# Supplementary Materials

"Don't Look Now: Audio/Haptic Guidance for 3D Scanning of Landmarks"

## Contents

# 1 Frequently asked questions

**You give a lot of numbers, and the audio/haptic app worked better in *some* ways, but *on average* is it better?** The final study included 50 participants. The null-hypothesis would state that the two final apps have the same effectiveness across that population. But the experiment showed that the choice of guidance app actually *does* make a difference. Yes, there are users who make good or even great scans with the visual app. But more users are better off with the audio/haptic interface in terms of engagement (which contains several criteria), and in terms of the quality of the resulting scan. Statistical significance tests are designed specifically to measure such situations (and we are happy to present the numbers in Bayesian terms instead of Frequentist statistics if needed). To prevent carryover effects, we used a between-subjects design. Visual inspections and ANCOVA/MANCOVA analyses were performed before analyzing specific variables (e.g. engagement) to reduce the chance of Type I errors. We used a validated survey to measure engagement (UES-SF) [35]. See Section 13 for full statistical test results.

**Did people *really* enjoy the audio/haptic app more than the visual app?** Besides the quantitative engagement results, we received many positive comments about the audio and haptics in our scanning apps, some of which are captured by the pilot study qualitative analysis in Section 9, and the anonymous comments from the final study in Section 14. For many people, it feels satisfying.

**You were motivated to improve safety by using audio/haptic instead of visual guidance, but I don't see safety-related results.** Although we initially set out to improve the safety of scanning, we did not want to put any of our participants in significant danger in order to measure it. Instead, we measured safety proxies (e.g. stress) during a challenging—but not dangerous—scanning task. As shown in Section 13, we did not find any significant results with respect to these proxies, despite measuring stress in multiple ways.

**Why don't the reconstructed meshes look better?** The meshes shown in the main paper are from single scans. In practice, multiple scans are often used to make reconstructions that capture different weather and lighting conditions. By aggregating and aligning the meshes from different scans, it is possible to compute a single overall and superior reconstruction, along with (non-visual) characteristics that support other tasks such as relocalization, as shown in Google's Maps Tiles API. There are also advanced reconstruction methods (e.g. NERFs) that use priors and regularization, but the design choices there are a separate area of research.

**How are your scanning apps different from those on the market?** We present a table comparing our scanning apps with those on the market in Section 6.

**What about scanning POIs that aren't the Agatha Christie Memorial? What if I can't go 360° around the object I want to scan?** As shown in the Video Figure, the current application can be used to scan many other statues and objects you can walk in a circle around. For objects that cannot be circled, like murals, the app's Completion Guidance (see Section 5) could easily be modified to require users to scan only 180° or less. The app could also be modified such that users "follow" a moving Object Transform e.g. along a building's facade or through an open area. There was evidence users could track moving Object Transforms in our study, as we found that users who encountered drift could identify the Transform's drifted-to location (though intentional moving/tracking would be communicated clearly to avoid frustration).

**Doesn't scanning violate the privacy of strangers walking past?** Yes, laws like GDPR have successfully driven companies to anonymize footage of people's faces and license plates that are captured in public. For example, faces and plates are blurred out in Google StreetView. The same is true for images in academic 3D reconstruction datasets. If the proposed approach were packaged for wide-scale use, the resulting data should go through an anonymization (face-blurring) pipeline, such as the pipeline [24] used by Wayfarer (the app we used in the pilot). For this CHI paper, we blurred all the faces in figures and videos using Photoshop and code from GitHub. This code does occasionally detect and blur Agatha Christie's face however.

# 2 Scanning Greyfriars Bobby: A user experience story

To clarify the idea of scanning, as well as motivate our app's design, in this section we describe the user experience of a fictional character, "Vanessa". Vanessa uses a fictional scanning app, "Story Scan", which we imagined through combining various well-established visual scanning apps [15, 31, 13, 14, 27, 11, 30, 19, 32], as well as our own scanning app design. We illustrate this story in Figure 1 in the main paper.

> When Vanessa first set foot in Edinburgh, she didn't know much about the history of the town. One look at the ominous castle lit up in the night sky, though, and her mind was filled with medieval knights and bloody battles. A far cry from her small-town Canadian home she knew and loved so well.
>
> After a few weeks in the city, though, her classmate showed her a different side of Edinburgh: One with a small, brave Skye Terrier, Greyfriars Bobby, who patrolled the city, endeared its citizens, and guarded the grave of his owner for 14 years. The city so loved its loyal pup, they commemorated him with a statue in a garden of flowers.
>
> The garden quickly became one of Vanessa's favourite spots. She wanted to share the tale of Greyfriars Bobby with her family back home—and when she heard she could create a virtual, 3D version of the statue and garden with the app, "Story Scan", she immediately downloaded it to her phone. **The app allowed her to scan the area, add a virtual storytelling-pup to the scene, and share the experience with anyone online.** Vanessa set out to complete her first scan.
>
> Upon arrival, she was a bit unsure she was up to the task. People were scattered around the park and **she didn't want them to think she was capturing footage of them**—even though Story Scan promised all footage from the scan would be anonymized. She also had no idea what scanning entailed, and how she could create a virtual mesh that actually looked good.
>
> When she opened the app though, her fears were at least partially quelled. Through a tutorial, she learned that **all she needed to do was to slowly walk around the statue in a circle while aiming the phone's camera at the statue**. Sounded easy enough!
>
> As she walked, she saw what looked like a transparent, green, bumpy terrain appearing over the ground, flowers, and statue. She was focused on making the green mesh fit more tightly on the statue, when suddenly, she collided with a tree. Stumbling a bit, she looked over her shoulder to see if anyone had noticed. It seemed no one had, so she jogged ahead, trying to keep her phone camera aimed at Bobby.

**After completing a full 360° of the statue, her virtual mesh appeared on-screen.** The first half of the statue looked fantastic: a detailed and realistic representation of Bobby. The second half though—the half where she was moving too fast to keep her eyes on green mesh's alignment—looked jagged and smeared. She guessed she would have to scan once more, **moving a bit more slowly**—and making sure she didn't run into any trees!

# 3 Related Work: Reconstructing geometry from scans

As discussed in the main paper, we present a walk-through of smartphone-related scanning considerations.

To compare the accuracy of various scanning systems, mesh reconstructions are needed. Almost all reconstruction pipelines start with estimating camera pose. This usually requires image feature extraction, feature matching, point triangulation, and a global optimization step. This can be done offline as in SfM [39] or online as in SLAM [8]. It is helpful to have an inertial measurement unit (IMU) [5], as in modern smartphones. Ideally, for accurate poses, images in a scan would balance having sufficient visual overlap with sufficient inter-frame *baseline*. When no overlap exists or if the images are featureless—like in images of a sidewalk or a plain wall—it becomes harder and near impossible to estimate pose. While IMUs can help fill these gaps by feeding in information on instantaneous acceleration and gyroscopic orientation, such dead reckoning will eventually lead to camera pose *drift*.

With accurate camera pose, the next step is estimating dense geometry, usually represented as meshes or point clouds. Conventionally, per-image depths are first estimated using Multi-View Stereo (MVS). There, image similarity is used to triangulate points densely across images [40], usually with the help of clever sampling [3] or CNNs [12]. Depth sensors [9] can also provide per-frame depth estimates with some confidence, as is in some iPhones. Depths can then be fused in real-time and online using TSDF fusion [16] after which meshes representing the geometry of a scene can be extracted via Marching Cubes [23]. Although tangential and purely a novel-view synthesis method, a neural radiance field [26] can also provide geometry estimates with extra effort. For any of these methods, the most accurate geometry is obtained with accurate poses, sufficient coverage of the subject with *many* overlapping views (a 360 loop is preferred), blur-free images, good camera baseline, and poses close enough to the subject. The last point is especially important for low power depth sensors on consumer devices outdoors; if the camera is too far, the sensor would not be able to estimate depth accurately.

Meshes are considered accurate when they contain all the geometry in the scene, and do not contain any extraneous geometry not in the scene. Ideally, straight walls in the scene would not be curved, and fine detail would be present without noise. A ground truth mesh or point cloud is often required for mesh evaluation, usually obtained via professional-grade hardware with sufficient time and expertise [18]. Mesh evaluation often involves sampling a dense point cloud of the mesh under evaluation, and computing distances between a candidate mesh's point cloud and the ground truth [18]. *Completeness* (smaller is better in meters) and the thresholded ratio version *recall* (higher is better) measure how much of a scene's geometry is captured in the reconstruction. *Accuracy* (counter-intuitively, smaller is better in meters) and the thresholded ratio version *precision* (higher is better) measure the extent of extraneous geometry in the reconstruction. These are averaged arithmetically in *Chamfer Distance* (CD) (smaller is better) and harmonically in *F-score* (higher is better). The best reconstructions will maximize F-score and minimize CD. For example, if a user neglects to capture images of a subject from a particular angle or if they are far enough such that

a depth sensor performs poorly, then the reconstruction will have missing regions leading to large completeness and CD distances and a low recall and F-score.

# 4 Scanning app conceptual design

Many popular apps with scanning features, including Snapchat, Pokémon Go, and Ingress, crowd-source their scans from hundreds of thousands of non-expert users [36, 15]. As described in the main paper, learning to scan well, however, is not a simple task, and non-expert users typically need study facilitators or in-app feedback to help [33, 1, 44, 25]. Further, for our audio/haptic scanning app (described later), users would need to learn to interpret the feedback signals, which may not be intuitive as the majority of apps today are visual-first rather than audio/haptic-first [22, 2]. Thus, we utilized Jackson's Theory of Conceptual Design, which aims to facilitate development of straight-forward, easy-to-learn systems.

Below, we outline the conceptual design of our scanning guidance system, which includes the system's *operational principles* (OP) and *concepts* (C)—or the essential ideas we want to communicate to users through our design [17]. We show our concept graph in Figure 1.

C: *Augmented Reality (AR)*[1]

- OP: The concept of having virtual content (*e.g.* characters, objects, sound effects) co-located in the real-world.

C: *Feature to Scan* (abstract concept)

- C: *Point of Interest (POI)*[2]
  - OP: A real-world object (*e.g.* statue) that could be incorporated into an AR experience.
- C: *Object Transform*
  - OP: When a user identifies a POI to the system, the system estimates its position, orientation and scale (*i.e.* the *Object Transform*), which allows the system to provide feedback on the user's scanning behaviour.

