AR HMD Guidance for Controlled Hand-Held 3D Acquisition

Category: Research

Figure 1: Top left: Our acquisition guidance system, made up of an AR HMD and a handheld camera rig tracked using a fiducial marker. Top middle: Operator AR view of scene, with virtual overlay of camera rig, automatically-generated suggested views (white icons), and suggested acquisition path (blue lines). Top right: Interactive guidance (green rectangle) for precise 6-DOF alignment of camera rig with suggested view. Second row: Photogrammetric reconstruction of scene from images captured during guided acquisition. Third row: AR view of suggested views on outdoor scene, and photogrammetric reconstruction from acquired views.

Abstract

Photogrammetry is a popular method of 3D reconstruction that uses conventional photos as input. This method can achieve high quality reconstructions so long as the scene is densely acquired from multiple views with sufficient overlap between nearby images. However, it is challenging for a human operator to know during acquisition if sufficient coverage has been achieved. Insufficient coverage of the scene can result in holes, missing regions, or even a complete failure of reconstruction. These errors require manually repairing the model or returning to the scene to acquire additional views, which is time-consuming and often infeasible. We present a novel approach to photogrammetric acquisition that uses an AR HMD to predict a set of covering views and to interactively guide an operator to capture imagery from each view. The operator wears an AR HMD and uses a handheld camera rig that is tracked relative to the AR HMD with a fiducial marker. The AR HMD tracks its pose relative to the environment and automatically generates a coarse geometric model of the scene, which our approach analyzes at runtime to generate a set of human-reachable acquisition views covering the scene with consistent camera-to-scene distance and image overlap. The generated view locations are rendered to the operator on the AR HMD. Interactive visual feedback informs the operator how to align the camera to assume each suggested pose. When the camera is in range, an image is automatically captured. In this way, a set of images suitable for 3D reconstruction can be captured in a matter of minutes. In a user study, participants who were novices at photogrammetry were tasked with acquiring a challenging and complex scene either without guidance or with our AR HMD based guidance. Participants using our guidance achieved improved reconstructions without cases of reconstruction failure as in the control condition. Our AR HMD based approach is self-contained, portable, and provides specific acquisition guidance tailored to the geometry of the scene being captured.

Index Terms: Human-centered computing—Mixed / augmented reality; Human-centered computing—User interface programming

1 Introduction

3D acquisition and reconstruction of real-world objects and scenes is an important technology with a wide range of applications. Inspection and maintenance of industrial facilities is made more efficient by capturing reliably precise 3D models of machinery. Capturing high-fidelity models of living spaces for easy interactive viewing online helps sellers in the real estate industry more efficiently attract interested buyers. Analysis of crime scenes by law enforcement is made more robust by 3D acquisition, by tracing bullet paths or determining visibility of different areas of the scene. The digital humanities and archaeology is enhanced by acquiring metrically-accurate models of artifacts both for preservation and for analysis by researchers around the world. The usefulness of these applications relies on the ability to acquire data efficiently with the goal of

achieving a high-quality reconstruction.

Acquisition based on digital photography (photogrammetry) is a popular method due to the ubiquity of high-quality cameras in smartphones, and due to recent improvements in photogrammetric reconstruction algorithms implemented in consumer-level photogrammetric software, which can more robustly detect salient features between photos captured from nearby views and can match these features to generate 3D models of real-world scenes.

Despite its reliance on relatively time-intensive offline reconstruction, photogrammetry is able to achieve high quality results, due to its emphasis on crisp, feature-rich images captured in a dense arrangement around the target scene. However, the high quality results depend on a high quality acquisition process, which depends on several factors that an operator must simultaneously keep in mind while acquiring the scene. The scene should be captured from views that are approximately the same distance from the scene so that the amount of detail is consistent across the model. Neighboring views should also have a minimum amount of image overlap so that scene features are imaged multiple times for reduced uncertainty. The placement of views in space should be adapted based on the shape of the scene. Each image captured should be blur-free so that feature matching has the highest chance of success. Coverage of all regions of interest in the scene is needed for an optimal reconstruction.

Poor acquisition leads to poor reconstruction results, such as blurry textures or low geometric resolution in some regions, or holes where geometry could not be reconstructed. In the extreme case, entire sections of the geometry may be missing, or only a small fraction of input images may be matched, leading to a catastrophic failure to reconstruct any model. Such acquisition problems are not immediately obvious until after the acquisition session, during the computationally intensive reconstruction stage, at which point returning to the scene to acquire more images may be impractical or even impossible.

Achieving high quality acquisition is a challenging task for a human operator, as it requires a quantitative analysis of the scene and precise measurement of camera position and orientation in order to assume the desired poses. An operator must keep track of which views should be captured, as well as which views have and have not yet been captured during acquisition.

While prior work has explored the problem of interactive guidance during 3D acquisition, such approaches have relied on external cloud-based computation or hand-held smartphone tracking that is limited to simple scenes (e.g. approximating the scene as a hemisphere on a flat surface) and that requires slow device movements. What is needed is a self-contained method of automatically generating view suggestions for complex scenes, as well as a method of intuitively visualizing the suggestions to the operator and tracking which views have or have not yet been captured.

