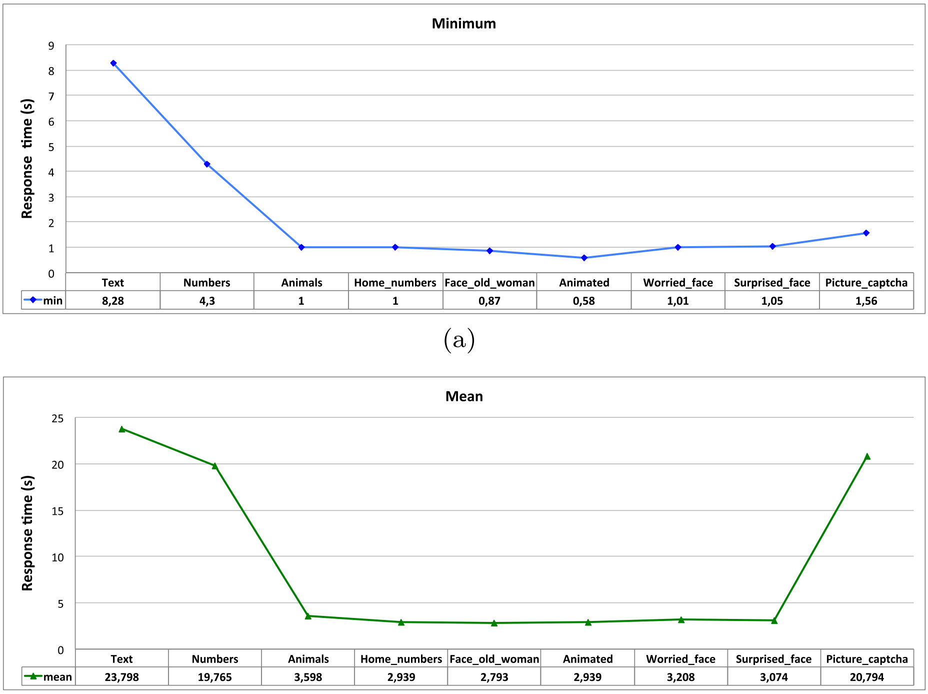
In this case, the 52 users are with a higher education level, while the 48 users are with a secondary education level. We can observe that (i) home numbers, (ii) worried face, and (iii) picture of the CAPTCHA show a statistically significant difference in the values of the response time between the groups (p-values < 0.05). Because all the image-based CAPTCHAs have no p-value < 0.05, this hypothesis is not proved. However, we can consider each image-based CAPTCHA as a separate element due to its differences. Hence, in the other types of CAPTCHA having p-values ≥ 0.05, there is no real evidence that the response time is different between the groups of users with a higher and secondary education. In particular, we notice that the response time is statistically higher for the group of Internet users with a secondary education level, in the case of home numbers and picture of the CAPTCHA. A difference between the groups is visible from the absolute value of Z, which is always above 1.96, and from the mean rank, increasing of around 22 from higher to secondary education in the case of home numbers (from 39.920 to 61.960), and of around 32 from higher to secondary education in the case of the picture of the CAPTCHA (from 35.080 to 67.210). Hence, for these two types of image-based CAPTCHAs, this part of hypothesis is proved. Also, in the worried face, the absolute value of Z is higher than 1.96, indicating the presence of some differences in the distribution of the response time values between the groups. However, this corresponds to a mean rank which is lower for the secondary than for the higher education level, with a decrement of around 16 (from 58.130 to 42.240). Hence, for the worried face, this part of the second hypothesis is disproved.

To validate the third hypothesis, we compute two statistical measures, i.e. the mean and minimum, of the response time to solve the text and image-based types of CAPTCHA, when the Internet users are divided into male and female groups. Table 3 shows mean and minimum response time of the text and image-based CAPTCHAs for male and female users. Figure 10 graphically illustrates these results. In fact, the third hypothesis assumes that the group of male users determines a faster response time to solve the image rather than the text-based types of CAPTCHA. In Table 3 we can observe that the male group is composed of 50 users who provided a solution to the text-based CAPTCHA, and of 49 users who provided a solution to the image-based CAPTCHA. On the contrary, the female group has 52 users who solved the text-based CAPTCHA, and 51 users who solved the image-based CAPTCHA. We notice that the minimum response time determined by the group of male users is higher for the text than for the image-based CAPTCHA. A clear trend of the

Table 3 Hypothesis 3: Descriptive statistics of the mean and minimum response time to solve the text and image-based CAPTCHAs by the groups of male and female users

		Text	Num.	Anim.	Home	Old	Anim.	Worr.	Surpr.	Pict.
				wild	num.	wom.	face	face	face	
M	*n*	50	50	49	49	49	49	49	49	49
	Avg.	23.79	19.76	3.59	2.94	2.79	2.94	3.21	3.07	20.79
	Min.	8.28	4.30	1.00	1.00	0.87	0.58	1.01	1.05	1.56
F	*n*	52	52	51	51	51	51	51	51	51
	Avg.	24.82	21.6	3.93	2.91	2.71	2.87	3.16	3.02	20.20
	Min.	3.00	3.00	1.00	0.56	0.87	0.58	1.00	1.00	1.56

The size of each group is represented by *n*



(a)

(b)

Fig. 10 Hypothesis 3: (**a**) minimum response time and (**b**) mean response time of the group of male users to solve the text and image-based CAPTCHAs

minimum and mean response time for the different CAPTCHAs is shown in Fig. 10. Again, it is visible that the male group determines a higher minimum response time to solve the text-based CAPTCHA w.r.t. the image-based types of CAPTCHA. However, looking at the mean values, we observe that the response time is higher only for the text CAPTCHA (see Fig. 10b). In fact, the picture of the CAPTCHA exhibits a higher mean response time than the number CAPTCHA. Hence, the third hypothesis is disproved, but in all parts except the picture of the CAPTCHA it is partially proved.

Table 4 shows the results of the Mann-Whitney U test for the fourth hypothesis, in order to investigate the dependence of the response time to solve the number and picture of the CAPTCHA from the gender groups. It assumes a statistically significant difference between the two gender groups in the response time to solve the number CAPTCHA and the picture of the CAPTCHA. The dependent variable is the response time to solve the text and image-based types of CAPTCHA. The independent variable is the gender, determining the two groups of male and female users. We can observe that there is no statistically significant difference of the response time between the gender groups for both text and image-based types of CAPTCHA. In particular, the results are not statistically significant for number and picture of the CAPTCHA (p-values > 0.05). Hence, there is no real evidence that the response time to solve the number and picture of the CAPTCHA is different between male and female groups. Consequently, the fourth hypothesis is disproved.