C: *Scan*

- OP: If the system obtains images and pose information (*i.e. Scan* information) from many different angles around a POI, the system can reconstruct a virtual version (*i.e. Mesh Reconstruction*) of the POI.

C: *Mesh Reconstruction*

- OP: The system can create a virtual version (*Mesh Reconstruction*) of a POI using information from a scan. This mesh (often alongside many other meshes) can then be used to create localised, persistent, and high-fidelity AR experiences.

In [17], Jackson describes how rapid prototyping and extensive user involvement are synergistic with conceptual design. Thus, to improve upon our initial design, we gathered user feedback in a pilot study, as described in the main paper. The results from the pilot user study heavily influenced our final design.

[1]Note that because scanning apps are used to create AR experiences, we assume the user already understands this concept. For the purposes of this study, we explain AR when onboarding the participant.

[2]Note that because scanning apps are used to scan POIs, we assume the user already understands this concept. For the purposes of this study, we explain POIs when onboarding the participant.

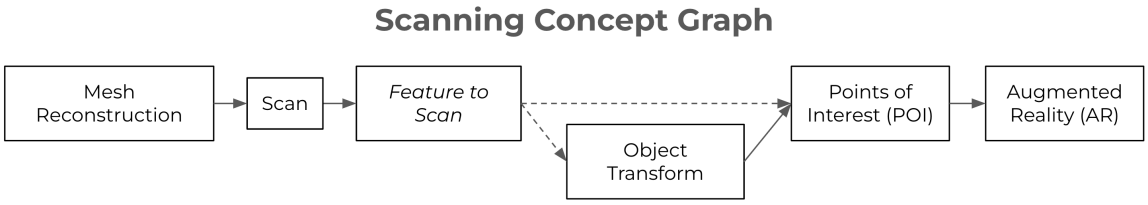


Figure 1: The concept graph for scanning, which can be described as follows: A *Mesh Reconstruction* depends on having a *Scan*, which depends on having a *Feature to Scan*. *Features to Scan* can either be *Object Transforms* or a *general POIs*. (Note that an *Object Transform* must be present to provide feedback on users' scanning behaviour, which is the purpose of the app in question.) The concept of a *POI* depends on having the concept of *AR*. In the figure, italicised concepts are abstract [17].

# 5 Guidance algorithms

To guide users in the scanning task, we implemented the following algorithms: Distance Guidance (Listing 1), Framing Guidance (Listing 2), Speed Guidance (Listing 3), and Completion Guidance (Listing 4). The instances where the algorithms differ between the pilot study audio/haptic app, full-scale study audio/haptic app, and the full-scale study visual app are labelled with square bracket tags (e.g. [PILOT AUDIO/HAPTIC:]). We provide further Listings in other sections. For example, our 3D bounding box algorithm is in Section 10, and our drift-reduction algorithm, "Landmark Tracker", is in the main paper.

Listing 1: Pseudo-code for our Distance Guidance (F1) algorithm.

```
1. User identifies the object they want to scan, which provides the system with the "Object
     Transform" (i.e. position/scale/orientation of the object)
2. User starts the scan
3. Obtain the distance from the Object Transform to the camera, and set the "Ideal Distance" to
     this value
   a. (Note that in the user study, participants were told to start at a specific position)
4. Set the "Closest Ideal Distance" and "Furthest Ideal Distance" radii, where participants are
     said to be within the ideal distance range (and will not be warned to move)
   a. E.g. Set Closest Ideal Distance to Ideal Distance minus some preset "Buffer Distance" (e.g.
      0.5m)
   b. E.g. Set Furthest Ideal Distance to Ideal Distance plus the Buffer Distance
5. Set any other Guidance Feedback variables
   a. [PILOT AUDIO/HAPTIC AND FULL AUDIO/HAPTIC:] Set the "Min Distance" and "Max Distance",
     which are closer and further to the Object Transform than the "Closest" and "Furthest" ideal
      distances respectively, and which represent the distances past which the maximum-strength
     guidance occurs
   b. [PILOT AUDIO/HAPTIC AND FULL AUDIO/HAPTIC:] Set the Lowest Volume, Highest Volume, Low Pass
      Filter Max Frequency, and Low Pass Filter Min Frequency variables
   c. [FULL AUDIO/HAPTIC:] Set the Default Pitch variable
6. [PILOT AUDIO/HAPTIC AND FULL AUDIO/HAPTIC:] Play music
   b. IF the user is within the ideal distance range: Keep the music at the Highest Volume
   ELSE:
       i. Get the user's current distance from the Object transform
       ii. IF the user is closer to the Object Transform than the Closest Ideal Distance:
           1. Set the music volume by linearly interpolating between the Highest Volume & Lowest
     Volume with respect to the user's current distance, and the Closest Ideal Distance & Min
     Distance
           2. Set the low pass filter cutoff frequency by linearly interpolating between the Min
     Frequency & Max Frequency with respect to the user's current distance, and the Closest Ideal
      Distance & Min Distance
       ELSE:
           1. Set the music volume by linearly interpolating between the Highest Volume & Lowest
     Volume with respect to the user's current distance, and the Furthest Ideal Distance & Max
     Distance
```

```
        2. Set the low pass filter cutoff frequency by linearly interpolating between the Min
     Frequency & Max Frequency with respect to the user's current distance, and the Furthest
     Ideal Distance & Max Distance
6. [FULL AUDIO/HAPTIC:] IF the user is within the ideal distance range: Keep the music at the
     Default Pitch
ELSE IF the user is closer to the Object Transform than the Closest Ideal Distance:
   a. Obtain a Pitch Value between 0 and 1 by linearly interpolating with respect to the user's
     current distance, and the Closest Ideal Distance & Min Distance
   b. Set the music pitch to 2^(-1 * Pitch Value)
ELSE:
   a. Obtain Pitch Value between 0 and 1 by linearly interpolating with respect to the user's
     current distance, and the Furthest Ideal Distance & Max Distance
   b. Set the music pitch to 2^(Pitch Value)
6. [FULL VISUAL:] Generate an AR blue ring, "Track", centered around the Object Transform, which
     has an inner radius of Closest Ideal Distance and outer radius of Furthest Ideal Distance
    a. IF the user is within the ideal distance range, they will be within the Track
       ELSE they will be outside the Track
7. Repeat Step 6 until the scan completes
```

Listing 2: Pseudo-code for our Framing Guidance (F2) algorithm.

```
1. User identifies the object they want to scan, which provides the system with the "Object
     Transform" (i.e. position/scale/orientation of the object)
2. User starts the scan
3. IF the Object Transform is fully captured by the camera view:
   a. Do not provide any Framing Guidance
ELSE:
   a. Calculate how far away the camera angle is from the Object Transform (e.g. in degrees)
   b. [PILOT AUDIO/HAPTIC:] Play a beeping noise that increasingly reverberates as the camera
     angle increasingly points away from the Object Transform. Additionally, vibrate the phone at
      a steady rate.
   b. [FULL AUDIO/HAPTIC:] Vibrate the phone with increasing strength and frequency as the camera
      angle increasingly points away
   b. [FULL VISUAL:] Display on-screen arrows with increasing opacity as the camera angle
     increasingly points away
4. Repeat step 3 until the scan completes
```

Listing 3: Pseudo-code for our Speed Guidance (F3) algorithm.

```
1. User identifies the object they want to scan, which provides the system with the "Object
     Transform" (i.e. position/scale/orientation of the object)
2. User starts the scan
3. Set the "Maximum Speed" variable (e.g. 5 degrees/sec)
3. Store the most recent X (e.g. 5) number of angular velocities around the Object Transform (e.g
     . by calculating the angular velocity over T milliseconds, e.g. 250ms)
4. IF the median velocity is greater than Maximum Speed:
   a. [PILOT AUDIO/HAPTIC AND FULL AUDIO/HAPTIC:] Play an audio file with voiced feedback (e.g. "
     Too fast!")
   a. [FULL VISUAL:] Show an on-screen text warning (e.g. "Slow down")
5. Repeat Steps 3 and 4 until the scan completes
```

Listing 4: Pseudo-code for our Completion Guidance (F4) algorithm.

```
1. User identifies the object they want to scan, which provides the system with the "Object
     Transform" (i.e. position/scale/orientation of the object)
2. User starts the scan
3. Set the Start Angle to the user's current angle in the horizontal plane around the Object
     Transform
4. Set a number of "Checkpoints" around the Object Transform by dividing 360 degrees into X
     number of parts (e.g. four parts would provide the Checkpoints: 90, 180, 270, 360 degrees)
5. IF the user's current angle is greater than the first Checkpoint:
   a. Set the User's Direction to +1 and mark the first Checkpoint as complete
ELSE IF the user's  current angle is less than -1 times the first Checkpoint
   b. Set the User's Direction to -1 and mark the first Checkpoint as complete
   c. Multiply all of the Checkpoints by -1
6. IF the User's Direction has been set:
   a. IF the User's Direction is +1 AND the user's current angle is greater than the next
     uncompleted Checkpoint: Set this Checkpoint as complete
   ELSE IF the User's Direction is -1 AND the user's current angle is less than the next
     uncompleted Checkpoint: Set this Checkpoint as complete
```

```
7. IF all Checkpoints are complete:
   a. Mark the scan as complete
   b. [PILOT AUDIO/HAPTIC AND FULL AUDIO/HAPTIC:] Play an audio file with voiced feedback (e.g. "
     You're done!") and process the scan
   b. [FULL VISUAL:] Show on-screen feedback and process the scan
ELSE: Repeat Steps 5-7
```

# 6 Comparison of current scanning apps

Table 1 compares the guidance features and visualisations of various current scanning apps on the market and our scanning apps. See the main paper for further descriptions of the scanning concepts/features, "Object Transform", "Distance Guidance", "Framing Guidance", "Speed Guidance", and "Completion Guidance".