Augmented reality head-mounted displays (AR HMDs), which couple an augmented overlay of visual information onto the real world with real-time acquisition of rough geometric models for localization purposes, can help address this challenge. The combination of these two properties allows for geometric analysis of a target scene for photogrammetric acquisition, while also providing the tracking and visualization needed to show an operator which views should be acquired and to guide the operator to capture imagery from precise locations that fulfill the acquisition criteria needed for good coverage.

In this paper we present a method for AR HMD-based guidance to enable efficient photogrammetric acquisition with guarantees. Specifically, our approach provides guidance to achieve consistent coverage of a target region of interest, consistent image overlap between neighboring views, and consistent distance between views and the scene.

Our system is made up of an AR HMD and a smartphone camera rig (Fig. 1, top left). Using the onboard camera on the AR HMD, we track the 6-degree-of-freedom (6-DOF) pose of the camera from a fiducial marker rigidly attached to the camera rig. The AR HMD automatically tracks its own pose within the operator's environment and generates a low-polygon approximate geometric model using active onboard sensors. The operator can select within the scene a region of interest to be acquired. Upon selection, we generate a set of views that surround the region of interest and that are suitable for photogrammetric acquisition (Fig. 1, top middle). The suggested views are rendered as AR overlays onto the operator's view of the real world scene. When the operator places the camera rig near the suggested view, visual feedback is provided to guide the operator to assume a precise 6-DOF pose, and a photo is automatically captured when the camera is held stably in place (Fig. 1, top right). The operator moves from view to view until a set of hundreds of images has been acquired suitable for offline photogrammetric reconstruction (Fig. 1, middle row). Our approach is self-contained and portable and works both indoors and outdoors in overcast/cloudy weather, making it well-suited to the use case of photogrammetry (Fig. 1, bottom row). We refer the reader to our accompanying video, which demonstrates our approach in use on a variety of scenes.

There are two central research problems addressed by our work. First is how to define a set of acquisition poses that enforce coverage guarantees specific to our reconstruction method of photogrammetry; our solution is based on signed distance functions and iterative mesh refinement that relies on representing overlap between images as the length of edges in a mesh. Second is how to provide visual guidance to allow a user to capture an image of a real-world scene from a specified 6-DOF pose; our approach hinges on adaptive visual guidance offset from the physical camera that allows the user to position the camera intuitively for each one of hundreds of acquisition poses.

We have validated our approach in a user study ($N = 10$) in which participants acquired imagery of a real-world scene either with a conventional smartphone camera (Control condition) or with our AR HMD based guidance system (Experimental condition). When compared against a ground truth reference reconstruction of the scene, images acquired in the Experimental condition led to a significantly more complete model, with approximately 95% of the scene reconstructed to within 5cm of ground truth as oppose to only 60% of the scene reconstructed in the Control condition.

2 Prior Work

In this section, we touch on some relevant prior work, particularly in the areas of view planning and the use of AR for acquisition guidance.

There has been much research into the problem of view planning, with the goal of defining a set of views to acquire that efficiently and completely cover a region of interest [7]. Ahn et al. examined a method of planning the placement of 3D scanners for large outdoor historical sites [3]. Wakisaka et al. used a voxel occupancy classification approach to define an optimal placement of terrestrial laser scanners for acquisition of industrial spaces [28]. While these approaches are suitable for the task of capturing a large scale scene with a capture device that is inconvenient to move regularly, the approach requires pre-labeling an existing aerial map which is less suited for more casual photogrammetric acquisition. Additionally, these approaches focus on the 2D placement of a scanner within an environment, while handheld photogrammetry usually relies on the additional views that specific 3D, 6-DOF poses can add to a reconstruction.

Augmented reality has been explored for its potential to improve the act of view planning during 3D acquisition and to integrate visual feedback into the capture process. Pan et al. presented an early approach to this problem by interactively reconstructing a handheld object while presenting a video AR overlay of the current reconstruction, as well as guidance arrows to tell the operator to

manipulate the object to view all sides [21]. This approach is suited for small-scale objects that can be easily approximated as a sphere for outside-in capture, but a different approach for visualization and analysis is needed for larger and more arbitrary geometry.

Some approaches to interactive guided 3D acquisition focus on online reconstruction, where each acquired image is integrated into an increasingly-improving model [14]. To some degree, our approach makes use of such incremental reconstruction by relying on the rough geometry generated by the AR HMD. However, our approach makes the assumption that a truly high quality photogrammetric reconstruction (of the sort that justifies the use of photogrammetry) is computationally intensive and not well suited for a fully self-contained and portable platform. By ensuring that our acquisition approach is self-contained, we retain an advantage of photogrammetry in that external computation or an always-on broadband Internet connection is not needed, and that *in situ* capture can be done even in austere environments such as archaeological sites.

This disadvantage of online reconstruction is illustrated by the work of Langguth and Goesele, in which a robust incremental reconstruction is achieved with next-best-view guidance to a user, but the processing time between each view limits feasible capture to only a few dozen images and not the hundreds that usually needed for high quality photogrammetric reconstruction of larger objects [16]. Another example of online reconstruction for guided AR acquisition was shown by Locher et al., where a smartphone interface displayed a next-best-view map to a user to encourage acquisition of a scene from uncaptured views [18]. However, each view was sent to a remote server for incremental reconstruction and recalculation of the next best view, a process took almost 3 minutes per view. This approach is suitable for a mass of users each acquiring a few additional images while walking past a famous landmark; it is less suitable for single-session acquisition by one or a few users.