To understand if there exists a statistically significant difference between age groups in the response time to solve the CAPTCHAs (fifth hypothesis), we run the Mann-Whitney

Table 4 Hypothesis 4: Results of the Mann-Whitney *U* test for the text and image-based CAPTCHA

CAPTCHA type	Gender	*n*	Mean Rank	Z	Asymp. Sig. (2-tailed)
Text	Male	50	51.690	−0.064	0.949
	Female	52	51.320		
Numbers	Male	50	51.470	−0.010	0.992
	Female	52	51.530		
Animals in the wild	Male	49	52.520	−0.683	0.495
	Female	51	48.560		
Home numbers	Male	49	54.640	−1.400	0.162
	Female	51	46.520		
Face of an old woman	Male	49	52.670	−0.734	0.463
	Female	51	48.410		
Animated face	Male	49	54.420	−1.324	0.185
	Female	51	46.740		
Worried face	Male	49	52.490	−0.672	0.501
	Female	51	48.590		
Surprised face	Male	49	52.880	−0.803	0.422
	Female	51	48.220		
Picture of the CAPTCHA	Male	49	52.370	−0.631	0.528
	Female	51	48.710		

The size of each group is represented by *n*

U test and show the results in Table 5. In particular, Table 5a shows the results of the test for the text-based types of CAPTCHA, while Table 5b shows the results of the test for the image-based types of CAPTCHA. The dependent variable is the response time to solve the text and image-based types of CAPTCHA. The independent variable is the age, dividing the Internet users into below 35 and above 35 years. The number of users below 35 years is 52 for the text and image-based types of CAPTCHA. On the contrary, the number of users above 35 years is 50 for the text and 48 for the image-based types of CAPTCHA.

From a first observation, we can conclude that the response time to solve the text and image-based types of CAPTCHA is statistically significant between the groups of users below and above 35 years. It is visible from the asymptotic significance (2-tailed) p-values, which are below 0.05 for all the types of CAPTCHA. Furthermore, the results obtained by the text-based types of CAPTCHA suggest that the response time is statistically higher for the group of users above 35 years than for the group of users below 35 years. The difference is also captured by the absolute value of Z, which is above 1.96. Also, the mean rank of the group of users above 35 years has an increment of around 39 w.r.t. the group of users below 35 years for the text CAPTCHA (from 32.540 to 71.220) and of around 40 for the number CAPTCHA (from 31.900 to 71.880). A similar condition can be observed for all the image-based types of CAPTCHA. In fact, the group of users above 35 years determines an absolute value of Z which is always higher than 1.96, and a mean rank which is around 30-40 values higher than the mean rank of the group of users below 35 years (see Table 5b). Hence, the fifth hypothesis is proved.

Table 5 Hypothesis 5: Results of the Mann-Whitney U test for: (a) text-based CAPTCHA, and (b) image-based CAPTCHA

CAPTCHA type	Age	*n*	Mean Rank	Z	Asymp. Sig. (2-tailed)
(a)					
Text	Below 35	52	32.540	−6.600	**0.000**
	Above 35	50	71.220		
Numbers	Below 35	52	31.900	−6.821	**0.000**
	Above 35	50	71.880		
(b)					
Animals in the wild	Below 35	52	33.050	−6.262	**0.000**
	Above 35	48	69.410		
Home numbers	Below 35	52	37.870	−4.533	**0.000**
	Above 35	48	64.190		
Face of an old woman	Below 35	52	32.400	−6.493	**0.000**
	Above 35	48	70.100		
Animated face	Below 35	52	30.080	−7.329	**0.000**
	Above 35	48	72.630		
Worried face	Below 35	52	34.860	−5.613	**0.000**
	Above 35	48	67.450		
Surprised face	Below 35	52	32.330	−6.520	**0.000**
	Above 35	48	70.190		
Picture of CAPTCHA	Below 35	52	52.870	−2.929	**0.003**
	Above 35	48	80.630		

Statistically significant results are marked in bold. The size of each group is represented by *n*

The sixth hypothesis involves the level of Internet experience of the users, which is the independent variable, ranging in the interval between 1 and 9 years. Also, the dependent variable is the response time to solve the text and image-based types of CAPTCHA. This hypothesis assumes that a low experience in the Internet use is related to a long response time to solve the CAPTCHAs. Table 6 shows the results of the Pearson's correlation coefficient test for the text and image-based types of CAPTCHA. We can observe that the asymptotic significance (2-tailed) p-value is always below 0.05. Hence, the results are statistical significant. Also, the correlation coefficient is negative for all the types of CAPTCHA. It indicates that the response time is statistically higher for users with a low Internet experience for all the types of CAPTCHA (see (3)). However, the strength of this correlation is quite different, depending on the type of CAPTCHA. In fact, we can observe that the number CAPTCHA exhibits the strongest correlation, with a value of -0.448. On the contrary, the animals in the wild has the weakest correlation, with a value of -0.211. In general, we can notice that the text-based types of CAPTCHA have a stronger correlation than the image-based types of CAPTCHA. This means that the response time is statistically more correlated to a low Internet experience for the text rather than for the image-based types of CAPTCHA. Considering that the negative correlation is pretty weak (values between -0.30 and -0.50), the sixth hypothesis is proved, but the level of linear dependence is not strong.

Table 6 Hypothesis 6: Results of the Pearson's correlation coefficient test for the text and image-based CAPTCHAs

Years of Internet use v.s. response time

	Text CAPTCHA		Image CAPTCHA						
	Text	Num.	Anim.	Home num.	Old wom.	Anim. face	Worr. face	Surpr. face	Pict.
r	−0.390	−0.448	−0.211	−0.324	−0.301	−0.319	−0.255	−0.320	−0.367
Sig.	**0.000**	**0.000**	**0.035**	**0.001**	**0.002**	**0.001**	**0.011**	**0.001**	**0.000**

Statistically significant results are marked in bold

For extending the fourth hypothesis involving only the number and picture of the CAPTCHA, the last hypothesis investigates if there exists a statistically significant difference between gender groups in the average response time to solve the CAPTCHA. For this reason, the Mann-Whitney U test is conducted on the population of Internet users and the results are presented in Table 7. This hypothesis is similar to the fourth hypothesis. The difference is that the average response time and all the CAPTCHA types are considered here. Again, the users have been divided into male and female groups. It implies that the independent variable is the gender. For the text-based CAPTCHA, the male group is composed of 50 Internet users, while the female group consists of 52 Internet users. For the image-based CAPTCHA, the two groups are composed respectively of 49 and 51 Internet users. About the dependent variable, it is the average response time to solve the text and image-based CAPTCHAs. For each CAPTCHA (text or image), it is computed as the response time of each Internet user averaged on the different text or image-based types of CAPTCHA. From these results, it is visible that the asymptotic significance (2-tailed) p-values are above 0.05 for both text and image-based CAPTCHAs. Hence, there is no statistical evidence that a gender difference determines a variation in the average response time to solve the CAPTCHA. For this reason, the seventh hypothesis, as well as the fourth one, is disproved.

