Sensors and Materials, Vol. 36, No. 10 (2024) 4585–4606
MYU Tokyo

S & M 3819

Mobile Augmented Reality Interface for Instruction-based Disaster Preparedness Guidelines

Sergio De León Aguilar,[1] Yuki Matsuda,[2] and Keiichi Yasumoto[1*]

[1]Graduate School of Information Science, Nara Institute of Science and Technology,
Takayama-cho 8916-5, Ikoma City, Nara 630-0192, Japan
[2]Faculty of Environmental, Life, Natural Science and Technology, Okayama University
Tsushimanaka 3-1-1, Kita-ku, Okayama City, Okayama 700-0082, Japan

(Received June 30, 2024; accepted October 21, 2024)

Keywords: guidelines, augmented reality, disaster preparedness, object recognition, user interface, knowledge transfer

Disaster preparedness guidelines help citizens protect themselves against disasters. Nonetheless, the general public has been found not to read them. Augmented reality (AR) interfaces are known to improve knowledge transfer in studies of education, industry, and elderly assistance. However, this is achieved this by creating specific interfaces for users, not the general public. To test the performance of these interfaces for general public guidance, we designed and implemented a novel AR-assisted disaster prevention guideline that leverages object detection models to identify targets of disaster preparedness advice. We then had a diverse-age audience compare our design against a real traditional paper-based preparedness guide in a room arranged as a common remote work bedroom. By testing their usability, task load, and capacity to make users aware of their environmental hazards, we gained important insights into the performance of different age groups following media developed for the general public. Regardless of different age groups achieving similar usability scores, we found minors improving their performance scores with our novel interface and adults from 20 to 49 years old seemingly performing better than other age groups. In this study, we highlight the importance of guidance alternatives for the young and the less-technology-aware population, contributing to the under-explored area of AR interfaces for the general public.

1. Introduction

Governments around the world produce disaster preparedness guidelines to increase the readiness of their citizens against natural or manmade disasters.[1–3] They intend to inform the public regarding actions that can be taken before, during, and after a disaster has occurred. These media give advice through easy-to-understand language, simple examples, and clear illustrations.

The consequences of an uninformed, ill-informed, or unprepared population experiencing a natural disaster such as an earthquake can range from increases in the severity of private

*Corresponding author: e-mail: yasumoto@is.naist.jp
https://doi.org/10.18494/SAM5215

ISSN 0914-4935
https://myukk.org/

property and business damage[4] to increases in incidences of minor or grave injuries due to secondary hazards, such as falling objects or fires.[5] This issue is relevant since a clear correlation has been observed between the media and activities produced on these programs and the increase in the levels of preparedness in the population.[6,7]

This situation can be framed as a knowledge transfer problem as found in organizational theory, in which "sticky information" refers to technical knowledge that is difficult to transfer to a different organizational unit.[8] In this study, we investigated challenges in organizational knowledge transfer, with a focus on the shift of technical expertise from governments to citizens and the subsequent application of this knowledge.

To assess these challenges, we developed a novel medium to express the knowledge from disaster preparedness guidelines. This new medium takes advantage of the literature on the effectiveness of augmented reality (AR) applications in transmitting common and technical knowledge or skills to a user. Furthermore, our application takes advantage of recent machine learning object detection models to contextualize the advice in guidelines and improve their usability and comprehension.

By testing our design on a sample of 16 subjects ranging from minors (<20 years old) to seniors (>50), implementing it on a ubiquitous platform with restricted processing power such as a mid-range smartphone, and benchmarking its performance against a printed earthquake preparedness guideline in a real-life-like environment, we were able to gain important insights into AR interfaces for guidelines intended for the general public, as follows:

- We confirmed the assumption that minors perform better with an application rather than with a traditional paper guide.
- We found that seniors perform worse than adults while following disaster preparedness guidelines.
- It is clear that ages further from the adult median have greater difficulty following disaster preparedness guidelines, regardless of the type of medium used.

This paper thoroughly extends on our previous work[9] in the following points:

- We provide a comprehensive description of the design choices taken during the implementation of our application. We elaborate on the development platform chosen, the device characteristics, and accessibility considerations on our interface's virtual elements.
- In the technical area, we discuss further the deep learning model used, the considerations we took to achieve practical performance on object lock-on and 3D placement accuracy, and the reasoning of the elements of our experimental setup.
- Furthermore, we include a new summary of users' written feedback, behavior, and opinions that we consider fundamental to contextualizing our findings, discussion, and future work.

2. Related Work

Preparedness guidelines focus on the actions the general public can apply in relation to their surroundings. They are meant to reduce injuries and damage to their belongings or properties. Previously, a clear correlation has been proven between the media and the activities produced on programs and the increase in the levels of preparedness in the population.[6,7] Nonetheless, it has

been observed that even if such knowledge can be substantially transmitted through such means, the effective level of preparedness in citizens' houses is not necessarily substantially increased.[10] Xiao and Peacock revealed that such levels of preparedness are relevant,[4] observing that businesses engaging in preparedness activities against hurricanes can substantially reduce damage to their properties and assets.

Preparedness activities found in printed guidelines can take diverse forms, such as fixing furniture to a wall, preparation of emergency rations, moving valuables to a higher ground level, clearance of emergency exit ways, or acquisition of useful knowledge for safety procedures during and after a disaster. Regardless of their proven usefulness, previous studies have shown that only a small portion of the general public is actively reading them or applying their advice. Particularly in one of our previous survey studies,[11] we found that elderly people experience a comparatively greater difficulty in understanding them as shown in Fig. 1.

AR applications have been extensively studied and used in classrooms. Ishimaru *et al.* demonstrated an application capable of identifying a student's focus on a virtual textbook and overlaying extra resources that the student can read further.[12] These AR interfaces have also been found capable of conveying better instructions for laboratory equipment usage, making students commit fewer mistakes while handling them.[13] However, game-like features such as progress indicators have been found to promote competitiveness among students, making them focus on completing the lesson faster rather than paying attention to the topic of study.[14]

In the industry context, Obermair *et al.* observed that AR technology can significantly reduce misidentifications of objects in maintenance tasks.[15] They showed that workers' mental load can be significantly reduced in these activities and that less experienced workers could reduce their number of mistakes the most. This is relevant for us since disaster preparedness guidelines purposefully target those inexperienced in maintenance tasks. Regardless of the benefits, Eder *et al.* found inherent problems when these applications are used on handheld devices.[16] One-hand

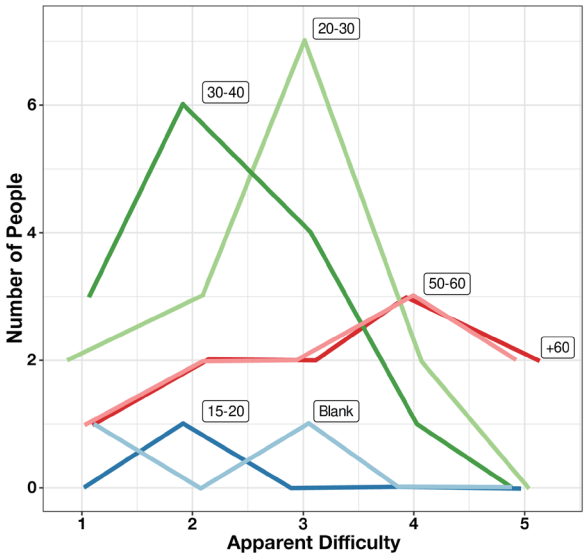


Fig. 1. (Color online) Compared with younger groups, older generations regard Disaster Preparedness Guidelines as more difficult to understand (figure taken from Ref. 11).

manipulation, for example, was shown to impede semi-complex double-handed tasks, which are common in maintenance processes.[16]

To understand the suitability of interfaces for the elderly, Kong *et al.* explored the impact of the familiarity of virtual interfaces as instructive manuals for daily life device handling, e.g., ATM usage and self-service coffee kiosks.[17] They also recognized that even the method of presenting instructions should be considered when addressing the elderly; because of occasionally decreased cognitive capacity, it is preferable to break down instructions and divide them into easy-to-follow and simple actions. Furthermore, Leonardi *et al.* demonstrated how using familiar elements in an interface design can reduce the perception of technological difficulty and promote technology acceptance in seniors.[18]

Regardless of the proven benefits of AR in the previous areas, the proposals in the literature achieve their performance through designs targeting a particular sector. Moreover, while addressing design guidelines for applications for disaster response, Tan *et al.* emphasized the lack of studies on interface design for smartphone applications for the general public.[19]

As presented, when contextualized on an application intended for the general public, we identify the following limitations in current studies on AR interfaces for guidance:

- They achieve improvements in task completion, error reduction, and academic performance; only by exploring designs that target a specific public.
- Their implementations may use AR or virtual reality (VR) devices that are not familiar to the general public.
- The interpersonal differences of their sampled population are mainly approached from an expertise context, not their age diversity.

