# AgeCAPTCHA: an Image-based CAPTCHA that Annotates Images of Human Faces with their Age Groups

**Jonghak Kim, Joonhyuk Yang, and Kwangyun Wohn**
Graduate School of Culture Technology, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
[e-mail: defigner@kaist.ac.kr, joony.yang@kaist.ac.kr, wohn@kaist.ac.kr]
*Corresponding author: Jonghak Kim

*Received May 7, 2013; revised January 15, 2014; accepted March 2, 2014; published March 31, 2014*

## *Abstract*

Annotating images with tags that describe the content of the images facilitates image retrieval. However, this task is challenging for both humans and computers. In response, a new approach has been proposed that converts the manual image annotation task into CAPTCHA challenges. However, this approach has not been widely used because of its weak security and the fact that it can be applied only to annotate for a specific type of attribute clearly separated into mutually exclusive categories (e.g., gender). In this paper, we propose a novel image annotation CAPTCHA scheme, which can successfully differentiate between humans and computers, annotate image content difficult to separate into mutually exclusive categories, and generate verified test images difficult for computers to identify but easy for humans. To test its feasibility, we applied our scheme to annotate images of human faces with their age groups and conducted user studies. The results showed that our proposed system, called AgeCAPTCHA, annotated images of human faces with high reliability, yet the process was completed by the subjects quickly and accurately enough for practical use. As a result, we have not only verified the effectiveness of our scheme but also increased the applicability of image annotation CAPTCHAs.

***Keywords:*** Age estimation, CAPTCHA, human computation, image annotation, usability, Web application

**http://dx.doi.org/10.3837/tiis.2014.03.021**

## 1. Introduction

With the rapid development of digital imaging and storage devices, the number of images on the Web has been increasing exponentially. For example, Flickr[1] has more than 5 billion images, and Facebook[2] hosts more than 50 billion images [1]. This increasing number of images has brought about the need to develop effective methods for searching through such image databases. One effective way to facilitate image retrieval is to annotate images with tags that describe the content of the images or that provide additional contextual and semantic information [2].

Unfortunately, considering the large number of images available on the Web today, it is almost impossible for humans to manually annotate every image. Furthermore, humans rarely annotate the images because it is laborious and the future benefits are unanticipated [3]. In response, researchers have developed various image recognition algorithms to automatically annotate a large number of images. However, the algorithms have yet to match the performance of humans or even to reach a level of precision that would be adequate for automatic image annotation [4]. As a result, image annotation remains a challenging task for both humans and computers.

To solve this problem, a new approach has been proposed, which converts the manual image annotation task into CAPTCHA challenges [5,6]. A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is an automated test that humans can pass but current computer programs cannot, such as recognizing distorted characters (text-based CAPTCHAs) or objects in images (image-based CAPTCHAs) [7,8]. CAPTCHAs have been used by many websites to tell whether their user is a human or a computer, and, in this way, they prevent abuse from malicious bots (automated programs) that are usually written to generate spam.

However, as optical character recognition algorithms have developed, text-based CAPTCHAs have become less secure [9]. As an alternative, various image-based CAPTCHAs have been proposed because image recognition is more challenging than text recognition for computers, in general [8]. Considering the gap between the image recognition capabilities of humans and computers, we expect that image annotation is an excellent task for CAPTCHA challenges. In addition, by making the annotation task a natural part of the activities that humans already perform, this approach is highly effective in encouraging humans to continuously participate in the task.

There are two types of image annotation CAPTCHAs. The first type asks users to input a word that best describes an image [5]. However, due to its dependence on language, this is fundamentally difficult to be used widely. The second type requires users to classify a specific object in the images by using a given set of categories, such as distinguishing between fruits and electronic products for images annotated with “Apple” [6]. Unfortunately, this type has also not been widely used, for two main reasons. First, it can be applied to only identify a specific type of attribute clearly separated into mutually exclusive categories (e.g., gender). Second, we cannot

[1] http://www.flickr.com
[2] http://www.facebook.com

be sure of its security and usability because there is no mechanism to verify that its test images are suitable for CAPTCHA challenges.

As an extension of the second type, the aim of this study is to propose a practical CAPTCHA scheme for image annotation that can be applied even if a target attribute (for annotation) is not clearly separated into mutually exclusive categories. To achieve this aim, we first develop a novel method for collecting user responses (two-layered response structure), whereby each response category (button) is valid for the image annotation. However, for CAPTCHA purposes, our method internally considers two or more frequently confused categories as a single category, which is expected to extend the applicability of the image annotation CAPTCHAs by increasing the human success rate while maintaining the quality of the image annotation. In addition, we propose a method of generating verified test images that are truly difficult for computers to identify but easy for humans, which enables the practical use of our CAPTCHA.

To test the feasibility, we applied our CAPTCHA scheme to annotate images of human faces with their age groups; this was done for two reasons. The first reason is that, to ensure the security of the CAPTCHA challenges, the test images should be generated and updated automatically [10]. In this regard, human faces are very suitable as test images because they can be automatically detected and cropped from images by face detection [11]. The second reason is that person-related information is one of the most useful cues for image retrieval [12]. Among such information, the age group (approximate age) is an attribute that humans can effortlessly extract from facial images [13]. However, since age is not clearly separated into mutually exclusive groups, it is difficult to assess by means of existing image annotation CAPTCHAs. Moreover, the current age estimation algorithms are less than 60% accurate for real-life facial images [14,15]. Thus, annotating human faces with their age groups, especially faces whose ages cannot be estimated by the existing algorithms (age-indeterminate faces), was considered to be an excellent candidate for applying and testing our CAPTCHA scheme. Hereafter, we call this system AgeCAPTCHA.

The remainder of this paper is organized as follows: Section 2 surveys the related work. Section 3 describes the design of AgeCAPTCHA. Section 4 clarifies the criteria for operating AgeCAPTCHA challenges. Section 5 presents our user study design and its result. Finally, Section 6 provides a conclusion.

## 2. Related Work

### 2.1 CAPTCHA Requirements

Image annotation CAPTCHAs were designed to produce tags that describe image content as a byproduct of CAPTCHA challenges [5,6]. However, as mentioned above, these CAPTCHAs have not been widely used because of their weaknesses in terms of the security and usability. For CAPTCHAs to be useful for practical applications, several requirements should be considered. In this subsection, based on the guidelines of Rui and Liu [10], we further investigate the requirements for designing practical CAPTCHAs.

(1) **Automation and gradability:** Above all, CAPTCHA challenges must be generated and graded automatically by computers [8,10,16].

(2) **Difficult for computers but easy for humans:** Considering the attack cost, a single CAPTCHA challenge should not permit a computer success rate of greater than 1 in 10,000 (0.01%), whereas the human success rate should approach 90% [17]. In addition, CAPTCHAs should be completed quickly by humans because a challenge that takes longer than 30 s is less useful for any practical purpose [10].

(3) **Universality:** To be widely accepted, CAPTCHAs should be independent of a user's background (e.g., the physical location or level of education) [10,18]. This requirement becomes more important for websites that have international users.