From the aforementioned analysis of the obtained results, the decision of the examined hypotheses is the following:

- Hypothesis 1 (Proved): The group of internet users with higher education level will have a faster response time in solving the text-based CAPTCHA,

Table 7 Hypothesis 7: Results of the Mann-Whitney U test for the text and image-based CAPTCHAs

Gender v.s. Avg. response time in solving CAPTCHA

CAPTCHA type	Gender	*n*	Mean Rank	Z	Asymp. Sig. (2-tailed)
Text CAPTCHAs	Male	50	51.610	−0.037	0.971
	Female	52	51.390		
Image CAPTCHAs	Male	49	52.800	−0.776	0.438
	Female	51	48.290		

The size of each group is represented by *n*

- Hypothesis 2 (Disproved, but only proved in partial): The response time of the image-based CAPTCHA is the longest one for Internet users with secondary education level,
- Hypothesis 3 (Disproved, but only proved in partial): A faster response time to solve the CAPTCHA is presumable for image-based compared to text-based CAPTCHA in the case of male users,
- Hypothesis 4 (Disproved): There exists a statistically significant difference between gender groups in solving the text-based CAPTCHA with only numbers and the picture of the CAPTCHA,
- Hypothesis 5 (Proved): There exists a statistically significant difference between age groups in solving the CAPTCHA,
- Hypothesis 6 (Proved): Subjects who use the Internet less time will have a longer response time in solving the CAPTCHA,
- Hypothesis 7 (Disproved): There exists a statistically significant difference between gender groups in the average response time to solve the CAPTCHA.

From all aforementioned, we can make the following important observations: (i) Internet users with a higher education level have the tendency to quickly solve the CAPTCHA types, (ii) Internet users with a secondary education level require more time to solve the home numbers and the picture of the CAPTCHA and less time to solve the worried face, (iii) male users spend less time to solve the image w.r.t. the text-based types of CAPTCHA, except the picture of the CAPTCHA, (iv) there is no real evidence that a gender difference influences the response time to solve the text-based CAPTCHA with only numbers and the picture of the CAPTCHA, neither the average response time to solve the text and image-based CAPTCHAs, (v) an age difference influences the response time to solve the CAPTCHA, (vi) there exists a weak correlation between a low Internet experience and a long time to solve the CAPTCHA.

5.1 Comparison results

To demonstrate the significance of the obtained results, we compare our method with other two state-of-the-art methods for analyzing the CAPTCHA's usability in terms of the response time: (i) the method proposed in [26], and (ii) the method proposed in [5, 6] (see Section 2.1). Also, we confirm the correctness of the results obtained by the Man-Whitney U test by performing the Kruskal-Wallis H test on the same data [25]. Next, we will provide further details about each of these methods and we will discuss about similarities and differences with our method.

The method proposed in [26] analyzed the correlation of age and distortion types with the response time in solving the text-based CAPTCHA, error rate, and users' parameters connected to their objective visual fatigue and subjective workload during the experiment of solving the CAPTCHA. Considered text-based CAPTCHAs contain only text (without distortion) or five different types of distortions concerning noise added to the characters or the background. The experiment involved a population of 24 Internet users, who was required to solve the different types of CAPTCHA, divided into two age groups: (i) elder group (aged 50-56 years), and (ii) young group (aged 23-24 years). Response time in solving the CAPTCHAs, error rate, and the other users' parameters, have been registered during the experiment. Then, a statistical analysis was performed on collected data to evaluate the effect of age groups and distortion types on the response time to solve the text-based CAPTCHA, error rate, and users' parameters (subjective workload, error rate and objective visual fatigue). In particular, the Friedman's and Man-Whitney U tests have been used for analyzing the effect of respectively distortion type and age group (independent variables) to the response time, error rate, and users' parameters (dependent variables) [16, 45].

Some differences can be found between the method in [26] and our method. Firstly, our method only analyzes the response time to solve the CAPTCHA. We do not collect the error rate, the objective visual fatigue and the subjective workload of the users during the experiment. Furthermore, we also consider different image-based types of CAPTCHA, and not only the text-based CAPTCHAs with different backgrounds. Hence, we provide an extension of the analysis in [26]. Secondly, we use a much larger population of Internet users for the text-based CAPTCHA (102 users) and for the image-based CAPTCHA (100 users). The users are of age between 18 and 52 years old, divided into two groups: (i) below 35 years (from 18 to 35 years), and (ii) above 35 years (from 36 to 52 years). Hence, the users below 23 years and above 24 years until 35 years are also considered as well as the users from 35 to 52 years as elder group. Furthermore, different demographic factors are analyzed in our experiment (gender, age, education level and Internet experience), and not only the age. Hence, our analysis is more general in that sense.

Table 8 shows the mean response time to solve the text and image-based types of CAPTCHA, together with the standard deviation given in brackets, obtained by our experiment. Table 9 shows the mean response time for age groups (above 35 years and below 35 years). The mean response time is reported in milliseconds (ms), to make easier the comparison with the values of Ref. [26]. However, we can only compare the response times in solving the text and number CAPTCHAs, because the image-based CAPTCHAs are not considered in [26].

In terms of CAPTCHA's elements, in the test we use text-based types of CAPTCHA of 8 characters instead of 6 ones used in [26], making more complex our experiment. Also, we have a combination of small and capital letters which further extended the response time to solve a CAPTCHA. Finally, we register the response time from the beginning, when the user approaches the CAPTCHA, until the end, when the answer to the test is given. On the contrary, in [26] the response time does not include the phase of answering the CAPTCHA. That is why we receive a mean response time of 24319 and 20700 ms for text and number-based CAPTCHAs (see Table 8) instead of a value between 4398.54 and 9354.17 ms (see [26]). If we compare the response time registered for age groups, we have values of 34143.00 and 14872.88 in the text CAPTCHA, and of 30462.60 and 11312.88 in the number CAPTCHA, for respectively the groups of age above (36-52 years) and below (18-35 years) 35 years (see Table 9). On the contrary, in [26] the response time to solve

Table 8 Mean response time to text and image types of CAPTCHA (in milliseconds), together with the standard deviation (in brackets)

Mean (SD)	Response time in ms
Text	24319.00 (15320.10)
Numbers	20700.00 (15324.11)
Animals in wild	3768.40 (4075.07)
Home numbers	2910.70 (2832.88)
Face of an old woman	2714.00 (2434.87)
Animated face	2869.90 (2758.03)
Worried face	3158.60 (2497.75)
Surprised face	3018.30 (2569.05)
Picture of CAPTCHA	20200.9 (16685.01)

Table 9 Mean response time to text and image types of CAPTCHA (in milliseconds) for age groups, together with the standard deviation (in brackets)

Mean (SD)	Age	Response time in ms
Text	above 35	34143.00 (15276.25)
	below 35	14872.88 (7416.22)
Numbers	above 35	30462.60 (15503.21)
	below 35	11312.88 (7130.42)
Animals in wild	above 35	5565.42 (5232.20)
	below 35	2109.61 (1077.93)
Home numbers	above 35	3907.50 (3682.42)
	below 35	1990.58 (1132.83)
Face of an old woman	above 35	3647.71 (3101.80)
	below 35	1852.11 (1031.50)
Animated face	above 35	4036.70 (3552.14)
	below 35	1792.88 (820.93)
Worried face	above 35	4108.54 (3222.18)
	below 35	2281.73 (953.44)
Surprised face	above 35	4099.58 (3262.71)
	below 35	2020.19 (940.91)
Picture of CAPTCHA	above 35	24583.96 (17282.68)
	below 35	16155.00 (15177.86)

the text-based CAPTCHA ranges between 6104.62 and 11974.85 in the senior group (50-56 years), and between 2692.46 and 6733.49 in the young group (23-24 years). It is worth noting that the standard deviation of the given results in both studies has a similar percentage range compared to the given results. From all aforementioned, it is quite clear that we receive higher response time due to the following reasons: (i) use of text-based CAPTCHAs with more elements, (ii) use of text-based CAPTCHAs with differentiation between capital and small letters, (iii) response time including the answer phase to CAPTCHA, and (iv) analysis of a population which is less accustomed to a given CAPTCHA because of different demographic characteristics. Still, we used a much wider population with smaller differences in the age groups, which reveals a higher generality of our results.