While most prior work on AR view guidance has focused on smartphones or tablets, there has been some investigation into the secondary perspective that an AR HMD can provide. Andersen and Popescu proposed an AR HMD based method of guiding a user to acquire a dense set of panoramas for the purpose of image-based modeling [4, 5]. The user wore an AR HMD with a panoramic camera attached, and walked with consistent head height through an indoor scene; an AR interface displayed a top-down view of the room, divided into grid cells that were marked as either captured or not yet captured. While this approach is suitable for capturing a large and dense set of panoramas, it is not suitable for photogrammetric capture which requires a set of views that is both varied in height and concentrated around a consistent distance to the target.

Our work is inspired by automated capture of outdoor scenes by aerial drones. Such approaches rely on an *explore-and-exploit* strategy in which a rough, low-detail reconstruction is made of a scene by acquiring imagery from a known safe altitude, and then determining a set of additional views that are both navigable and achieve a higher quality reconstruction [27]. Huang et al. implemented a next-best-view acquisition of scenes using a toy drone, for the purposes of image-based modeling [15]. Roberts et al. proposed a method of drone-based refined acquisition by voxelizing the scene, determining an optimal camera orientation for each voxel, and using an additive approximation of a coverage criterion to select views and an efficient path [26]. Our approach to view generation is perhaps most similar to the work of Peng and Isler, who simplify the 3D search space of finding views from voxels into a 2D search space by defining points on a manifold that wraps around the target scene; however, our approach uses signed distance fields rather than relying solely on extrusion of mesh points along a surface normal due to the noisy nature of the rough geometry from our AR HMD [22, 23].

Some recent research has focused on fully autonomous scene reconstruction by robotic operators [1]. For example, Liu et al. investigated a method of both autonomously exploring a scene without prior human input while also scanning individual objects within the scene [17]. While autonomous devices hold great promise for the future of 3D acquisition, especially for large scenes that would be tedious to acquire manually, it is still the case that robotic systems are expensive, bulky, and have difficulty navigating many cluttered environments. In contrast, humans are extremely effective at navigating the sorts of human-designed environments that contain many interesting acquisition subjects.

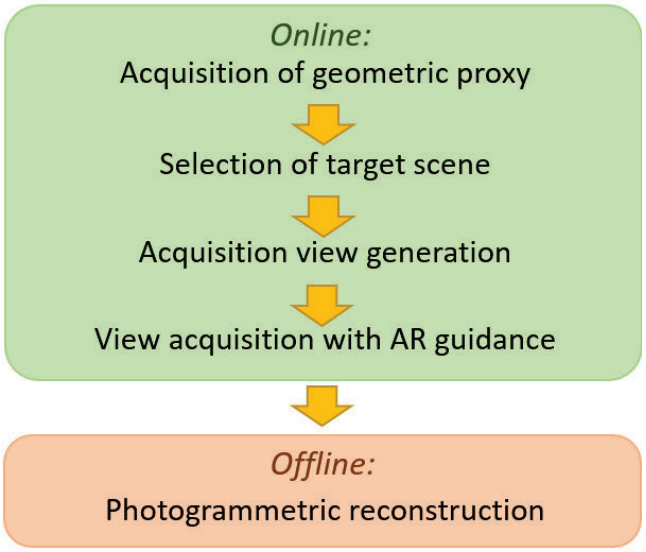

Figure 2: Pipeline of our approach.

We also wish to distinguish our approach from a recent work by Dong and Höllerer that uses an AR HMD to capture color textures of a scene and apply them to a reconstructed mesh [9]. The work is focused on augmenting the AR HMD's rough geometric model of the scene (which we also use in our work) with color aligned from the AR HMD's onboard RGB camera. The goal of their work is to operate under a constrained memory and performance footprint and to achieve good texturing of that rough geometry in real time. However, their approach only results in a textured mesh at the same level of quality as the original rough geometric model, with holes or artifacts still present. Regions of missing geometry cannot be repaired after acquisition because image data is not preserved. Our approach acknowledges that a high quality reconstruction is only feasible with offline processing, and so focuses only on guidance and saving all input data during acquisition for later processing.

3 Overview

In this section, we provide an overview of our approach. An operator uses our AR acquisition interface (Sect. 4), which is made up of an AR HMD and a tracked handheld camera rig. The AR HMD visualizes a set of suggested acquisition views, and provides interactive guidance to the operator to assume each view with the handheld camera rig.

Fig. 2 illustrates our pipeline, which can be divided into an online acquisition stage and an offline reconstruction stage. In our work, we focus exclusively on the online acquisition stage in order to increase the coverage and density of acquired images, and we use conventional off-the-shelf 3D reconstruction approaches to generate a 3D model from the acquired images.

Acquisition of geometric proxy: Our method first requires the acquisition of a rough geometric proxy of the scene. We rely on modern AR HMDs' ability to provide rough geometry of the wearer's environment using onboard sensors. An operator can simply wear the

AR HMD and walk around a target scene and generate a geometric proxy in a matter of seconds.

