Crowdsourcing Design Guidance for Contextual Adaptation of Text Content in Augmented Reality

John J. Dudley
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
jjd50@cam.ac.uk

Jason T. Jacques
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
jtj21@cam.ac.uk

Per Ola Kristensson
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
pok21@cam.ac.uk

ABSTRACT

Augmented Reality (AR) can deliver engaging user experiences that seamlessly meld virtual content with the physical environment. However, building such experiences is challenging due to the developer's inability to assess how uncontrolled deployment contexts may influence the user experience. To address this issue, we demonstrate a method for rapidly conducting AR experiments and real-world data collection in the user's own physical environment using a privacy-conscious mobile web application. The approach leverages the large number of distinct user contexts accessible through crowdsourcing to efficiently source diverse context and perceptual preference data. The insights gathered through this method complement emerging design guidance and sample-limited lab-based studies. The utility of the method is illustrated by re-examining the design challenge of adapting AR text content to the user's environment. Finally, we demonstrate how gathered design insight can be operationalized to provide adaptive text content functionality in an AR headset.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; **Mixed / augmented reality**.

KEYWORDS

Augmented Reality, Crowdsourcing, Privacy

ACM Reference Format:
John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2021. Crowdsourcing Design Guidance for Contextual Adaptation of Text Content in Augmented Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445493

1 INTRODUCTION

The emergence of head-worn augmented reality (AR) represents an enormous opportunity for ubiquitous computing. Despite the popularity of early examples of AR games and experiences, such as Pokémon GO [25], nascent AR designers lack the guidance, solution principles, and analytical approaches required to create aesthetic and seamless user experiences.

We can expect that design guidance and solution principles will coalesce over time both through developer trial-and-error and research. However, the knowledge derived from developers can be difficult to generalize while the findings of researchers may have poor applicability. In particular, lab-based studies typically introduce high levels of control that ultimately detract from the external validity of the findings. Such studies are also laborious and costly to execute.

We propose a blended model through which design guidance for AR can be efficiently gathered via crowdsourcing. This approach is particularly well suited to investigating AR design problems due to the key role that the user's own environment plays in an AR setting. While crowdsourcing has been widely used before as a research tool, we specifically seek to leverage the access it provides to a large number of distinct user contexts. This is to address a fundamental challenge encountered in AR design: *unknowable deployment contexts*, that is, the inability for the developer to foresee the environment in which their application will be deployed.

An example of the influence of context on AR interface design is the presentation of virtual text in the physical environment. Depending on the use-case, the designer may wish this text to either subtly blend content with the physical environment or explicitly attract the attention of the user. Clearly an awareness of the user's physical context is necessary to deliver this behavior.

This paper demonstrates how crowdsourcing can be leveraged to obtain a greater understanding of AR context dependence. We develop and deploy a low-fidelity AR experience as a mobile web application to prompt crowdworkers to capture images of their local environment while also obtaining feedback on the visual qualities of virtual elements overlaid on that context. The ubiquity of mobile devices and the increasing capabilities of web-based frameworks allow simple AR experiences to be quickly prototyped and rapidly deployed to a large number of users. This approach therefore allows large-scale testing and diverse dataset collection not afforded by lab-based studies. The collection of data from anonymous crowdworkers, particularly locations, images or video, does, however, expose potential privacy concerns. The method presented in this paper accommodates these concerns by limiting data collection and providing a user-driven obfuscation and acceptance protocol for sharing images.

As a demonstration of the proposed method, we use it to investigate how virtual text content might be dynamically styled in AR given the physical setting. In its current form, the method leverages an AR experience delivered on a smartphone as a necessary

CHI 2021, May 8–13, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445493

consequence of the currently limited scale of deployment of true head-mounted AR. Nevertheless, it is anticipated that many findings are transferable between low and high fidelity implementations of AR. At the end of this paper, the collected data and findings from the mobile-based investigation are transferred to a high-fidelity head-mounted AR prototype application to illustrate the efficacy of the approach.

The two key novel contributions of the paper are:

- A privacy-conscious method for conducting AR experiments in the end-user's own context via crowdsourcing.
- A demonstration of this method on the design challenge of providing contextually adaptive virtual text content.

2 RELATED WORK

The primary contribution or this paper is the research method for conducting AR design studies via crowdsourcing. This aspect of the paper draws upon the established body of work utilizing crowdsourcing as a research tool. A secondary focus of this paper is the exploration of contextually adaptive text content. In part, we choose this case study for focus given the available body of lab-based studies addressing the same topic. This literature serves as a reference against which we can compare and validate the comparable crowdsourced findings. Due to this split foci, we divide our coverage of the related work into conducting research with the crowd and contextually-adaptive text content in AR.

2.1 Conducting Research with the Crowd

Crowdsourcing offers a potential strategy for greatly expanding the range of contexts in which perceptual issues in AR can be explored. Prior research has also previously demonstrated that crowdsourcing studies can replicate in-the-lab human perception studies [1, 11] and professional assessments [21]. Crump et al. [1] replicate a range of cognitive behavioral experiments on Amazon Mechanical Turk (AMT) and find good agreement with laboratory results. Ma et al. [18] carefully curated a panel of AMT workers with access to a Virtual Reality (VR) device and conducted behavioral experiments involving three different VR illusions. There is also precedence in applying crowdsourcing to facilitate interface feature design in mobile-based AR. Previous work has demonstrated how crowdsourcing and probabilistic optimization strategies can be combined to efficiently refine interactions [4].

Prior crowdsourcing work primarily exploits the fact that crowdsourcing provides access to a large number of individuals. Once recruited into a study such as the one by Ma et al. [18], however, the real context of the crowdworker plays no part in the experiment. The approach presented in this paper offers an important extension in highlighting that crowdsourcing can be leveraged to not only reach a large number of participants but to also conduct experiments *within* a large number and varied range of real contexts.

The privacy considerations in crowdworking have been examined from various perspectives. Daniel et al. [2] provide a survey of quality related issues in crowdsourcing and potential mitigation strategies. As an outcome of this survey, Daniel et al. [2] define a quality model for crowdsourcing tasks which notably includes privacy as a potential factor influencing quality. Legion:AR [15] is a framework for augmenting activity recognition models by allowing crowdworkers to label uncertain cases while preserving privacy. The faces of people in the videos to be labeled by crowdworkers are obscured by auto-generated 'veils'. Beyond merely individual privacy concerns, Lasecki et al. [15] suggest that reducing the resolution of video or image data is a reasonable strategy to avoid sharing sensitive information contained in the scene. The influence of blurring on the accuracy of crowdworkers performing behavioral coding of people in videos has also been investigated by Lasecki et al. [14]. These approaches examine the preservation of privacy for people who appear in crowdsourced tasks but do not provide insight on how to manage the privacy of the crowdworkers themselves. The objectives of McDuff et al. [22] and Tan et al. [28] are similar to this work in that they operate at the uncomfortable nexus of information capture and potential intrusions into privacy. McDuff et al. [22] solicited webcam footage of people watching commercials to generate a dataset of facial responses. Privacy was managed using an opt in approach. Tan et al. [28] proposed a game suited to crowdsourcing for capturing user images in order to construct a diverse dataset of facial expressions. Its approach for handling privacy is to allow users to only send facial feature locations as opposed to raw images.

The literature suggests, therefore, two guiding principles of: i) limiting information capture to strictly what is necessary; and ii) giving users ultimate control over what is shared. In this paper, we seek to apply these principles in developing a privacy-sensitive protocol that, we argue, has good generalizability beyond the specific investigation of context-dependence of textual content.

2.2 Contextually-Adaptive Text Content in AR

Many applications of AR are likely to involve the display of textual content. Wither et al. [32] propose a detailed taxonomy of annotation in an AR setting and highlight that there are two key components of an AR annotation: the spatially dependent component (i.e. the association between the physical and virtual world) and the spatially independent component (i.e. the attributes related to its appearance).