Table 1: Features of various commercial (or in-development) scanning apps, as well as our final apps from the main paper.

| Scanning app | Currently available? | Visualisation of scanning | C1: Object transform positioning (to enable guidance features) | F1: Distance guidance | F2: Framing guidance | F3: Speed guidance | F4: Scan completion guidance |
|---|---|---|---|---|---|---|---|
| **Snap Custom Landmarker Creator [15]** | Yes | Feature-points, mesh or procedurally generated objects on physical object | None | None | None | None | Progress bar: Minimum/ maximum number of perspectives to capture |
| **Immersal [13]** | Yes | Feature-points on physical object | None | None | None | None | None |
| **Wayfarer [31]** | Yes | Darker pixels/feature-points disappear when scanned | None | None | None | Dynamic text ("Slow down") | Static text ("Scans must be longer than 20 secs") |
| **Scaniverse [32]** | Yes | Stripes disappear when scanned | None | Stripes do not disappear if too far | None | Dynamic text ("Slow down","Avoid fast turns") | None |
| **Pokemon Go Pokestop Scanning [11]** | Yes | None | None | None | None | Dynamic text ("Movement not detected. Keep moving slowly around the Pokéstop.") | Progress bar: Fills with time |
| **Ingress Portal Scanning [30]** | Yes | Feature-points on physical object | None | None | Static text ("Keep the Portal within the frame") | Static text ("... slowly walk around the Portal, if possible.") | Progress bar: Fills with time |
| **Apple's Object Capture WWDC'23 presentation [14]** | No (likely will be available with iOS 17) | Feature-points on separate virtual mesh | Direct manipulation: Requires user to walk around the object before scanning | Dynamic text ("Move closer") | Arrow pointing towards object | Dynamic text ("Slow down") | Progress bar ("dial"): Fills as user scans |
| **Our full-study visual app** | Ours | Mesh similar to Snap!'s, as well as feature-points on virtual object | Automatic placement via interactive segmentation and 3D bounding-box estimation | Visual ring around the object | Dynamic text ("The camera isn't aimed at the object") and blinking arrows | Dynamic text ("Slow down") | Dynamic text ("You did it!") for full 360 degrees |
| **Our full-study audio/haptic app** | Ours | N/A | Automatic placement via interactive segmentation and 3D bounding-box estimation | Music adjustments (volume, pitch, low-pass filter) | Haptic signals that increase in frequency and intensity as the camera angle deviates further from the POI | Voiced audio ("Too fast", "Slow down please", etc.) | Voiced audio ("You're done!", "You did it!", etc.) for full 360 degrees |

# 7 Pilot study: Additional information about participants

Six participants (n=6) completed the study in London, with appointments between March 29 and April 3, 2023. On arrival, participants signed research consent forms and were provided with anonymous codenames. Recruited participants' ages ranged from 20 to 27 ($\bar{x}$=23.83, SD=2.79) with one participant being female and five being male.

Prior experience with 3D scanning was not a requirement for recruits, and we recruited participants through various University College London email lists. We asked interested individuals to complete an availability form and then contacted them to confirm an appointment. Participants received a £25 gift card and complimentary merchandise (*e.g.* stickers).

# 8 Pilot study: Semi-structured interview questions

Listing 5 contains the semi-structured interview questions we asked participants in the pilot study.

Listing 5: Semi-structured interview questions from the pilot study.

```
--- Semi-Structured Interview Questions: No-Screen Scanning Pilot Study ---
Notes to participants:
"These questions are very broad/general on purpose, as we don't want to bias any of your answers,
     so please feel free to answer however you want and with whatever comes to mind. Also note
    that we're not judging you or your scanning ability, but rather, judging the app because we
    want to improve it."

- Semi-Structured Interview Questions -
Codename: ______________

When considering the two different scanning processes...

1. How did you hold the phone in each scanning experience?
   a. What felt most natural / socially acceptable?

2. Was there anything that was particularly stressful in either scanning experience?

3. Did you feel like you were aware of your surroundings, comparing between the two scanning
     experiences?
   a. Did you feel like you were walking safely?
   b. Did you trip or bump into things?

4. Imagine that your job was to make good scans. Did you feel confident you were scanning
     correctly, comparing between the two scanning experiences?

5. Did you feel like you were aware of the scanning task/process, comparing between the two
     scanning experiences?

6. What was most difficult about each app experience?
   a. Was there anything you would change about it?

7. Was there anything that was especially enjoyable or rewarding?

8. Do you feel like there were pros/cons to either/each scanning app?

9. Any other comments?
```

# 9 Pilot study: Thematic analysis results

As discussed in the main paper, we present a deeper analysis of participants' sentiments from the pilot semi-structured interviews here. This includes stress-related themes, organized into the three categories outlined in [38]: Social-evaluative, cognitive, and physical stress. It also includes rewarding aspects of our audio/haptic app, and other opportunities for our app.

## 9.1 Stressors and stress reducers

We discuss stress-related themes next.

**Social-evaluative.** Five of six participants mentioned stressors involving other people's perceptions of them. For instance, P1 mentioned that scanning felt awkward because "people [were] looking at [her] ... because [she] was walking around [the statue] so many times". She used strong language in her description, mentioning she felt like bystanders were thinking, "What the f*** are you doing?". She also mentioned, however, that other bystanders may not care, stating, "Some people were just like, 'Oh yeah, it's London. It's completely normal [for odd things to happen here]' ".

P2, P3, P4 and P6 had similar concerns about social perceptions, but regarding the loud app sounds when using the audio/haptic app in a public space. For example, P2 mentioned "I think I definitely got way more 'looks' with the [audio/haptic app] because my phone was beeping", and P6 mentioned "I guess having headphone[s] attached to the phone while [using the audio/haptic app] would be better because the sounds might be too loud". Thus, for the final study, we incorporated headphones, and recommend **minimizing external audio in public** (F7 in Table 1 in the main paper).

P3 and P4 mentioned the visual scanning condition felt more invasive of people's privacy because it was obvious they were recording. For instance, P4 said, "I think I saw people look at me, suspiciously thinking, 'What is this person doing, walking around the set here [while videoing]?' ". P3 said, "For the video app, I saw people duck away a couple of times [...] because they really thought I was videoing. [...] The [audio/haptic app] was a little bit more socially acceptable, mainly because if [bystanders] happened to look at my screen, it didn't seem like I was recording".

Participants had various views on how the busyness of the area affected them. For instance, P1 mentioned "maybe it helps that it's London and that everyone [regularly experiences] weird stuff, and because there are so many people around, no one will give you much attention' ". P5, however, felt having many people around was a stressor because when "the street is quite crowded, even if you stand still, you feel like you're blocking someone". On the other hand, P1 mentioned that having fewer people around could increase stress because, "if it's a quieter area and [...] you have your phone volume turned on, people might notice".

Interestingly, participants mentioned that anonymity, or conversely, celebrity status, could reduce feelings of social-evaluative stress. For example, P6 mentioned wearing a mask made him feel comfortable scanning in public, whereas P5 mentioned that he felt "quite accepted" because he felt like an "Instagram" or "TikTok star".

**Cognitive.** Through the interviews, we identified a bug in the audio/haptic app, where users would not see the completion screen after finishing their scan. Users who encountered it (P1, P2, and P5) described this as increasing their stress. E.g. P5 said, "I think the only stressful thing was that I had to [scan with the audio/haptic app] twice because it didn't say [I finished]" (P2). We also observed that five of six participants mentioned that—at a certain point during the scan—the audio/haptic feedback stopped aligning with the location of the actual POI. For example, P2 described how the system was providing a lot of negative feedback ("buzzing"), which "could have been because it shifted place in the middle [of the scan] or something". We suspected this was because the AR session had lost track of the environment (*e.g.* due to SLAM drift [42]), and the *Object Transform* had drifted away from the physical POI location. Thus, we decided we needed to update our app by **minimizing drift and eliminating any bugs** (F5 in Table 1 in the main paper).

Some users found the visual app added cognitive stress because it did not provide scanning guidance. For instance, P1 said, "The audio one was way more like, 'Good job, you're doing the right thing!', but with the [visual app] sometimes I couldn't see the object". Similarly, P5 mentioned he was "more confident" using the audio/haptic app because it gave him "immediate feedback if [he was] not in the correct position". Thus, we recommend **providing scanning guidance to increase feelings of confidence**.

Nonetheless, some participants felt their cognitive load decreased with the visual app because they had prior experience with similar apps. For instance, P2 mentioned that it was "more natural [...] because [he had] taken video of things before and [he was] used to it". There was no "training [required] for normal people". Similarly, P6 mentioned, "not looking at your at the screen while scanning is a bit counterintuitive in terms of our daily experience [with visual apps]".

**Physical.** Five of six participants mentioned they felt like they may collide with something, especially when looking at the phone screen. For example, P4 stated, "The video [app] makes you more focused and you're less aware of your surroundings. [...] With the video [app], I bumped into somebody just because I had to focus on the camera". Similarly, P3 described how when "there was a guy walking towards" him, he could navigate around "him more easily [with the audio/haptic app] than with the video one".

Participants also mentioned feeling like people may steal from them. For example, P3 mentioned he had "a slight worry that someone would grab the phone and start running". P1 described similar feelings, and how she could see how the audio/haptic app could allow users to "look around more [...] for pickpockets [...] so no one [could] snatch your phone". Because hiding the video feed seemed to increase user awareness (and reduce users' social-evaluative stress, as described in Section 9.1), we recommend **hiding the video feed to increase users' feelings of awareness and reduce social-evaluative stress**.

Despite how many participants felt more aware when using the audio/haptic app, two participants also mentioned they felt like they needed to grip the phone more strongly with the audio/haptic app because they were not looking at it. Because of this, and the potential for thieves, we provided participants with phone grippers in the full-scale study, and recommend **improving users' grip, when possible** (F6 in Table 1 in the main paper).

## 9.2 Rewarding aspects

Participants also mentioned rewarding aspects of the audio/haptic scanning experience. For example, P2 mentioned, "I feel like a lot of people play Niantic's [location-based AR] games, and they normally listen to music as they're going along. So I think that's quite awesome [that they could listen and play at the same time]." P5 mentioned that the audio/haptic app "is quite a fun. I feel happy after [using it]", in comparison with the visual app. P1 and P3 mentioned the voiced feedback was rewarding, "I thought the audible feedback at the end was nice. Like, 'You're done!' It feels like I accomplished something" (P4). Thus, we recommend **utilising music and voiced feedback as rewarding features**.