In our case, we can only compare in terms of response time. However, we cannot compare according to the distortion types, because we do not have a similar classification. Furthermore, in [26] the experiment is repeated 10 times for each CAPTCHA, which is an assumption for using the Friedman's test. We do not repeat the experiment, for avoiding bias of the past experience in the response time to solve the CAPTCHA. In conclusion, we cannot apply the Friedman's test on our data. Nonetheless, we can compare the results obtained in [26] by the Man-Whitney U test with our results obtained by the same statistical test (see Section 5). In particular, the comparison will be performed in terms of p-value (see [26]), with significance threshold $\alpha = 0.01$. Hence, we are using the same experimental setting as given in Ref. [26]. Table 5(a) shows the result of the test for the text and number CAPTCHAs. The independent variable is the age, while the dependent variable is the response time. We may observe that the p-values are below 0.01 (0.0000 in both cases). It indicates that the age groups are statistically significant in terms of response time. In particular, users of age above 35 years take more time in solving the CAPTCHA w.r.t. users of age below 35 years, as explained by hypothesis 5. Ref. [26] came to the same conclusions (see [26]), but with a much smaller population and with homogeneous age groups, which might compromise the reliability of the results.

The second method has been introduced in [6] for the text-based CAPTCHA, and in [5] for the image-based CAPTCHA. It proposed an advanced statistical analysis using the association rules for evaluating the dependence of the response time to the CAPTCHA from the co-occurrence of the demographic factors: (i) age, (ii) education level, and (iii) Internet experience. The first part of the experiment was similar to the first part of our proposed experiment. A population of Internet users was required to solve different types of CAPTCHA. For each user, the demographic factors, together with the solution time to the CAPTCHAs, were collected. However, the dataset of users' data was subjected to association rule mining by the Apriori algorithm for the extraction of the association rules. Then, a filtering phase selected the only ARs whose consequent was the response time. An Association Rule (AR) was characterized by the antecedent, composed of a combination of values of the independent variables (age, education level, Internet experience), and by the consequent, composed of the response time values (dependent variable). The statistical significance of the AR was quantified by the support. The conditional probability of the consequent given the antecedent was measured by the confidence [5, 6]. The correlation degree between antecedent and consequent was measured by the lift [5, 6].

In order to compare our method with the method proposed in [5, 6], we extract the ARs from our datasets by the Apriori algorithm. We have one dataset for the text-based and another dataset for the image-based CAPTCHA. The first one is composed of 102 instances (the number of Internet users solving the text-based CAPTCHAs), while the second one contains 100 instances (the number of Internet users solving the image-based CAPTCHAs). In our case, the gender is added to the set of independent variables, too. Then, we split the Internet experience into three ranges by employing the Equal-Width discretization algorithm, and the response time into three ranges by using the K-Medians algorithm (see Refs. [5, 6]). Table 10 shows the (eventually discretized) values of each feature of the datasets.

Table 10 Possible values and corresponding intervals for each feature of the dataset

Feature	Name	Value/interval
gender	male	1
	female	2
age	below 35	1
	above 35	2
education level	higher education	1
	secondary education	2
number of years of Internet use (years)	high Internet use	]6, +∞[
	middle Internet use	]3, 6]
	low Internet use	[0, 3]
response time in solving the text-based CAPTCHAs (seconds)	high response time	]32.36, +∞[
	middle response time	]10.22, 32.36]
	low response time	]0.00, 10.22]
response time in solving the image-based CAPTCHAs (seconds)	high response time	]32.36, +∞[
	middle response time	]10.22, 32.36]
	low response time	]0.00, 10.22]

In order to extract the ARs, similarly to [5, 6] we set the *minsupport* and *minconfidence* thresholds respectively to 10% and 50% for both the datasets, by which we obtain the highest variability in the consequent values. It allows to examine all the cases of low, middle and high response time.

Tables 11, 12 and 13 show the ARs extracted for the different types of CAPTCHA, in terms of their antecedent and consequent, together with their support, confidence and lift values. In particular, the ARs corresponding to home numbers, animated character, worried face, and surprised face CAPTCHA are almost identical. Consequently, they are condensed into Table 13 (left). All the ARs are considered as statistically significant, because we set the *minsupport* threshold and only consider the ARs whose support value is higher than *minsupport*. Our aim will be to prove or disprove the 7 research hypotheses by employing the ARs, then compare the result with that obtained by our method.

About the hypothesis 1, we may observe it is confirmed by both methods. In fact, for the text and number CAPTCHA, a higher education level is only associated with a low and middle response time, with a high lift value between 1 and 5. In particular, when a higher education level is associated with an age below 35 years, a low response time is obtained (see 9, 14 and 16 in Table 11 (left) and 1, 2, 3, and 4 in Table 11 (right)). Otherwise, when a higher education level co-occurs with an age above 35 years and a middle Internet experience, a middle response time is obtained (see 23, 24 in Table 11 (left), and 12, 17 in Table 11 (right)).

According to our method, hypothesis 2 should be partially proved for home numbers, the picture of the CAPTCHA and worried face. We can observe from Table 13 (left and right) that the hypothesis is disproved, because all the ARs including the secondary education are associated with a low response time. This is due to the discretization of the response time into three ranges, which does not capture its smaller variations between secondary and higher education. Hence, the discretization level influences the acceptance of the hypothesis, which is a limitation of this method. In the case of picture of the CAPTCHA, a secondary education level is associated with a middle response time (see 4, 6, 9, 10 in Table 13 (right)). Accordingly, the hypothesis 2 is partially proved by this method, too.

The hypothesis 3 is proved by our method for the text-based CAPTCHA with only text and all the image-based CAPTCHAs. However, it is not proved for the text-based CAPTCHA with only numbers and the picture of the CAPTCHA. In this case, we may observe that male users are only associated with a middle response time in solving the text CAPTCHA (see 5, 10, 15, 17, 19, 22, and 24 in Table 11 (left)). Lift values are between 1 and 1.6 and confidence values are between 0.60 and 0.92, indicating a positive correlation between male users and middle response time. For all the image-based CAPTCHAs, male users are always related to a low response time. Hence, the hypothesis is proved in this case, too. In the case of number CAPTCHA (see Table 11 (right)), most of the ARs involving male users are associated with a middle response time, except the AR number 4, having the low response time as the consequent. For the picture of the CAPTCHA, male users are only related to low response time (see 8 in Table 13 (right)). Accordingly, differently from our method, the hypothesis 3 is proved for text and number CAPTCHA. Again, this is due to the discretization of the response time, missing small variations between the response time of the number CAPTCHA and the picture of the CAPTCHA (see Table 3).