(4) **Robustness when the database is publicized:** If an attacker can easily break a CAPTCHA by acquiring its database, the practical use of the system becomes difficult. Therefore, the security of CAPTCHAs should not be based on the database's secrecy [10,19]. In addition, to reduce the risk of data theft, test images should be constantly updated and replaced [20].

(5) **Mobile-friendly interface:** In general, responding to CAPTCHA challenges on a mobile device is more time-consuming than on a traditional desktop computer with a keyboard and mouse [21]. Currently, as the number of mobile Internet users is increasing rapidly, CAPTCHAs should be designed to be more acceptable in mobile environments.

### 2.2 Image-based CAPTCHAs and their Test Images

As image-based CAPTCHAs typically require users to recognize a specific attribute of an object in the images, their security and usability are significantly constrained by the quality of the test images. Therefore, it is worth examining the development process of image-based CAPTCHAs, focusing on their test images.

In the simplest form of image-based CAPTCHAs, test images were manually constructed by human administrators, and thus, only a small number of test images were available. For example, the face recognition CAPTCHA of Misra and Gaj [22] displayed two sets of distorted human faces and asked users to match faces with the same person. As the UMIST face database[3], which was used as test images, consisted of only 564 facial images of 20 people, it was highly vulnerable to hacking activities. Recently, Kalsoom et al. [20] proposed a CAPTCHA that asks the users to identify five characteristics of human appearance (gender, hair type, hair color, ethnicity, and expression) from manually prepared facial images. They also mentioned the need to increase the number of test images and proposed a method for generating the test images based on the user responses. However, they did not implement and verify their ideas.

To acquire a sufficient number of test images, later CAPTCHAs were designed and implemented on the basis of particular large-scale human-generated image databases constructed by other applications. The PIX of von Ahn et al. [23] used an annotated image database generated by the ESP Game [24] to display four different images with the same label and asked the users to choose the label that is related to all of the images from a menu of 70 labels. Similarly, HotCAPTCHA[4] used images from HotOrNot.com, a website that invites users to rate others as attractive or not. During the challenge, each user was asked to select three attractive

[3] http://www.sheffield.ac.uk/eee/research/iel/research/face
[4] This system is no longer available since the website HotCAPTCHA.com was closed down in 2009.

people from nine different people in the images. The Asirra of Elson et al. [18] used cat and dog images in Petfinder.com that were manually annotated by volunteers at animal shelters. Its task was to identify cats in a set of 12 images of cats and dogs. However, this type of test image was not verified as being truly difficult for computers to identify, which permits a machine learning attack to break Asirra CAPTCHA with a probability of 10.3% [25]. In addition, as these CAPTCHAs were not able to update and replace their test images without the help of external applications, they were still vulnerable to data theft.

In response, more recent CAPTCHAs were designed to generate test images for themselves, and previous image annotation CAPTCHAs were included in this approach. For example, the TagCAPTCHA of Morrison et al. [5] and the iCAPTCHA of Khot et al. [6] displayed publicly available Web images to users with existing test images during their CAPTCHA challenges. TagCAPTCHA asked the users to input an English word that best describes the corresponding image. In contrast, iCAPTCHA users were asked to classify an object in each image into a given set of categories. Based on the user responses, the images were annotated and then used as test images. However, the generated test images were not verified as being difficult for the computers to identify and, at the same time, easy for humans. As a result, these CAPTCHAs have not been widely used for practical applications.

To develop practical CAPTCHAs, researchers have started to use image recognition algorithms to generate test images that are difficult for computers to identify. Gossweiler et al. [16] and Kim et al. [26] proposed similar CAPTCHAs that ask the users to identify the upright orientation of randomly rotated images. As they used images whose orientation is not confidently detected by automatic orientation detectors as test images, their CAPTCHAs maintained a high security level. However, some of the automatically generated test images could be difficult for humans to identify, and thus, they frequently lead to incorrect results. As a result, both CAPTCHAs required an additional process to filter out the error-prone test images.

### 2.3 Synthesis

As explained above, image-based CAPTCHAs that are useful for practical applications must themselves be able to generate double-verified test images that are difficult for computers to identify but easy for humans. Based on the investigation, AgeCAPTCHA was designed to automatically generate test images verified by both age estimation algorithms and its users. By virtue of the double verification, we expected AgeCAPTCHA to satisfy the CAPTCHA requirements and thus be secure and easy enough for practical use. In the next section, we describe the design and implementation of AgeCAPTCHA in detail.

## 3. AgeCAPTCHA: System Architecture

### 3.1 AgeCAPTCHA Input

AgeCAPTCHA can use all types of publicly available images, which allow others to remix, tweak, or build upon the original work, as input (e.g., Flickr images with Creative Commons Licenses[5]). From the input images, AgeCAPTCHA extracts age-indeterminate faces by using

[5] http://www.creativecommons.org

state-of-the-art face detection and age estimation algorithms (the detailed extraction method will be presented in Section 4).

Then, the extracted age-indeterminate faces are automatically cropped from the original images for two reasons. One is to make AgeCAPTCHA challenges more easily integrated into a webpage because the cropped images take up less screen space. The other is to minimize the exposure of additional image features that attackers might use to identify the age group of the age-indeterminate faces (e.g., body shape, clothing, and hairstyle). As there is no standard method for cropping the facial region, we decided to make the cropped facial images look similar to those in the FG-NET aging database[6], which is a standard facial image database for evaluating the performance of the age estimation algorithms.

The detailed process of image cropping is as follows: First, we make a virtual square that contains only the facial region by using a state-of-the-art face detector, as shown in **Fig. 1a**. Second, based on the length of each side of the square, we extend the image area by 50% from the top and bottom of the square and by 20% from the left and right. Thus, the square has been extended into a rectangle (**Fig. 1b**). Finally, the rectangle that contains the age-indeterminate face is cropped from the source image and presented to users during the AgeCAPTCHA challenges.

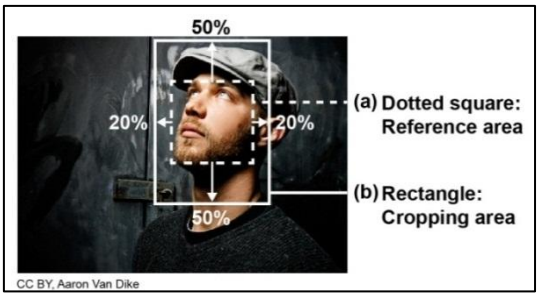


**Fig. 1.** Example of age-indeterminate face and cropped facial region

## 3.2 AgeCAPTCHA Design

### 3.2.1 Defining the Response Categories (Buttons)

When an AgeCAPTCHA challenge begins, age-indeterminate faces are sequentially displayed to a user along with predefined response categories. Then, the user is asked to identify the age group of each face by selecting one of the response categories. At this time, due to the limitations of existing face detection algorithms, various false positives are included in the extracted age-indeterminate faces. Hence, for the task of identifying the age, three different types of response categories are necessary.