As observed by Manghisi et al. [20], there are three distinct strategies for actively promoting text legibility in AR: i) adjust the text placement; ii) adjust the text appearance; and iii) place a panel behind the text. This first strategy of dynamic text placement has been widely explored in the literature [23, 24, 29, 30]. Tanaka et al. [29, 30] introduce a simple strategy for scoring slots in the field of view based on the scene background. They also seek to accommodate the importance of the virtual content and the degree to which movement should be limited to perform this assessment. Rather than adapting the content, they promote legibility by finding regions of dark, uniform texture on which to place text. Gabbard et al. [6, 7] examined three alternative schemes for actively modifying text color in AR: complement, maximum HSV (hue, saturation, value) complement, and maximum brightness contrast. Their active schemes did not perform well, however, and a simple solution of blue text on an opaque white background panel (also referred to as a 'billboard') yielded the best performance. Gabbard and Swan II [5] subsequently found that maximizing the luminance contrast ratio aids readability on billboards. Debernardis et al. [3] evaluated different presentation styles and billboard colors and suggest that

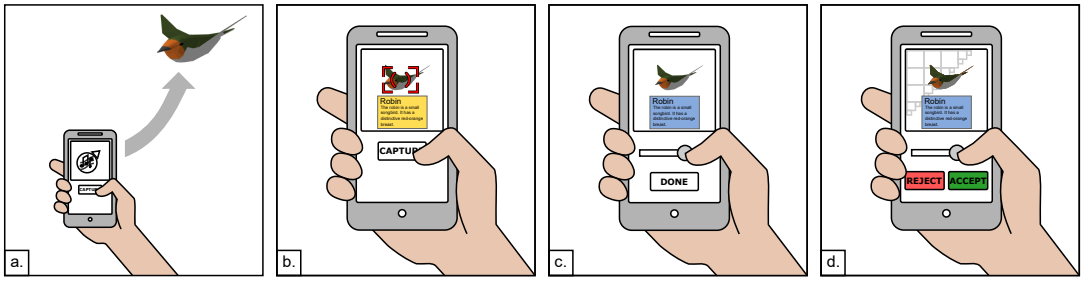

Figure 1: Storyboard illustrating the task procedure. (a.) The participant follows the hints to locate the bird. (b.) The participant holds the device steady with the view reticle centered on the bird until the capture button is enabled. The capture button is pressed to capture an image of the background. (c.) The participant adjusts the appearance of the textual content overlaid on the now static background image. (d.) The participant reviews the captured image (optionally applying pixelation) and chooses to accept or reject its transfer to the server.

white text on blue billboards yields good legibility. This result is reinforced by Kruijff et al. [13], who also examined preferences associated with the presentation of text labels in AR and found that blue background panels were overwhelmingly preferred.

The difficulty of adapting content to background context is exacerbated by several additional considerations, such as coherence, density, and imposition on the user's cognitive load. Grasset et al. [9] introduced a label placement strategy based on visual saliency and edge analysis. In an evaluation of their approach, Grasset et al. [9] found that locally adapting the coloration of text and billboards can be problematic for users due to negative effects on the apparent coherence of presented information. Tatzgern et al. [31] explored the challenge of managing the display of dense information in AR via clustering. Madsen et al. [19] examined the influence of temporal coherence and found that presenting labels in object space (as opposed to image space) is preferred. These various research efforts highlight the need for a structured approach to dynamic content adaptation.

A common theme in the literature is the complexity of robustly accommodating diverse background textures and colors. Human perception capabilities and preferences are difficult to isolate in even the most tightly controlled psychological study. It is therefore unsurprising that small-sample HCI studies in this area uncover numerous perplexing results. While perhaps beneficial as preliminary guidance, the fact that a particular design, for example, white text on a blue billboard, has good general performance provides limited real insight to designers. It also ignores many implementation considerations, such as a desire to maintain a common aesthetic in an application. For the design guidance to solidify in this space, it is essential that research work finds methods that enhance external validity.

3 METHOD: CROWDSOURCING AR DESIGN GUIDANCE

The crowdsourcing method is based on a low-fidelity AR experience delivered by a mobile web application. We pursue a web-based architecture for two key reasons. First, online tasks are more readily integrated into existing crowdsourcing platforms. Second, a web-based implementation minimizes the imposition on crowdworkers (that is, there is no requirement to install software) and reduces friction in the steps between recruitment and completion.

The user's rear-facing camera stream is fed directly into the web page frame and virtual content is overlaid on this stream to deliver a through-the-screen AR experience. A web framework for building VR experiences provides the functionality to ensure device movements produce corresponding changes in the virtual elements. Crowdworkers can then be instructed to perform specific activities or provide feedback on interface features in this setting.

For the specific use case of contextually-adaptive text content, we have four high-level design goals which we aim to satisfy in our crowdsourcing method. First, maximize the variety of background contexts in which feedback is captured. Second, maximize the quality of the background images captured. Third, maximize the richness of the feedback expressed by participants. And fourth, maximize the amount and quality of captured feedback. These design goals directly map into the formulation of the four stages involved in completing the data collection task: i) *search*; ii) *image capture*; iii) *appearance refinement*; and iv) *image review*. The overall procedure is illustrated in Figure 1 and the details of each stage are summarized below.

3.1 Stage 1: AR Search Task

To promote variety in the range of contextual information captured in the two experiments conducted, the application instructed participants to complete a series of target acquisition tasks. Participants located targets, styled as virtual birds, that were presented at semi-random locations within their local environment. The rear-facing device camera stream provides the background of this virtual environment, producing a low-fidelity through-the-screen AR experience.

The location of birds was quasi-randomized to promote spatial diversity in the context images captured from the participant's environment. For each instance of the target acquisition task (a participant performs five instances over the experiment), the new bird was located at between 60 and 100 degrees rotation from the current view azimuth. The sign of this offset was randomized. The bird was placed between -10 and 30 degrees elevation from the horizontal plane. An icon would appear at fixed time intervals in the center of the participant's view to indicate where they must look

to find the target. This strategy of subtly prompting view variation allows us to maximize the diversity of images captured in each environment.

The virtual component of the AR scene was implemented using A-Frame[1]. This framework manages the scene camera adjustment based on device orientation changes. It is important to note, however, that no translation motion of the device was reflected in the virtual scene: the position of the scene camera is fixed. For our purposes, allowing translation offers limited change in the captured view but demands additional sensor capabilities and potentially limits the participant pool. The lack of registration between the physical and virtual scene also means that the AR experience is imperfect. Nevertheless, it remains sufficiently convincing for simple experimental and data collection tasks. The decision to frame the target acquisition task as an exercise in locating and photographing 'birds' mediates the disruptive effect associated with imprecision in the virtual-physical alignment. Participants may reasonably expect a bird to move around whereas this same behavior may be more disruptive if the target is a fixed inanimate object. When the target is found, the participant enters the *image capture* stage of the task.

3.2 Stage 2: Image Capture

Once the target is found, the participant must hold the reticle (mimicking the viewfinder of a camera) fixed on the bird. This serves two purposes: 1) stabilizing the virtual scene; and 2) ensuring captured images do not inherently suffer from motion blur. An animation of the reticle indicates when the bird is in focus. After the required focus period, the capture button is enabled. The participant then simply presses the capture button and the background image from the rear-facing camera is temporarily recorded in memory on the client's device. The use of static images is established practice for AR design related research [6, 16]. Note that the image could be captured automatically after a timeout but it was considered preferable to make all image collection require a deliberate action from the user.

It is at this point, with the image now recorded on the client side, that potential privacy concerns emerge. These concerns, and our mitigating solution, are described later in subsection 3.4. Immediately following *image capture*, however, the web application presents an interface to allow participants to alter the visual appearance of the virtual content as part of the *appearance refinement* stage.

3.3 Stage 3: Appearance Refinement

The *appearance refinement* stage is the point in the task when feedback is collected from participants. The specific characteristics of the interface differ depending on the factor under investigation. The interface and interactions developed for the two experiments presented in this paper are described later in the context of the specific tasks performed. The general procedure applied is to allow the user to customize the appearance according to their preference or as per specified instructions.

In this paper, two key sub-problems related to text panel presentation in AR are investigated in two separate experiments: Experiment 1 focuses on panel coloration; and Experiment 2 focuses on panel placement. For example, in Experiment 1 the participants are asked to modify the appearance of virtual text panels overlaid on the background image. Their instruction for the task performed in Experiment 1 was to adjust the appearance of the text panel to improve visibility and readability of the text. Presenting this stage of the task as a pseudo-design exercise represents an engaging form of feedback collection, compared with, for example, assigning a subjective score. This choice stems from our specified design objective of maximizing the expressiveness of user feedback cycles.

[1]https://aframe.io/

3.4 Stage 4: Image Review

After completing the *image capture* and *appearance refinement* stages of the task, the participant is presented with the *image review* interface. This section introduces a protocol for managing customizable levels of user privacy for crowdsourced image capture tasks. This protocol was well-received by participants and exhibited a high acceptance rate in the data collection undertaken.