On the contrary, hypothesis 4 is disproved by our method and by this method, too. In fact, there is no statistically significant difference between gender groups in the response time of number and picture of the CAPTCHA (see Tables 11 and 13 (right)). In the case of number CAPTCHA, female users are associated with a low response time (see 1, 13 in Table 11 (right)), but also with a middle response time (see 27, 28 in Table 11 (right)). Male users

Table 11 Association rules whose consequent is the response time to solve: text (left), and number (right) CAPTCHA

ID	Antecedent	Consequent	Supp.	Conf.	Lift	ID	Antecedent	Consequent	Supp.	Conf.	Lift
1	F,below 35,Secondary educ.	Middle resp. time	0.127	1	1.729	1	F,below 35,Higher educ.	Low resp. time	0.127	1	3.187
2	F,Secondary educ., Middle internet use	Middle resp. time	0.108	1	1.729	2	below 35,Higher educ., Middle internet use	Low resp. time	0.157	0.941	3
3	below 35,Secondary educ., Middle internet use	Middle resp. time	0.108	1	1.729	3	below 35,Higher educ.	Low resp. time	0.235	0.923	2.942
4	below 35,Secondary educ.	Middle resp. time	0.245	0.961	1.662	4	M,below 35,Higher educ.	Low resp. time	0.108	0.846	2.697
5	M,below 35,Secondary educ.	Middle resp. time	0.118	0.923	1.596	5	M,below 35,Secondary educ.	Middle resp. time	0.108	0.846	1.726
6	below 35,Low internet use	Middle resp. time	0.108	0.917	1.585	6	Secondary educ.,Middle internet use	Middle resp. time	0.167	0.809	1.651
7	below 35,Secondary educ., Low internet use	Middle resp. time	0.108	0.917	1.585	7	below 35,Secondary educ.	Middle resp. time	0.196	0.769	1.569
8	Secondary educ., Middle internet use	Middle resp. time	0.186	0.905	1.564	8	above 35,Low internet use	High resp. time	0.118	0.75	3.825
9	F,below 35,Higher educ.	Low resp. time	0.108	0.846	4.795	9	M,above 35,Middle internet use	Middle resp. time	0.118	0.75	1.53
10	M,Secondary educ.	Middle resp. time	0.176	0.750	1.297	10	above 35,Middle internet use	Middle resp. time	0.196	0.714	1.457
11	above 35,Low internet use	High resp. time	0.118	0.750	3.06	11	M,Secondary educ.	Middle resp. time	0.167	0.708	1.445
12	Secondary educ.	Middle resp. time	0.363	0.740	1.279	12	above 35,Higher educ., Middle internet use	Middle resp. time	0.118	0.667	1.36
13	F,Secondary educ.	Middle resp. time	0.186	0.731	1.263	13	F,below 35	Low resp. time	0.167	0.654	2.084
14	below 35,Higher educ., Middle internet use	Low resp. time	0.118	0.706	4	14	below 35,Middle internet use	Low resp. time	0.176	0.643	2.049
15	M,below 35	Middle resp. time	0.176	0.692	1.197	15	Secondary educ.	Middle resp. time	0.314	0.64	1.306
16	below 35,Higher educ.	Low resp. time	0.176	0.692	3.923	16	M,above 35	Middle resp. time	0.147	0.625	1.275
17	M,above 35,Middle internet use	Middle resp. time	0.108	0.687	1.189	17	above 35,Higher educ.	Middle resp. time	0.157	0.615	1.255
18	above 35,Middle internet use	Middle resp. time	0.186	0.678	1.173	18	F,Higher educ.	Low resp. time	0.147	0.577	1.839

Table 11 (continued)

ID	Antecedent	Consequent	Supp.	Conf.	Lift	ID	Antecedent	Consequent	Supp.	Conf.	Lift
19	M,Middle internet use	Middle resp. time	0.196	0.667	1.152	19	F,Secondary educ.	Middle resp. time	0.147	0.577	1.177
20	below 35	Middle resp. time	0.323	0.635	1.097	20	M,Middle internet use	Middle resp. time	0.167	0.567	1.156
21	Middle internet use	Middle resp. time	0.343	0.625	1.080	21	M	Middle resp. time	0.274	0.56	1.142
22	M	Middle resp. time	0.304	0.62	1.072	22	above 35	Middle resp. time	0.274	0.56	1.142
23	above 35,Higher educ., Middle internet use	Middle resp. time	0.108	0.611	1.056	23	below 35	Low resp. time	0.284	0.558	1.778
24	M,Higher educ., Middle internet use	Middle resp. time	0.118	0.6	1.037	24	Middle internet use	Middle resp. time	0.294	0.536	1.093
25	F,above 35	High resp. time	0.127	0.5	2.04	25	Higher educ.	Low resp. time	0.265	0.519	1.655
26	above 35,Secondary educ.	High resp. time	0.118	0.5	2.04	26	M,below 35	Middle resp. time	0.127	0.5	1.02
						27	F,above 35	Middle resp. time	0.127	0.5	1.02
						28	F,Middle internet use	Middle resp. time	0.127	0.5	1.02
						29	above 35,Secondary educ.	High resp. time	0.118	0.5	2.55
						30	above 35,Secondary educ.	Middle resp. time	0.118	0.5	1.02

Each AR is identified by an ID and represented in terms of antecedent and consequent. Support, confidence and lift values are also reported

Table 12 Association rules whose consequent is the response time to solve: animals in the wild (left), and old woman (right) CAPTCHA

ID	Antecedent	Consequent	Supp.	Conf.	Lift	ID	Antecedent	Consequent	Supp.	Conf.	Lift
1	below 35	Low resp. time	0.52	1	1.087	1	below 35	Low resp. time	0.52	1	1.020
2	M,below 35	Low resp. time	0.26	1	1.087	2	Higher educ.	Low resp. time	0.52	1	1.020
3	M,Middle internet use	Low resp. time	0.28	1	1.087	3	Middle internet use	Low resp. time	0.53	1	1.020
4	F,below 35	Low resp. time	0.26	1	1.087	4	High internet use	Low resp. time	0.18	1	1.020
5	below 35,Higher educ.	Low resp. time	0.26	1	1.087	5	M,below 35	Low resp. time	0.26	1	1.020
6	below 35,Secondary educ.	Low resp. time	0.26	1	1.087	6	M,Higher educ.	Low resp. time	0.26	1	1.020
7	below 35,Middle internet use	Low resp. time	0.28	1	1.087	7	M,Middle internet use	Low resp. time	0.28	1	1.020
8	Secondary educ.,Middle internet use	Low resp. time	0.18	1	1.087	8	F,below 35	Low resp. time	0.26	1	1.020
9	M,Higher educ.,Middle internet use	Low resp. time	0.2	1	1.087	9	F,Higher educ.	Low resp. time	0.26	1	1.020
10	below 35,Higher educ., Middle internet use	Low resp. time	0.17	1	1.087	10	F,Middle internet use	Low resp. time	0.25	1	1.020
11	Middle internet use	Low resp. time	0.51	0.962	1.046	11	below 35,Higher educ.	Low resp. time	0.26	1	1.020
12	M,Secondary educ.	Low resp. time	0.22	0.956	1.039	12	below 35,Secondary educ.	Low resp. time	0.26	1	1.020
13	Higher educ.,Middle internet use	Low resp. time	0.33	0.943	1.025	13	below 35,Middle internet use	Low resp. time	0.28	1	1.020
14	M	Low resp. time	0.46	0.939	1.020	14	above 35,Higher educ.	Low resp. time	0.26	1	1.020
15	Secondary educ.	Low resp. time	0.45	0.937	1.019	15	above 35,Middle internet use	Low resp. time	0.25	1	1.020
16	M,Higher educ.	Low resp. time	0.24	0.923	1.003	16	Higher educ.,Middle internet use	Low resp. time	0.35	1	1.020
17	F,Secondary educ.	Low resp. time	0.23	0.92	1	17	Higher educ.,High internet use	Low resp. time	0.15	1	1.020
18	F,Middle internet use	Low resp. time	0.23	0.92	1	18	Secondary educ.,Middle internet use	Low resp. time	0.18	1	1.020
19	above 35,Middle internet use	Low resp. time	0.23	0.92	1	19	M,Higher educ.,Middle internet use	Low resp. time	0.2	1	1.020
						20	F,Higher educ.,Middle internet use	Low resp. time	0.15	1	1.020
						21	below 35,Higher educ., Middle internet use	Low resp. time	0.17	1	1.020
						22	above 35,Higher educ., Middle internet use	Low resp. time	0.18	1	1.020
						23	F	Low resp. time	0.5	0.980	1.000