Five of six participants mentioned the visualisation of the resulting mesh reconstruction in the audio/haptic app (as shown in the main paper) was rewarding. For instance, P2 stated, "Looking around at the object [3D mesh] with the the audio app was pretty cool. And I was really surprised how well [it made it]." P6 was so impressed with the reconstruction, he "wanted to do a second scan!". Thus, we recommend **providing a mesh reconstruction (even low-fidelity) as a rewarding feature**.

Another rewarding feature included using haptic feedback (rather than sight) to view-find. E.g. P4 said it was especially rewarding "when you at some point figure out how to angle the camera with the [haptic feedback]", and P1 mentioned, "Finding the correct orientation doing audio scan was probably the [most] rewarding part". Thus, we recommend **utilising haptic feedback for view-finding as a rewarding feature**.

### 9.3 Other opportunities

Participants also mentioned opportunities where the audio/haptic app would be useful, and how to improve it. For example, P2 mentioned audio/haptic scanning could be useful for "people who are vision impaired". P5 mentioned the app was democratising scanning technology: "If you allow people to use [the audio/haptic app], anyone could use their phone and scan", as opposed to how "Google" requires "professional machines to make a scan".

P2, P3 and P6 mentioned an opportunity to improve the concept of the *Object Transform*. P6 described how it would be better if the *Object Transform* "fully encapsulated" the statue; for example, with a "rectangle instead of a small character". P3 mentioned it would "be nice if people could maybe take a still [picture] and then mark the area. So it can just detect the object [automatically]". Thus, we recommend **utilising computer vision to detect what the user wants to scan**, and **better conveying the *Object Transform* concept by using a neutral virtual object (*e.g.* bounding-box), and fully enveloping the physical object** (C1 in Table 1 in the main paper).

P6 mentioned difficulty locating the "correct distance" using the audio/haptic app because when he "was moving either away or towards the object, the volume drop[ped]". He also mentioned this could be solved by "changing the tone" when "moving in [a particular] direction". Similarly, P3 mentioned that "it could be nice to work with tones or if [the music] somehow changes a little bit". Thus, we recommend **utilising tone changes to improve music guidance** (F1 in Table 1 in the main paper).

## 10 Full-scale study app: 3D bounding-box estimation algorithm

As described in the main paper, in the full-scale study app, the user taps the object they want to scan, and the app uses an interactive segmentation model [4] to identify the pixels of the object the user tapped. After this, the app estimates the 3D location, scale and orientation of the object (i.e. the *Object Transform*). We implemented this using the algorithm in Listing 6. We recognize there are likely other 3D bounding-box estimation algorithms that are more robust to alternative POI shapes; however, this algorithm worked for the purposes of our study.

Listing 6: Pseudo-code for the 3D bounding-box estimation of the object being scanned.

```
1. The user taps on-screen to identify the object to scan
    a. Using the tap's location, segment the object pixels to obtain a "pixel mask"

2. For each pixel in the mask, obtain a depth value (e.g. from a LIDAR point cloud, if
available, or image depth estimation)

3. Estimate two 3D points on the object's front face by calculating the nth (e.g. 10th) and
mth (e.g. 15th) percentile depth values in the camera's forward direction (Note: percentiles
 are used to filter out extremely close/far depth values)
    a. Create a line of best fit between the two points for the object's front face
    b. In the horizontal plane, obtain the length of the line segment that would represent
the front face using the points furthest away from each other in the direction of the line
of best fit
```

```
4. Repeat step 3 in the "right" and "up" directions with respect to the front face's
   orientation to obtain a line segment of best fit for the right and top faces

5. Create a bounding-box ("Object Transform") based on the front, right, and top face
   estimations
```

# 11 Full-scale study: Questionnaires

The following section contains references to the validated questionnaires we used in the full-scale study (which can be found in the literature cited), as well as the questionnaires we developed ourselves.

## 11.1 Pre-/post-survey

In the pre- and post-surveys, we included emotional affect questions using the Self-Assessment Manikins (SAM) in [28]. This was to determine users' arousal and valence before and after scanning. Additionally, we included questions we developed based on observations from the pilot study. You can find these in in Listing 7.

Listing 7: Additional questions in the pre/post-surveys. These were presented in randomized order within each subsection.

```
--- On both the pre- and post-survey ---
1. Please click the facial expression that best matches how you feel right now. [Multiple choice:
     SAM - valence facial expressions.]

2. Please click the facial expression that best matches how you feel right now. [Multiple choice:
     SAM - arousal facial expressions.]

3. Any anonymous comments? [Long answer]

--- Only on the pre-survey ---
1. How much are the following 'stressors' contributing to your current stress levels, if at all?
    [Likert scale, 1-5, for each stressor:]
  a. Social stress, like being watched or judged, etc.
  b. Cognitive stress, like having too much to think about, etc.
  c. Physical stress, like feeling like someone might bump into you or steal from you, or feeling
     like you might bump into something, etc.

--- Only on the post-survey ---
1. How much did the following 'stressors' contribute to your stress levels during the scan, if at
     all? [Likert scale, 1-5, for each stressor:]
  a. Social stress, like being watched or judged, etc.
  b. Cognitive stress, like having too much to think about, etc.
  c. Physical stress, like feeling like someone might bump into you or steal from you, or feeling
     like you might bump into something, etc.

2. During the scan, did you bump into anything or did anything bump into you? [Multiple choice:]
  a. No
  b. Yes, at least once
  c. Yes, multiple times
```

## 11.2 Final survey

The final survey included two validated questionnaires: the State-Trait Anxiety Inventory [41] and the User Engagement Scale - Short Form (UES-SF) [35]. As in the pre-/post-surveys, we included questions we developed based on observations from the pilot study, as shown in Listing 8.

Listing 8: Additional questions in the final survey. These were presented in randomized order within each subsection.

```
--- Prior experience ---
1. Approximately how many times have you scanned outdoors before? This could include scanning
     using an app, like scanning in Pokemon Go, or taking a video of an object outside while
     circling around it.

2. Prior to this study, how much did you know about scanning to create 3D meshes/objects? [
     Multiple choice:]
  a. I didn't know what scanning and/or 3D meshes/objects were
  b. I had basic knowledge of what scanning is
  c. I knew about scanning and how to improve scans
  d. I had extensive knowledge about how to scan well, and could have taught others how to
     improve scans

3. Does your profession or any of your major hobbies involve using a camera? If so, what is this
     hobby/profession? Include any past professions or hobbies related to cameras too. [Short
     answer]

--- Heart rate related demographics ---
4. What is your gender? (We ask this as it could affect your heart rate) [Short answer]

5. Has your gender changed? (We ask this as it could affect your heart rate) [Short answer]

6. How old are you?

--- State-Trait Anxiety Inventory ---
Read each statement and select the appropriate response to indicate how you generally feel.
There are no right or wrong answers. Do not spend too much time on any one statement, but give
     the answer which seems to describe how you generally feel. [Likert scale: State-Trait
     Anxiety Inventory.]

--- User engagement scale ---
The following statements ask you to reflect on your experience engaging with the scanning app.
     For each statement, please use the following scale to indicate what is most true for you. [
     Likert scale: UES-SF]

--- Stressors ---
7. Check the boxes next to anything that contributed to your stress levels (if at all). Check "
     None" if none apply. [Checkboxes:]
  a. Navigating around people
  b. Navigating around objects
  c. Physically holding the phone
  d. The potential for thieves
  e. Feeling like the activity wasn't socially acceptable
  f. Feeling like I wasn't providing other people enough privacy
  g. Feeling like the activity was noisy or distracting for other people
  h. Running into a bug in the app
  i. The app visuals (if using the video app)
  j. The app sounds/vibrations (if using the audio app)
  k. Feeling like there's too much information at once
  l. None
  m. Other... [Short answer]

8. Select the item that contributed *most* to your stress levels (if at all).
  a. Navigating around people
  b. Navigating around objects
  c. Physically holding the phone
  d. The potential for thieves
  e. Feeling like the activity wasn't socially acceptable
  f. Feeling like I wasn't providing other people enough privacy
  g. Feeling like the activity was noisy or distracting for other people
  h. Running into a bug in the app
  i. The app visuals (if using the video app)
  j. The app sounds/vibrations (if using the audio app)
  k. Feeling like there's too much information at once
  l. None
  m. Other... [Short answer]

  --- Drift ---
```

```
9. Did you feel like the location where you were supposed to point your phone "drifted" off of
     the actual object location (i.e. to get the app to tell you that you were "scanning
     correctly", you had to point the phone in a different direction than the statue)? [Multiple
     choice:]
  a. Yes, I felt like the location "drifted"
  b. No, I didn't feel like the location "drifted"
  c. Other... [Short answer]

10. If you felt like the location where you were supposed to point your phone "drifted" off of
     the actual object location, were you able to "re-locate" it (i.e. were you able to move the
     phone so the app said you were "scanning correctly" even if you weren't necessarily pointing
      at the statue)? [Multiple choice:]
  a. The object didn't drift for me
  b. Yes, I was able to relocate where I was supposed to point my phone
  c. No, I wasn't able to relocate where I was supposed to point my phone
  d. Other... [Short answer]

11. Please describe any bugs. [Short answer]

--- Comments ---
12. Any anonymous comments? [Long answer]
```

# 12 Full-scale study: Procedure for scanned mesh comparisons

In this section, we describe how we generated accuracy scores for our scanned meshes.

**Ground Truth reconstruction.** We densely captured high resolution images of the statue on a clear day with no traffic using a modern DSLR. We then used COLMAP's Structure-from-Motion [39] and Multi-View-Stereo [40] to obtain a high quality reconstruction of the Agatha Christie statue. Since this reconstruction is only accurate up to scale, we measured the statue at multiple points and used these measurements to scale our reconstruction.

**Scan alignment and pose refinement.** Each scan lives in a different coordinate system to that in the GT scan, and each scan's poses suffer from significant drift due to heavy static scene occlusion and moving objects—notably from people and bikes. To remedy this and produce the best possible reconstructions from each scan, we register the images in each scan to the GT with COLMAP to compute new poses for every scan. During registration, we carefully mask pixels containing moving objects (people, bikes, cars) for feature extraction. We use computed poses from this step for the remainder of the pipeline.

**Mesh reconstruction and TSDF fusion.** We leverage the iPhone's LiDAR depth sensor to reconstruct meshes from each user study scan. We fuse each scan's depths into a TSDF volume using the TSDF fuser from [37]. We extract a mesh using marching cubes.