As previously discussed, we chose an architecture that ensured image data remained on the client until it was approved in order to accommodate user privacy concerns. Only after approval would the image be sent to the server and saved in the database. Reflecting the hypothesis that workers would be generally unwilling to share personal image data, effort was taken to forestall the situation in which the majority of images were rejected. To this end, we included an obfuscation layer in the review protocol. We elected to use pixelation (also known as mosaicing) for obfuscation. As part of the image review stage, the worker may increase or decrease the level of pixelation. To counter overuse of pixelation, the instruction given to users was: "Please share as much image detail as you are willing." Pixelation was chosen for two key reasons: i) it is broadly familiar to a non-technical audience; and ii) it produces non-recoverable information loss.

In the *image review* stage, the user may adjust the level of pixelation applied to the raw image by setting the sub-block size. Sub-blocks in the image are averaged and the average color is used to replace all the pixels in the sub-block. Increasing the size of the sub-block removes more information from the image. The default sub-block size upon presentation of the image review page was $s = 1$ (no pixelation). The pixelation control was presented as a range slider with sub-block sizes: 1, 2, 4, 6, 8, 10, 12, 16 and 20. This range of values was chosen as they are factors of the default image resolution setting (480×480 pixels). After setting the level of pixelation, the worker may then choose to either approve or reject sharing the image. Figure 2 illustrates the obfuscation achieved with a subset of the pixelation levels available.

3.5 Deployment

We consolidated the described components of the web application as a Human Intelligence Task (HIT) and deployed it on the Amazon Mechanical Turk service. Prior to full-scale deployment, the task was subjected to rigorous sandbox testing and small-scale pilot testing as a quality control measure. The web application was deployed on our own server to ensure a high level of control over the experience and data collection. In order to commence the HIT, participants had to visit the Mechanical Turk listing using a mobile device. Upon accepting the HIT, participants reviewed a short

Figure 2: Effect of sub-block size, s. No pixelation is $s = 1$ up to a maximum value of $s = 20$ (pixelation at 20×20 pixels).

description of the task and its purpose. This included the fact that images of their environment may be captured but would only be recorded after explicit approval from them. Participants were then required to explicitly express consent in order to complete the task. More detailed instructions were then provided on the role of the device camera and the image review process. The first bird capture activity was guided and then participants repeated the activity a further four times without explicit guidance.

4 EXPERIMENT 1: PANEL COLORATION

To make this investigation concrete, we selected a text panel design use-case where participants were asked to refine the appearance of a billboard-style virtual text panel appearing in the environment. When a bird was in focus (i.e. inside the view reticle), the bird name and a short description appeared below on a colored panel (see Figure 3). The coloration of the billboard was randomly initialized but was always shown at 50% opacity to approximate the appearance of AR content on an optical see-through head-mounted display (OST HMD). Once the image was captured, participants were instructed to refine the appearance of the text panel.

To highlight the flexibility of this approach, participants were primed with the intentionally qualitative instruction: "Choose a color that you think is best given the background. Please try to maximize visibility and readability of the text." This use case exposes an interesting and subtle interplay between subjective user impressions related to aesthetics and practical concerns relating to legibility. The interface for customizing the description panel appearance in Experiment 1 is illustrated in Figure 3 (left). The top slider adjusts the hue while the bottom slider adjusts the lightness. The hue slider was initialized with a random rotation applied to the standard hue circle and the initial midpoint value was used as the initial panel color. The lightness slider (where lightness is defined according to the hue, saturation, lightness (HSL) color model) was always initialized to the midpoint value. A toggle was available to change the text color between black and white. The toggle state was randomly initialized. The random initialization of hue and text color was done to prompt participants to make color changes when required. The image review interface is shown in Figure 3 (right). Here the user can choose to adjust the pixelation level and approve or reject transmitting the image to the server.

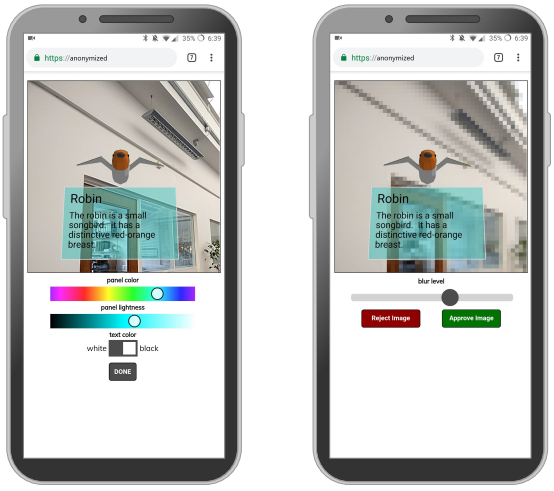

Figure 3: The appearance refinement interface (left) and the image review interface (right).

4.1 Results

A total of 200 participants (113 male, 84 female, 3 unspecified, 32.4 mean age) from 16 different countries were recruited through Amazon Mechanical Turk for the study. Each received US$1 as compensation for their time. The mean completion time for the task was 8.5 minutes (including training and instructions).

4.1.1 Approval Rate and Participant Behavior. With five task instances per participant and 200 participants there were a potential dataset of 1,000 images. The approval rate was very high with only five images rejected in total by five different participants (an approval rate of 99.5%).

The degree of pixelation (sub-block size) was left unchanged in 54.6% of approved images. Recall that the sub-block range slider had a default initial value of $s = 1$ (no pixelation). In an additional 4% of images, participants raised the degree of pixelation before returning it to $s = 1$. The distribution of pixelation levels employed by participants is summarized in Table 1. Table 1 appears to show three distinct modes. The no pixelation default, $s = 1$, dominates

Table 1: Participant usage (%) of pixelation sub-block sizes, s, in Experiments 1 and 2. Note that no pixelation, $s = 1$, is the default and also the most frequently used setting. The usage results show three distinct modes at $s = 1$, 6 and 20 across both experiments.

Exp.	1	2	4	6	8	10	12	16	20
1	58.5	3.1	4.2	6.8	6.5	5.5	5.1	3.4	6.7
2	54.0	3.7	5.6	6.6	6.1	3.7	4.8	3.4	12.0

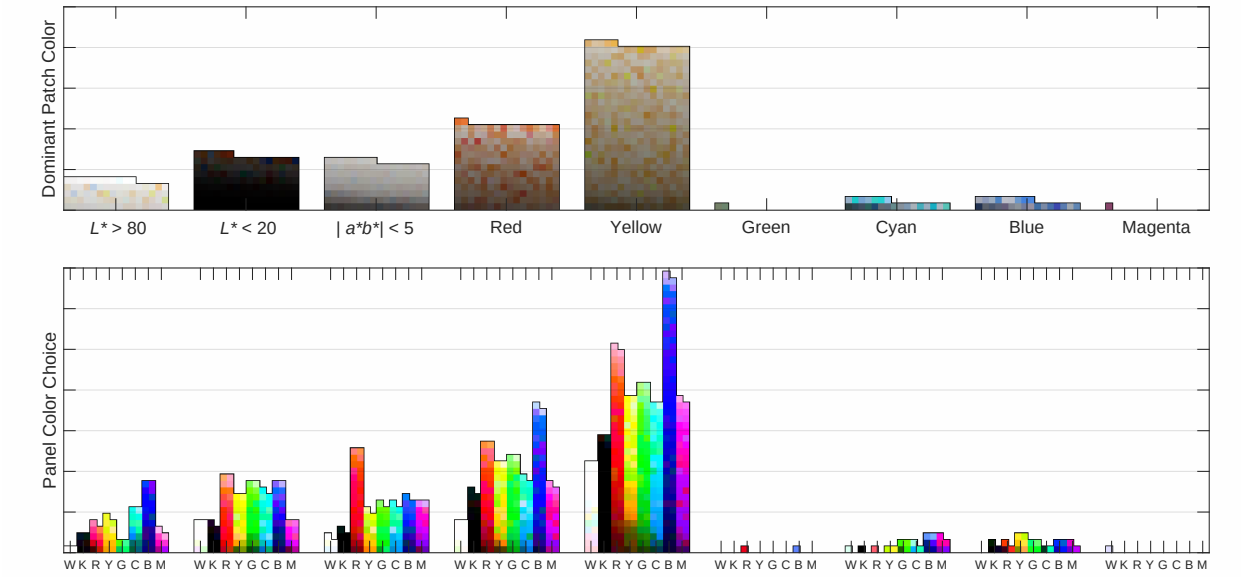

Figure 4: The top plot shows the collected samples, binned based on hue, lightness (L^*) and saturation (a^*b^*). Each pixel shows the dominant color of the image patch (on which basis they are binned). The bottom plot shows the distribution of selected panel colors for each of the top groups, binned according to panel hue and lightness (W: $L < 0.1$, K: $L > 0.9$, R: red, Y: yellow, G: green, C: cyan, B: blue, M: magenta). A preference for blue and red panel coloration is observable, particularly in the Red and Yellow patch groups.