Selection of target scene: The operator uses an interface on the AR HMD to select a region of interest and a selection radius. The subset of the rough geometric proxy that is within the selection radius is copied and used for generating suggested acquisition views.

Generation of suggested views: Given an input mesh taken from the AR HMD's rough geometry, our approach automatically generates a set of acquisition views based on heuristics that are widely used in the photogrammetric community to ensure a high quality reconstruction. First, imagery should be captured from a consistent distance from the scene, so as to prevent cases where some regions of interest have widely varying resolution or detail from other regions of interest. Second, a large amount of images overlap (usually $50\% - 70\%$) between adjacent photos is important, so that a single feature point is imaged in multiple views and can be reconstructed accurately. A third, and implicit, design requirement is that any views presented to the operator should be physically reachable by a human without excessive difficulty. Sect. 5 provides detail about our method of automatic view generation.

View acquisition with AR guidance: Once a set of views has been generated, they are visualized to the operator as an AR overlay superimposed onto the target scene. The operator then is tasked with physically placing the handheld camera rig at each view, matching both position and orientation. The AR interface provides interactive visual feedback so that the operator can precisely align the camera with the suggested view. Once the camera has been placed near the suggested view, the camera automatically captures an image. The operator repeats this process until all views have been captured.

Photogrammetric reconstruction: The output of acquisition is a set of photos that cover the scene, along with a rough estimate of the camera's pose based on the AR HMD's tracking of the camera rig. We input these to a conventional structure-from-motion system, which extracts and matches features of nearby photos, generates a sparse point cloud, and then creates a textured 3D mesh of the target scene.

4 AR Acquisition Interface

In this section, we describe our AR HMD-based acquisition interface. We explain how an operator can use our approach to select a region of the environment for acquisition, to visualize a set of acquisition views suitable for photogrammetric reconstruction, and to use AR-based guidance to precisely place a handheld camera at each suggested view.

During acquisition, the operator wears an AR HMD and uses a handheld camera rig to capture imagery of the scene from multiple views (Fig. 1, top middle). The camera rig and AR HMD are wirelessly networked together. The AR HMD contains active sensors that track the headset's position/orientation relative to the environment as the operator walks around the scene. The AR HMD also uses these active sensors to generate a rough geometric model of the scene. A forward-facing RGB camera onboard the AR HMD is calibrated prior to operation to determine the camera intrinsics and is used to track the 6-DOF pose of a fiducial marker attached to a handheld camera rig. The camera rig is made up of a smartphone with a scene-facing RGB camera and a rigidly-mounted fiducial marker. Fig. 1, top right, illustrates the operator's view of the tracked camera rig as seen through the AR HMD. The frame of the smartphone (white rectangles) and the position of the smartphone's camera (blue dot in top left of frame) are highlighted to the operator as AR overlays.

Given a real-world scene to be acquired, the operator first points at the center of the scene with an AR cursor visible on the AR HMD, and defines a radius of interest to select a subset of the AR HMD's rough geometric model. A user interface on the camera rig allows the operator to specify a desired target image overlap $\omega \in [0.0, 1.0]$ and desired camera-to-scene distance d. Our automatic view generation approach defines, in a matter of seconds, a set of camera poses that cover the region of interest at a consistent distance from the scene and with the desired image overlap. We next describe the details of our automatic view generation approach.

5 Generation of Suggested Views

Input: mesh M, camera-to-scene distance d, image overlap ω_h, image sidelap ω_v, camera horizontal FOV θ_h, camera vertical FOV θ_v, user height h_u, floor height h_f
Output: set of 6-DOF views

$b_h \leftarrow (1-\omega_h)\left(2d\tan\frac{\theta_h}{2}\right); b_v \leftarrow (1-\omega_v)\left(2d\tan\frac{\theta_v}{2}\right)$
$sdf \leftarrow SignedDistance(M)$
$IC \leftarrow IsoContour(sdf, d)$
$M_{offset} \leftarrow MarchingCubes(IC)$
$M_{refined} \leftarrow M_{offset}$
// Baseline enforcement
for *k iterations* **do**
  foreach *edge* $e_{ij} = (v_i, v_j)$ *in* $M_{refined}$ **do**
    $n_i \leftarrow v_i.normal$
    $view_i \leftarrow LookAt(v_i, -n_i, up)$
    $(x_j, y_j) \leftarrow Project(view_i, v_j)$
    $b_{ij} \leftarrow lerp(b_h, b_v, atan2(y_j, x_j)\frac{2}{\pi})$
    if $e_{ij}.length < b_{ij} - \varepsilon$ **then** collapse e_{ij}
    else if $e_{ij}.length > b_{ij} + \varepsilon$ **then** split e_{ij}
  end
end
$M_{clipped} \leftarrow Clip(M_{refined}, h_u, h_f)$
// View definition
$views \leftarrow \{\}$
foreach *vertex* v_i, *normal* n_i *in* $M_{clipped}$ **do**
  $view_i \leftarrow LookAt(v_i, -n_i, up)$
  $views \leftarrow views \cup \{view_i\}$
end
// View augmentation
foreach *vertex* v_i, *normal* n_i *in* M **do**
  $p \leftarrow v_i + d * n_i$
  if $dist(p, NearestNeighbor(p, views)) > max(b_h, b_v)$ **then**
    $v \leftarrow LookAt(p, -n_i, up)$
    $views \leftarrow views \cup \{v\}$
  end
end
return *views*

Algorithm 1: Our method of generating acquisition views.