Each AR is identified by an ID and represented in terms of antecedent and consequent. Support, confidence and lift values are also reported

Table 13 Association rules whose consequent is the response time to solve: home numbers, animated character, worried face, surprised face CAPTCHA (left), and the picture of the CAPTCHA (right)

ID	Antecedent	Consequent	Supp.	Conf.	Lift	ID	Antecedent	Consequent	Supp.	Conf.	Lift
1	below 35	Low resp. time	0.52	1	1.031	1	below 35,Higher educ.	Low resp. time	0.24	0.923	2.367
2	Higher educ.	Low resp. time	0.52	1	1.031	2	below 35,Higher educ., Middle internet use	Low resp. time	0.15	0.882	2.262
3	Middle internet use	Low resp. time	0.53	1	1.031	3	F,Higher educ.	Low resp. time	0.2	0.769	1.972
4	High internet use	Low resp. time	0.18	1	1.031	4	F,Secondary educ.	Middle resp. time	0.18	0.72	1.846
5	M,below 35	Low resp. time	0.26	1	1.031	5	Higher educ.	Low resp. time	0.37	0.711	1.824
6	M,Higher educ.	Low resp. time	0.26	1	1.031	6	Secondary educ.,Low internet use	Middle resp. time	0.18	0.667	1.709
7	M,Middle internet use	Low resp. time	0.28	1	1.031	7	Higher educ.,Middle internet use	Low resp. time	0.23	0.657	1.685
8	F,below 35	Low resp. time	0.26	1	1.031	8	M,Higher educ.	Low resp. time	0.17	0.654	1.676
9	F,Higher educ.	Low resp. time	0.26	1	1.031	9	below 35,Secondary educ.	Middle resp. time	0.17	0.654	1.676
10	F,Middle internet use	Low resp. time	0.25	1	1.031	10	Secondary educ.	Middle resp. time	0.31	0.646	1.656
11	below 35,Higher educ.	Low resp. time	0.26	1	1.031	11	Low internet use	Middle resp. time	0.18	0.621	1.591
12	below 35,Secondary educ.	Low resp. time	0.26	1	1.031	12	below 35,Middle internet use	Low resp. time	0.15	0.536	1.374
13	below 35,Middle internet use	Low resp. time	0.28	1	1.031						
14	above 35,Higher educ.	Low resp. time	0.26	1	1.031						
15	above 35,Middle internet use	Low resp. time	0.25	1	1.031						
16	Higher educ.,Middle internet use	Low resp. time	0.35	1	1.031						
17	Higher educ.,High internet use	Low resp. time	0.15	1	1.031						
18	Secondary educ.,Middle internet use	Low resp. time	0.18	1	1.031						
19	M,Higher educ.,Middle internet use	Low resp. time	0.2	1	1.031						
20	F,Higher educ.,Middle internet use	Low resp. time	0.15	1	1.031						
21	below 35,Higher educ.,Middle internet use	Low resp. time	0.17	1	1.031						
22	above 35,Higher educ.,Middle internet use	Low resp. time	0.18	1	1.031						
23	M	Low resp. time	0.48	0.979	1.010						

Each AR is identified by an ID and represented in terms of antecedent and consequent. Support, confidence and lift values are also reported

are related to a middle response time (see 5, 9, 11, 16, 20, 21, and 26 in Table 11 (right)). However, there is one AR where male users are associated with a low response time (see 4 in Table 11 (right)), with a high lift and a confidence value respectively of 2.70 and 0.85. In the case of the picture of the CAPTCHA, a low response time is related to both female and male users (see 3 and 8 in Table 13 (right)), with a lift value up to 2 in both cases. Also, female users are associated with a middle response time, too (see 4 in Table 13 (right)).

About the hypothesis 5, for text and number CAPTCHA, we may observe that users below 35 years are associated with a low response time (see 9, 14 in Table 11 (left), and 1, 2 in Table 11 (right)), but also with a middle response time (see 1, 3 in Table 11 (left), and 5, 7 in Table 11 (right)). Users above 35 years are related to a middle response time (see 18, 23 in Table 11 (left), and 22, 30 in Table 11 (right)), but also to a high response time (see 11, 26 in Table 11 (left), and 8, 29 in Table 11 (right)). Because both age groups are "partially" associated with a middle response time, there is no clear distinction between the groups, due to the discretization of the response time. This is a limitation of the method which partially proves the hypothesis. In the case of image-based CAPTCHAs (except the picture of the CAPTCHA), the hypothesis 5 is disproved. In fact, there is no statistically significant difference between age groups in terms of response time (see Tables 12–13 (left) where a low response time is associated with both age groups). This is another consequence of the discretization of the response time. For the picture of the CAPTCHA, the age group above 35 years does not appear (see Table 13 (right)). Consequently, there is no statistically significant difference between age groups in terms of response time. Hence, the hypothesis 5 is disproved by this method, and proved by our method.

The hypothesis 6 is only partially proved by this method, while it is fully proved by our method. In fact, for the text CAPTCHA we may observe that a low Internet experience is related to a high response time (see 11 in Table 11 (left)), with a lift value of 3.06 and confidence value of 0.75. However, a low Internet experience, as well as a middle Internet experience, is also associated with a middle response time (see 2, 3, 6, and 7 in Table 11 (left)). The aim is to investigate if a lower Internet experience is related to a higher response time. If a low and middle Internet experience are both associated with a middle response time, there is no more distinction between the Internet experience levels. Accordingly, the hypothesis is partially proved. For the number CAPTCHA, a low Internet experience is only associated with a high response time (see 8 in Table 11 (right)), with a high lift value of 3.82. Hence, the hypothesis is proved. For the image-based CAPTCHAs (except the picture of the CAPTCHA), a high Internet experience is associated with a low response time, as well as a middle Internet experience, with the same lift and confidence values. Consequently, there is no distinction between high and middle experience, which is due to the discretization of the response time. Also, there are no ARs with a low Internet experience. Hence, the hypothesis 6 is disproved. Finally, for the picture of the CAPTCHA, a low Internet experience is related to a middle response time (see 6 in Table 13 (right)), while a middle Internet experience is associated with a low response time (see 2, 7, and 12 in Table 13 (right)). Accordingly, the hypothesis 6 is proved.