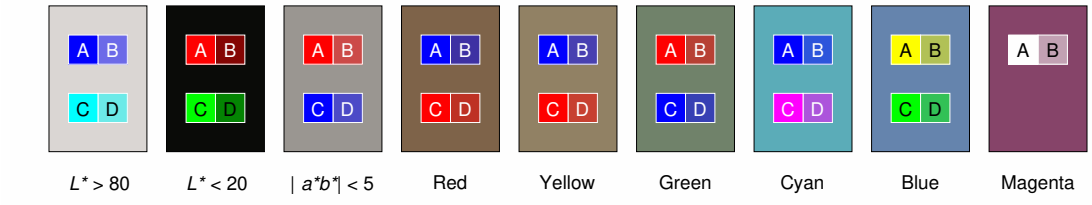

Figure 5: Most (A) and second most (C) frequently selected billboard color overlaid on the median color of the corresponding groups in Figure 4. Billboards at 50% opacity shown by (B) and (D). Note that the Magenta group only includes a single sample and so no secondary billboard color is shown.

(58.5%) but there is a second peak at $s = 6$ (6.8%) and a third peak at the other extreme, $s = 20$ (6.7%). This result suggests some stratification in the behavior of participants. Additionally, a small number of the images appeared to be provided as extreme close-ups or with the camera lens covered, yet with no change to the pixelation level. This suggests that these users were further, conscientiously, trying to ensure their privacy. While this mechanism is entirely carried out on-device, these users may have been taking active steps to ensure images were not being surreptitiously captured without their consent.

The proportion of approved images in which the panel and text color was altered provides a proxy measure for task engagement. The panel color was adjusted in 75.7% of images and the text color was adjusted in 47.6%. Note that a participant may choose not to change the panel color if they consider its initial value to be appropriate given the background. Nevertheless, we can conservatively estimate that at least three quarters of participants were actively engaged in the appearance refinement activity as instructed.

4.1.2 Billboard Color Choice. This section describes the process of mining the collected context image and appearance refinement dataset for common patterns that inform the billboard color selection problem. It is reasonable to anticipate that the dataset suffers from various noise factors, such as individual user preferences, user apathy[2], and interpretation differences. These types of factors appear in lab studies but their effect is more extreme in crowdsourcing due to the fact that only limited and unsupervised training can be provided in a web-based task. Nevertheless, these effects can be mitigated by collecting large volumes of observations. To demonstrate the potential of a more complete dataset, this section shows that useful information, on par with similar lab studies, can be extracted

[2]A small minority of crowdworkers are known to race through tasks providing nonsensical data in order to minimize completion time [8].

from our 200 participant dataset. In addition to the noise factors described above, there are also aspects of the signal that frustrate simple analysis techniques. For an identical background context there are likely to be multiple billboard color choices that yield similar legibility and aesthetics from the user's perspective.

Accepting that the dataset likely contains both of these noise and signal effects, we seek to uncover any summative patterns reflected in the data. This analysis strategy involves i) identifying informative groupings of similar background contexts; and ii) identifying common panel colors selected for these groupings. To do this, we first extract the sub-region or patch of the full image upon which the description panel was displayed. The dominant color of this patch is then extracted by taking the mode of the hue histogram and the mean of the S (saturation) and V (value) values in HSV space. We hypothesized that the patch hue is unlikely to influence panel color selection at high (i.e. white) and low (i.e. black) lightness values and at low saturation (i.e. grey) values. To group on low saturation we place a threshold on the vector a^*b^* (<5) of the dominant color in CIE 1976 $L^*a^*b^*$ color space. High and low lightness values are grouped by thresholds on L^* (>80 and <20 respectively). The remaining ungrouped patches are then binned based on their hue value. Binning was performed according to standard 60° segments around the hue circle (with 'red' on the interval -30° to 30°). The resultant groups are illustrated in the top of Figure 4. Each patch is represented by a pixel based on the patch's dominant background color.

The groupings shown in Figure 4 are highly illustrative of typical background contexts to be encountered in AR. A review of the collected images indicate that the vast majority (95%) of images were captured indoors (more detailed results are presented later in Section 4.1.5). White and off-white are common interior colorings. Similarly, the large groups for red and yellow correspond well with the large number of wood paneling and brick backgrounds captured. Far less common are background contexts with prominent green, cyan, blue and magenta coloring. The prevalence of black is largely due to images captured in low light.

The billboard colors chosen by participants corresponding to each of these background contexts were then grouped. To support summative review, the selected hue value was binned into its corresponding segment on the hue circle (again 60° segments, with 'red' on the interval -30° to 30°). Billboard colors with extreme lightness, L, values (recall participants could modulate lightness using the slider) were separated into black ($L < 0.1$) and white ($L > 0.9$) bins. Figure 4 (bottom) shows the panel color selection based on this binning. An interpretation, therefore, of Figure 4 is that it shows the distribution of billboard color choice given the dominant background color.

Figure 5 summarizes the results of the background groupings by overlaying billboards of the most (A) and second most (C) frequent color choices on the median color of the clustered backgrounds. Also shown are the billboards at 50% opacity (B and D respectively). Clearly these groupings are sensitive to small datasets but the exploratory results are promising. Figure 5 shows a consistent preference for red and blue panels despite diverse background settings. Figure 6 shows a plot similar to Figure 4, grouped solely based on lightness. The corresponding most and second most frequent billboard color choices are shown in Figure 7. These plots again highlight the general preference for blue panels, except when the background is very dark, in which case bright colors, such as red and green, are preferred. This result shows good alignment with the lab-based findings of Debernardis et al. [3] and Kruijff et al. [13], which found a distinct preference for blue panels. In contrast to these studies, however, our data presents a much better picture of the sensitivity of this choice.

4.1.3 Text Color Choice. The second critical aspect for text billboard design is the assignment of text color. In the deployed web application, participants were allowed to toggle between black and white text. Therefore, the scope of this analysis is constrained to choosing between these two options.

Intuitively, black text is more legible on bright backgrounds while white text is more suitable on dark backgrounds. The World Wide Web Consortium (W3C) provides a simple recommended formula for calculating the perceived brightness of a color [26]. This yields a brightness value on the range 0 to 255. The W3C suggests a brightness difference of 125 promotes good visibility [26], essentially maximizing perceived contrast. Figure 8 (left) shows the boxplots of perceived billboard color brightness grouped according to the choice of black or white text in all samples. Figure 8 (right) shows the same boxplots, excluding samples in which the text color was unchanged. The median brightness for black text selection is significantly higher than that for white text selection, as expected. The spread of each group does, however, highlight the fact that there is no clear threshold indicating the point at which one is clearly perceived by our participants to be better than the other. Indeed there is limited evidence-based guidance on an appropriate choice of this threshold. From Figure 8 (right) it can be observed, however, that the interquartile range of white text selection does not overlap with the interquartile range of black text. Therefore, the range between white text $q_3 = 131.0$ and black text $q_1 = 147.7$ may suggest a reasonable region of transition.

4.1.4 Privacy Survey. The concerns of crowdworkers related to sharing images of their private settings was also investigated concurrently as part of Experiment 1. After capturing the last image, participants completed a short survey examining their privacy concerns. They were asked to respond to three questions on a five-point Likert scale. These questions and the allocation of responses to each are summarized in Figure 9. 58.5% of participants indicated that they were either not at all concerned or somewhat unconcerned about sharing images via a Mechanical Turk HIT from a privacy point of view. This result is remarkably consistent with the usage proportion of the default pixelation value. However, 79% of participants thought it was either somewhat or very important to be able to review their images. As a method for mediating privacy concerns it appears that the pixelation functionality was considered either very or somewhat useful by 76% of participants.

In summary, the findings related to privacy highlight that Mechanical Turk workers are generally willing to provide images of their local context. The ability to obfuscate or reject sensitive images appears to successfully accommodate those with stronger reservations. To maximize data acquisition while addressing participant concerns, the image review protocol presented appears to be an effective strategy.