In this section, we describe our method of generating a set of suggested acquisition views suitable for photogrammetric reconstruction. algorithm 1 provides an overview of our approach. The input to our algorithm is a triangle mesh M that is copied from the AR HMD's rough geometric model and which defines a region of interest in the scene (Fig. 3, top left). The output of our algorithm is a set of 6-degree-of-freedom poses (positions and orientations) that define the suggested views for the operator's handheld camera rig.

Besides the input mesh M, we also take as an input parameter a desired camera-to-scene distance d, as well as a desired image overlap as a value between 0.0 (0% overlap) and 1.0 (100% overlap). We consider both horizontal overlap (which we label ω_h), and vertical *sidelap* or ω_v. The handheld camera's field of view is an additional parameter and is a fixed property of the camera hardware. We label the camera's horizontal field of view as θ_h and the vertical field of view as θ_v. Our view generation method takes into account both the horizontal and vertical field of view of our acquisition device.

As explained in Sect. 3, our method must also ensure that the suggested views are reachable by a human operator. At runtime,

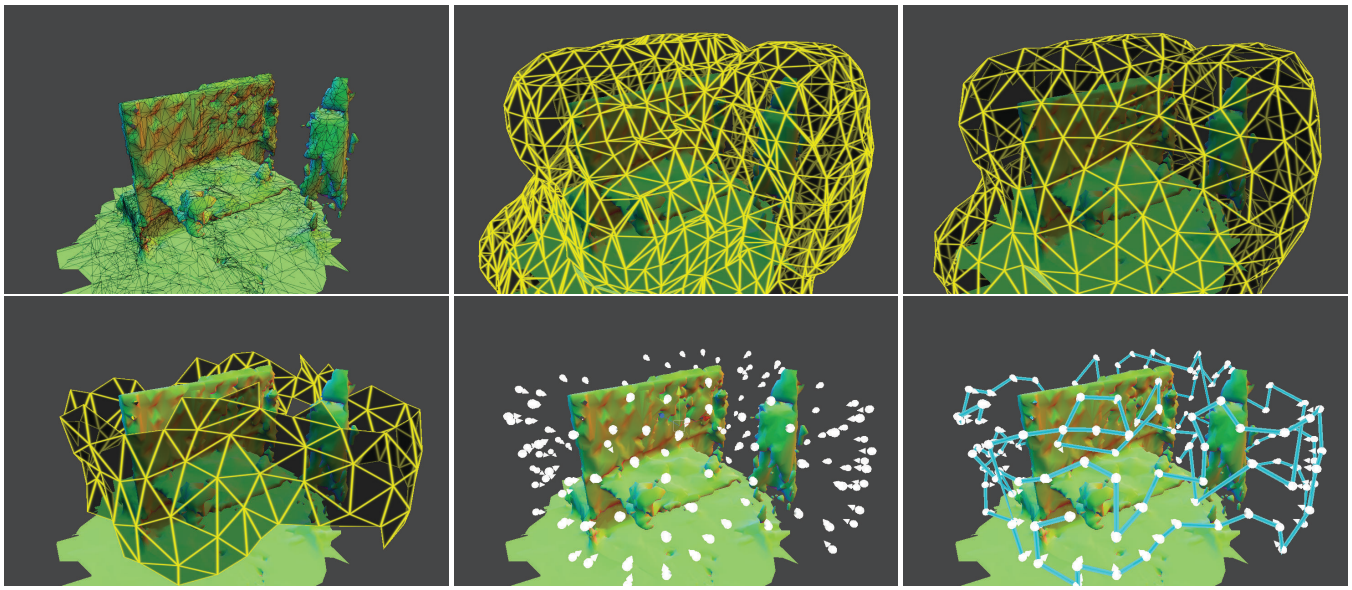

Figure 3: The stages of the view generation pipeline. Top left: input mesh M selected from the AR HMD rough geometry. Top center: mesh M_{offset} created with marching cubes from offset SDF isocontour. Top right: mesh $M_{refined}$ after iterative mesh refinement. Bottom left: mesh trimmed to remove human-unreachable views. Bottom center: generated views (white arrows). Bottom right: suggested acquisition path between views (blue lines).

we raycast from the operator's HMD position directly downward to the floor, and estimate the operator's height h_u and the height of the floor plane h_f. We include h_f and h_u as input parameters to our view generation.

Given values of d, ω_h, ω_v, θ_h, and θ_v we define a desired horizontal camera baseline b_h and a desired vertical camera baseline b_v. That is, if two views are horizontally adjacent to each other, their baseline should be b_h to achieve an image overlap of ω_h; if two views are vertically adjacent (one directly above the other), the baseline should be b_v to achieve an image sidelap of ω_v. We assume that the scene can be locally approximated as a plane imaged by adjacent cameras oriented opposite the plane's normal. algorithm 1 details how the baseline values are computed. For example, with an FOV $\theta_h = 65^\circ$, a camera-to-scene distance $d = 0.5$m, and an overlap of $\omega_h = 0.67$, the horizontal baseline b_h is approximately 0.21m.