The first type is used to annotate real human faces with their age groups, which can be formed by dividing the age into several ranges without the omission of any age range. The number of age groups and the age range of each group can be changed according to whichever age

[6] http://www.fgnet.rsunit.com

classification scheme is applied. In this study, we adopted the verified age classification scheme that has been used to annotate the MIRFLICKR-25000 image collection[7], which is one of the most widely used image databases for research purposes. As a result, five response categories were determined, namely, Baby, Child, Teenager, Adult, and Elderly.

At this point, the age ranges of the Adult and Elderly categories were more than several decades, whereas those of Baby, Child, and Teenager were less than a decade. When the age ranges are narrow, it can be difficult for users to select one of the response categories. Therefore, it was necessary to clearly present the criteria for classifying the age groups. One of the usual ways to clarify the criteria was to state the actual age ranges of each age group, such as Baby (0-2 years) or Child (3-12 years). However, in many Asian countries, the aging scheme is different from that in Western countries. In Korea, for example, on the day that a baby is born, its age is one. This small difference could have a large influence on the selection of the response button, particularly in the Baby and Child categories, for which the age ranges are very narrow. Therefore, instead of age ranges, we decided to use the school system as a standard of classification. Accordingly, each age group was presented as follows: Baby, Child (kindergartners and elementary school students), Teenager (middle and high school students), Adult, and Elderly.

The second type of response category is required to identify the false positives. There are two types of false positives to consider. One type consists of images that do not contain any human-related information, such as images of mannequins, animals, buildings, vehicles, and so on. The other type consists of images that contain human body parts except for faces. To filter these false positives, we added Not Human and Body Part to our response categories. In particular, Body Part was also included in the 2010 version of the list of visual concepts for image annotation[8], which is used to annotate the MIRFLICKR-25000 image collection, and hence, the need for that category underwent verification once again.


**Fig. 2.** Interface of AgeCAPTCHA

[7] http://press.liacs.nl/mirflickr
[8] http://imageclef.org/system/files/concepts.txt

Finally, the age groups of some of the human face images can be difficult for humans to identify. Moreover, the users sometimes cannot even be sure whether an image contains a human face due to poor image quality or occlusion. Therefore, the third type of response category is required to give users an opportunity to skip the unidentifiable images. We named such a category Not Sure. As a result, a total of eight response categories (Baby, Child, Teenager, Adult, Elderly, Not Human, Body Part, and Not Sure) were defined for the task of identifying age-indeterminate faces and were presented to users during AgeCAPTCHA challenges (**Fig. 2**).

### 3.2.2 The Composition of a Single AgeCAPTCHA Challenge

Following the approach of von Ahn et al. [7], two different types of images with age-indeterminate faces are used in a single AgeCAPTCHA challenge. The first type consists of images for which we already know the correct responses. These are called the test images, and they play a role in discriminating between humans and computers. By identifying the age group of all of the test images correctly, users can be confirmed as humans. Considering the attack cost, the probability of passing a CAPTCHA challenge by random guessing should be less than 1 in 10,000 [17]. With respect to AgeCAPTCHA, this requirement can be expressed as $(1 / c^n) \leq 0.0001$, where $c$ is the number of valid response categories and $n$ is the number of test images used in a single AgeCAPTCHA challenge. Thus, once $c$ is given, the smallest natural number that satisfies the expression is taken to be the minimum number of test images.

The second type of images with age-indeterminate faces is directly extracted from publicly available images. Therefore, the annotation results (age groups) for these facial images are not known in advance. These are called candidate images, and they are not used to influence the user's success in solving AgeCAPTCHA challenges; instead, they are used to gather the user responses. After most of the user responses, the candidate images are annotated with one of the eight response categories. Later, some of the candidate images that can be consistently identified by the users are converted to test images. In each AgeCAPTCHA challenge, at least one candidate image is presented to a user without informing the user that it is a candidate image.

The following are the details of the process behind an AgeCAPTCHA challenge:

(1) When an AgeCAPTCHA challenge begins, a test image is randomly selected and then presented to a user along with the eight response categories. The user is asked to identify the test image by selecting one of the response categories.

(2) Whenever the user correctly identifies a test image, the next test image is presented. In contrast, if the user fails to give the correct response, AgeCAPTCHA terminates the challenge and presents another challenge with a new set of test images. This is required to minimize the exposure of the test images. The timing of the terminations should be carefully determined, because a bot can attempt to collect test images from the test image database. If AgeCAPTCHA is designed to terminate its challenge immediately when a user fails to identify a test image correctly, the appearance of the next test image reveals that the user gave the correct response. In this case, the bot can easily find out the answer to the test image. For this reason, even if a user fails to give the correct response, the challenge should not be terminated until at least the third or fourth test image has been classified by the user.

(3) During a challenge, a candidate image is randomly presented either last or second-last in order. This arrangement minimizes the exposure of the candidate images, to allow them to later be used as test images.

(4) The AgeCAPTCHA challenge is completed when all of the test images are correctly identified. Subsequently, a completion message is displayed to the user.

(5) Based on the user responses, each candidate image is annotated with one of the eight response categories. Some qualified candidate images are selected as test images. Conversely, if an existing test image induces frequent errors, AgeCAPTCHA filtered it out from the test image database.

### 3.3 AgeCAPTCHA Output

The output of AgeCAPTCHA is as follows:

(1) **Differentiating between humans and computers:** By correctly identifying all of the test images presented in an AgeCAPTCHA challenge, the users can be confirmed as human.

(2) **Annotating candidate images:** After most of the user responses, the candidate images are annotated with one of the eight response categories (Baby, Child, Teenager, Adult, Elderly, Not Human, Body Part, or Not Sure), which can be used to facilitate the image retrieval (e.g., searching for images that include a baby) and to train face detection and age estimation algorithms to increase their performance. In addition, by identifying the false positives (candidate images annotated as Not Human or Body Part), the number of human faces in each image can be confirmed. This is a powerful filter to narrow the search space (e.g., searching for images that contain 3–6 people) as well as one of the most useful cues for image retrieval [12].

(3) **Generating verified test images:** Some candidate images that users consistently identify are selected and then used in later AgeCAPTCHA challenges as test images.

## 4. User Study 1: Establishing Criteria for Operating AgeCAPTCHA

In this section, we describe an experiment conducted with three objectives to determine specific criteria for operating AgeCAPTCHA.

First, a total of eight response categories (buttons) are determined in the previous section. Each category can be selected by users during the AgeCAPTCHA challenges and thus is valid for image annotation purposes. However, there can be a set of categories often confused with one another, which means that these categories are not readily distinguished and can significantly decrease the human success rate of AgeCAPTCHA. To increase the human success rate, these categories should be considered to be a single category. Therefore, by measuring the consistency of the responses in each category, we aimed to identify which categories are valid for CAPTCHA purposes. Through this process, the minimum number of test images that should be used in a single AgeCAPTCHA challenge was determined.

Second, AgeCAPTCHA was designed to annotate age-indeterminate faces (candidate images) with one of the eight response categories based on the predominant type of response by users. As there was a trade-off between the speed and reliability of the image annotation, it was necessary

to determine the minimum number of user responses needed to reliably annotate the candidate images. In addition, we attempted to establish a criterion for selecting candidate images that users consistently identify, which we can use as test images in later AgeCAPTCHA challenges.