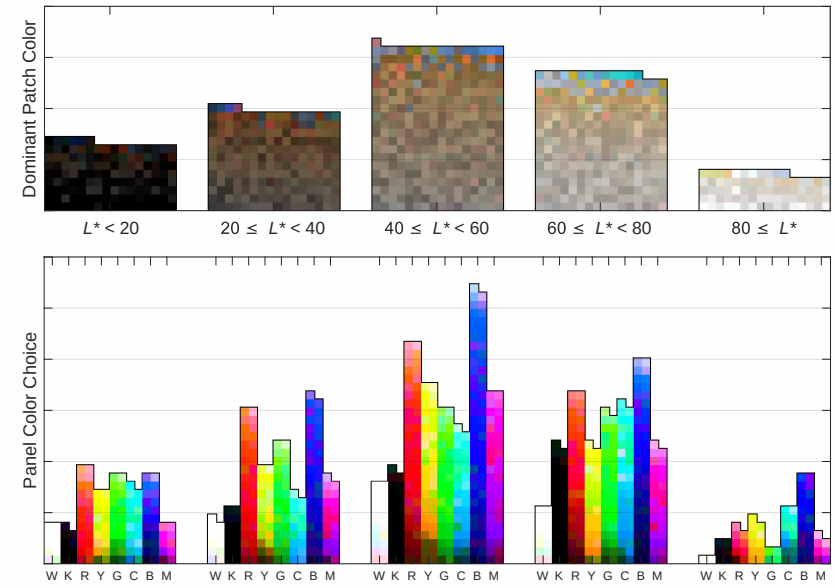

Figure 6: The top plot shows the collected samples, binned based on lightness (L^*). Each pixel shows the dominant color of the image patch (on which basis they are binned). The bottom plot shows the distribution of selected panel colors for each of the top groups, binned according to panel hue and lightness (W: $L < 0.1$, K: $L > 0.9$, R: red, Y: yellow, G: green, C: cyan, B: blue, M: magenta). Again, a preference for blue and red panel coloration is observable, particularly in the $20 \leq L^* < 40$ and $40 \leq L^* < 60$ groups.

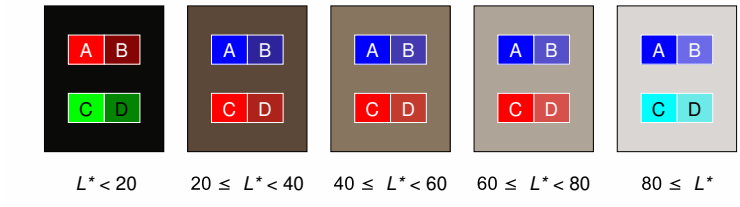

Figure 7: Most (A) and second most (C) frequently selected billboard color overlaid on the median color of the corresponding groups in Figure 6. Billboards at 50% opacity shown by (B) and (D).

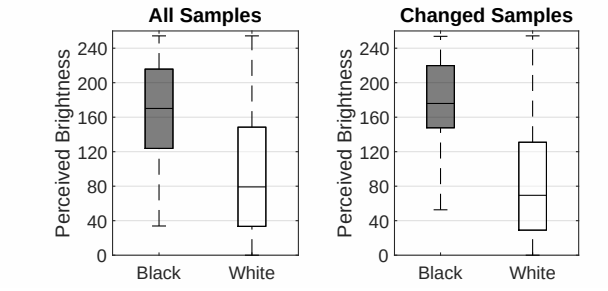

Figure 8: Boxplots of the perceived brightness of the chosen billboard color with grouping based on the user's selection of black or white text. The left plot contains all samples while the right plot contains only samples where the font color was changed by the user.

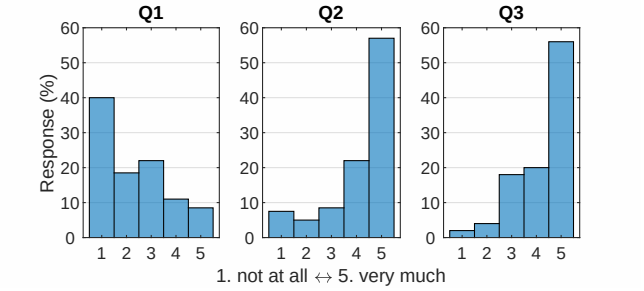

Figure 9: Responses to survey questions from 1—not at all to 5—very much. Q1. Do you have any privacy concerns about sharing images of your workspace via a Mechanical Turk HIT? Q2. Do you think it is important to be presented with your images for review before sharing? Q3. Did you find the blurring capability useful for removing private detail from captured images?

4.1.5 Range of Contexts Captured. The range of locations in which images were captured by crowdworkers yields an appreciation of how representative the data capture is of anticipated usage contexts of AR. Table 2 summarizes the variety of locations in which images were captured for both Experiment 1 and 2. These locations were categorized by the authors and were not self-reported by the participants. The assignment of images to these categories was based on the presence of objects or features clearly indicative of a particular location. Images not clearly falling into one of the indoor categories were assigned to 'Other'. The variety of contexts captured highlights the value of our method over an equivalent lab-based study. It is particularly interesting to note that approximately 5% of images in Experiment 1 and approximately 8% in Experiment 2 were collected outdoors. This analysis suggests that Mechanical Turk workers are less tied to their computer than typically assumed. Based on this finding, future work may examine the feasibility of prompting crowdworkers to enter prescribed contexts.

5 EXPERIMENT 2: PANEL PLACEMENT

Experiment 2 investigated the placement of text panels within the environment. Specifically, this experiment captured contextualized user feedback on the preferred placement of text panels, accounting for coloration, given the physical background. The process of capturing the initial image was identical to Experiment 1. Upon targeting the bird, however, rather than the full description panel shown in Experiment 1, only a label of the bird name was shown. This label was placed randomly around the bird but within the view frame and a leader line connected the bird model and the label. Once the image was captured, the participant was instructed to: "Place the label so as to maximize visibility and readability of the text." The label could be moved by simply touching on the screen within the image frame and/or by modifying the apparent depth of the label in the scene using a slider. Note that label color was randomized and text color was randomly assigned to be either black or white. Participants were still given the opportunity to review, reject or pixelate their images as required, however, the survey examining privacy concerns was removed.

5.1 Results

As with Experiment 1, 200 participants (125 male, 74 female, 1 unspecified, 31.2 mean age) from 18 different countries were recruited through Amazon Mechanical Turk. Each received US$1 as compensation for their time. Participants were only permitted to complete the task once (participants from Experiment 1 were not prevented from completing Experiment 2). The mean completion time for the task was 7.2 minutes (including training and instructions).

Table 2: Summary of participant image locations (as interpreted by the authors) in both experiments.

Location		Exp. 1	Exp. 2
Indoors	Home Office / Workplace	41	39
	Bedroom	2	4
	Kitchen	7	6
	Bathroom	11	2
	Other	880	863
Outdoors	Garden / Balcony	54	66
	Car		15

5.1.1 Approval Rate and Pixelation Behavior. With each participant again capturing five images, there were a potential 1,000 total images from 200 participants. The approval rate was again very high with only five images rejected in total by five different participants (an approval rate of 99.5%). The distribution over the usage of different pixelation levels is also roughly consistent with Experiment 1 (see Table 1). No pixelation, $s = 1$, again dominates (54.0%) but with secondary peaks at $s = 6$ (6.6%) and $s = 20$ (12.0%).

5.1.2 Label Placement Behavior. Figure 10 illustrates the initial and final label placement centers. The target (virtual bird) is always centered in this window. Recall that the initial label location was randomized relative to the bird target. The left plot in Figure 10 shows the initial randomized position of the label relative to the bird. The right plot in Figure 10 reflects the distribution of the final placement locations across all samples. Notable in this plot is the frequency of label placements above and below the bird model while also avoiding overlap with the model itself. This behavior suggests a label placement preference that is, in part, independent of the background context. Note that this observed label placement behavior provides empirical grounding to the related strategy Lindlbauer et al. [17] use to assess image sub-regions for label placement. It is important to note, however, that the display orientation and label sizing clearly has an effect on constraining appropriate label placement locations and isolating this influence for non-standard virtual objects may require specific investigation.

5.2 Influence of Background Texture and Color on Placement

The approaches taken in the literature of scoring label placement locations based on texture and coloration suggest exploring whether this behavior is observable in the dataset. First we test the hypothesis that a highly colorful background region will be avoided when placing the label. Hasler and Suesstrunk [10] introduce a simple *colorfulness* metric that can be computed based on an image's RGB color space. Hasler and Suesstrunk [10] define the colorfulness metric, M, to provide correspondence with human judged attributes of an image ranging from *not colorful* to *extremely colorful*. The colorfulness metric M first requires collapsing the color channels into: $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$. These are then transformed into a representative mean, $\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$, and standard deviation, $\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}$. Finally, M is computed using the formula: $M = \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb}$.