We create a signed distance field (SDF) from our input region of interest mesh M. We define an offset isocontour from the SDF at distance d, and convert the offset implicit function to a triangulated mesh M_{offset} using marching cubes [19]. Fig. 3, top center, shows M_{offset}.

We next iteratively refine the offset mesh. The vertices and surface manifold of M_{offset} are approximately distance d from the surface of M, but the initial distance between the vertices is arbitrary and derived only from the resolution of the marching cubes algorithm. We use a modified version of Garland and Heckbert's Quadric Error Metric (QEM) algorithm for iterative mesh refinement to split and join faces based on a target edge length. [10].

By default, the QEM algorithm adjusts meshes based on a single target edge length. In our case, we desire a differing edge length for each edge depending on whether camera views centered at the two vertices of the edge, and oriented opposite the vertices' normals, would be horizontally or vertically aligned with each other. During each iteration of mesh refinement, we use the *LookAt* function to compute a 6DOF view for each endpoint of an edge e_{ij} in the mesh. The position of the view is at the vertex position, the forward direction of the view is opposite the vertex's normal, and the input up direction is the world up direction in the AR HMD coordinate system (+Y). We then project the vertex position of one edge endpoint onto the other endpoint's view. Depending on how horizontal vs vertical the projected point is, we linearly interpolate between b_h and b_v to get a baseline b_{ij}, which we set as the target edge length for this edge during this iteration of mesh refinement. To prevent the refined mesh from oversmoothing and changing its shape radically from the original surface manifold defined by the SDF, we project the updated vertex positions after each iteration onto the nearest triangle of the original M_{offset}. We refine the mesh over k iterations (we use $k = 20$ in our experiments) until all edge lengths are approximately b_{ij}. After iterative mesh refinement, we have a mesh $M_{refined}$ where all vertices are approximately d from the surface of M while also having adjacent vertices separated by a baseline between b_h and b_v, depending on how horizontally or vertically aligned the adjacent vertices are (Fig. 3, top right).

The resulting mesh $M_{refined}$ may include regions that are difficult or impossible for a human operator to reasonably reach (e.g. vertices under the floor, or too high to reach with one's arms). Using the defined input parameters h_f and h_u for the height of the floor and the height of the user, we clip $M_{refined}$ of all faces and vertices that are above a certain offset from h_u or below a certain offset from h_f. In our experiments, we set the lower cutoff to 0.6m above the floor height h_f, and we set the upper cutoff to 1.05 times the estimated height of the operator h_u. Fig. 3, bottom left, shows the resulting trimmed mesh.

We now convert the generated mesh into a set of camera positions and camera forward vectors. We take each vertex V_i and associated normal N_i of $M_{refined}$ and define a camera pose C_i with position V_i and forward direction $-N_i$. To avoid regions of the scene being undersampled in a complex scene, we augment the set of views by taking the original vertices of our input mesh M, extruding them d along their normal directions, and adding them iteratively to our set of camera poses if (1) they are in between our cutoff heights, (2) they are not within a collision distance of M, and (3) they are not within $max(b_h, b_v)$ of any other camera in the list of camera poses.

Now that we have our list of positions and forward directions for our cameras, we create full 6-DOF poses for each by using the *LookAt* method with the world up-direction $(+Y)$ as the input up direction. This direction assures that most poses will keep the

handheld camera rig in a consistent orientation during acquisition, which is most comfortable for a human user. Fig. 3, bottom center, shows the generated views.

At runtime, the operator is shown a suggested acquisition path from one view to the next (Fig. 3, bottom right). This is computed by starting at the closest view to the operator, then selecting the immediate nearest neighbor view repeatedly in a greedy approach until all currently-unacquired views are in the path. The path is recomputed each time the user acquired a new view.

6 View Acquisition with AR Guidance

Once the set of suggested acquisition views has been generated, they are rendered on the AR HMD as floating icons to indicate both the position and orientation of the views (Fig. 1, top middle). We also calculate and present to the operator a suggested acquisition path, which starts at the nearest view to the user's current position and connects each subsequent closest neighbor view in a greedy algorithm. At this point, the operator can begin photogrammetric acquisition of the scene. The goal for the operator is to physically position the camera rig such that the smartphone's camera matches the position and orientation of the suggested view, and then to capture an image of the scene from the view.

Our AR HMD interface provides interactive adaptive guidance to help the operator more precisely guide the camera rig into the proper location, as well as an automatic photo capture feature to ensure that acquired images are free from motion blur. When the camera on the handheld rig is placed near an acquisition view, the view's icon (Fig. 4, a) expands into a world-space-aligned yellow frame (Fig. 4, b). The operator aligns the camera rig such that the rectangular frame of the smartphone is aligned with the suggested view's frame. As the camera rig is aligned, the frame changes color gradually from yellow to green (Fig. 4, c). Once the camera rig's position and orientation are both within a desired threshold (in our experiments we set the position threshold to 5cm and the rotation threshold to 15°), the AR HMD indicates to the smartphone that a photo should be taken. The smartphone tracks its own acceleration using onboard accelerometers, and once the phone has been held stably in place at the suggested view, a photo is automatically captured and the view is removed from the AR visualization (Fig. 4, d). We define the smartphone to be held stably in place if its linear acceleration remains below $1 m/s^2$ for 300ms.