Finally, CAPTCHAs that take too much time to complete are difficult to be used in practice [10]. Therefore, by measuring the time required to respond to each age-indeterminate face, we wanted to test the feasibility of AgeCAPTCHA.

## 4.1 Experiment Design

### 4.1.1 Subjects

We recruited a total of 528 South Korean subjects (331 males and 197 females) through an Internet advertisement. Their ages ranged between 14 and 79 years, with an average age of 26.18 years (SD = 5.92 years).

### 4.1.2 Materials

As raw material, we used the MIRFLICKR-25000 image collection. This collection was one of the most widely used image databases for research purposes and, thus, had proven its suitability for the intended use. In addition, as the database consisted of 25,000 Flickr images supplied under the Creative Commons License by 9,862 Flickr users, which included human faces of various races, it provided a realistic environment for the experiment. Additionally, the images could be used without any concern about permission. Thus, we decided to extract age-indeterminate faces from the MIRLICKR-25000 image collection.

In principle, age-indeterminate faces should be extracted using multiple face detection and age estimation algorithms because the robustness of AgeCAPTCHA to external attacks is determined by whether the extracted images are truly difficult for computers to identify. However, as the main focus of this study was on implementing AgeCAPTCHA and testing its feasibility, we used a single face detection and age estimation algorithm for convenience in the implementation and experiments. After due consideration, we decided to use the face recognition engine of Face.com[9].

With respect to convenience, the Face.com engine could perform both face detection and age estimation. In addition, it offered an application program interface (API) that enabled us to set up our experiment on the Internet and have stable access to a large number of subjects. In terms of the performance, this engine was state-of-the-art in face detection [27]. When we applied the engine to the FG-NET aging database (1,002 facial images), 985 faces (98.3%) were successfully detected, and the mean absolute error (MAE)[10] for the age estimation was 5.51 years, which was superior to the human ability (MAE = 6.28 years) for controlled frontal face images [13].

When the Face.com engine was applied to facial images, it returned a degree of confidence that was in the range from 0% to 100% with respect to whether a detected face was truly human. In addition, the engine presented an estimated age, maximum age, and minimum age for each face. By using the four properties, we defined age-indeterminate faces as those for which the

[9] http://developers.face.com

[10] MAE is defined as the average of the absolute errors between the estimated values and the ground-truth values.

degree of confidence in face detection was greater than 50% but for which all the age-related properties were unidentified. As a result, a total of 1,761 age-indeterminate faces were extracted from the MIRLICKR-25000 image collection and then were automatically cropped from the original images according to the aforementioned method. We used 1,000 images in this experiment and left 761 images for the next experiment.

#### 4.1.3 Implementation and Procedure

We built our experimental website using HTML, PHP, and JavaScript. To participate in this experiment, our subjects connected to the website at their own location (e.g., home or office). An introduction to the experiment was presented on the main page of the website. After reading this introduction, the subjects were asked to input their personal information (gender and year of birth), and then, a start button was displayed on the screen. When the subject clicked the start button, the experiment began.

Once the experiment had been started, an age-indeterminate face was randomly selected from the 1,000 prepared images, and it was displayed on the screen with the eight response categories, as shown in **Fig. 2**. Then, the subject was asked to identify the age group of the displayed face by selecting one of the response categories. Whenever the subject selected a response category, the response and response time were both recorded, and another age-indeterminate face was displayed on the screen. Once this task had been repeated 30 times, the experiment was terminated.

### 4.2 Results and Discussion

#### 4.2.1 Criteria for Annotating Age-indeterminate Faces

As some subjects quit the experiment before completing the whole trial (30 tasks), a total of 15,611 responses were collected. As a result, the 1,000 age-indeterminate faces were presented to the subjects a minimum of 14 times and a maximum of 23 times, for an average of 15.61 times. Based on the responses of our subjects, the 1,000 age-indeterminate faces were annotated by a majority vote. However, if there was no majority, the image was annotated as Not Sure. The annotation results are listed in **Table 1**, and **Fig. 3** shows examples of the images.

**Table 1.** Annotation results of the 1,000 age-indeterminate faces

| Annotation results | Number of images | Types of images |
|---|---|---|
| Baby | 16 | Human faces (77.6%) |
| Child | 48 | |
| Teenager | 56 | |
| Adult | 615 | |
| Elderly | 41 | |
| Body Part | 26 | False positives (19.0%) |
| Not Human | 164 | |
| Not Sure | 34 | Unidentified (3.4%) |
| **Total** | **1000** | **100%** |


**Fig. 3.** Examples of the annotated age-indeterminate faces

Most of the age-indeterminate faces (776 images) were human faces (Baby, Child, Teenager, Adult, and Elderly), whereas Not Human and Body Part images (false positives) were relatively few. This was expected because the age-indeterminate faces were extracted by the state-of-the-art face recognition engine. Only 34 images could not be identified by our subjects and thus were annotated as Not Sure. Among the human faces, the Adult faces were the largest in number, followed by Teenager, Child, Elderly, and Baby. However, as there was no ground truth for the age-indeterminate faces, it was difficult to verify the accuracy of the annotation results. Alternatively, we evaluated the annotation reliability by measuring the consistency of the responses to each image. **Table 2** summarizes the results.

**Table 2.** Response consistency for the 1,000 age-indeterminate faces

| | Number of inconsistent responses | | | | | |
|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **≥ 5** |
| **Number of images** | 217 | 177 | 131 | 92 | 78 | 305 |
| **Cumulative total (%)** | 21.7 | 39.4 | 52.5 | 61.7 | 69.5 | 100 |

The average response consistency for the 1,000 age-indeterminate faces was 80.19% (12,520 from a total of 15,611 responses). More specifically, for almost 70% of the age-indeterminate faces (695 images), the number of inconsistent responses was less than or equal to four. In particular, more than half of the age-indeterminate faces (525 images) received fewer than or equal to two inconsistent responses. Considering that we collected 15.61 responses per image on average, we concluded that most of the age-indeterminate faces were reliably annotated.

For convenience, we called this annotation result, which is based on the average of 15.61 responses, the standard annotation. Then, to determine the minimum number of responses required to reliably annotate each age-indeterminate face, we reduced the number of responses per image by decrements of one. Based on the smaller number of responses, we annotated each image again and compared the annotation result with the standard annotation. The results are listed in **Table 3**. The concordance rate here refers to the proportion of the images that received their standard annotation.

**Table 3.** Changes in the number of age-indeterminate faces that received their standard annotations and the number of images annotated as Not Sure

| | Number of responses | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Concordance rate (%)** | 69.5 | 88.0 | 84.5 | 89.1 | 89.7 | 92.3 | 91.3 | 94.1 | 93.2 |
| **Number of images annotated as Not Sure** | 307 | 55 | 112 | 59 | 72 | 42 | 60 | 33 | 53 |

As the number of responses increased, the rate of concordance also tended to increase. However, in some cases, when the number of responses was an even number, the concordance rate was lower than for the previous odd number of responses. This trend occurred because any image that had no majority response was annotated as Not Sure. **Table 3** shows that the concordance rate was affected by the changes in the number of Not Sure images. In general, when the images were annotated based on at least five responses, a concordance rate of over 89.1% was achieved.