# 13 Full-scale study: Additional analyses and results

This section contains detailed output from the post-hoc and MANCOVA/ANCOVA linear regression analyses.

## 13.1 Engagement analysis

The results of the post-hoc engagement anslyses are shown in Listing 9. Note how the p-values with respect to engagement, reward factor, and perceived usability are significant ($p<.05$), meaning the

audio/haptic conditions were more engaging, rewarding and usable.

Listing 9: Post-hoc results from analysing engagement.

```
--- Comparison between audio/haptic & visual engagement (overall) ---
Normal? True -> t-test

Significant? True
p-value: 0.014935564235268971
t(49): 2.238363378186459

Engagement means:
audio/haptic: 3.76
visual: 3.40


--- Comparison between audio/haptic & visual focused attention ---
Normal? True -> t-test

Significant? False
p-value: 0.06336467987593615
t(49): 1.554102371160823

Focused attention means:
audio/haptic: 3.37
visual: 3.04


--- Comparison between audio/haptic & visual reward factor ---
Normal? True -> t-test

Significant? True
p-value: 0.04109324601514395
t(49): 1.7753146379037608

Reward factor means:
audio/haptic: 3.95
visual: 3.60


--- Comparison between audio/haptic & visual perceived usability ---
Normal? False -> Mann-Whitney U test

Significant? True
p-value: 0.04997801503072596
U: 396.0

Perceived usability means:
audio/haptic: 3.95
visual: 3.57
```

Listing 10 shows the results of our ANCOVA analysis for engagement. Engagement was transformed to satisfy the ANCOVA requirements, as described in [6]. Note how the model, and the variable, app type (audio/haptic vs. visual), are significant.

Listing 10: ANCOVA results from analysing engagement.

```
Model: (log_1.2(engagement))^3 ~ C(appType) + timesScanning + busyness

                            OLS Regression Results
==============================================================================
Dep. Variable:             engagement   R-squared:                       0.159
Model:                            OLS   Adj. R-squared:                  0.104
Method:                 Least Squares   F-statistic:                     2.898
Date:                Tue, 22 Aug 2023   Prob (F-statistic):             0.0450
Time:                        13:14:26   Log-Likelihood:                -307.32
No. Observations:                  50   AIC:                             622.6
Df Residuals:                      46   BIC:                             630.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
```

```
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept              441.5688     67.656      6.527      0.000     305.384     577.753
C(appType)[T.video]    -69.9171     33.598     -2.081      0.043    -137.546      -2.288
timesScanning           15.0202     12.420      1.209      0.233      -9.981      40.021
busyness              -183.7114    126.382     -1.454      0.153    -438.106      70.683
==============================================================================
Omnibus:                        5.107   Durbin-Watson:                   2.348
Prob(Omnibus):                  0.078   Jarque-Bera (JB):                3.959
Skew:                           0.594   Prob(JB):                        0.138
Kurtosis:                       3.700   Cond. No.                         22.1
==============================================================================
```

## 13.2 Safety proxy analyses

This section contains detailed analyses of safety proxy variables from the full study, including:

- the number of times users collided with obstacles,
- incremental HR as a proxy for stress [21, 20, 7],
- emotional affect via the circumplex model to determine whether users felt more stressed after scanning [28, 38], and
- social-evaluative, cognitive, and physical stressors via Likert scales [38, 10].

As described in the main paper, none of the analyses indicated either condition (audio/haptic or visual) was less safe or more stressful than the other.

**Emotional affect.** To determine whether people felt differently before and after scanning, we utilized the circumplex model, which measures affect in terms of valence and arousal. The valence/arousal scale allows researchers to identify how participants feel, as shown in Figure 2 [28]. Specifically, we wanted to determine whether participants leaned towards "stress" in either condition.

From Figures 3 and 4, it would seem that audio/haptic-app users' affect shifted towards happiness/contentedness (ignoring outliers), whereas visual-app users' affect shifted away from happiness/-contentedness. After modelling affect using MANCOVA ($F_{2,37}$=3.69, Pillai'sTrace=.166, p=.035), however, we did not find evidence that app type significantly predicted affect.

Listing 11 shows the results of our MANCOVA analysis for emotional affect based on the circumplex model (arousal/valence). Note how the model is significant, but app type (audio/haptic vs. visual) is not.

Listing 11: MANCOVA results from analysing emotional affect.

```
Model: valenceAfter + arousalAfter ~ C(appType) + valenceBefore + arousalBefore + profession +
    busyness
                Multivariate linear model
============================================================

------------------------------------------------------------
       Intercept        Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.8337 2.0000 37.0000  3.6902 0.0346
         Pillai's trace 0.1663 2.0000 37.0000  3.6902 0.0346
 Hotelling-Lawley trace 0.1995 2.0000 37.0000  3.6902 0.0346
    Roy's greatest root 0.1995 2.0000 37.0000  3.6902 0.0346
------------------------------------------------------------

------------------------------------------------------------
       C(appType)       Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
```

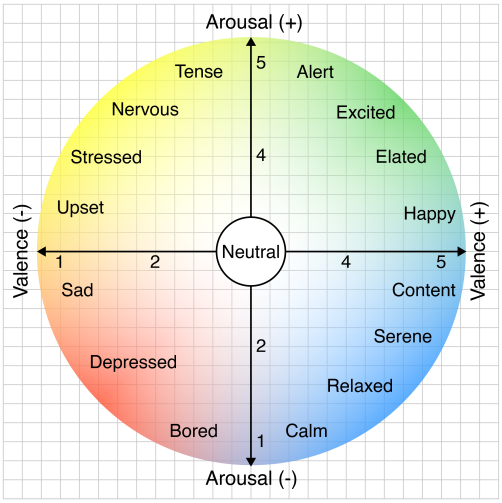


Figure 2: The circumplex model with emotions based on [43] (adapted from [29]). The arousal/valence scales (1-5) are based on the Likert scale SAM questions [28] we used in our study.

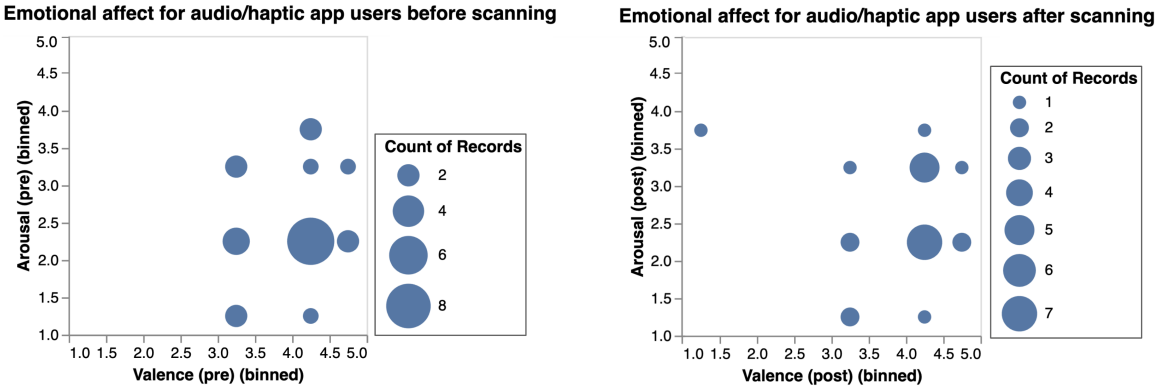


Figure 3: Audio/haptic-app users' affect before and after scanning (which map to Figure 2). Note how (ignoring the outlier)—after scanning—the left-most bins decrease in size.

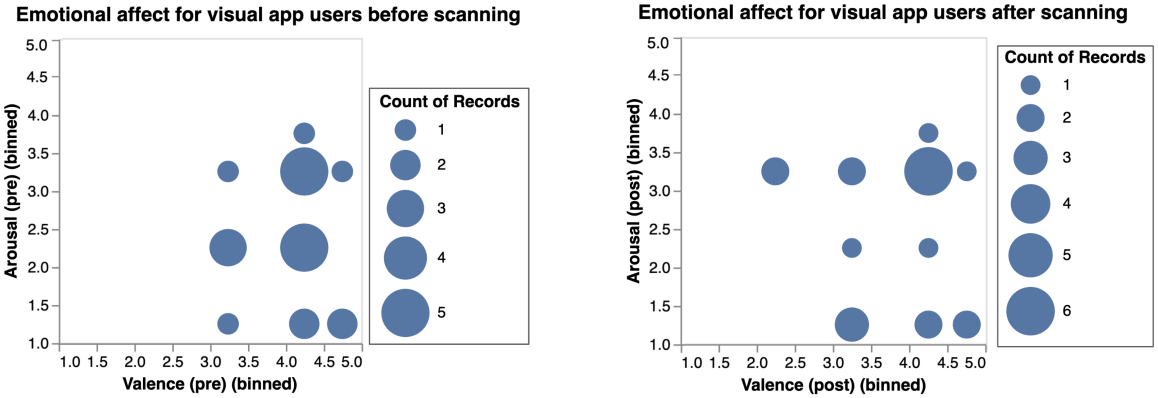


Figure 4: Visual-app users' affect before and after scanning (which map to Figure 2). Note how—after scanning—additional leftward bins appear or increase in size.