The operator repeats this process view by view, until all desired views have been captured by the handheld camera rig. The output of the acquisition is a set of RGB images that achieve coverage of the scene subject to the input parameters of camera-to-scene distance and image overlap. The set of acquired images can then be processed offline by conventional structure-from-motion software for photogrammetric 3D reconstruction. We additionally save at the time of each photo capture the estimated poses of the camera rig in the AR HMD's coordinate system, and we use these initial poses as input during camera alignment.

7 Results and Discussion

In this section, we provide implementation details of our prototype AR HMD guided acquisition system, we present results of several acquired scenes, and we detail a user study conducted to validate our approach.

7.1 Implementation overview

We implemented a prototype system for our AR-HMD-guided photogrammetric acquisition method. The AR HMD we used was the first version of the Microsoft HoloLens, and we used a Google Pixel 3 smartphone (resolution: 4032 x 3024, FOV: 65deg x 49deg) in our handheld camera rig [12,20]. The handheld camera rig was mounted with ArUco fiducial markers and was tracked by the HoloLens using the HoloLensARToolkit library [6, 11, 25]. The marker was tracked by the AR HMD at 30fps, which is also the frame rate of the HoloLens' onboard RGB camera. Wireless communication between the AR HMD and the camera is achieved with a socket connection and using the smartphone as a Wi-Fi hotspot, which makes the system completely self-contained and portable even in austere environments without any Internet connection. Our application runs on the AR HMD at 60fps. Given an input region of interest about 2.5m to a side (represented in the AR HMD rough geometry with about 15000 triangles) our system generates a set of suggested views in about 13 seconds, as computed locally on the HoloLens.

Table 1: Summary of acquired scenes.

Scene (Figs.)	Distance d (m)	Overlap ω	Time (sec)	Num. images
StackedRocks (1, bottom row)	0.5	0.50	430	93
Pentagon (5)	0.5	0.50	510	146
Turbine (1, middle row)	0.5	0.66	667	170

7.2 Reconstruction results

We captured sets of images from several scenes using our AR HMD guidance method. Table 1 summarizes each scene's input parameters, number of pictures taken, and acquisition time. As can be seen, high quality reconstructions can be achieved by following the guidance provided by our AR HMD approach in just a matter of minutes.

7.3 User study

To validate our approach and to gain formative feedback on our user interface, we conducted a user study in which ten participants each acquired imagery of a medium-sized complex scene (2.0m x 1.6m x 1.7m) under both unguided (Control) and AR HMD guided (Experimental) conditions.

Participants: 10 participants (8 male, 2 female; age: 28.9 ± 4.4) were recruited. In a pre-session 5-point Likert scale questionnaire, participants self-reported their prior experience levels in various skills related to the task. Participants reported some knowledge of augmented reality (2.6 ± 1.1), head-mounted displays (2.9 ± 1.4), and 3D reconstruction (2.4 ± 1.1), but were generally unfamiliar with photogrammetry itself (1.5 ± 0.7).

Task: Participants were tasked with acquiring images of a target scene with the goal of achieving a complete 3D reconstruction. Fig. 6 shows the target scene set up for participants to acquire. The scene is feature-rich but is geometrically complex, with challenges such as a thin barrier bisecting the scene that could result in reconstruction failure if the edges are insufficiently acquired.

Prior to acquisition, participants were given a short tutorial about the principles of photogrammetry: how images are matched, and the importance of consistent distance to the scene, image overlap, and scene coverage. For each condition, participants were asked to acquire imagery while keeping approximately 0.5m from the scene (camera-to-scene distance) and with approximately 50% overlap. Participants were not told the specific number of images they should take to achieve a good coverage of the scene, as this would require prior knowledge specific to the scene rather than general knowledge of photogrammetric principles.

Conditions: Participants were randomly placed into one of two groups in a 2x2 counterbalanced design. All participants acquired the scene twice: one group acquired the scene first under the Control condition and second under the Experimental condition; the other group acquired imagery with the Experimental condition first and then the Control condition.

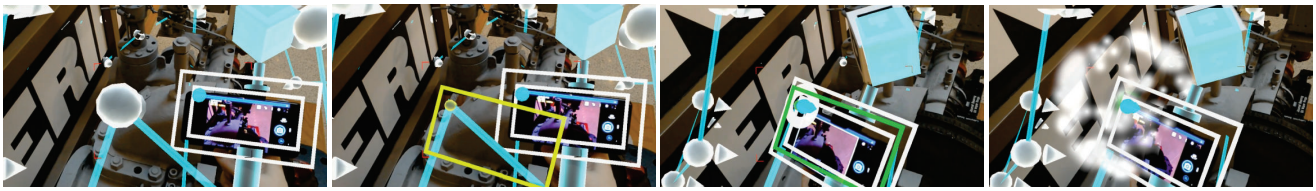

Figure 4: AR interactive guidance to place camera at suggested view. (a): before approaching view. (b): as camera approaches, view icon transforms (yellow rectangle). (c): camera in range with visual feedback (green rectangle and loading circle). (d): feedback immediately after automatic capture (view disappears with animation).