### 4.2.2 Valid Response Categories for Passing AgeCAPTCHA Challenges

As mentioned above, it was necessary to identify which response categories are valid for CAPTCHA purposes, except for the Not Sure category to skip unidentifiable images. Hence, we examined the response consistency for each category, as shown in **Table 4**. The response consistencies for the images annotated as Baby, Child, Adult, Elderly, Body Part, and Not Human ranged from 71.5% to 86.06% (darkly shaded columns), which greatly exceeding the percentages of the respective second most frequent response categories (lightly shaded columns). As a result, those six response categories were expected to be useful independently for both image annotation and CAPTCHA purposes.

**Table 4.** Response consistency for each category

| Annotation results | Number of responses in each category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Baby** | **Child** | **Teenager** | **Adult** | **Elderly** | **Body Part** | **Not Human** | **Not Sure** |
| **Baby** | 208 (85.53%) | 32 (12.85%) | 0 | 2 (0.8%) | 0 | 3 (1.2%) | 3 (1.2%) | 1 (0.4%) |
| **Child** | 51 (6.85%) | 574 (77.15%) | 78 (10.48%) | 20 (2.69%) | 3 (0.4%) | 0 | 3 (0.4%) | 15 (2.02%) |
| **Teenager** | 0 | 61 (0.706%) | 528 (61.11%) | 234 (27.08%) | 0 | 11 (1.27%) | 8 (0.93%) | 22 (2.55%) |
| **Adult** | 2 (0.02%) | 30 (0.31%) | 866 (9.02%) | 7964 (82.98%) | 262 (2.73%) | 32 (0.33%) | 103 (1.07%) | 338 (3.52%) |
| **Elderly** | 0 | 1 (0.16%) | 0 | 104 (16.12%) | 508 (78.76%) | 1 (0.16%) | 25 (3.88%) | 6 (0.93%) |
| **Body Part** | 0 | 5 (1.23%) | 9 (2.21%) | 38 (9.34%) | 0 | 291 (71.5%) | 39 (9.58%) | 25 (6.14%) |
| **Not Human** | 22 (0.86%) | 37 (1.44%) | 16 (0.62%) | 79 (3.08%) | 15 (0.58%) | 45 (1.75%) | 2210 (86.06%) | 144 (5.61%) |
| **Not Sure** | 1 (0.19%) | 14 (2.61%) | 68 (12.66%) | 157 (29.24%) | 40 (7.45%) | 23 (4.28%) | 68 (12.66%) | 166 (30.91%) |

In contrast, the response consistency for the Teenager images was the lowest (61.11%), and 27.08% of the responses were Adult. This finding indicates that the Teenager category was often confused with the Adult category. Therefore, it appears to be reasonable to conclude that the Teenager category should not be used independently as a valid response category for CAPTCHA purposes. Instead, considering that the second most frequent response to the Adult images was Teenager (9.02%), we decided to unify Teenager and Adult into a single response category named ‘Teenager or Adult’. However, each still existed independently for image annotation purposes (**Fig. 4**).

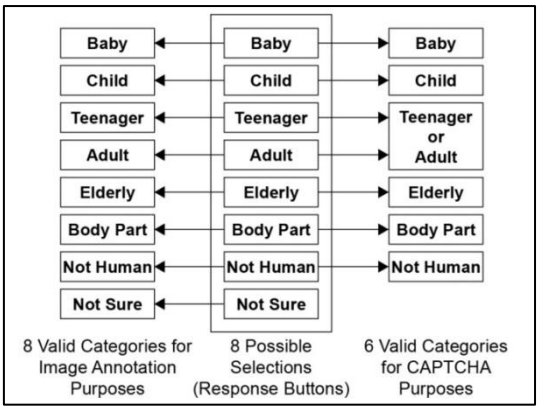


**Fig. 4.** Valid response categories for image annotation and CAPTCHA purposes

In summary, for CAPTCHA purposes, Baby, Child, 'Teenager or Adult', Elderly, Body Part, and Not Human were identified as valid response categories. As there were six valid response categories, at least six test images were required for a single AgeCAPTCHA challenge. With six test images and six valid response categories, the probability of passing an AgeCAPTCHA challenge by random guessing is 1 in 46,656 (0.002%), which clearly exceeds the required security standard (0.01%) [17]. As a result, the minimum AgeCAPTCHA challenge is composed of seven age-indeterminate faces (six test images and one candidate image).

### 4.2.3 Criteria for Generating Test Images

To be used in practice, the human success rate for a CAPTCHA challenge should approach at least 90% [17]. For AgeCAPTCHA, the criterion can be expressed as $p^n \geq r$, where $p$ is the probability that a user will identify a test image correctly, $n$ is the number of test images used for each AgeCAPTCHA challenge, and $r$ is the human success rate. We set $n$ to 6 and $r$ to 0.9. As a result, $p$ is required to be greater than 0.9826, which means that the response consistency for age-indeterminate faces in test images should be at least 98.26%.

However, in practice, it was difficult to spend enough time to obtain sufficient user responses to determine whether each age-indeterminate face could be consistently identified by users. Therefore, it was necessary to establish an efficient criterion that enables us to select age-indeterminate faces suitable for test images based on as few responses as possible. To solve this problem, we investigated the relationship between the number of consecutive identical responses (from the first response) and the average response consistency for the selected images. **Table 5** gives the results.

**Table 5.** Changes in the average response consistency for the age-indeterminate faces selected as test images

| | Number of consecutive identical responses | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Number of images selected** | 724 | 668 | 628 | 586 | 562 | 536 | 514 | 498 |
| **Average response** | 94.15% | 95.65% | 96.51% | 97.41% | 97.87% | 98.3% | 98.7% | 98.94% |

| **consistency** | (10,632/ 11,293) | (9,964/ 10,417) | (9,446/ 9,788) | (8,903/ 9,140) | (8,580/ 8,767) | (8,216/ 8,358) | (7,910/ 8,014) | (7,681/ 7,763) |
|---|---|---|---|---|---|---|---|---|
| **Anticipated human success rate for 7-image AgeCAPTCHA** | 69.64% | 76.59% | 80.78% | 85.42% | 87.87% | 90.23% | 92.46% | 93.83% |

When we selected the images of age-indeterminate faces that obtained at least eight consecutive identical responses, 536 images were selected, and their response consistency was 98.3% (lightly shaded columns). Since the human success rate for a 7-image AgeCAPTCHA challenge was expected to exceed 90%, eight consecutive identical responses were regarded as the minimum requirement for generating test images. However, to achieve a stable human success rate of greater than 90%, we decided to increase the required number of consecutive identical responses from eight to nine. By this new criterion, 514 images were selected as test images, and the anticipated human success rate was increased to 92.46% (darkly shaded columns). Hereafter, we call the 514 images Group A, and the other 486 images Group B.