The interface design proposed in this paper intends to serve as a new platform to study further this under-documented application of mobile technology in the context of a diverse audience.

3. Proposed Method

We created a novel design for disaster preparedness guidelines in the form of a smartphone AR software platform. This platform consists of an open database of curated disaster preparedness advice and an AR application that shows brief advice on screen. As the scope of this paper, we limited our experiment to the evaluation of the novel AR-based interface implemented as an earthquake preparedness guideline.

In traditional guidelines, expert advice and judgment are encoded in simple graphics and phrases. However, restrictions such as limited space and limited examples put a strict limit on the knowledge that can be effectively transmitted. In our design, expert advice is extracted from guidelines and expressed through concrete, one-line actions that the user should take to secure a particular object identified by a machine-learning object detection model. In our system, as shown in Fig. 2, this detection model expresses the expert judgment that governments and private institutions want to transfer to the public.

After identifying the objects, our application provides custom checklists with short suggestions. These checklists are placed near the identified objects in the AR 3D world

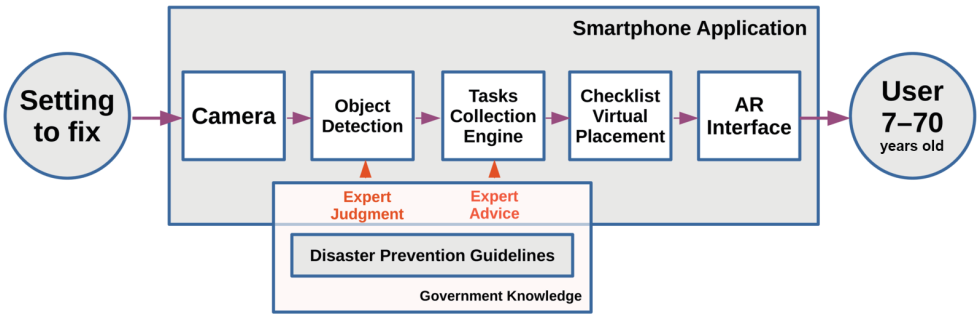


Fig. 2. (Color online) Interface design and information integration flow.

coordinate system to achieve "near-transfer learning" as explained in organizational theory: the similarity between a teaching example and the real context where the learner will use it.[20] At this point, the user can interact with the checklists, and a visual aid shows them their progress as seen in Fig. 3.

To achieve an experience comparable to traditional guidelines, we use smartphones as a familiar interactive platform. Nonetheless, running AR technology and object detection models in processing-power-restricted devices comes with some challenges, such as bounding box clutter during detection and an unreliable depth estimation in mid-range smartphones that lack depth sensors of any kind.

3.1 Smartphone application design

To decrease technology rejection by elderly people and reach the general public, we are selecting a platform they are already familiar with: consumer-grade smartphones. The commercial and widespread adoption of smartphones by all age ranges of the population makes them a promising distribution platform to reach the general public. Furthermore, their processing power, both in computing and graphical capabilities, and the recent progress in machine learning algorithms in edge devices allow us to bring a smooth and accurate enough experience for mainstream users.

However, the downsides of smartphones as an AR application platform are substantial. Eder *et al.* observed, in the industrial maintenance context, the limitation of always having to use at least one hand to carry them, and two hands to operate them while following maintenance protocols.[16] Nevertheless, while comparing AR-based instruction applications with traditional instructions, the relative mental load was decreased with their AR application, since the information is overlaid over the real image of the target object.[16]

Head-mounted displays are a popular technology for AR experiences. These devices can enhance the potential of an AR application by providing screens with an immersive wide field of vision and stereo sound, which can be used to improve the location of objects through audio cues. The reason for not using head-mounted displays for this research is our pragmatic approach to implementing this application in reality. The reasoning is as follows:

- A consumer survey publicly released by the survey aggregator Statista, Inc. revealed that the current hardware and software users of AR experiences have grown to 9.7 million in 2022.[21]

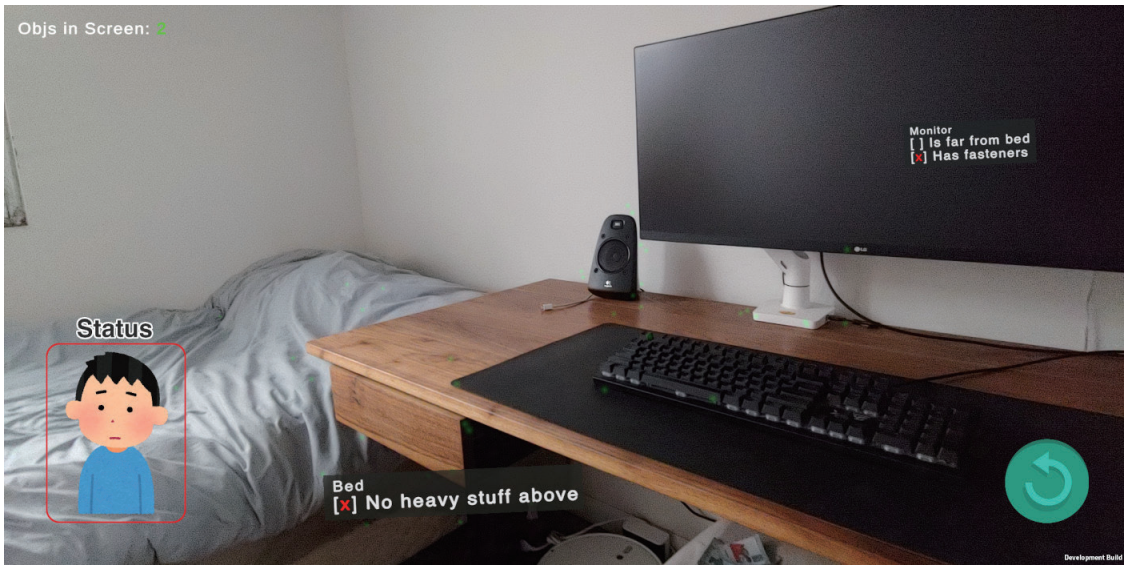


Fig. 3. (Color online) Real example of our application interface. A typical bedroom is seen through the screen of our application. Advice in the form of checklists is placed in a 3D world near the bed and PC monitor. Checklists' completeness is represented with the visual status on the bottom-left corner.

This contrasts to the over 107 million users of smartphones only in Japan.[22]

- The audience we could reach would decrease if we implement our design on head-mounted displays.

- Owing to the current low penetration of Smart Glasses in the elderly population, this public sector is still not familiar with their manipulation.
  - Our platform could face an increased technology rejection from part of our intended audience.
- It is expected the comparatively novel and different input interface found in head-mounted displays will require user training, which is not required with smartphones.
  - New users could face an increased entry barrier to use such platform.