The hypothesis 7 evaluates the differences in the mean response time between the gender groups. We will omit this analysis which requires again the extraction of all the ARs with the average response time of text and image-based CAPTCHAs as the consequent. Hence, differently from our method, the analysis proposed by this method is not modular.

Finally, we performed the Kruskal-Wallis H statistical test on the original users' data [25], to confirm the correctness of the results obtained by the Man-Whitney U test. The Kruskal-Wallis H test is an extension of the Man-Whitney U test, which may process more than two groups of the independent variable. It is a nonparametric test based on the ranks for evaluating the differences of two or more groups in terms of the dependent variable. Similarly to the Man-Whitney U test, the Kruskal-Wallis H test is based on four assumptions: (i) the dependent variable should be interval-based or continuous, (ii) the independent variable should be composed of two or more independent groups, (iii) the observations in each group and among the groups should be independent, and (iv) the distributions in each group could not have the same shape. Again, our data satisfy each of these assumptions. By the Kruskal-Wallis H test we computed the same statistics that we presented in Section 5, which were computed by the Man-Whitney U test. We also used the Internet experience as the independent variable (containing more than two groups) and the response time as the dependent variable. The results were perfectly consistent with those obtained by the Man-Whitney U test and by the Pearson's correlation coefficient test, in terms of mean ranks, r and p-values. Hence, we will omit to report them. This confirms the robustness of our analysis.

From all above, it is worth noting that some demographic characteristics can influence the solution time of the CAPTCHA. Basically, the analysis showed that a faster solution to the CAPTCHA is sure for Internet users that are younger as well as with higher Internet experience. Furthermore, there is no evidence that any gender plays a role in the solution time of the CAPTCHA. Still, there exist different conclusions concerning the solution time of the text-based (with text and with numbers) CAPTCHA and the image-based CAPTCHA. Essentially, the image-based CAPTCHA mainly has a much lower solution time, but an exception is the picture of the CAPTCHA. It is proved that the image-based CAPTCHA can be solved in lower time than the text-based CAPTCHA in circumstances when the included image is understandable to the Internet users.

If we compare the both techniques that are used for the analysis, it is worth noting that the statistical analysis is more supervised than the association rule mining. Hence, if the hypotheses, which are completely supervised, are wrongly set, the results can be less valuable. In the contrast, ARs are less supervised, which leads to more valuable results, if they are correctly interpreted. However, association rule mining is supervised according to some level because the threshold level of the support and confidence should be defined, too. Still, both techniques can be used in the same analysis as complement to each other. In this scenario, the first one can be analyzed by the association rule mining. Extracting and correctly interpreting the relations obtained by ARs can be suitable for establishing the proper hypotheses to be used in the statistical analysis.

In the conclusion, it is worth noting that statistical analysis is more adapted to solve a problem with known statistical characteristics which are more oriented to scale compared to interval variables, which are important for ARs extraction. Hence, in our case, the statistical analysis shows a better fitting to solve the given problem of the response time to which this type of analysis is more accustomed. The previously showed results proved this statement.

6 Conclusion

This paper presented a new study of the response time of Internet users to different CAPTCHA types. Accordingly, 9 different CAPTCHAs divided between text and image-based CAPTCHAs were evaluated. The experiment was conducted on a population of 102 Internet users for solving the text-based CAPTCHA, and on a population of 100 Internet users for solving the image-based CAPTCHA. Users were differentiated by their age, gender, education level and number of years of Internet use. The statistical analysis showed that the education level is an important indicator of the response time to solve the CAPTCHA. On the contrary, the gender difference is not a significant criterion to evaluate the response time. Again, the results indicate that young Internet users are more accustomed to quickly solve the CAPTCHA. Finally, the Internet experience weakly influences the response time to solve the CAPTCHA.

Future research work will include the creation of a prediction model of the user's response time, in order to predict the typical user's response to specific types of CAPTCHA. In this way, this prediction can be used for creating a loop that consists of different types of CAPTCHA. The loop of CAPTCHA can give to the users a choice of the most suitable CAPTCHA, which will emulate a "good" CAPTCHA.

Acknowledgements The authors are fully grateful to Sanja Petrovska for collecting the data, and to anonymous users for providing their data. This study was partially funded by the Grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia, as a part of the project TR33037 within the framework of the Technological development program.

Compliance with Ethical Standards

Conflict of interests Author Darko Brodić declares that he has no conflict of interest. Author Alessia Amelio declares that she has no conflict of interest. Author Radmila Janković declares that she has no conflict of interest.

Ethical approval This article does not contain any dangerous study with human participants or animals performed by any of the authors.

Funding This study was partially funded by the Grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia, as a part of the project TR33037 within the framework of the Technological development program. The receiver of the funding is Dr. Darko Brodić.

References

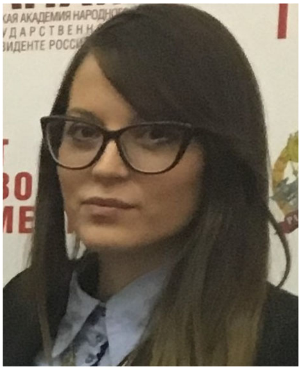
1. Baecher P, Fischlin M, Gordon L, Langenberg R, Lutzow M, Schroder D (2010) CAPTCHAS: the good, the bad, and the ugly. In Proc. of GI-Sicherheit, lecture notes in informatics, vol 170, pp 353–365
2. Baird HS, Riopka T (2005) Scattertype: a reading CAPTCHA resistant to segmentation attack. In: Proc. document recognition and retrieval XII, SPIE-IS&T electronic imaging, vol 5676. SPIE, pp 197–207
3. Belk M, Germanakos P, Fidas C, Spanoudis G, Samaras G (2013) Studying the effect of human cognition on text and image recognition CAPTCHA mechanisms. Proc HAS/HCII Lect Notes Comput Sci 8030:71–79
4. Belk M, Fidas C, Germanakos P, Samaras G (2015) Do human cognitive differences in information processing affect preference and performance of CAPTCHA? Int J Human-Comput Stud 84:118
5. Brodić D, Amelio A (2016) Analysis of the human-computer interaction on the example of image-based CAPTCHA by association rule mining. In: Proc. of 5th international workshop on symbiotic interaction, lecture notes in computer science, vol 9961. Springer, pp 38–51
6. Brodić D, Amelio A, Draganov IR (2016) Response time analysis of text-based CAPTCHA by association rules. In: Proc. of 17th international conference on artificial intelligence: methodology, systems, applications AIMSA, lecture notes in computer science, vol 9883. Springer, pp 78–88
7. CAPTCHA. http://www.captcha.net
8. Chellapilla K, Larson K, Simard P, Czerwinski M (2005) Designing human friendly human interaction proofs (HIPs). In: Proc. of SIGCHI conf. on human factors in computing systems, pp 711–720
9. Cui J, Liu Y, Xu Y, Zhao H, Zha H (2013) Tracking generic human motion via fusion of low- and High-Dimensional approaches. IEEE Trans Syst Man Cybern Syst 43(4):996–1002
10. Cumming G (2012) Understanding the new statistics: effect sizes confidence intervals and meta-analysis. Routledge, New York
11. DICE CAPTCHA. http://dice-captcha.com/
12. Dhamija R, Tygar J (2005) Phish and HIPs: human interactive proofs to detect phishing attacks. In: Proc. of Human interactive proofs: second international workshop (HIP), pp 127–141