| | | | | | |
|---|---|---|---|---|---|
| Wilks' lambda | 0.9487 | 2.0000 | 37.0000 | 1.0003 | 0.3775 |
| Pillai's trace | 0.0513 | 2.0000 | 37.0000 | 1.0003 | 0.3775 |
| Hotelling-Lawley trace | 0.0541 | 2.0000 | 37.0000 | 1.0003 | 0.3775 |
| Roy's greatest root | 0.0541 | 2.0000 | 37.0000 | 1.0003 | 0.3775 |

| valenceBefore | Value | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|
| Wilks' lambda | 0.6613 | 2.0000 | 37.0000 | 9.4736 | 0.0005 |
| Pillai's trace | 0.3387 | 2.0000 | 37.0000 | 9.4736 | 0.0005 |
| Hotelling-Lawley trace | 0.5121 | 2.0000 | 37.0000 | 9.4736 | 0.0005 |
| Roy's greatest root | 0.5121 | 2.0000 | 37.0000 | 9.4736 | 0.0005 |

| arousalBefore | Value | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|
| Wilks' lambda | 0.4737 | 2.0000 | 37.0000 | 20.5527 | 0.0000 |
| Pillai's trace | 0.5263 | 2.0000 | 37.0000 | 20.5527 | 0.0000 |
| Hotelling-Lawley trace | 1.1110 | 2.0000 | 37.0000 | 20.5527 | 0.0000 |
| Roy's greatest root | 1.1110 | 2.0000 | 37.0000 | 20.5527 | 0.0000 |

| profession | Value | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|
| Wilks' lambda | 0.6435 | 2.0000 | 37.0000 | 10.2473 | 0.0003 |
| Pillai's trace | 0.3565 | 2.0000 | 37.0000 | 10.2473 | 0.0003 |
| Hotelling-Lawley trace | 0.5539 | 2.0000 | 37.0000 | 10.2473 | 0.0003 |
| Roy's greatest root | 0.5539 | 2.0000 | 37.0000 | 10.2473 | 0.0003 |

| busyness | Value | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|
| Wilks' lambda | 0.9993 | 2.0000 | 37.0000 | 0.0125 | 0.9875 |
| Pillai's trace | 0.0007 | 2.0000 | 37.0000 | 0.0125 | 0.9875 |
| Hotelling-Lawley trace | 0.0007 | 2.0000 | 37.0000 | 0.0125 | 0.9875 |
| Roy's greatest root | 0.0007 | 2.0000 | 37.0000 | 0.0125 | 0.9875 |

**Social-evaluative, cognitive and physical stressors.** We also wanted to directly investigate self-reports of the three stress types. After a visual (see Figure 5) and MANCOVA ($F_{3,35}$=7.52,

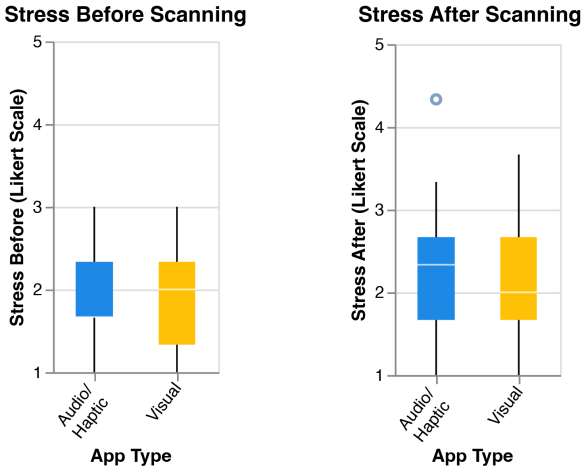


Figure 5: Stress distributions on a 5-point Likert Scale before and after the scanning experience. Notice the similarity between the audio/haptic and visual conditions in each plot.

Pillai'sTrace=.392, p<.001) investigation of the data (see Listing 12), there was no evidence that app type significantly affected changes in social-evaluative, cognitive or physical stress levels. Conducting an ANCOVA (see Listing 13) with an average of the three stressors ($F_{4,39}$=14.0, $R^2$=.589, p<.000) also did not result in any evidence for app type affecting stress levels.

Listing 12 shows the results of our MANCOVA analysis for stress. Note how the model is significant, but app type (audio/haptic vs. visual) is not.

Listing 12: MANCOVA results from analysing stress levels.

```
Model: socialstressafter + cognitivestressafter + physicalstressafter ~ socialstressbefore +
     cognitivestressbefore + physicalstressbefore + C(appType) + anxietyTrait + profession
                 Multivariate linear model
============================================================

------------------------------------------------------------
       Intercept         Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.6080 3.0000 35.0000  7.5227 0.0005
         Pillai's trace 0.3920 3.0000 35.0000  7.5227 0.0005
 Hotelling-Lawley trace 0.6448 3.0000 35.0000  7.5227 0.0005
    Roy's greatest root 0.6448 3.0000 35.0000  7.5227 0.0005
------------------------------------------------------------

------------------------------------------------------------
       C(appType)        Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.9176 3.0000 35.0000  1.0479 0.3836
         Pillai's trace 0.0824 3.0000 35.0000  1.0479 0.3836
 Hotelling-Lawley trace 0.0898 3.0000 35.0000  1.0479 0.3836
    Roy's greatest root 0.0898 3.0000 35.0000  1.0479 0.3836
------------------------------------------------------------

------------------------------------------------------------
   socialstressbefore   Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.7299 3.0000 35.0000  4.3179 0.0108
         Pillai's trace 0.2701 3.0000 35.0000  4.3179 0.0108
 Hotelling-Lawley trace 0.3701 3.0000 35.0000  4.3179 0.0108
    Roy's greatest root 0.3701 3.0000 35.0000  4.3179 0.0108
------------------------------------------------------------
```

```
------------------------------------------------------------
 cognitivestressbefore  Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
         Wilks' lambda 0.5647 3.0000 35.0000  8.9939 0.0001
        Pillai's trace 0.4353 3.0000 35.0000  8.9939 0.0001
 Hotelling-Lawley trace 0.7709 3.0000 35.0000  8.9939 0.0001
    Roy's greatest root 0.7709 3.0000 35.0000  8.9939 0.0001
------------------------------------------------------------

------------------------------------------------------------
  physicalstressbefore  Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
         Wilks' lambda 0.8091 3.0000 35.0000  2.7524 0.0571
        Pillai's trace 0.1909 3.0000 35.0000  2.7524 0.0571
 Hotelling-Lawley trace 0.2359 3.0000 35.0000  2.7524 0.0571
    Roy's greatest root 0.2359 3.0000 35.0000  2.7524 0.0571
------------------------------------------------------------

------------------------------------------------------------
      anxietyTrait      Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
         Wilks' lambda 0.9050 3.0000 35.0000  1.2245 0.3154
        Pillai's trace 0.0950 3.0000 35.0000  1.2245 0.3154
 Hotelling-Lawley trace 0.1050 3.0000 35.0000  1.2245 0.3154
    Roy's greatest root 0.1050 3.0000 35.0000  1.2245 0.3154
------------------------------------------------------------

------------------------------------------------------------
       profession       Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
         Wilks' lambda 0.6772 3.0000 35.0000  5.5615 0.0031
        Pillai's trace 0.3228 3.0000 35.0000  5.5615 0.0031
 Hotelling-Lawley trace 0.4767 3.0000 35.0000  5.5615 0.0031
    Roy's greatest root 0.4767 3.0000 35.0000  5.5615 0.0031
============================================================
```

Our ANCOVA analysis with an average of the three stressors, shown in Listing 13, also did not result in any evidence for app type affecting stress levels.

Listing 13: ANCOVA results from analysing stress levels.

```
Model: stressafter ~ stressbefore + C(appType) + anxietyTrait + profession
                            OLS Regression Results
==============================================================================
Dep. Variable:            stressafter   R-squared:                       0.589
Model:                            OLS   Adj. R-squared:                  0.547
Method:                 Least Squares   F-statistic:                     13.97
Date:                Sat, 26 Aug 2023   Prob (F-statistic):           3.69e-07
Time:                        13:25:55   Log-Likelihood:                -30.429
No. Observations:                  44   AIC:                             70.86
Df Residuals:                      39   BIC:                             79.78
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  2.8537      0.614      4.644      0.000       1.611       4.096
C(appType)[T.video]       -0.0332      0.158     -0.210      0.835      -0.354       0.287
stressbefore               0.8554      0.154      5.544      0.000       0.543       1.167
anxietyTrait              -0.1155      0.183     -0.633      0.531      -0.485       0.254
profession                -2.1588      0.525     -4.111      0.000      -3.221      -1.097
==============================================================================
Omnibus:                        1.681   Durbin-Watson:                   2.291
Prob(Omnibus):                  0.431   Jarque-Bera (JB):                1.277
Skew:                          -0.195   Prob(JB):                        0.528
Kurtosis:                       2.262   Cond. No.                         30.8
==============================================================================
```

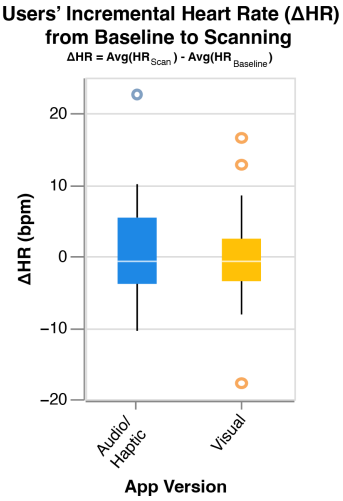


Figure 6: Users' incremental heart rates between the two app types. Notice their similarity.

**Heart rate.** When we initially performed ANCOVA on incremental HR (see Listing 15), the normality assumption was not satisfied. As recommended by Osborne and Overbay, we removed one outlier from the incremental heart rate data based on three standard deviations from the mean [34]. This resulted in the normality assumption being satisfied and the model being significant ($F_{5,37}$=2.78, $R^2$=.273, p=.031). The ANCOVA results after removing outliers (see Listing 14) resulted in the same variables having significant p-values: age (t(37)=-2.27, p=.029) and the anxiety trait (t(37)=2.13, p=.040). In both cases, the variable of interest, app type, was not significant. Furthermore, the distributions between the audio/haptic and visual conditions look similar, as shown in Figure 6.