By using smartphones as our application's platform, we expect to improve user engagement while following disaster preparedness guidelines, tackling the current concern of low citizen participation in disaster preparedness activities.[11]

3.2 Object detection

We decided to use common machine learning object detection models instead of object tracking models owing to their advanced state-of-the-art performance in terms of both accuracy and efficiency. Since our platform runs on a mobile device that must handle a deep learning model on top of Unity's framework and its AR 3D world-building mechanisms, we were inclined to use a simple, accurate, quick, and sufficiently efficient option.

The downside of using common object detection models is that they work on a frame-by-frame basis. This makes it harder to discern between false positives (FPs) and true positives

(TPs) across time whenever FPs cross the confidence threshold for only a few frames. This is seen on the screen as confusing detections for the user that rapidly appear and disappear. To address these transient misidentifications, we implemented a time-wise non-max suppression algorithm to select the best candidates for detected objects. Here, we count consecutive detections of the same object in the same area of the screen and remove detections with lesser object detection confidence. An illustrative example of this algorithm is shown in Fig. 4.

The inference time interval multiplied by the number of consecutive detections needed to achieve an acceptable inference confidence determines the object detection time interval the user experiences. The algorithm we developed to solve this issue is an improvement from the algorithm found in the example code by Lei *et al.*,[23] achieving multi-target confidence lock-on and virtual elements anchoring in under 800 ms in the best cases.

3.3 Checklist virtual placement

In the transfer learning literature, the concept of near and far transfer learning refers to the similarity between the example that the teaching tool is using, to the situation where the person applying their new knowledge wants to apply it. Near-transfer learning implies that the subject can apply the learned knowledge through similar steps and in a similar situation. The relevance of near-transfer learning resides in the general higher expectation of knowledge transfer to the subject as compared with far-transfer learning.[20]

To achieve an efficient transfer of knowledge through near-transfer learning, we are placing our disaster preparedness advice as virtual elements directly over the actual properties of the

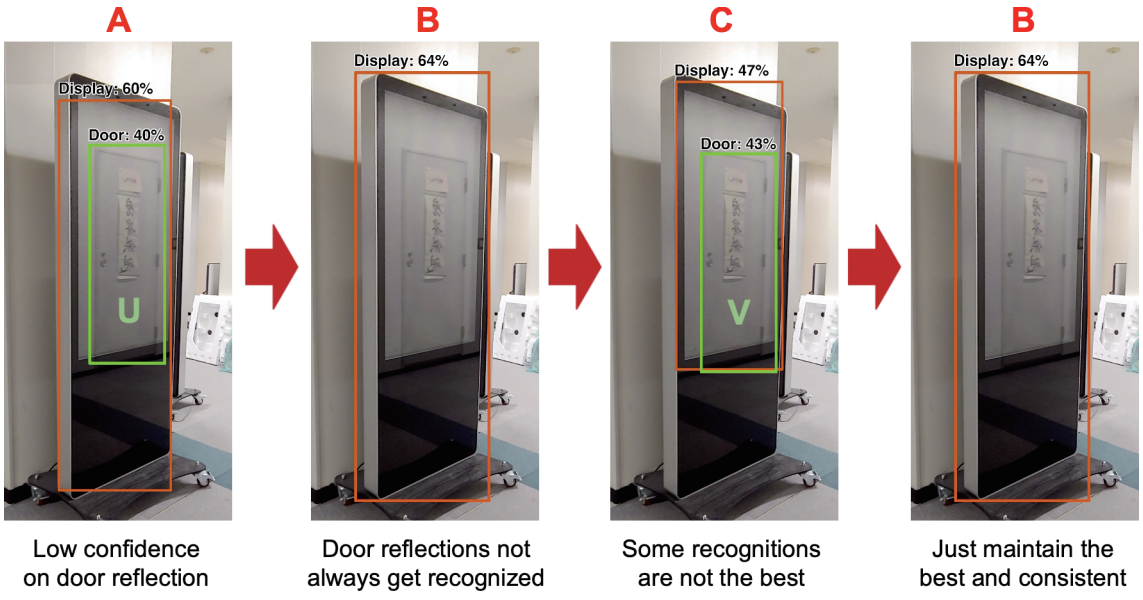


Fig. 4. (Color online) Illustration of our time-wise non-max suppression algorithm. Each bounding box is individually identified with a letter. Correct detections (A to C) are highlighted in red and intermittent spurious detections (U and V) are highlighted in green. Intermittent recognitions are discarded and the ones with the highest confidence levels are preserved. The red "B" bounding box seen in the second frame is locked on as the correct bounding box at the end of the algorithm.

user. Through this method, we expect to reduce the burden of advice interpretation by non-experts of said advice.

Unity's AR Foundation framework provides us with a collection of feature points that can be used to anchor virtual elements into the 3D world mapping of the environment seen through the camera. These feature points are derived from processing all data available regarding the environment. Regrettably, most common smartphones come without depth recognition capabilities (LiDAR or time-of-flight sensors), which forces Unity 3D's libraries to derive depth information from common image processing techniques such as motion parallax.

Empirical testing revealed that these techniques frequently produce poor-quality anchor points over objects that contain large surfaces without clear patterns, such as displays or cardboard boxes. These poor-quality anchor points would result in virtual objects placed at random distances from the smartphone's camera, missing the right placement over the desired object. On the other hand, areas possessing more visual information or patterns, such as bezels or control panels, were found to produce feature points with more accurate depth information and producing them in a higher number.

We leverage this grouping of higher-quality feature points over visually complex surfaces, to select a feature point capable of placing our virtual element at the right distance. Using minimal computational burden, we created an algorithm that identifies the location of the biggest cluster of feature points and utilizes one of them as our virtual object anchor.

In this algorithm, first, the area of the bounding box estimated with the object recognition model is compared against the screen's total area. From the area ratio, a bounding box size-relative-to-screen classification is calculated as described in Table 1.

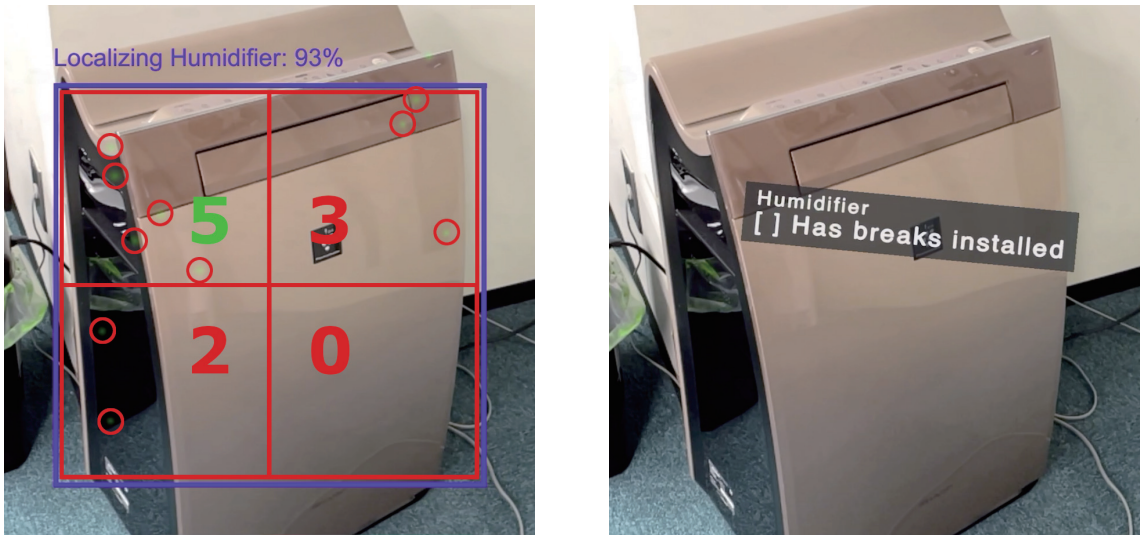
Each bounding box size is divided into smaller subsegments according to the designed *Divisor* number shown in Table 1. For example, a *Small* bounding box has a *Divisor* equal to 1; in turn, it will not be subdivided (sides divided by 1). However, a *Big* bounding box's *Divisor* equals to 3, so it will be subdivided into 9 smaller segments (sides divided by 3).