Figure 11 (left) shows boxplots of the change in colorfulness ΔM between the initial label placement region and the final label placement region for three groups of initial region colorfulness. The change in colorfulness ΔM will be negative when the label is moved from a colorful region to a less colorful region. Figure 11 suggests that when the initial region is *not colorful* ($M < 15$), users typically find a region that is similarly flat in color. When the initial region is *slightly colorful* ($15 \geq M < 33$), there is some sign of a general preference for placement in regions yielding a negative ΔM.

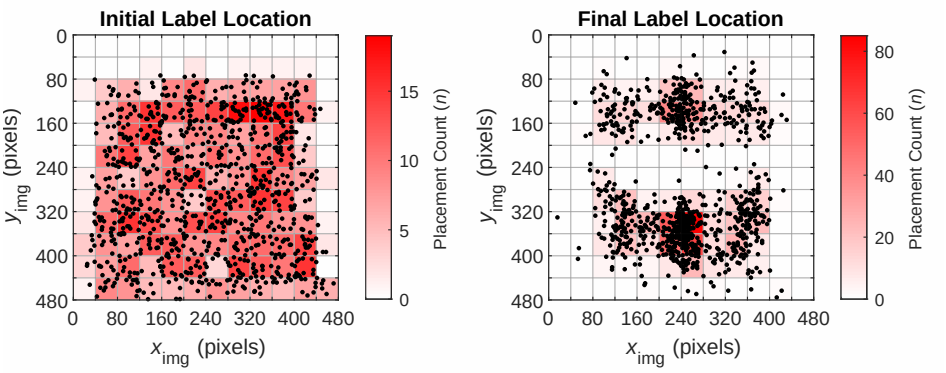

Figure 10: Black dots denote the center of the initial (randomized) label placement location (left) and final label placement location (right) within the captured image window. Frequency of placement within image sub-regions (regular 40×40 pixel blocks) is represented by the coloration.

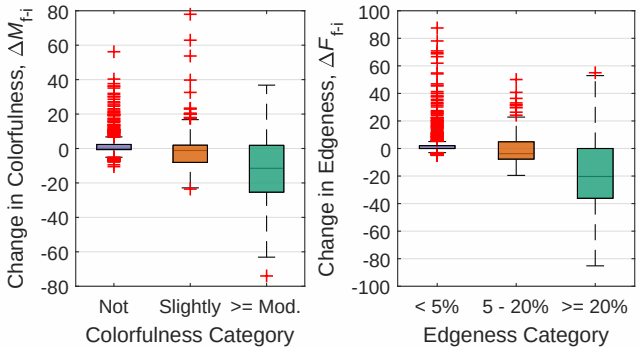

Figure 11: (left) Boxplots of change in *colorfulness*, ΔM between the initial and final label placement region, grouped based on initial region *colorfulness* (*not colorful, slightly colorful* and *moderately colorful* or more). (right) Boxplots of change in *edgeness per unit area*, ΔF, between the initial and final label placement region, grouped based on initial region *edgeness*.

When the initial region is *moderately colorful* or more ($M \geq 33$), there is a definite bias towards a negative ΔM. This suggests that less colorful regions are preferred for label placement.

Another informative point of analysis is the influence of background clutter on label placement. *Edgeness per unit area*, F, is a simple metric for quantifying the degree of texturing or 'busyness' of an image [27]. F is computed for a region of N pixels by counting the number of pixels, p, for which the gradient magnitude, $\text{Mag}(p)$, exceeds threshold, T. More concisely: $F = \frac{|\{p|\text{Mag}(p)\geq T\}|}{N}$. The change in edgeness between the initial label patch and the final label patch, ΔF, provides an indication of the effect of background 'busyness' on placement behavior. Figure 11 shows boxplots of ΔF over three groupings of initial patch edgeness ($T = 100$). Moving from a patch with high edgeness to a patch with less texture will yield a negative ΔF. Figure 11 (right) suggests that when the initial patch has low edgeness (<5%) the ΔF is likely to be close to zero. As the edgeness of the initial patch increases, however, participants increasingly relocate the label to less textured regions (yielding a negative ΔF).

In summary, Experiment 2 highlights several key determinants of label placement preference: offset, colorfulness and edgeness. Recall that users were unable to set the panel color in Experiment 2 and so placement related concerns are expected to dominate. The behaviors observed in participants are consistent with the algorithmic strategies for dynamic label placement proposed by Tanaka et al. [29], Orlosky et al. [23] and others. This agreement between the empirical evidence gathered via crowdsourcing and the design solutions evolved by others provides confidence in the crowdsourcing methodology and its value in delivering data-driven guidance.

6 A PREFERENCE MODEL FOR DYNAMIC TEXT PANELS ON AN OPTICAL SEE-THROUGH HMD

The purpose of this case study is to highlight the viability of the described crowdsourcing experimental method to inform head-mounted AR interface design. To confirm the design guidance obtained is useful and implementable, we demonstrate a high-fidelity AR application solution for contextually adaptive text panels. This application, designed for use with the Microsoft HoloLens OST HMD, provides dynamic placement and coloration of billboard style *tooltips*. We now present the design of the dynamic text panel procedure derived from the collected data.

6.1 System Design

Formalizing the color selection and placement problem for text panels in AR necessitates the consideration of three sub-problems: i) billboard color choice; ii) text color choice; and iii) billboard placement. A simple strategy for dynamic text appearance adaptation can be derived from the collected data using a compounding probabilistic approach. The approach converts the frequency responses observed for color choice and placement (in terms of offset, coloration and edgeness) into probabilities. It then combines them to

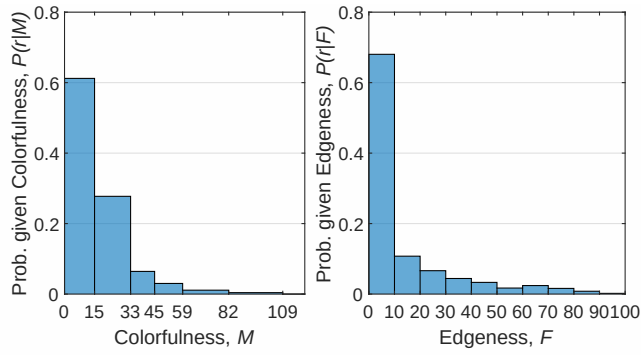

Figure 12: Estimated probability distributions of the likelihood of sub-region selection given colorfulness, M, (left), and edgeness, F (right). Note that the binning of colorfulness, M, is based on the groupings defined by Hasler and Suesstrunk [10].

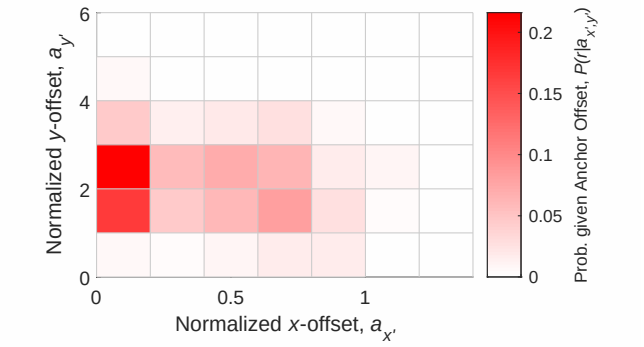

Figure 13: Estimated probability distribution of the likelihood of sub-region given normalized x,y offset (symmetric about the x and y axis).

yield a mixture distribution estimating the preferred placement sub-region, r, in an image and the preferred color, c, given that region. The estimated distributions for placement region given colorfulness and edgeness are presented in Figure 12. The estimated distribution for placement offset (normalized based on the billboard width for x and height for y in image coordinates) is presented in Figure 13. This procedure requires transforming the tooltip anchor location (i.e. the point referred to by the tooltip) into the image coordinate system and the selected tooltip location in image coordinates back into the world frame. Accepting that designers typically wish to provide an interface with a consistent color palette, a final uniform distribution is applied over a set of predetermined colors. This distribution serves to bias the color selection towards selecting only from within the palette, but informed by the preference model. The entire procedure is summarized in Algorithm 1.

The estimated likelihood of selecting region, r, given edgeness (line 5) for an example tooltip target location is illustrated in Figure 14. The combined mixture distribution for this same target location (the summing log probabilities step at line 7 in Algorithm 1) is illustrated in Figure 15. The resulting tooltip placement and coloration for this target location is illustrated in Figure 16.