Listing 14: ANCOVA results from analysing incremental heart rate. Note that this analysis has an outlier removed, which enables the data to satisfy the normality assumption.

```
Model: incHR ~ C(appType) + age + anxietyTrait + busyness + timesScanning
                            OLS Regression Results
==============================================================================
Dep. Variable:                   incHR   R-squared:                       0.273
Model:                             OLS   Adj. R-squared:                  0.175
Method:                  Least Squares   F-statistic:                     2.783
Date:                 Mon, 21 Aug 2023   Prob (F-statistic):             0.0313
Time:                         16:25:39   Log-Likelihood:                -133.94
No. Observations:                   43   AIC:                             279.9
Df Residuals:                       37   BIC:                             290.4
Df Model:                            5
Covariance Type:             nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                  -1.9451      6.545     -0.297      0.768     -15.206      11.316
C(appType)[T.video]         0.9476      1.817      0.522      0.605      -2.734       4.629
age                        -0.2427      0.107     -2.272      0.029      -0.459      -0.026
anxietyTrait                3.7578      1.763      2.132      0.040       0.186       7.330
busyness                    7.3427      6.948      1.057      0.297      -6.735      21.420
timesScanning              -0.3271      0.681     -0.480      0.634      -1.707       1.053
==============================================================================
Omnibus:                        0.770   Durbin-Watson:                   1.815
Prob(Omnibus):                  0.680   Jarque-Bera (JB):                0.181
```

Table 2: The number of participants who collided with objects or people.

| App Type | Number Participants | Collided Once | Collided Multiple Times | Collided (at all) |
|---|---|---|---|---|
| Audio/haptic | 26 | 5 (19.23%) | 1 (3.85%) | 6 (23.08%) |
| Visual | 24 | 3 (12.50%) | 2 (8.33%) | 5 (20.83%) |

```
Skew:                          -0.068   Prob(JB):                         0.913
Kurtosis:                       3.287   Cond. No.                         324.
==============================================================================
```

Listing 15: ANCOVA results from analysing incremental heart rate. Note that this analysis includes an outlier, such that the normality assumption is not satisfied.

```
Model: incHR ~ C(appType) + age + anxietyTrait + busyness + timesScanning
                            OLS Regression Results
==============================================================================
Dep. Variable:                    incHR   R-squared:                       0.314
Model:                            OLS   Adj. R-squared:                  0.223
Method:                 Least Squares   F-statistic:                     3.474
Date:                Thu, 24 Aug 2023   Prob (F-statistic):             0.0111
Time:                        13:35:35   Log-Likelihood:                -140.78
No. Observations:                  44   AIC:                             293.6
Df Residuals:                      38   BIC:                             304.3
Df Model:                           5
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     -2.6613      7.105     -0.375      0.710     -17.045      11.723
C(appType)[T.video]            0.3428      1.960      0.175      0.862      -3.625       4.311
age                           -0.2708      0.116     -2.343      0.024      -0.505      -0.037
anxietyTrait                   4.4991      1.894      2.376      0.023       0.666       8.332
busyness                       9.9702      7.480      1.333      0.190      -5.171      25.112
timesScanning                 -0.4493      0.738     -0.609      0.546      -1.944       1.045
==============================================================================
Omnibus:                        1.558   Durbin-Watson:                   1.728
Prob(Omnibus):                  0.459   Jarque-Bera (JB):                0.765
Skew:                           0.264   Prob(JB):                        0.682
Kurtosis:                       3.372   Cond. No.                         324.
==============================================================================
```

We took two additional alternative baseline HR measurements: one while participants walked indoors and once while they scanned indoors. As with the baseline taken during an outdoor walk, the distributions of the incremental HRs looked similar between the two app conditions. We show this in Figure 7. Thus, we did not expect there to be any significant differences and did not pursue further analyses.

**Collisions.** Self-reports of collisions were similar between the audio/haptic and visual conditions (see Table 2). Overall, 22% of participants collided with at least one object or person.

## 13.3 User scan accuracy and scan length

The results of the post-hoc analyses of user scan accuracy and scan length are shown in Listing 16. Note how the p-value with respect to scan length is significant, meaning users engaged with the audio/haptic scanning process longer than with the visual app.

Listing 16: Post-hoc results from analysing user scan accuracy and scan length.

```
--- Comparison between audio/haptic & visual: percentTimeOnScreen ---
```

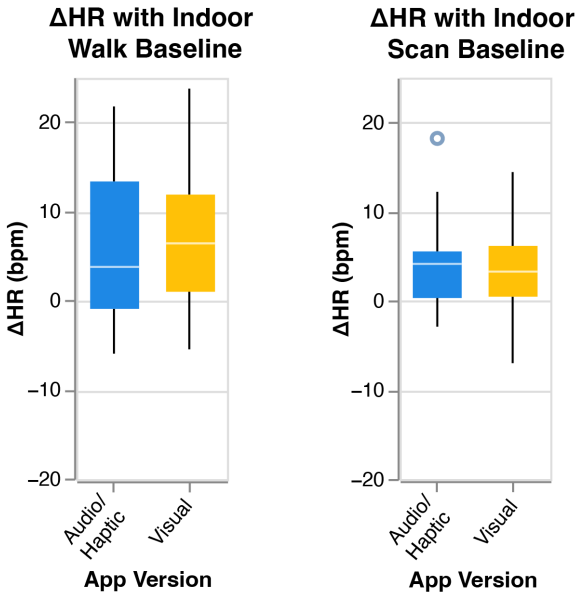


Figure 7: Incremental heart rate distributions measured using alternative baselines. On the left, the baseline measurement was taken while the participant walked indoors. On the right, the baseline measurement was taken while the participant scanned indoors. Notice the similar distributions between the audio/haptic condition and visual condition in each plot.

```
Normal? False -> Mann-Whitney U test

Significant? False
p-value: 0.08222109934077182
U: 240.0

percentTimeOnScreen means:
audio/haptic: 0.79
visual: 0.84

--- Comparison between audio/haptic & visual: percentTimeWithinDist ---
Normal? False -> Mann-Whitney U test

Significant? False
p-value: 0.0636218221836183
t(49): -1.551948734234481

percentTimeWithinDist means:
audio/haptic: 0.67
visual: 0.77

--- Comparison between audio/haptic & visual: tooFastCount ---
Normal? False -> Mann-Whitney U test

Significant? False
p-value: 0.24699165224461167
U: 297.0

--- Comparison between audio/haptic & visual: totalTime ---
Normal? False -> Mann-Whitney U test

Significant? True
p-value: 0.023274266598568395
U: 415.0

totalTime (seconds) means:
audio/haptic: 59.16
visual: 45.03
```

Listing 17 shows the results of our MANCOVA analysis for user scan accuracy and scan length. Note how the model and app type are significant.

Listing 17: MANCOVA results from analysing user scan accuracy and scan length.

```
Model: percentTimeOnScreen + percentTimeWithinDist + tooFastCount + totalTime ~ C(appType) +
     drift + busyness + timesScanning
                 Multivariate linear model
============================================================

------------------------------------------------------------
       Intercept        Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.3712 1.0000 42.0000 71.1612 0.0000
         Pillai's trace 0.6288 1.0000 42.0000 71.1612 0.0000
 Hotelling-Lawley trace 1.6943 1.0000 42.0000 71.1612 0.0000
    Roy's greatest root 1.6943 1.0000 42.0000 71.1612 0.0000
------------------------------------------------------------

------------------------------------------------------------
       C(appType)       Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.8389 1.0000 42.0000  8.0663 0.0069
         Pillai's trace 0.1611 1.0000 42.0000  8.0663 0.0069
 Hotelling-Lawley trace 0.1921 1.0000 42.0000  8.0663 0.0069
    Roy's greatest root 0.1921 1.0000 42.0000  8.0663 0.0069
------------------------------------------------------------

------------------------------------------------------------
         drift          Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.7626 1.0000 42.0000 13.0754 0.0008
```

Table 3: The number of participants who moved too quickly and received speed warnings. Note how participants in the visual condition did not seem to slow down after receiving their first warning.

| App Type | Number Participants | Participants with one or more speed warning | Participants with multiple speed warnings |
|---|---|---|---|
| Audio/haptic | 26 | 1 (3.85%) | 0 (0%) |
| Visual | 24 | 2 (8.33%) | 2 (8.33%) |

```
         Pillai's trace 0.2374 1.0000 42.0000 13.0754 0.0008
 Hotelling-Lawley trace 0.3113 1.0000 42.0000 13.0754 0.0008
    Roy's greatest root 0.3113 1.0000 42.0000 13.0754 0.0008
------------------------------------------------------------

------------------------------------------------------------
     timesScanning      Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.8250 4.0000 42.0000  2.2267 0.0824
         Pillai's trace 0.1800 4.0000 42.0000  2.3048 0.0741
 Hotelling-Lawley trace 0.2060 4.0000 42.0000  2.1631 0.0898
    Roy's greatest root 0.1728 4.0000 42.0000  1.8139 0.1441
------------------------------------------------------------

------------------------------------------------------------
        busyness        Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
          Wilks' lambda 0.9446 1.0000 42.0000  2.4610 0.1242
         Pillai's trace 0.0554 1.0000 42.0000  2.4610 0.1242
 Hotelling-Lawley trace 0.0586 1.0000 42.0000  2.4610 0.1242
    Roy's greatest root 0.0586 1.0000 42.0000  2.4610 0.1242
============================================================
```

### 13.3.1 Speed warnings

Seeing as participants needed to move 360° around the statue to complete the scan, and the visual-app users completed this faster, they also likely moved faster. Nonetheless, only two participants received speed warnings in the visual condition (compared to only one in the audio/haptic condition). That being said, as shown in Table 3, the participants in the visual condition did not seem to pay attention to the warning, as they received multiple additonal warnings after the first. In the audio condition, the participant who received the warning did not receive another.

## 13.4 Scan mesh accuracy

Our post-hoc and ANCOVA analyses of scan mesh accuracy (via chamfer distance) are in Listing 18 and 19. Note how the ANCOVA model, and the relationship between app type and chamfer distance are significant (Listing 19). The chamfer distance for the audio/haptic condition is also significantly smaller (i.e. more accurate) than the in the visual condition.