Finally, through raycasting from the center point of each subsegment, we identify how many feature points are inside each subsegment. The feature point closest to the center point of the subsegment with the most feature points is then selected as the object's checklist virtual anchor.

Through this algorithm, we recognize the most feature-point-dense areas of our bounding boxes. Consequently, we can anchor our checklists to areas with the most probable accurate depth information even on phones without depth sensors. This is particularly helpful when dealing with objects with big smooth flat surfaces, which do not provide enough information, such as computer screens or televisions. Visualization of the segmentation and feature point counting is shown in Fig. 5.

Table 1
Bounding box relative size classification. Each size has a corresponding "Divisor" that divides each side of the bounding box into inner smaller subdivisions called "Subsegments".

Bounding box classification	Divisor	Number of subsegments	Area ratio (%)
Small	1	1	$r \leq 20$
Medium	2	4	$20 < r \leq 40$
Big	3	9	$40 < r \leq 70$
Large	4	16	$70 < r$



(a) (b)

Fig. 5. (Color online) Example of bounding box segmentation process and anchor point selection. (a) Subsegments of a bounding box (red) and the feature points (green) inside red circles markers. The upper-left box is selected since it has the greatest count of feature points. (b) Finally, the closest feature point to the center of the selected subsegment is used as the anchor point for our virtual element.

3.4 AR interface design

Technology acceptance is improved when familiar interfaces are presented to elderly people.[18] In light of this, we selected checklists as a medium to present the expert knowledge extracted from our guideline selections. We think the checklists are familiar enough to the public in all age ranges since they are regularly used in traditional and digital formularies. Thus, we expected their usage not to be impaired by any learning curve for all the age ranges of the general public.

Disaster preparedness advice was deemed a suitable choice for our checklists since they are usually already shown in a simplified and illustrated form in disaster preparedness guidelines. Similar options such as recommended practices and behavior during disasters, were deemed unsuitable for our experimental setup since influencing factors such as stress or the sense of urgency would be more complex to simulate.

To simplify our application guidance, we avoided long phrases, e.g., the shortest phrase is 15 characters long in English (*"Is far from bed"*) and the longest is 30 (*"Nothing will roll and block it"*.) Short phrasing is achieved by relying on the contextual information from both the name of the object in the title of the text box and the spatial information provided by the AR environment field of vision. As seen in Fig. 3, the *"No heavy stuff above"* assertion relies on the "Bed" text box title to answer readers' question "above what?" and the full-bedroom visualization to acknowledge the lack of items in the space above the bed.

Both written guidelines and our application were provided in Japanese and English to our subjects. The original paper guideline was translated into English, and the application text and advice were translated into Japanese with the supervision of native speakers. In this Japanese

translation of the advice, the effect of Chinese-based ideograms that abstract entire concepts into one character resulted in even shorter phrases that readily fulfilled our requirement of avoiding long phrases.

To improve readability, we opted for bold white simple typefaces over a translucent black background. For simple and easy-to-read typefaces, we chose a popular font that is well-regarded as readable. For English, we selected the Helvetica family, which the "ICT for Information Accessibility in Learning" project recommends for accessible textual information.[24] For the Japanese language, we selected the Google accessibility standard Noto Sans JP font.[25]

Considering the low diversity of accessibility needs in our planned sample, we focused on widely available fonts that were deemed accessible. However, by opting for familiarity and readability, this choice does not account for particular accessibility issues such as dyslexia-friendliness.

To improve engagement in minors, we implemented a visual progress indicator consisting of a cartoon character showing three states of emotions: fear, worried, and relaxed. To reflect the number of checkboxes ticked, the emotion shown on the visual progress indicator changes as described in Fig. 6.

4. Experiment

We chose to implement our interface in a widely available smartphone and benchmark its performance by comparing it with a printed earthquake preparedness guideline. Our experimental setup would be defined to answer the following research questions.

Can an AR-based disaster preparedness guideline, when compared with a traditional written one do the following:

Q.1. Reduce users' burden of expert knowledge interpretation?
Q.2. Be more engaging?
Q.3. Help users to successfully complete a disaster preparedness guideline?

We used an iPhone 12 as the platform for our application owing to its comprehensive library of AR functions with ARKit. The application was authored with Unity3D, which allows us to export the experience to multiple mobile targets.

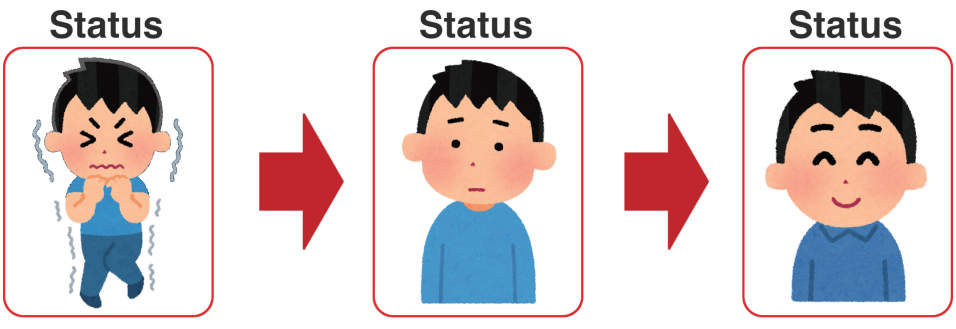


Fig. 6. (Color online) Emotion sprites shown in the visual progress indicator. The sprite changes according to the checklists' completion progress, from a fear state to a relaxed state.

The printed guideline selected for the experiment consisted of a copy of the "Securing your Home" section of an official printed disaster prevention guideline. For this purpose, two pages were extracted from the guideline "Think by yourself and protect your life. Preparedness Textbook", a guideline intended for children to read.[26] This guide was selected owing to the clarity of its instructions and the number of explicit objects described on its pages. For non-Japanese residents, an English-translated version of the guide was provided.

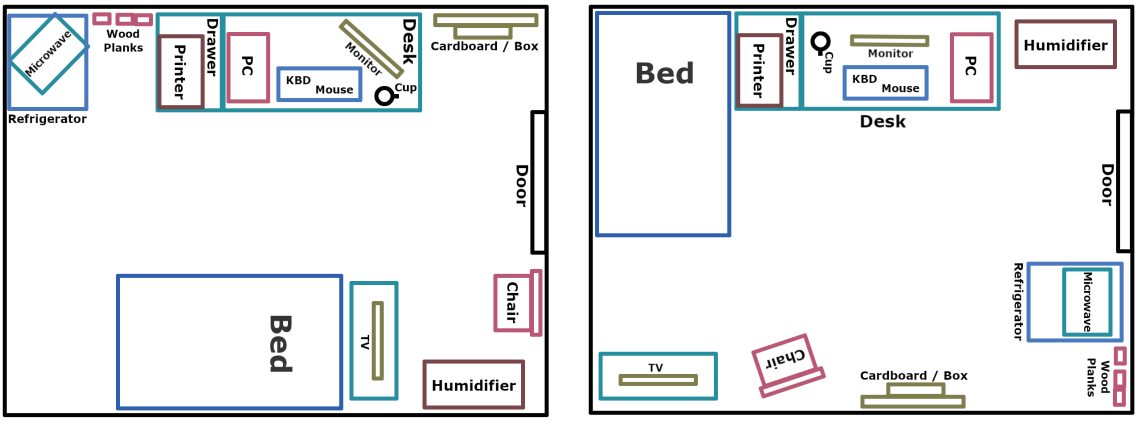
To identify the targets of the guideline, we trained by transfer learning an Ultralytics' YoloV5s object detection model[27] using a custom dataset created from the objects to be used in the experiment. We extended the base code[23] of Lei *et al.* to enable it to run current YOLO object detection models, include our interface elements, and bring UX performance to practical levels.