For Group A, the number of test images in each valid response category was proportional to the number of annotated images (**Table 1)**. Accordingly, the 'Teenager or Adult' test images were the most prevalent (413 images), followed by Not Human (73 images), Child (13 images), Elderly (11 images), and Baby (4 images). None of the Body Part images were selected as test images. Due to this imbalance in the ratios of the generated test images, there was a possibility of breaking AgeCAPTCHA challenges by selecting the most frequent test image type. To maintain the security of AgeCAPTCHA, properly inserting the rarer types of test images into any single AgeCAPTCHA challenge should be considered.

In summary, the criteria determined for generating test images were as follows:

(1) Each candidate image is displayed as many as nine times.

(2) Candidate images that receive nine consecutive identical responses are converted into test images.

(3) If the user responses become inconsistent before a candidate image has been displayed nine times, the candidate image is displayed as many times as necessary for reliable annotation but cannot be used as a test image.

### 4.2.4 Response Time

As this was an uncontrolled experiment, there were several outliers (e.g., when a subject chatted with a friend during the experiment). Therefore, we calculated the median response time (RT) for each age-indeterminate face, because this measure is robust against outliers. The average median RT for the 1,000 age-indeterminate faces was 2.36 s (SD = 0.66 s). More specifically, the average median RT for Group A was 2.18 s (SD = 0.57 s), and that for Group B was 2.55 s (SD = 0.69 s). The RT difference between the groups was statistically significant ($p < 0.001$, Student's t-test). Based on these results, we can anticipate that, on average, users would spend less than 16 s to complete a 7-image AgeCAPTCHA challenge.

Within Group A, which can be used as test images, the Not Human images (false positives) had an RT significantly shorter than that for the human faces ($p < 0.01$, Student's t-test), as shown in **Table 6**. Then, we used the Kruskal-Wallis test to investigate the RT differences

between the types (response categories) of Group A images. As statistical significance was demonstrated by a value of $p < 0.001$, the Mann-Whitney test was subsequently applied to investigate the significance of the RT difference between each type of human face and the Not Human images.

**Table 6.** Average median response time for each type of Group A image

| Types of images | | Average median response time (s) | |
|---|---|---|---|
| Human faces | Baby | 2.01 (SD = 0.53) | 2.21 (SD = 0.57) |
| | Child | 2.12 (SD = 0.34) | |
| | Teenager or Adult | 2.23 (SD = 0.58) | |
| | Elderly | 1.47 (SD = 0.13) | |
| False positive | Not Human | 2.00 (SD = 0.5) | |
| **Total** | | **2.18 (SD = 0.57)** | |

The RT differences between the Baby, Child, and Not Human images were not significant. However, as the RT for the 'Teenager or Adult' images was significantly longer than that for the Not Human images ($p < 0.001$), we reconfirmed that it was difficult for the subjects to distinguish between the 'Teenager' and 'Adult' categories. In contrast, the RT for the Elderly images was significantly shorter than that for the Not Human images ($p < 0.001$). This finding could be interpreted that the Elderly images were much easier for the subjects to identify than the other categories. Accordingly, the use of Elderly test images was expected to reduce the time required to complete the AgeCAPTCHA challenges.

## 5. User Study 2: Usability Testing of AgeCAPTCHA

Once our first experiment determined the criteria for operating AgeCAPTCHA, we conducted a second experiment to evaluate the usability of AgeCAPTCHA. More specifically, we wanted to measure the human success rate as well as the time required to complete a single 7-image AgeCAPTCHA challenge. We also attempted to determine whether AgeCAPTCHA can reliably annotate candidate images and then generate new test images.

### 5.1 Experiment Design

#### 5.1.1 Subjects

We recruited a total of 267 South Korean subjects (131 males and 136 females) through an Internet advertisement. Their ages ranged between 18 and 64 years, with an average age of 26.16 years (SD = 8.96 years).

#### 5.1.2 Materials

As mentioned in Section 4.1.2, 1,761 images of age-indeterminate faces were extracted from the MIRFLICKR-25000 image collection. In the previous experiment, we used 1,000 images and left 761 images. In this experiment, the remaining 761 images were used as candidate images, and Group A (514 images) was used as test images.

### 5.1.3 Implementation and Procedure

The experimental task was to complete a 7-image AgeCAPTCHA challenge three times. To accomplish this goal, similar to the previous experiment, we built an experimental website and asked our subjects to connect to the website at their own location. After reading an introduction to this experiment and providing the required information (gender and year of birth), the subjects participated in the experiment.

During the experiment, each age-indeterminate face was displayed individually, with the eight response categories shown in **Fig. 2**. Then, one of the response categories was selected according to how the subject identified each image. However, when a subject found it difficult to identify an age-indeterminate face, the Not Sure category could be selected. If the unidentifiable image was a test image, it was skipped and replaced with another test image. Conversely, if it was a candidate image, the response was considered to be valid and was used to annotate the candidate image.

In each AgeCAPTCHA challenge, the first six images were test images, and a candidate image was displayed at the end. To balance the test image ratios, the test images were uniformly selected from each valid response category with no overlaps. In contrast, candidate images were presented in a predefined order to elicit the minimum number of responses needed to annotate the images and to select the test images as quickly as possible. A candidate image was not presented to a subject more than once.

In practice, to minimize the exposure of the test images, AgeCAPTCHA should stop the challenge when a user does not correctly respond to a test image. However, as we wanted to analyze the errors that can occur during AgeCAPTCHA challenges, the challenge was not stopped, regardless of how the subject had responded. After each AgeCAPTCHA challenge was finished, the time taken by the subject to finish each challenge and the result (pass or fail) were recorded. Then, a completion message and a button labeled Next were displayed. The next challenge was not started until the subject pressed the Next button. This step was to focus the subject's concentration and to accurately measure the completion time.

## 5.2 Results and Discussion

### 5.2.1 Completion Time

In this experiment, 267 subjects finished 797 AgeCAPTCHA challenges, with an average of 2.99 challenges per subject. As this was also an uncontrolled experiment, there are several outliers that make it difficult to measure the average completion time. Alternatively, we calculated the median completion time, which was 15.75 s (interquartile range 11.32 s to 20.18 s). This is longer than for an ordinary 7-letter text-based CAPTCHA, which takes 13.51 s on average [7]. However, considering that the Microsoft's Asirra CAPTCHA [18] takes approximately 15 s to complete and had been used in practice, we concluded that the completion time of the 7-image AgeCAPTCHA challenges was acceptable for practical use.

### 5.2.2 Human Success Rate

As our subjects failed to pass 94 of the 797 challenges, the human success rate for a 7-image AgeCAPTCHA challenge was 88.21%, which is close to the recommended criterion (90%) [17].

For a total of three challenges per subject, 191 subjects (71.54%) passed all three challenges, 60 subjects (22.47%) passed two challenges, and 14 subjects (5.24%) passed one challenge. Only two subjects could not pass any of the challenges. This means that 99.25% of our subjects could pass an AgeCAPTCHA challenge in three attempts.