Algorithm 1: Contextually Adaptive Tooltips

1 **Function** `AdaptTooltip`(I, a)
  Input : Background image, I, and tooltip anchor location, $a_{x,y,z}$
  Output : Tooltip position, $t_{x,y,z}$, billboard color, c_b, and text color, c_t
2 Transform anchor position, $a_{x,y,z}$, into its equivalent position in image coordinates, $a_{x',y'}$
3 **foreach** *Sub-region, r, of the background image, I* **do**
4 Lookup $P(r|M)$, probability of selecting r given colorfulness, M
5 Lookup $P(r|F)$, probability of selecting r given edgeness, F
6 Lookup $P(r|a)$, probability of selecting r given offset from anchor position, $a_{x',y'}$
7 Combine $P(r|M)$, $P(r|F)$ and $P(r|a)$ to yield mixture distribution, $H(r)$
8 **end**
9 Choose sub-region, r_{max}, corresponding to the maximum of the mixture distribution, $H(r)$
10 Extract dominant patch color, c_p, and patch lightness, l_p, from image region r_{max}
11 **foreach** *Billboard color group, g, in billboard color groupings* **do**
12 Lookup $P(g|c_p)$, probability of selecting g given patch color, c_p
13 Lookup $P(g|l_p)$, probability of selecting g given of patch lightness, l_p
14 Lookup $P(g|palette)$, probability of selecting g given defined color palette
15 Combine $P(g|c_p)$, $P(g|c_p)$ and $P(g|palette)$ to yield mixture distribution, $G(g)$
16 **end**
17 Choose group, g_{max}, corresponding to the maximum of the mixture distribution, $G(g)$
18 Choose billboard color, c_b, corresponding to g_{max} in palette
19 Choose text color, c_t, based on threshold of the perceived brightness of color c_b
20 Transform image coordinates of the centre of r_{max} to equivalent world position, $t_{x,y,z}$
21 **end**

7 DISCUSSION

This paper serves as a vehicle for demonstrating the value of crowdsourced AR evaluation datasets. In this investigation, context and physical-virtual dependence information was captured across heterogeneous settings. This information was readily operationalized to build a prototype application delivering contextually-adaptive text content on an early commercially available AR OST HMD. Below we discuss limitations of the investigation and promising future avenues.

The dynamic text panel case study presented in Section 6 serves to illustrate how the crowdsourced design guidance can be translated into practical use. The derived preference model satisfies this

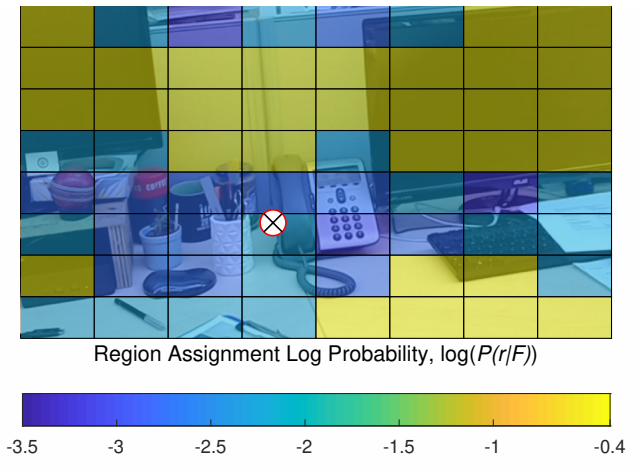

Figure 14: Tooltip placement probabilities given edgeness of sub-regions in the background image. Yellow regions indicate high probability. Blue regions indicate low probability. The tooltip anchor center is indicated by the red circle.

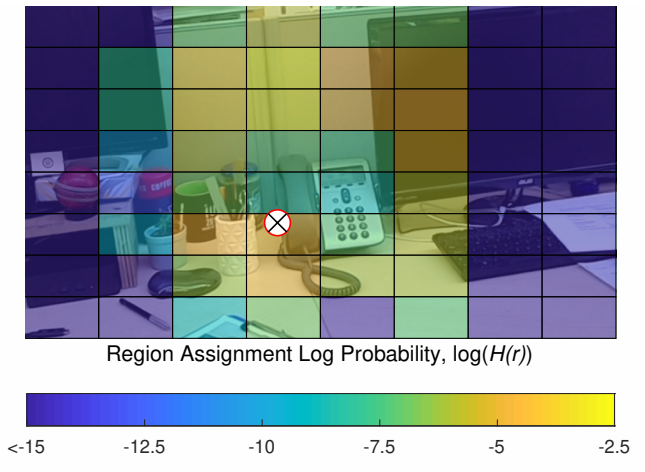

Figure 15: Resultant tooltip placement mixture probabilities of sub-regions in the background image. Yellow regions indicate high probability. Blue regions indicate low probability. The tooltip anchor center is indicated by the red circle.

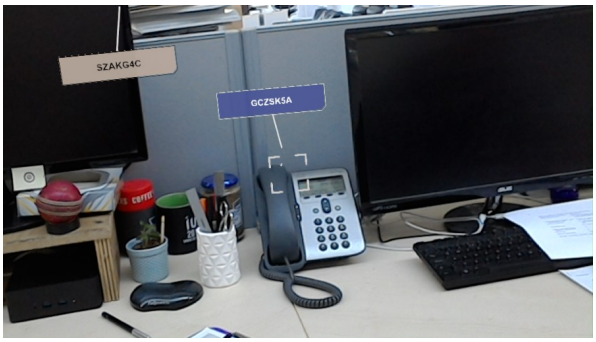

Figure 16: Two example tooltips generated by the preference model as viewed in the Microsoft HoloLens.

objective and demonstrates that the design guidance can be utilized. However, it does not strictly validate its efficacy in improving text panel legibility. Further work is required to probe the bounds of transferability and efficacy for data captured in this manner.

A limitation of this work is the confined set of design controls given to users for changing the description panel appearance. Only billboard hue, luminance and text color could be varied in Experiment 1, while only panel placement could be modified in Experiment 2. This constrained interaction space was designed to not overwhelm users and to avoid excessive 'twiddling' behavior. However, future work will explore how additional design control can be provided to users without these disruptive effects. Further, in terms of balancing task quality and engagement, it may also be feasible to leverage the model for text panel preferences derived in Section 6 to introduce quality assurance checks into the task. For example, crowdworkers can periodically be presented with a control task involving a standard background context for which there exists a clear set of acceptable design choices based on the preference model.

The perception of color is sensitive to a great number of factors. In this investigation, we do not enforce any calibration or specific display settings on the device. This means that color rendering differences between devices may introduce noise into the user feedback. We made the experimental choice not to control for this factor since this is more realistic of an actual user's experience with a simple application: it is useful to have a model for dynamic adaptation that works for most users in most cases. Nevertheless, there is an opportunity for a strict investigation of how device variation might influence design choices and what experimental controls can be applied to address this factor.

The task and its framing are likely to have influenced participant attitudes towards privacy and consent. The image capture aspect of the task was intentionally embedded within the bird finding activity. The plausible reason for capturing contexts was designed to positively bias participants towards the task, but was not specifically investigated as a factor. Related to this is the task introduction, which made explicit mention of the university affiliation and stated that data will be anonymized and images will not be published. Tasks requested by researchers may positively influence trust. Such aspects of participant behavior have been previously explored [12] but are worthy of examination in this context.

There is active work in streamlining mobile AR frameworks for use on the web. The WebXR Device API[3] is a working draft for supporting VR and AR on the web. This standard outlines support for six degrees of freedom (DOF) pose tracking with mobile devices. This presents a significant opportunity for enhancing the fidelity of the mobile AR experience presented to crowdworkers. With six DOF tracking, the experiments described in this paper could examine a wider range of additional factors influencing AR content presentation. Furthermore, while the example experiments were limited to static image capture, the approach, where sufficient bandwidth is available, could be extended to real-time video capture

[3]https://www.w3.org/TR/webxr/

and processing. This would enable the investigation of temporal and spatial coherence of virtual content.

8 CONCLUSIONS

This study demonstrates that crowdsourcing context information for adaptive AR is not only feasible but also efficient. Web-based AR offers the ability to rapidly access a large corpus of users and environments, today. Even well-equipped labs, with a large number of AR devices, would struggle to capture the diverse range of environments recorded by the participants in this study in such a short time period. For only US$440 (including platform fees), two context-dependent AR user studies were conducted with 400 users spanning 22 countries to assemble a dataset of almost 2,000 images and user-defined billboard preference profiles. Crowdworkers were willing to engage with a low-fidelity AR experience and share images of their local environment.

We have demonstrated that the user preference data captured via this low-fidelity mobile AR experience can be readily transferred to deliver contextually-adaptive functionality on an OST HMD. Overall, this paper highlights new avenues for investigating and evaluating contextually-informed AR applications using crowdsourcing. Considering the inherent complexity in AR user interface design, crowdsourcing is a promising complementary method to assist evolving new data-driven designs that are difficult to achieve using traditional lab studies. The potential improvements in external validity offered by the AR crowdsourcing method for obtaining emerging design guidance is a crucial contribution at this early stage of AR user interface development.