4.1 Environment setup

To compare the written guideline (*WG*) and our application (*App*), 16 participants of three categories of age ranges (3 minors ≤20 years old, 9 adults >20 to <50, and 4 seniors ≤50) were divided into two groups of 8 participants: *A* and *B*. Each participant from these groups explored a room twice, and each time, the room was arranged in one of two different layouts (layout *X* for *WG*, and layout *Y* for *App*) to reduce cross-learning. Both layouts of the room are shown in Fig. 7.

The room setup was intended to mimic a typical bedroom with an office desk for remote work. We consider it an environment both familiar to many of our subjects after the COVID-19 pandemic, and versatile enough to naturally and realistically accommodate a diversity of objects found in bedrooms and workplaces.

The sample population recruited was meant to represent a general population, which included four different groups: minors, adults, elders, and recent foreign residents. Our final sample included 16 participants: three minors, having lived in Japan for over 5 years, five Japanese



(a) (b)

Fig. 7. (Color online) General layout of the two different settings prepared for each medium. (a) Layout *X* used with the traditional paper guideline. (b) Layout *Y* used with our application.

adults, four non-Japanese adults who have been living in Japan for less than 2 years, and four seniors who have lived in Japan for over 5 years and are over 49 years old. All the participants were compensated with 1000 yen in cash for their cooperation. No subject was particularly experienced with virtual or AR applications.

For the exploration round using *WG*, the subjects would be given 10 min to study an excerpt from an earthquake preparedness guideline before entering the room. For *App*, we would only explain to them the basic controls of the application.

During each exploration round, without modifying anything inside the room, the subjects would write down the actions they would personally take to prevent accidents caused by the objects in the room during a hypothetical earthquake. The subjects were encouraged to write down any idea they could think of on the basis of the medium handed to them. We also encouraged them to write down comments about the medium being used.

4.2 Evaluation

The subjects responded to usability and task load questionnaires for each round. Between each session, we gave them a break of 15 min to answer the questionnaires and changed the layout of the room.

According to Lewis and Sauro, items of the System Usability Scale (SUS) questionnaire can be dropped to simplify it and reduce confusion among the participants. They found a minimal deviation from the results of a SUS questionnaire with one question dropped and found it viable to drop a second item of the questionnaire.[28] Owing to the application only presenting a single functionality shown to the user and following their rationale, we dropped items 5 and 6 of the questionnaire and produced English and Japanese versions with adapted phrasing for our application.

A printed NASA Task Load Index was used to measure the six dimensions of task load on the subjects.[29] A simple Japanese version was produced on the basis of the online version of Egawa and Vertanen.[30]

The room was set to contain objects commonly found in bedrooms equipped for remote work. Among those, 15 were selected to be counted as targets to be addressed by the user's comments. The number of addressed targets comprised the Completion Score. The time the user would take to feel satisfied with the number of targets found and exit the room was considered the Time of Completion. We estimated 12 min as a reasonable time to find the targets in the room and considered it the time limit for the experiment. Post-experiment feedback confirmed that 12 min was sufficient for the subjects to explore and feel satisfied about the targets they found.

We chose some targets not to be identified by the object detection model nor be part of the traditional written guideline. To identify them, we expected the subjects to recognize them from related advice (far-transfer learning) regarding similar objects, e.g., advice addressing the chance of a humidifier blocking an evacuation path, may apply to a chair that is close to an exit door. Regardless, compared with the written guide, we added more targets and advice in the application to express the application's extended information capacity when used in real-use

cases. A picture of the simulated bedroom and how targets were particularly positioned can be seen in Fig. 8.

Additionally, we preserved four circumstances between the two room layouts *X* and *Y*. These persistent circumstances (PCs) were meant to provide targets that should be discovered if the user would extrapolate the knowledge from the guideline, as in the case when a subject consciously analyzes the environment and perceives an environmental hazard not explicitly mentioned in the guideline. For example, an expert in disaster preparedness might advise moving a printer from a table beside a bed, since during an earthquake, it may fall over a sleeping person (Fig. 9). The recognition of these circumstances was meant as a signal of transfer learning. For this experiment, if the relationship between the objects or circumstances was not found to be expressed in any way on the subjects' notes, it would not be counted in their PC score.

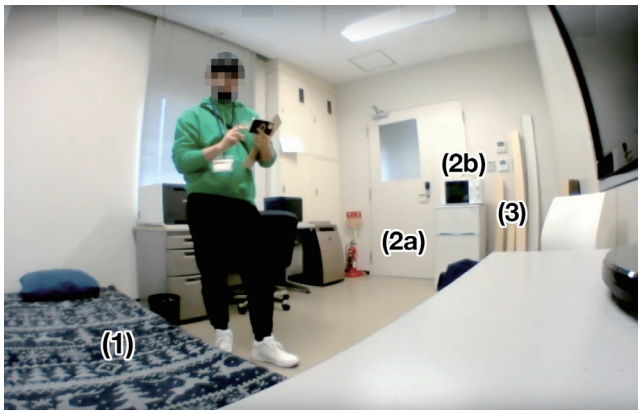


Fig. 8. (Color online) Subject exploring the experimental setup with our application. (1) A target found in the guidelines. (2a & 2b) A potential hazard only recognized by reasoning on the objects' interaction. (3) An object not found in the guidelines.

Fig. 9. (Color online) Example of a PC: a heavy object (a printer in *X*, a TV in *Y*) should not fall over the bed.

5. Results

In this study to evaluate our method, we chose a significance level of 0.05 for all our metrics.

5.1 Subjective metrics

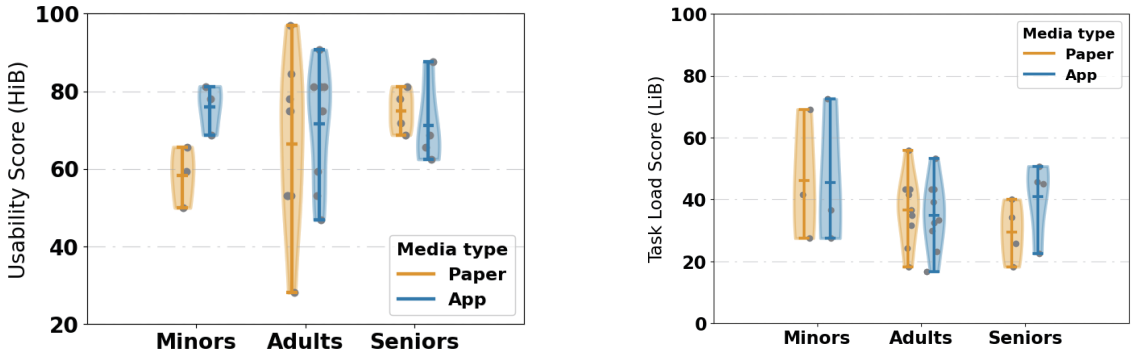
On the usability metric, the mean score of our application was 72.27, whereas that of the written guideline was 66.99. Regardless, with a score above 70, our application can be reported as having an "acceptable" performance as per wide-range usability surveys by Bangor *et al.*[31] As seen in Table 2, we found no significant difference between both media usability scores ($p = 0.3343$) as per their Kruskal–Wallis H test results. Although the usability of our application is comparable to the industry's mean, both media being tested may be considered to have similar usability levels.

However, when grouped by age, we found an apparent difference for minors, as shown in Fig. 10(a). As shown in Table 3, the Kruskal–Wallis H tests of different age groups show a significant difference between the usability scores of our application and the written guideline in the minors age group ($p = 0.0495$*). However, the small sample ($n = 3$) suggests that this value is not reliable and there is a need to increase the number of participants in the study.