13. Exact and asymptotic p-value. Available online. https://analyse-it.com/docs/user-guide/101/exactasymptoticpvalues
14. Field A (2009) Discovering statistics using SPSS. SAGE Publications Ltd, Los Angeles
15. First Workshop on Human Interactive Proofs (2002). http://www2.parc.com/istl/groups/did/HIP2002
16. Friedman Test in SPSS Statistics. Available online. https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php
17. Goswami G, Powell BM, Vatsa M, Singh R, Noore A (2014) FaceDCAPTCHA: face detection based color image CAPTCHA. Futur Gener Comput Syst 31(2):59–69
18. Hernandez-Castro CJ, Ribagorda A (2010) Pitfalls in CAPTCHA design and implementation: the math CAPTCHA, a case study. Comput Secur 29(1):141–157
19. Hubbard R (2004) Blurring the distinctions between p's and a's in psychological research. Theory Psychol 14(3):295–327
20. IBM SPSS software. http://www.ibm.com/analytics/us/en/technology/spss/
21. Kalsoom S, Ziauddin S, Abbasi AR (2012) An image-based CAPTCHA scheme exploiting human appearance characteristics. KSII Trans Internet Inf Syst 6(2):734–749
22. Khan M, Shah T, Batool SI (2016) A new implementation of chaotic S-boxes in CAPTCHA. SIViP 10(2):293–300
23. Kim JW, Chung WK, Cho HG (2010) A new image-based CAPTCHA using the orientation. Vis Comput 26(6):1135–1143
24. Kim J, Yang J, Wohn K (2014) AgeCAPTCHA: an image-based CAPTCHA that annotates images of human faces with their age groups. KSII Trans Int Inf Syst 8(3):1071–1092
25. Kruskal-Wallis H Test using SPSS Statistics. Available online. https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php
26. Lee YL, Hsu CH (2011) Usability study of text-based CAPTCHAs. Displays 32(2):81–86
27. Li Q (2015) A computer vision attack on the ARTiFACIAL CAPTCHA. Multimed Tools Appl 74(13):4583–4597
28. Lillibridge MD, Abadi M, Bharat K, Broder A (2001) Method for selectively restricting access to computer systems. US Patent 6,195,698. http://www.google.com/patents/US6195698
29. Liu Y, Cui J, Zhao H, Zha H (2012) Fusion of low-and high-dimensional approaches by trackers sampling for generic human motion tracking. In: Proc. of the 21st international conference on pattern recognition, pp 898–901
30. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2activity: recognizing complex activities from sensor data. In: Proc. of the 24th international conference on artificial intelligence. AAAI Press, pp 1617–1623
31. Liu Y, Zhang L, Nie L, Yan Y, Rosenblum DS (2016) Fortune teller: predicting your career path. In: Proc. of the Thirtieth AAAI conference on artificial intelligence. AAAI Press, pp 201–207
32. Liu Y, Nie L, Liu L, Rosenblum DS (2016) From action to activity: sensor-based activity recognition. Neurocomputing 181:108–115
33. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum DS (2016) Recognizing complex activities by a probabilistic interval-based model. In: Proc. of the Thirtieth AAAI conference on artificial intelligence. AAAI Press, pp 1266–1272
34. Liu Y, Zhang X, Cui J, Wu C, Aghajan H, Zha H (2010) Visual analysis of child-adult interactive behaviors in video sequences. In: Proc 16th International conference on virtual systems and multimedia, pp 26–33
35. Liu Y, Liang Y, Liu S, Rosenblum DS, Zheng Y (2016) Predicting urban water quality with ubiquitous data. CoRR abs/1610.09462
36. Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016) Urban water quality prediction based on multi-task multi-view learning. In: Proc. of IJCAI. IJCAI/AAAI Press, pp 2576–2581
37. Lu Y, Wei Y, Liu L, Zhong J, Sun L, Liu Y (2017) Towards unsupervised physical activity recognition using smartphone accelerometers. Multimed Tools Appl 76(8):10701–10719
38. Madathil GF, Alapatt JS, Greenstein JS, Madathil KC (2010) An investigation of the usability of image-based CAPTCHAs. Proc. Human Factors Ergon Soc Annual Meeting 54(16):1249–1253
39. Mann-Whitney U Test using SPSS Statistics Laerd Statistics. https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php
40. Moran TP (1981) The command language grammar: a representation for the user interface of interactive computer systems. Int J Man-Mach Stud 15(1):3–50
41. Preotiuc-Pietro D, Liu Y, Hopkins D, Ungar L (2017) Beyond binary labels: political ideology prediction of Twitter users. In: Annual meeting of the association for computational linguistics
42. Research Methods I: SPSS for Windows part 3, Nonparametric tests. Available online. http://www.discoveringstatistics.com/docs/nonparametric.pdf

43. Rui Y, Liu Z (2004) ARTiFACIAL: automated reverse Turing test using FACIAL features. Multimed Syst 9(6):493–502
44. Rusu A, Govindaraju V (2005) Visual CAPTCHA with handwritten image analysis. In: Proc. of HIP, lecture notes in computer science, vol 3517. Springer, pp 42–52
45. The Mann-Whitney U-test – Analysis of 2-Between-Group Data with a Quantitative Response Variable. Available online. http://psych.unl.edu/psycrs/handcomp/hcmann.PDF
46. Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460
47. Von Ahn L, Blum M, Langford J (2004) Telling humans and computers apart automatically. Commun ACM 47(2):57–60

Darko Brodić received his BEE and MEE from the Faculty of Electrical Engineering, University of Sarajevo in 1987 and 1990, as well as Ph.D. in electrical engineering from the Faculty of Electrical Engineering, University of Banja Luka in 2011. Now he is Associate Professor of computer science at the Technical Faculty in Bor, University of Belgrade, Serbia. His current research interests include different aspects of signal processing, natural image processing, document image processing, pattern recognition, artificial intelligence and sensor data measurement. He is the author of more than 40 SCIE journal papers, and over 80 conference papers.

Alessia Amelio received her BSc and MSc in Computer Science Engineering from University of Calabria in 2005 and 2009, as well as Ph.D. in computer science engineering and systems from the Faculty of Engineering, University of Calabria in 2013. Now she is Research Fellow of computer science at the Department of Computer Science Engineering, Modeling, Electronics and Systems, University of Calabria, Italy. Her current research interests include different aspects of image processing, document classification, pattern recognition, social network analysis, data mining for sensor data measurement and web applications. She is the author of more than 50 scientific papers.



Radmila Janković received her BSc and MSc from the Technical Faculty in Bor, University of Belgrade in 2015 and 2016, respectively. Now, she is a PhD student at the Technical Faculty in Bor, University of Belgrade. Her current research interests include different aspects of statistical analysis and e-business implementation.