Listing 18: Post-hoc (t-test) results from analysing scan mesh accuracy scores. Note how the audio/haptic chamfer distance is significantly smaller than the visual chamfer distance. Since decreasing chamfer distance is associated with increasing accuracy, the audio/haptic scans are significantly more accurate than the visual scans with respect to chamfer distance.

```
--- Comparison between audio/haptic & visual: Chamfer Distance ---
Normal? True

Significant? True (One-tailed: audio/haptic less than visual)
p-value: 0.021246919335621808
t(49): -2.0841366628664324
```

```
chamfer means:
audio/haptic: 0.0549
visual: 0.0666
```

Listing 19: ANCOVA results from analysing chamfer distance, a mesh accuracy score. Note how the overall model and the "app type" variable are both significant.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                chamfer   R-squared:                       0.172
Model:                            OLS   Adj. R-squared:                  0.118
Method:                 Least Squares   F-statistic:                     3.188
Date:                Thu, 14 Sep 2023   Prob (F-statistic):             0.0324
Time:                        08:30:15   Log-Likelihood:                 128.70
No. Observations:                  50   AIC:                            -249.4
Df Residuals:                      46   BIC:                            -241.7
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        0.0395      0.010      3.934      0.000       0.019       0.060
C(appType)[T.video]              0.0113      0.005      2.071      0.044       0.000       0.022
drift[T.True]                    0.0121      0.007      1.714      0.093      -0.002       0.026
age                              0.0004      0.000      1.426      0.161      -0.000       0.001
==============================================================================
Omnibus:                        0.898   Durbin-Watson:                   0.932
Prob(Omnibus):                  0.638   Jarque-Bera (JB):                0.939
Skew:                           0.291   Prob(JB):                        0.625
Kurtosis:                       2.667   Cond. No.                         121.
==============================================================================
```

# 14 Full-scale study: Unprompted comments

In the full-scale study, we provided users with text boxes for "anonymous comments" on each of the surveys. Although we did not expect many users to provide additional comments, around 30% of participants did. Below, we describe the comment highlights. Afterwards, we list all of the comments.

Note that users entered these comments *before* experiencing the other app (although some users had previously used visual scanning apps, and thus describe the audio/haptic app with respect to those experiences).

**Audio/haptic use case highlights.** Three users described use cases for the audio/haptic app:

- "Audio cue[s] would suit a city environment for user[s] who usually have headphones"
- "The use of the audio is very inclusive"
- "I think it's great for safety measures, can be used to improve AR features in current games"

**Learning and enjoyment of the audio/haptic app and mesh reconstruction highlights.** A number of users mentioned enjoying the audio/haptic app:

- "As a new feature it does take few minutes do adapt! But it was fun!!!"
- "Great experience! I had fun!"
- "all was good for me (:"
- "Very good experience, thank you for having me be a part of it. Loved seeing the scan output and feel that others would benefit from [this]"
- "The instant result of being able to see the meshed scan is very beneficial"

**Privacy and confidence highlights.** One user mentioned that using audio/haptics "greatly enhanced [their] scanning experience!" They also stated, "having the screen off and not displaying a live image help[ed] [them] feel less concerned that others perceived [them] as invading their privacy", and that they felt "much more confident scanning while the screen remain[ed] dark" because they could tell they were "still producing a good scan using audio cues".

Two users of the visual app mentioned privacy concerns:

- "Biggest concerns were what people thought when pointing camera at them whilst scanning, due to other pedestrians had to get quite close to some people."
- "I had a few people looking at me oddly thinking that I was recording them so had to explain that I was actually recording the statue (object) which made me feel a bit uncomfortable."

### All comments from audio/haptic app users.

- i think its great for safety measures, can be used to improve AR features in current games
- Using audio cues greatly enhanced my scanning experience! Not only was I able to hear/feel whether my scans are capturing the correct and proper mesh that is needed to generate the representation of the object, but having the screen off and not displaying a live image helps me to feel less concerned that others perceived me as invading their privacy. I feel much more confident scanning while the screen remains dark and I was still producing a good scan using audio cues.
- The "bounding-box" not quite matching up over the object was quite frustrating, and I felt like I had to click it a few times before I was happy with where it was. Wondering if a big flat rectangle isn't the best way of expressing the "thing" I'm meant to be pointing at? It felt a little bit unclear or unsatisfying.
- I felt a little insecure about being theft in the street, as I saw a person that looked like he had malicious intentions, and walking around holding up your phone does not feel secure. Anyway, nothing happened, but it stressed me a little bit.
- Very good experience, thank you for having me be a part of it. Loved seeing the scan output and feel that others would benefit from seeing what they input eg Ingress Portal Scans
- No idea of what I really scanned ,
- the use of the audio is very inclusive although 'good job' feels a little childlike, dependent on game program this would benefit being changed to suit audience. That said the audio was more beneficial than the vibration as it was very light taps and more difficult to ascertain direction to point the phone The instant result of being able to see the meshed scan is very beneficial and would be a great feature or side feature for game players
- I am on a prescribed medication that increases my heart rate
- Unsure if the audio/distance was related to the POI or the location where I was standing when starting the scan. The overriding feel was that this was my physical start position.
- Great experience! I had fun!
- If others stand in front of the object it blocks the scan
- Audio cue would suit a city environment for user who usually have headphones in or a vibrate from your questions it seems these are availible in different versions
- Was asked by a tour group to stop filming. Had to explain that I was not actually filming. Done! Audio etc tends to blend in with background voices. A little difficult to decipher
- I think it went okay!!!
- From scanning [before] in Ingress its more natural to look at the poi when scanning landmarks, from playing Niantic games for approx 9 years i have no concern in public using devices or walkng/waiting for public to be able to move again
- Just hoping it all goes well first time
- all was good for me (:

**All comments from visual app users.**

- A lot of people were taking photos with the statue after I started the scan, which was a little nerve wracking
- Some people weren't sure if I was trying to record them or something but usually just moved out of the way
- It took a little while to get the object to properly match before scanning, but otherwise it was fun :)
- In a busy area it's a bit hard scanning with a wide radius. If the radius can be smaller that helps and the angle of the phone makes it tricky. Would be good if you can leverage wide angles on phones.
- It is cool but i dont understand the point
- I dislike the possibility of submitting video of strangers when uploading scans, so I probably won't be scanning this Agatha Christie monument in game
- As with scanning in game I would sooner avoid talking to strangers about the activity. I felt apprehensive before doing the scanning about scanning this object in such a busy area in august
- Tutorial needs work, especially when taking steps to left or right
- It was quite busy with people who apologised for getting in the way
- i dont understand the point of scanning objects
- We can think of ways to improve the bugs as it's quite laggy while moving.
- It felt weird to point at other peoples faces with a phone camera especially kids
- A group of people was posing for a picture and they got nervous with me "filming". I said I was not filming and I offered to take a picture for them but they were uneasy and left.
- pikachu :)
- Biggest concerns were what people thought when pointing camera at them whilst scanning, due to other pedestrians had to get quite close to some people.
- I had a few people looking at me oddly thinking that I was recording them so had to explain that I was actually recording the statue (object) which made me feel a bit uncomfortable.

# References


[1] Daniel Andersen, Peter Villano, and Voicu Popescu. Ar hmd guidance for controlled hand-held 3d acquisition. *IEEE Transactions on Visualization and Computer Graphics*, 25(11):3073–3082, 2019.

[2] Inc. Apple. App store, 2023.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

[4] Valentin Bazarevsky and Ben Hahn. Interactive image segmentation task guide, Jul 2023.

[5] Carlos Campos, Richard Elvira, Juan J. Gomez, Jose M. M. Montiel, and Juan D. Tardos. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[6] Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*. Chapman and Hall/CRC, 2017.

[7] Ching Kit Chen, Barbara Cifra, Gareth J. Morgan, Taisto Sarkola, Cameron Slorach, Hui Wei, Timothy J. Bradley, Cedric Manlhiot, Brian W. McCrindle, Andrew N. Redington, Lee N. Benson, and Luc Mertens. Left ventricular myocardial and hemodynamic response to exercise in young patients after endovascular stenting for aortic coarctation. *Journal of the American Society of Echocardiography*, 29(3):237–246, 2016.

[8] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[9] Silvio Giancola, Matteo Valenti, and Remo Sala. *A survey on 3D cameras: Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies*. 2018.

[10] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, 73:159–170, 2017.

[11] Pokémon GO. Scanning a pokéstop, 2021.

[12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2020.

[13] Hexagon. Immersal app - your personal metaverse, 2023.

[14] Apple Inc. Meet object capture for ios, 2023.

[15] Snap Inc. Bringing locations to life with ar - snap!'s custom landmarker creator, 2022.

[16] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *UIST '11 Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, October 2011.

[17] Daniel Jackson. Towards a theory of conceptual design for software. In *2015 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!)*, Onward! 2015, page 282–296, New York, NY, USA, 2015. Association for Computing Machinery.

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[19] Samuli Laato and Thomas Tregel. Into the unown: Improving location-based gamified crowdsourcing solutions for geo data gathering. *Entertainment Computing*, 46:100575, 2023.

[20] Taija MM Lahtinen, Jukka P Koskelo, Tomi Laitinen, and Tuomo K Leino. Heart rate and performance during combat missions in a flight simulator. *Aviation, space, and environmental medicine*, 78(4):387–391, 2007.

[21] Yung-Hui Lee and Bor-Shong Liu. Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviation, space, and environmental medicine*, 74(10):1078–1084, 2003.

[22] Google LLC. Google play, 2023.

[23] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, aug 1987.

[24] Niantic Ltd. Evolving niantic ar mapping infrastructures, 2023.

[25] Hanuma Teja Maddali and Amanda Lazar. Understanding context to capture when reconstructing meaningful spaces for remote instruction and connecting in xr. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[27] Edward Miller. Building the 'ar-cloud': Why we're building a realtime, 3d map of the world that machines can understand, 2017.

[28] Jon D Morris. Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, 35(6):63–68, 1995.

[29] MrAnmol. Circumplex model of emotion, 2023.

[30] Inc. Niantic. Ingress portal scanning, 2022.

[31] Inc. Niantic. Niantic wayfarer, 2023.

[32] Inc. Niantic. Scaniverse: Capture life in 3d, 2023.

[33] J. Ortiz-Sanz, M. Gil-Docampo, T. Rego-Sanmartín, M. Arza-García, and G. Tucci. A pbel for training non-experts in mobile-based photogrammetry and accurate 3-d recording of small-size/non-complex objects. *Measurement*, 178:109338, 2021.

[34] Jason W Osborne and Amy Overbay. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1):6, 2004.

[35] Heather L. O'Brien, Paul Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28–39, 2018.

[36] Victor Prisacariu and Pierre Fite-Georgel. Lightship vps: Building our 3d map from crowd-sourced scans, Nov 2022.

[37] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–19, Cham, 2022. Springer Nature Switzerland.

[38] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. Wearable-based affect recognition—a review. *Sensors*, 19(19), 2019.

[39] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[41] Petros Skapinakis. *Spielberger State-Trait Anxiety Inventory*, pages 6261–6264. Springer Netherlands, Dordrecht, 2014.

[42] Sebastian Thrun. *Simultaneous Localization and Mapping*, pages 13–41. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[43] Daniel Tompkins, Dimitra Emmanouilidou, Soham Deshmukh, and Benjamin Elizalde. Multi-view learning for speech emotion recognition. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, June 2023.

[44] Yi-Chin Wu, Liwei Chan, and Wen-Chieh Lin. Tangible and visible 3d object reconstruction in augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 26–36, Beijing, China, 2019. IEEE.