To compare the perceived burden of interpreting the advice found in each medium, an unweighted (raw) NASA TLX scoring was used. In this test, six dimensions—mental, physical,

Table 2
Metrics comparison between our *App* and a written guideline. Mean and standard deviation in parentheses. HiB or LiB indicates "Higher/Lower is Better". C. Time and C. Score refer to completion time and completion score, respectively. PCs refers to persistent circumstances.

	SUS (Ind. $\bar{x}$: 68, HiB)	NASA-TLX (LiB)	C. Time (LiB)	C. Score (max: 15, HiB)	PCs (max: 4, HiB)
WG	66.99 (16.81)	36.67 (13.27)	9.28 (2.8) min	5.56 (2.63)	2.19 (0.91)
App	72.27 (12.34)	38.48 (13.88)	10.84 (1.86) min	6.44 (2.87)	2.38 (1.09)
K.W.H. *p*	0.3343	0.6237	1	0.2825	0.5023



(a) (b)

Fig. 10. (Color online) (a) System Usability Scale scores by age group. (b) NASA Task Load Index scores by age group. Data points are shown in gray. Middle lines represent the mean of each group.

Table 3
Kruskal–Wallis H test on usability and mental load scores between *WG* and *App* usage by each age group.

	Usability Scores			NASA-TLX Scores		
	WG	*App*	*p*	*WG*	*App*	*p*
Minors	58.3	76.0	**0.0495***	46.1	45.6	1.0000
Adults	66.3	71.5	0.5916	36.7	35.0	0.6896
Seniors	75.0	71.1	0.3094	29.6	41.0	0.1489
Mean	66.99	72.27		36.67	38.48	

temporal, performance, effort, and frustration—are averaged to produce a dimensionless task load score. With two tasks evaluated with this test, a lower score in one task implies that the user perceives it as easier to complete than the other. With means of 36.67 for *WG* and 38.48 for *App*, a Kruskal–Wallis H test confirmed no significant difference in both task load indexes ($p = 0.6237$), as inferred from their variance in Table 2.

Although an apparent difference can visually be seen in Fig. 10(b) between media tested by the senior category, no significant difference was found ($p = 0.1489$) as seen in Table 3. With task load indexes between media for adults and minors not showing any significant difference, further study with bigger samples might be worth considering.

5.2 Objective metrics

To measure the efficiency of transfer learning for each medium, we counted the time the subjects took to finish the room exploration. A lower time of completion would imply a medium that is better at transmitting expert knowledge to the user. With means of 9.28 min for the written guideline and 10.84 min for our application, a Kruskal–Wallis H test confirmed that there was no significant difference in the time the users would take to complete the tasks with either medium ($p = 1$). Age grouping would not find a significant difference between the completion time of the groups with each medium.

The high variance was notable for both media as shown in Table 2, most likely caused by several users preferring to use all the allowed time to find more targets and a subject that had trouble following the experiment protocol alone.

To measure the effectiveness of each medium in helping its users secure their environment, their notes would be interpreted to recognize how many of the 15 targets the subject could identify. As seen in the completion scores' means shown in Table 2 (5.56 for *WG* and 6.44 for *App)*, a Kruskal–Wallis H test discards any significant difference in both media completion scores ($p = 0.2825$).

Considering the language of the guideline used by adults may have an impact on completion scores, we run Kruskal–Wallis H tests between the completion scores of each medium in a particular language (Japanese: $p = 0.0765$, English: $p = 0.1512$), and between the completion scores of each language in a particular medium (*WG* $p = 0.8977$, *App* $p = 0.2663$). The results did not provide any clear insight into the effect of the language used by the users on their target score performance.

When grouped by age, the completion scores of each medium show a contrasting difference when minors and seniors use one or the other, as shown in Fig. 11(a). A Kruskal–Wallis H test showed a significant difference ($p = 0.0450$*) between the three age groups' mean scores in the *App* case and only an apparent difference ($p = 0.0613$) for the *WG* completion scores.

In light of these differences, post hoc pairwise Dunn's tests with Bonferroni correction were run for both cases, to determine whether a group used WG or the App. In the case of the traditional written guideline, only an apparent difference ($p = 0.0842$) between completion score means was found between adults ($\bar{x} = 6.22$) and minors ($\bar{x} = 2.67$). By this metric, minors seem to have a harder time than adults finding the targets of the study while using a paper guideline.

On the other hand, Dunn's test between the different age groups using our application shows a significant difference ($p = 0.0473$*) between the scores of adults ($\bar{x} = 7.89$) and seniors ($\bar{x} = 3.75$). This implies that seniors had more problems than adults finding the targets when using the application. All pairwise p-values are shown in Table 4.

To find the factors affecting the completion score, we ran a two-way ANOVA test on Media and Age group factors. Age was found to affect the completion of the task with a p-value of 0.0251* for the completion score and for the PC score ($p = 0.0100$*) as shown in Table 5.

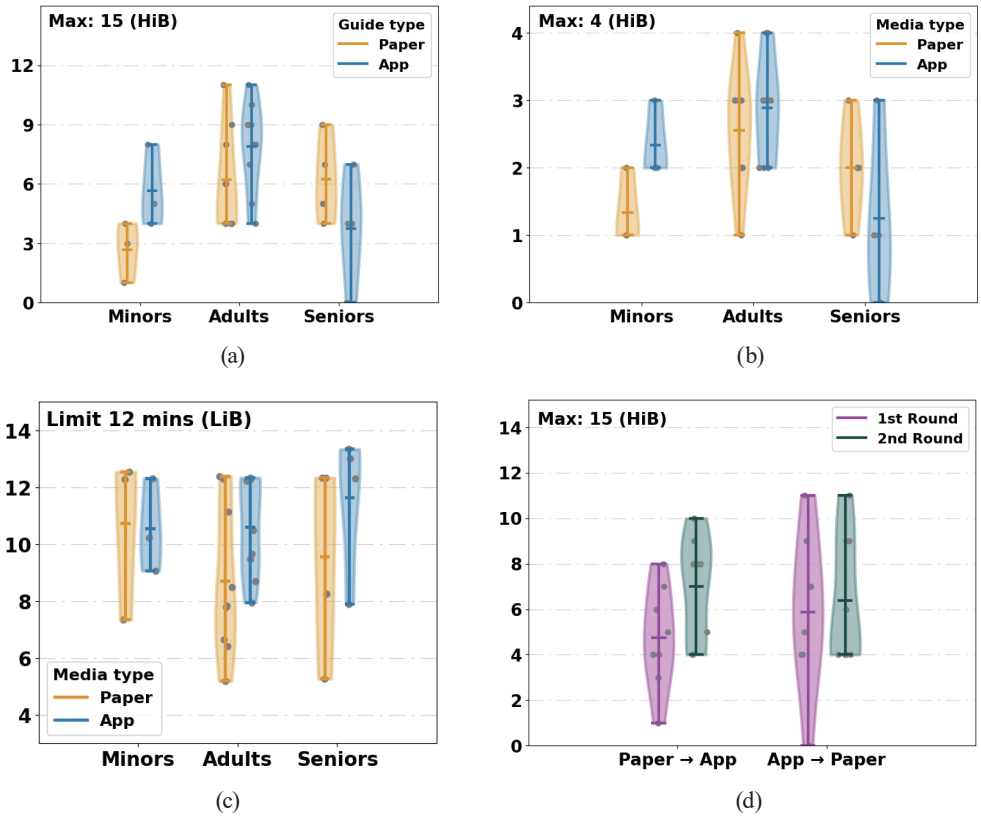


Fig. 11. (Color online) Objective metrics grouped by age group: (a) completion score (HiB), (b) PC score (HiB), and (c) completion time (LiB). (d) Completion score by order of medium used. Data points are shown in gray. Middle lines represent the mean of each group.