The result was similar to that for the Asirra CAPTCHA, since 99.5% of their subjects were able to pass an Asirra challenge in three attempts [18]. This was very encouraging because we achieved a similar human success rate with the specialized image-based CAPTCHA that was used in practice, annotating image content that was difficult to separate into mutually exclusive categories (age groups). From these results, we could expect our AgeCAPTCHA to be used in practice as well as verify the effectiveness of the proposed two-layered response structure.

In the 94 failed challenges, a total of 105 incorrect responses (errors) occurred, as shown in **Table 7**. Among these, 52 errors (49.52%) were induced by the Child test images. In particular, our subjects seemed to have trouble distinguishing between the Baby and Child categories, because the Child test images were frequently misclassified as Baby.

**Table 7.** Number of incorrect responses for each type of test image

| Types of test images | Number of incorrect responses in each category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Baby** | **Child** | **Teenager** | **Adult** | **Elderly** | **Not Human** | **Body Part** | **Total** |
| Baby | - | 7 | 0 | 1 | 0 | 0 | 0 | 8 |
| Child | 31 | - | 14 | 5 | 0 | 1 | 1 | 52 |
| Teenager or Adult | 1 | 6 | - | | 5 | 4 | 0 | 16 |
| Elderly | 0 | 0 | 0 | 19 | - | 1 | 0 | 20 |
| Not Human | 2 | 1 | 1 | 2 | 1 | - | 2 | 9 |

To solve this problem and increase the human success rate of AgeCAPTCHA, two different approaches could have been considered. One approach was to increase the number of consecutive identical responses for generating the Child test images, which can make it more difficult for error-prone test images to be selected as test images. In this approach, the test image generation speed was decreased, since it was necessary to obtain more user responses. However, it was also possible to selectively reduce the required number of consecutive identical responses for several valid response categories that could be identified consistently by the users (e.g., Not Human). Therefore, we expected to increase the human success rate of AgeCAPTCHA while maintaining the speed of the test image generation.

The other approach was to consider Baby and Child as a single response category (‘Baby or Child’) for responding to AgeCAPTCHA challenges but using these independently for image annotation. Accordingly, the number of valid response categories was reduced from six to five (‘Baby or Child’, ‘Teenager or Adult’, Elderly, Not Human, and Body Part). For six test images with five valid responses, the probability of passing an AgeCAPTCHA challenge by random guessing was reduced to 1 in 15,625. However, this still fulfills the recommended security criterion (less than 1 in 10,000) [17]. When we recalculated the human success rate, assuming that there were five valid response categories, a total of 67 incorrect responses occurred in the 64 AgeCAPTCHA challenges. As a result, the human success rate was increased from 88.2% to 91.97%, which is enough for practical use.

### 5.2.3 Image Annotation and Test Image Generation

During the 797 AgeCAPTCHA challenges, a total of 107 candidate images were presented to the subjects. Of these, 100 images were annotated with one of the eight response categories based on at least 5 different subjects' responses, whereas the other 7 images were not annotated simply because the experiment was terminated before each image obtained five responses. Similar to in the previous experiment, most of the annotated images were human faces (**Table 8**). Specifically, the number of Adult images was the largest, since those have the widest age range. As the average response consistency of the 100 candidate images was 84.04% and only 6 of the images were annotated as Not Sure, we have reconfirmed the reliability and usefulness of AgeCAPTCHA in image annotation.

More than half of the annotated candidate images (53 images) obtained nine consecutive identical responses and could be used as test images in later AgeCAPTCHA challenges (**Table 8**). **Fig. 5** shows examples of the new test images generated by AgeCAPTCHA. Overall, as the number of test images in each category was proportional to the number of annotated images, the 'Teenager or Adult' test images were the largest in number, followed by Not Human, Child, and Baby images. In contrast, Elderly and Body Part test images were not generated. As a result, the imbalance in the generated test image ratios was continuously observed.

**Table 8.** Numbers of annotated images and generated test images in each response category

| Annotation results | Number of candidate images annotated | Number of test images generated |
|---|---|---|
| **Baby** | 1 | 1 |
| **Child** | 5 | 2 |
| **Teenager** | 9 | 42 (Teenager or Adult) |
| **Adult** | 56 | |
| **Elderly** | 2 | 0 |
| **Body Part** | 3 | 0 |
| **Not Human** | 18 | 8 |
| **Not Sure** | 6 | - |
| **Total** | **100** | **53** |

Although the number of test images in each category could be changed according to the raw materials (images) used to extract the candidate images, it was once again noticeable that none of the images annotated as Body Part were selected as test images. In our experiments, although the number of Body Part images was quite small, we could not completely deny the existence of the Body Part test images, and this matter should be further investigated. However, even in the worst case, if there were no Body Part test images, then the probability of passing an AgeCAPTCHA challenge by random guessing would be 1 in 15,625 (0.006%), which still fulfills the security requirement [17].


**Fig. 5.** Examples of the new test images generated by AgeCAPTCHA

## 6. Conclusions and Future Work

In this study, we have proposed a novel method that uses CAPTCHAs in image annotation. Our method can differentiate between humans and computers, annotate even image content that is difficult to separate into mutually exclusive categories, and generate verified test images suitable for CAPTCHA challenges. For an evaluation, we used images of faces whose ages cannot be automatically estimated by existing age estimation algorithms. We applied our method to annotate these facial images with their age groups and then conducted user studies. The results were very encouraging because our CAPTCHA annotated facial images with a high degree of reliability. At the same time, the process was completed by the subjects quickly and accurately enough for practical use. As a result, we not only verified the effectiveness of our method but also increased the applicability of CAPTCHAs for image annotation.

In addition, many relevance feedback methods have been developed in recent years, becoming widely used in relation to content-based image retrieval (CBIR) [28]. In a relevance feedback process, a user at the outset takes the image retrieval results initially returned from a given query and labels relevant images as positive-feedback samples and irrelevant images as negative-feedback samples. Then, based on the labeled feedback samples, CBIR systems refine all of the image retrieval results, iteratively. However, the performance of relevance feedback algorithms can be poor when the number of feedback samples is small. To solve this problem, many attempts have been made to embed a variety of different relevance feedback algorithms and to realize further improvements of the image retrieval performance [28-31]. Given this situation, an extension of this study will involve the use of the proposed method to annotate other types of information presented in images. Through this process, we expect to generate a large number of annotated images that have sufficiently high quality to be used as positive-feedback samples and thus contribute to the multimedia research field.

## References

[1] T. Xinmei and T. Dacheng, "Visual reranking: From objectives to strategies," *IEEE Multimedia*, vol. 18, no. 3, pp. 12-21, March, 2011. Article (CrossRef Link).

[2] B. Sigurbjörnsson and R.v. Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. of 17th Int. Conf. on Wrold Wide Web,* pp. 327-336, 2008. Article (CrossRef Link).

[3] M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," in *Proc. of SIGCHI Conf. on Human Factors in Computing Systems,* pp. 971-980, 2007. Article (CrossRef Link).