ACKNOWLEDGMENTS

This work was supported by EPSRC (grants EP/R004471/1 and EP/S027432/1). Supporting data for this publication is available at https://doi.org/10.17863/CAM.62931.

REFERENCES

[1] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE* 8, 3 (March 2013), e57410. https://doi.org/10.1371/journal.pone.0057410

[2] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys (CSUR)* 51, 1 (April 2018), 7. https://doi.org/10.1145/3148148

[3] Saverio Debernardis, Michele Fiorentino, Michele Gattullo, Giuseppe Monno, and Antonio Emmanuele Uva. 2014. Text Readability in Head-Worn Displays: Color and Style Optimization in Video versus Optical See-Through Devices. *IEEE Transactions on Visualization and Computer Graphics* 20, 1 (Jan. 2014), 125–139. https://doi.org/10.1109/TVCG.2013.86

[4] John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2019. Crowdsourcing Interface Feature Design with Bayesian Optimization. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 252:1–252:12. https://doi.org/10.1145/3290605.3300482

[5] Joseph L. Gabbard and J. Edward Swan II. 2008. Usability Engineering for Augmented Reality: Employing User-Based Studies to Inform Design. *IEEE Transactions on Visualization and Computer Graphics* 14, 3 (May 2008), 513–525. https://doi.org/10.1109/TVCG.2008.24

[6] Joseph L. Gabbard, J. Edward Swan II, and Deborah Hix. 2006. The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. *Presence: Teleoperators and Virtual Environments* 15, 1 (Feb. 2006), 16–32. https://doi.org/10.1162/pres.2006.15.1.16

[7] Joseph L. Gabbard, J. Edward Swan II, Deborah Hix, Robert S. Schulman, John Lucas, and Divya Gupta. 2005. An empirical user-based study of text drawing styles and outdoor background textures for augmented reality. In *IEEE Proceedings. VR 2005. Virtual Reality, 2005.* 11–18. https://doi.org/10.1109/VR.2005.1492748

[8] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1631–1640. https://doi.org/10.1145/2702123.2702443

[9] Raphaël Grasset, Tobias Langlotz, Denis Kalkofen, Markus Tatzgern, and Dieter Schmalstieg. 2012. Image-driven view management for augmented reality browsers. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 177–186. https://doi.org/10.1109/ISMAR.2012.6402555

[10] David Hasler and Sabine E. Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, Vol. 5007. International Society for Optics and Photonics, 87–95. https://doi.org/10.1117/12.477378

[11] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357

[12] Jason T. Jacques and Per Ola Kristensson. 2013. Crowdsourcing a HIT: Measuring Workers' Pre-Task Interactions on Microtask Markets. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[13] Ernst Kruijff, Jason Orlosky, Naohiro Kishishita, Christina Trepkowski, and Kiyoshi Kiyokawa. 2018. The Influence of Label Design on Search Performance and Noticeability in Wide Field of View Augmented Reality Displays. *IEEE Transactions on Visualization and Computer Graphics* (2018), 1–1. https://doi.org/10.1109/TVCG.2018.2854737

[14] Walter S. Lasecki, Mitchell Gordon, Winnie Leung, Ellen Lim, Jeffrey P. Bigham, and Steven P. Dow. 2015. Exploring Privacy and Accuracy Trade-Offs in Crowdsourced Behavioral Video Coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1945–1954. https://doi.org/10.1145/2702123.2702605 Seoul, Republic of Korea.

[15] Walter S. Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P. Bigham. 2013. Real-Time Crowd Labeling for Deployable Activity Recognition. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1203–1212. https://doi.org/10.1145/2441776.2441912 San Antonio, Texas, USA.

[16] Alex Leykin and Mihran Tuceryan. 2004. Automatic determination of text readability over textured backgrounds for augmented reality systems. In *Third IEEE and ACM International Symposium on Mixed and Augmented Reality*. 224–230. https://doi.org/10.1109/ISMAR.2004.22

[17] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32Nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. ACM, New York, NY, USA, 147–160. https://doi.org/10.1145/3332165.3347945 New Orleans, LA, USA.

[18] Xiao Ma, Megan Cackett, Leslie Park, Eric Chien, and Mor Naaman. 2018. Web-Based VR Experiments Powered by the Crowd. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 33–43. https://doi.org/10.1145/3178876.3186034

[19] Jacob Boesen Madsen, Markus Tatzqern, Claus B. Madsen, Dieter Schmalstieg, and Denis Kalkofen. 2016. Temporal Coherence Strategies for Augmented Reality Labeling. *IEEE Transactions on Visualization and Computer Graphics* 22, 4 (April 2016), 1415–1423. https://doi.org/10.1109/TVCG.2016.2518318

[20] Vito Modesto Manghisi, Michele Gattullo, Michele Fiorentino, Antonio Emmanuele Uva, Francescomaria Marino, Vitoantonio Bevilacqua, and Giuseppe Monno. 2017. Predicting Text Legibility over Textured Digital Backgrounds for a Monocular Optical See-Through Display. *Presence: Teleoperators and Virtual Environments* 26, 1 (2017), 1–15.

[21] Tara McAllister Byun, Peter F. Halpin, and Daniel Szeredi. 2015. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders* 53 (Jan. 2015), 70–83. https://doi.org/10.1016/j.jcomdis.2014.11.003

[22] Daniel McDuff, Rana El Kaliouby, and Rosalind W. Picard. 2012. Crowdsourcing Facial Responses to Online Videos. *IEEE Transactions on Affective Computing* 3, 4 (2012), 456–468. https://doi.org/10.1109/T-AFFC.2012.19

[23] Jason Orlosky, Kiyoshi Kiyokawa, and Haruo Takemura. 2013. Dynamic Text Management for See-through Wearable and Heads-up Display Systems. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 363–370. https://doi.org/10.1145/2449396.2449443 Santa Monica, California, USA.

[24] Jason Orlosky, Kiyoshi Kiyokawa, and Haruo Takemura. 2014. Managing Mobile Text in Head Mounted Displays: Studies on Visual Preference and Text Placement. *SIGMOBILE Mob. Comput. Commun. Rev.* 18, 2 (June 2014), 20–31. https://doi.org/10.1145/2636242.2636246

[25] Janne Paavilainen, Hannu Korhonen, Kati Alha, Jaakko Stenros, Elina Koskinen, and Frans Mayra. 2017. The PokéMon GO Experience: A Location-Based Augmented Reality Mobile Game Goes Mainstream. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2493–2498. https://doi.org/10.1145/3025453.3025871

[26] Chris Ridpath and Wendy Chisholm. 2000. *Techniques For Accessibility Evaluation And Repair Tools.* https://www.w3.org/TR/AERT

[27] George Stockman and Linda G. Shapiro. 2001. *Computer Vision* (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.

[28] Chek Tien Tan, Hemanta Sapkota, and Daniel Rosser. 2014. BeFaced: A Casual Game to Crowdsource Facial Expressions in the Wild. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 491–494. https://doi.org/10.1145/2559206.2574773

[29] Kohei Tanaka, Yasue Kishino, Masakazu Miyamae, Tsutomu Terada, and Shojiro Nishio. 2007. An Information Layout Method for an Optical See-through HMD Considering the Background. In *2007 11th IEEE International Symposium on Wearable Computers*. 109–110. https://doi.org/10.1109/ISWC.2007.4373791 ISSN: 2376-8541.

[30] Kohei Tanaka, Yasue Kishino, Masakazu Miyamae, Tsutomu Terada, and Shojiro Nishio. 2008. An Information Layout Method for an Optical See-through Head Mounted Display Focusing on the Viewability. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR '08)*. IEEE Computer Society, Washington, DC, USA, 139–142. https://doi.org/10.1109/ISMAR.2008.4637340

[31] Markus Tatzgern, Valeria Orso, Denis Kalkofen, Giulio Jacucci, Luciano Gamberini, and Dieter Schmalstieg. 2016. Adaptive information density for augmented reality displays. In *2016 IEEE Virtual Reality (VR)*. 83–92. https://doi.org/10.1109/VR.2016.7504691 ISSN: 2375-5334.

[32] Jason Wither, Stephen DiVerdi, and Tobias Höllerer. 2009. Annotation in outdoor augmented reality. *Computers & Graphics* 33, 6 (2009), 679 – 689. https://doi.org/10.1016/j.cag.2009.06.001