Table 4
Pairwise Dunn's test on completion scores for each age group. (a) Written guideline and (b) Application.

(a)

WG	Minors	Adults	Seniors
Minors	1	0.0842	0.1078
Adults	0.0842	1	1
Seniors	0.1078	1	1

(b)

App	Minors	Adults	Seniors
Minors	1	0.6245	1
Adults	0.6245	1	**0.0473***
Seniors	1	**0.0473***	1

Table 5
Two-way ANOVA on the effects of media and age group on completion time, completion score, and PC score.

	Completion Time		Completion Score		PC Score	
	F	*p*	*F*	*p*	*F*	*p*
Media	3.2332	0.0837	1.0812	0.3079	0.3809	0.5424
Age	0.6014	0.5554	4.2565	**0.0251***	5.5193	**0.0100***
Media & Age	0.4720	0.6289	2.8576	0.0755	1.9260	0.1659

Furthermore, both Age and Media combined seem to also affect the completion score with a *p*-value of 0.0755.

To study further the effect of Age on the aggregated completion scores (*WG* and *App*), we proceeded with a post-hoc pairwise Dunn's test with Bonferroni correction as shown in Table 6. In this test, a difference is suggested between adults ($\overline{x}$ = 7.06) and minors ($\overline{x}$ = 4.17), implying that minors might have more problems following disaster preparedness guidelines even when using a guideline designed for them or a platform they feel comfortable with.

The analysis of the PC score was considered a different metric from the completion score, since instead of measuring effectiveness, we try to measure the conscious use of the knowledge transmitted via the number of PC targets found by the subjects. With 4 as the maximum score and means of 2.19 for *WG* and 2.38 for *App*, we did not find any difference between both media when a Kruskal–Wallis H test was conducted (p = 0.5023).

When these scores were grouped by age group, we observed a variance of results similar to the completion scores, in which each medium by each age group contrasts when minors and seniors use each medium, as shown in Fig. 11(b). To follow up this analysis, a Kruskal–Wallis H test of the difference between the means of the PC scores of the three age groups was conducted, which showed an apparent difference between the groups' PC scores for *App* (p = 0.0753).

Then, pairwise Dunn's tests with Bonferroni correction were run for this (*App*) case, and we observed what seemed a performance difference (p = 0.0720) between adults ($\overline{x}$ = 2.89) and seniors ($\overline{x}$ = 1.25), probably indicating promising insights with bigger sample sizes. All pairwise *p*-values are shown in Table 7.

5.3 User comments and experiment notes

Most of the subjects' comments regarding the medium used focused on the application usage. A compilation of the main points and concerns raised is shown in Table 8.

Through this feedback, we observed factors that may be affecting the usability of AR applications for guidance, such as the low processing power of consumer smartphones and the accuracy of object recognition models intended for edge devices (compact models). It also shows

Table 6
Pairwise Dunn's test on the aggregated completion scores of each age group.

All CS	Minors	Adults	Seniors
Minors	1	0.0706	1
Adults	0.0706	1	0.3239
Seniors	1	0.3239	1

Table 7
Pairwise Dunn's test on PC completion scores of each age group when using App.

App	Minors	Adults	Seniors
Minors	1	1	0.9612
Adults	1	1	0.0720
Seniors	0.9612	0.0720	1

Table 8
Subjects' notes and direct feedback to the researchers about the application being tested.

Type	Comments
Positive points	It allows us to recognize dangerous objects.
	It is useful.
	The hints are easy to understand.
Negative points	Sometimes objects are not well identified.
	It is slow to recognize objects.
	Text in far objects is hard to read.
Recommendations	It should show dangerous areas.
	It should recognize small objects.
	It should start recognizing objects with a finger tap on the screen.

areas of opportunity to improve our application in future experiment settings: interactive guidance, dynamic virtual element anchoring for visualization of far objects, and more intuitive controls.

6. Discussion

We found the following inherent practical issues in our experimental setup that did not allow us to answer any of our research questions.

- (Q1 and Q3) Although the comparison of both media completion scores does not show any significant difference, in this iteration of our application, we only reached a usability median score of 72, close to the 68 commonly found in the industry (Table 2). Furthermore, considering that users' feedback reflects issues in our interface that can be revised in software, we believe that our solution has the potential to perform better than its printed counterpart.
- (Q2) By explaining to the user the goal of finding objects to address and by not supervising their exploration sessions, users tended to use most of the 12 minutes for each exploration round as shown in Fig. 11(c), which did not allow us to meaningfully compare users' engagement on the task and the medium teaching efficiency.

These issues are related to the limitations of our experimental setup:

- We tested a novel interface that integrates for the first time the design advice for different public targets.
- Consequently, our experiment was designed to compare two media, rather than to compare age groups' performances.

Nonetheless, we found differences in how different age groups perform with each medium. Adults in general seemed to perform better at using both media than minors and seniors.

Table 9
Kruskal–Wallis H tests for difference in completion score means grouped by order of use of the media. In column "Diff. in Order" the test *p*-value is shown, for the comparison between the first medium scores and the second. In row "CL effect" (cross-learning effect) we indicate the comparison *p*-value of the scores of the same medium when it was used first against when it was used second. Max score: 15.

	WG $\bar{x}$ (SD)	*App* $\bar{x}$ (SD)	Diff. in Order *p*
WG first	4.75 (2.25)	7.00 (2.33)	0.0697
App first	6.38 (2.88)	5.88 (3.40)	0.9568
CL effect *p*	0.3049	0.4566	

However, the way that minors and seniors respond to novel or traditional media seemed to differ. Minors appeared to be more comfortable and performed better with our application, whereas seniors appeared to show the opposite trends.

Additionally, while analyzing the effect of the order of media used in the completion scores, we obtained a suggestion ($p = 0.0697$, Table 9) that the application improved users' target recognition when it was used after the traditional guideline as shown in Fig. 11(d). On the other hand, we did not observe any significant difference between the means of completion scores of the media depending on if they were used first or second (i.e. cross-learning effect) as seen in Table 9.

Although the small scale of our study does not allow us to produce conclusive evidence, we think that it is worth scaling it up and that our experimental setup should be redesigned. The insight that minors and seniors underperform with guidelines intended for the general public is particularly relevant since these age groups are precisely the most vulnerable population that these guidelines take priority to protect. Furthermore, in a world with an increasingly elderly population and the need to assist in their healthy aging,[32] we think that it is worth exploring the mechanisms to improve mass media meant to support and protect them.

7. Conclusion

In this study, we explored the under-documented area of AR interface guidelines for the general public. By proposing a novel AR interface for earthquake preparedness guidelines and testing it in a controlled environment with real users, we gained insight into how vulnerable portions of the public are considerably affected, particularly, when compared with the public closer to the designer mindset: the adults.

Although no significant differences were found between the two media tested, we observed indicators of different performances between age groups using disaster preparedness guidelines: the minors' performance improved when they used our application, and the seniors' performance conversely decreased. Although these findings do not allow us to confirm the superiority of our application against traditional media in the dimensions that our research questions intended to explore, the preliminary results show us a promising path to follow to improve and produce valuable knowledge for the design of AR experiences for the general public and the factors that may affect their effectiveness in transmitting knowledge from the expert community to the general public.

Regardless of our application achieving industry median levels of usability, it is an indicator of how much our design can be improved. On the basis of the subjects' comments, in our future works, we will improve the checklists' readability when placed far from the camera plane. General optimizations in the user-space code and moving frame preprocessing to GPU shaders should also improve detection and virtual element anchoring times, in hopes of matching the expectation of immediacy of the senior group of the population.