[4] N. Kumar, A. Berg, P.N. Belhumeur and S. Nayar, "Describable visual attributes for face verification and image

search," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 10, pp. 1962-1977, March 10, 2011. Article (CrossRef Link).
[5] D. Morrison, S. Marchand-Maillet and É. Bruno, "TagCAPTCHA: annotating images with CAPTCHAs," in *Proc. of ACM SIGKDD Workshop on Human Computation,* pp. 44-45, 2009. Article (CrossRef Link).
[6] R.A. Khot and K. Srinathan, "iCAPTCHA: Image tagging for free," in *Proc. of 3rd Int. Conf. on Usable Software and Interface Design, pp. 1-6, 2009.*
[7] L. von Ahn, B. Maurer, C. McMillen, D. Abraham and M. Blum, "reCAPTCHA: human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465-1468, September 12, 2008. Article (CrossRef Link).
[8] M. Chew and J. Tygar, "Image recognition CAPTCHAs," in *Proc. of 7th Int. Information Security Conf.,* pp. 268-279, 2004. Article (CrossRef Link).
[9] K. Chellapilla, K. Larson, P.Y. Simard and M. Czerwinski, "Computers beat humans at single character recognition in reading based human interaction proofs," in *Proc. of 2nd Conf. on Email and and Anti-Spam*, 2005.
[10] Y. Rui and Z. Liu, "ARTiFACIAL: Automated reverse Turing test using FACIAL features," *Multimedia Systems*, vol. 9, no. 6, pp. 493-502, June, 2004. Article (CrossRef Link).
[11] M. Toews and T. Arbel, "Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1567-1581, September, 2009. Article (CrossRef Link).
[12] N. Naaman, S. Harada, Q.Y. Wang, H. Garcia-Molina and A. Paepck, "Context data in geo-referenced digital photo collections," in *Proc. of 12th ACM Int. Conf. on Multimedia,* pp. 196-203, 2004. Article (CrossRef Link).
[13] X. Geng, Z.H. Zhou and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2234-2240, December, 2007. Article (CrossRef Link).
[14] F. Alnajar, C. Shan, T. Gevers and J.M. Geusebroek, "Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions," *Image and Vision Computing*, vol. 30, no. 12, pp. 946-953, December, 2012. Article (CrossRef Link).
[15] J. Ylioinas, A. Hadid and M. Pietikainen, "Age classification in unconstrained conditions using LBP variants," in *Proc. of Int. Conf. on Pattern Recognition*, pp. 1257-1260, November 11-15, 2012.
[16] R. Gossweiler, M. Kamvar and S. Baluja, "What's up CAPTCHA?: a CAPTCHA based on image orientation," in *Proc. of 18th Int. Conf. on World Wide Web,* pp. 841-850, 2009. Article (CrossRef Link).
[17] K. Chellapilla, K. Larson, P. Simard and M. Czerwinski, "Designing human friendly human interaction proofs (HIPs)," in *Proc. of SIGCHI Conf. on Human Factors in Computing Systems,* pp. 711-720, 2005. Article (CrossRef Link).
[18] J. Elson, J.R. Douceur, J. Howell and J. Saul, "Asirra: a CAPTCHA that exploits interest-aligned manual image categorization," in *Proc. of 14th ACM Conf. on Computer and Communication Security,* pp. 366-374, 2007. Article (CrossRef Link).
[19] L. von Ahn, M. Blum, N.J. Hopper and J. Langford, "CAPTCHA: Using hard AI problems for security," *Adcances in Cryptology – EUROCRYPT 2003 in Lecture Notes in Computer Science*, vol. 2656, pp. 294-311, 2003. Article (CrossRef Link).
[20] S. Kalsoom, S. Ziauddin and A.R. Abbasi, "An image-based CAPTCHA scheme exploiting human appearance characteristics," *KSII Transactions on Internet and Information Systems,* vol. 6, no. 2, pp. 734-750, February, 2012. Article (CrossRef Link).
[21] R. Chow, P. Golle, M. Jakobsson, L. Wang and X.F. Wang, "Making captchas clickable," in *Proc. of 9th Workshop on Mobile Computing Systems and Applications,* pp. 91-94, 2008. Article (CrossRef Link).
[22] D. Misra and K. Gaj, "Face recognition CAPTCHAs," in *Proc. of Int. Conf. on Internet and Web Applications and Services,* pp. 122-122, February 19-25, 2006. Article (CrossRef Link).
[23] L. von Ahn, M. Blum and J. Langford, "Telling humans and computers apart automatically," *Communications of the ACM,* vol. 47, no. 2, pp. 56-60, February, 2004. Article (CrossRef Link).
[24] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. of SIGCHI Conf. on Human Factors in Computing Systems,* pp. 319-326, 2004. Article (CrossRef Link).
[25] P. Golle, "Machine learning attacks against the Asirra CAPTCHA," in *Proc of 15th ACM Conf. on Computer and Communications Security,* pp. 535-542, 2008. Article (CrossRef Link).
[26] J.W. Kim, W.K. Chung and H.G. Cho, "A new image-based CAPTCHA using the orientation of the polygonally cropped sub-images," *The Visual Computer*, vol. 26, no.6-8, pp. 1135-1143, June, 2010. Article (CrossRef Link).

[27] Z. Xiangxin and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. of IEEE Conf. on Biometric Compendium*, pp. 2879-2886, June 16-21, 2012. Article (CrossRef Link).
[28] D. Tao, X. Tang, X. Li and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1088-1099, July, 2006. Article (CrossRef Link).
[29] X. Tian, D. Tao and Y. Rui, "Sparse transfer learning for interactive video search reranking," *ACM Transactions on Multimedia Computing*, *Communications, and Applications,* vol. 8, no. 3(26), July, 2012.. Article (CrossRef Link).
[30] W. Meng, L. Hao, T. Dacheng, L. Ke and W. Xindong, " Multimodal graph-based reranking for web image search," *IEEE Transcations on Image Processing*, vol. 21, no. 11, pp. 4649-4661, November, 2012. Article (CrossRef Link).
[31] L. Weifeng and T. Dacheng, "Multiview hessian regularization for image annotation," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2676-2687, July, 2013. Article (CrossRef Link).

**Jonghak Kim** received the B.S. degree in Industrial and Media Design from Handong University and the M.S. and Ph.D. degree in Culture Technology from KAIST. He is currently a user experience (UX) specialist at Daum Communications Corp., an internet portal company in Republic of Korea. His research interests include Human-Computer Interaction (HCI) and human computation.

**Joonhyuk Yang** received the B.S. degree in Industrial and Management Engineering from POSTECH. He is currently working toward the Ph.D. degree at Graduate School of Culture Technology, KAIST. His research interests include entertainment marketing and cultural economics.

**Kwangyun Wohn** is Professor and Founding Dean of the Graduate School of Culture Technology, KAIST. Before he joined KAIST, he had been a lecturer at Harvard University and an Assistant Professor at University of Pennsylvania. At KAIST, he was in the Computer Science Department for 15 years before establishing a new graduate school dedicated to digital culture in 2005. While his research interest spans the broad range of digital culture-from theoretical aspects to practicalities-he focuses his research efforts on the application of virtual reality technology to various cultural artifacts such as museum exhibition and entertainment contents.