For our experiment protocol, in our future works, we expect to scale up the number of participants and address the issues inherent in our experimental setup. This will allow us to properly explore the performance and behavioral differences between our users' age groups. In this redesign, we will leverage user physiological signals to measure the effect and cognitive load of our interface's elements on performance and usability. Furthermore, simulating different scenarios and dividing the exploration rounds into two days will allow us to explore the environment's effect on user predisposition to finding hazards and reduce the apparent impact of the order of media used on each exploration round.

We believe these co-creation exercises with the affected portions of our user base will allow us to redirect our design strategy from the viewpoint of an adult technology expert, towards the expectations of the public that require our most attention: the young and the elderly demographics.

Acknowledgments

This study was supported in part by JST PRESTO under Grant No. JPMJPR2039.

References

1 Disaster Prevention Information: https://www.metro.tokyo.lg.jp/english/guide/bosai/index.html (accessed June 2024).
2 Prepare Before Hurricane Season: https://www.noaa.gov/prepare-before-hurricane-season (accessed July 2023).
3 O. of Disaster Management Dominica, "Prepare with perrie parrot: A guide to natural hazards for primary schools" https://www.preventionweb.net/quick/74549 (accessed July 2023).
4 Y. Xiao and W. Peacock: Nat. Hazard. Rev. **15** (2014) 3. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000137
5 C. Peek-Asa, J. Kraus, L. Bourque, D. Vimalachandra, J. Yu, and J. Abrams: Int. J. Epidemiol. **27** (1998) 459. https://doi.org/10.1093/ije/27.3.459
6 A. Ardalan, H. Mowafi, H. Malekafzali Ardakani, F. Abolhasanai, A.-M. Zanganeh, H. Safizadeh, S. Salari, and V. Zonoobi: Disaster Medicine and Public Health Preparedness **7** (2013) 481. https://doi.org/10.1017/dmp.2013.93
7 E. E. A. Bogdan, A. M. Roszko, M. A. Beckie, and A. Conway: Int. J. Disaster Risk Reduct. **55** (2021) 102060. https://doi.org/10.1016/j.ijdrr.2021.102060
8 v. H. Eric: Manage. Sci. **40** (1994) 429. https://doi.org/10.1287/mnsc.40.4.429
9 S. De Leon A., Y. Matsuda, and K. Yasumoto: 14th Int. Conf. Mobile Computing and Ubiquitous Networking (2023) 1–7. https://doi.org/10.23919/ICMU58504.2023.10412237
10 W. Adiyoso and H. Kanegae: J. Disaster Research **8** (2013) 1009. https://doi.org/10.20965/jdr.2013.p1009
11 S. De León Aguilar, Y. Matsuda, and K. Yasumoto: IPSJ SIG Technical Reports (ICS) **12** (2022) 1.
12 S. Ishimaru, K. Watanabe, N. Grossman, C. Heisel, P. Klein, Y. Arakawa, J. Kuhn, and A. Dengel: UbiComp '18 Adjunct (ACM, 2018) 357–360. https://doi.org/10.1145/3267305.3267667
13 S. Tzima, G. Styliaras, and A. Bassounas: Educ. Sci. **9** (2019) 2. https://doi.org/10.3390/educsci9020099
14 M. Dunleavy and C. Dede: Handbook of Research on Educational Communications and Technology (Springer, New York, 2014) 735–745. https://doi.org/10.1007/978-1-4614-3185-5_59
15 F. Obermair, J. Althaler, U. Seiler, P. Zeilinger, A. Lechner, L. Pfaffeneder, M. Richter, and J. Wolfartsberger: ICIEA'20 (IEEE, 2020) 942–947. https://doi.org/10.1109/ICIEA49774.2020.9102078

16 M. Eder, M. Hulla, F. Mast, and C. Ramsauer: 10th Conf. Learning Factories **45** (2020) 7–12. https://doi.org/10.1016/j.promfg.2020.04.030
17 J. Kong, A. Guo, and J. P. Bigham, : ASSETS'19 (ACM, 2019) 661–663. https://doi.org/10.1145/3308561.3354593
18 C. Leonardi, C. Mennecozzi, E. Not, F. Pianesi, and M. Zancanaro: Gerontechnology **7** (2008) 151. https://doi.org/10.4017/gt.2008.07.02.088.00
19 M. L. Tan, R. Prasanna, K. Stock, E. E. Doyle, G. Leonard, and D. Johnston: Prog. Disaster Sci. **7** (2020) 100118. https://doi.org/10.1016/j.pdisas.2020.100118
20 D. Perkins and G. Salomon: Transfer of learning, International encyclopedia of education (Pergamon Press, England, 1992) 2nd ed., 425–441.
21 AR hardware B2C market users 2022–2029: https://www.statista.com/forecasts/1337381/ar-hardware-b2c-market-users-worldwide (accessed December 2022).
22 Japan: number of smartphone users 2028: https://www.statista.com/statistics/275099/number-of-smartphone-users-in-japan (accessed December 2022).
23 D. Lei, A. Ashikhmin, and Unity Technologies: GitHub - derenlei/Unity_Detection2AR: Localize 2D image object detection in 3D Scene with Yolo in Unity Barracuda and ARFoundation https://github.com/derenlei/Unity_Detection2AR (accessed December 2022).
24 How to make your textual information accessible: https://www.ict4ial.eu/guidelines/making-text-accessible/how-to-make-textual-information-accessible (accessed June 2024).
25 Introducing accessibility in typography: https://fonts.google.com/knowledge/readability_and_accessibility/introducing_accessibility_in_typography (accessed June 2024).
26 MLIT Chubu Regional Development Bureau, "Think by yourself and protect your life. preparedness textbook," https://www.cbr.mlit.go.jp/kisokaryu/ (in Japanese, accessed February 2022).
27 Yolov5 by ultralytics https://github.com/ultralytics/yolov5/ (accessed December 2022).
28 J. Lewis and J. Sauro: J. Usability Studies **13** (2017) 38. http://uxpajournal.org/dropping-item-sus/
29 F. M. Calisto and J. C. Nascimento: ResearchGate (2018). https://doi.org/10.13140/rg.2.2.26978.79044
30 NASA-TLX in HTML and JavaScript: https://www.keithv.com/software/nasatlx/ (accessed December 2022).
31 A. Bangor, P. Kortum, and J. Miller: J. Usability Studies **3** (2009) 114. http://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/
32 J. R. Beard, A. Officer, I. A. de Carvalho, R. Sadana, A. M. Pot, J.-P. Michel, P. Lloyd-Sherlock, J. E. Epping-Jordan, G. M. E. E. G. Peeters, W. R. Mahanani, J. A. Thiyagarajan, and S. Chatterji: The Lancet **387** (2016) 2145. https://doi.org/10.1016/S0140-6736(15)00516-4

About the Authors

De León A. Sergio received his B.E. degree from the Technological Institute of Ciudad Madero, Mexico, in 2011 and his M.S. degree from the Nara Institute of Science and Technology, Japan, in 2023. From 2011 to 2021, he was a programmer and IT consultant in Mexico. Since 2023, he has been working on his Ph.D. degree in NAIST. His research interests are in user interfaces, knowledge transfer, disaster preparedness, and sensors.
(sergio.deleon@techoblige.com)

Yuki Matsuda received his B.E. degree in advanced course of mechanical and electronic system engineering from the National Institute of Technology, Akashi College, Japan, in 2015, and his M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016 and 2019, respectively. Since 2024, he has been a lecturer at the Convivial Computing Laboratory, Faculty of Environmental, Life, Natural Science, and Technology of Okayama University. His current research interests include participatory sensing, location-based information systems, wearable computing, and affective computing.
(yukimat@okayama-u.ac.jp)

Keiichi Yasumoto received his B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University, Osaka, Japan, in 1991, 1993, and 1996, respectively. He is currently a professor of the Graduate School of Science and Technology at Nara Institute of Science and Technology. His research interests include distributed systems, mobile computing, and ubiquitous computing. (yasumoto@is.naist.jp)