Figure 5: An example of our AR HMD guided acquisition functioning in an outdoor environment. Top: Suggested views and acquisition path. Bottom: High quality photogrammetric reconstruction.

Figure 6: The scene captured by user study participants.

For the Control condition, participants used a handheld camera rig with a smartphone, but without wearing an AR HMD or having a fiducial marker on the camera rig. A Bluetooth-connected shutter release button on the rig allowed the user to take pictures manually. Participants were told to take pictures of the scene until they felt confident that they had acquired enough for a good reconstruction.

For the Experimental condition, participants wore the AR HMD and a short tutorial (5-10 minutes) was given in the functionality of the AR HMD guidance system, during which participants were allowed to practice placement of the camera rig given a test set of acquisition poses to ensure they were comfortable with the automatic capture functionality. Approximately 150 acquisition locations were visualized by the system to the participant, and participants were asked to capture every view.

Metrics: After each acquisition session (Control or Experimental) was complete, participants filled out a questionnaire that included a NASA Task Load Index (NASA-TLX) workload assessment [13]. The NASA-TLX questionnaire contains 6 metrics on a 21-point scale: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants also answered an additional questionnaire, which asked their level of agreement with a series of seven statements on a 5-point Likert scale:

1. Using this approach was enjoyable
2. Using this approach was comfortable
3. I feel confident that the images I took will make a good 3D reconstruction
4. The method helped me learn how to do 3D capture
5. After using this approach, I am interested in doing 3D capture
6. It was easy for me to know which areas I should capture
7. It was easy for me to remember which areas I had already captured

Additionally, the number of captured images in each session and the scan session time were recorded.

The acquired images for each session were input into a conventional photogrammetric reconstruction software (Agisoft Metashape [2]). In the case of images acquired during the Experimental condition, the estimated poses of the camera rig at each capture timestamp (according to the AR HMD's coordinate system) were input to initialize cameras during SfM reconstruction. For all sets of images,

Table 2: Summary of results for our user study.

Metric	**Control**	**Experimental**	*p*
Images captured	91 ± 59.5	162.6 ± 15.5	0.005
Time (sec)	358.5 ± 178.4	1033.2 ± 310.3	0.0004
Time per image (sec)	4.5 ± 1.3	6.3 ± 1.5	0.003
% points reconstructed within 1cm	33.7 ± 20.3	64.6 ± 7.2	0.0011
% points reconstructed within 3cm	46.6 ± 25.3	84.0 ± 4.1	0.0009
% points reconstructed within 5cm	60.1 ± 26.4	95.9 ± 2.0	0.0021

reconstruction was completed using identical settings in the software and without manual cleanup of extraneous points in between stages.

To quantify the completeness of the participants' models, we separately acquired a "ground truth" model of the scene reconstructed from a highly dense set of 600 photos (over three times as many photos as captured by any participant). Each reconstructed model was manually aligned into the same coordinate system as the ground truth model. 10,000 sample points were uniformly selected from the surface of the ground truth model, and the distance between each ground truth sample point and the closest point on the reconstructed model was found. We computed the percentage of points that were within 1cm, 3cm, and 5cm of the ground truth model.

We also analyze whether or not there is a significantly different amount of blurriness in the images acquired in the Control or in the Experimental conditions. For each set of acquired images, we compute the frequency domain based image quality metric of De and Masilamani for each image [8].

7.3.1 User study results

All scan session results were considered to be in one of two populations, Control and Experimental; the independent variable was the use of either a conventional smartphone for acquisition or our AR HMD guidance system for acquisition. A paired two-tailed T-test was performed on our dependent variables. Dependent variables are: acquisition time, number of images captured, reconstruction quality as measured by percent of ground truth points in reconstruction, and the responses to our post-session questionnaires.

Table 2 summarizes the results of our metrics for the user study. Participants using our system captured significantly more images when using our approach. However, the acquisition time for the Experimental condition was far longer than for the Control condition, measured both in total time and in seconds per image captured. One cause may be the automatic capture feature of our system; several participants described issues in precisely aligning the camera rig with the AR guidance in order to trigger an automatic capture. In particular, the depth cues of aligning the camera frame with a rectangle signifying the suggested view were not very strong. While alignment in X and Y (left/right/up/down relative to the operator) was easier to achieve, alignment in Z (towards or away from the operator) was not clear, leading to participants holding out the camera rig and waiting for automatic capture while still being out of range. In future work we plan to investigate improved interfaces for intuitive 6-DOF alignment that address this ambiguity in the Z direction.

Examples of the models generated by participants can be seen in Fig. 7. The reconstructions under the Experimental condition (Fig. 7, right column) tend to be far more complete than those captured under the Control condition (Fig. 7, left column). Fig. 8 shows, for each reconstructed model, the results of our quantitative geometric analysis. The Experimental condition showed a statistically significant improvement in the percentage of the ground truth model that was reconstructed to within 1cm, 3cm, and 5cm. While a few

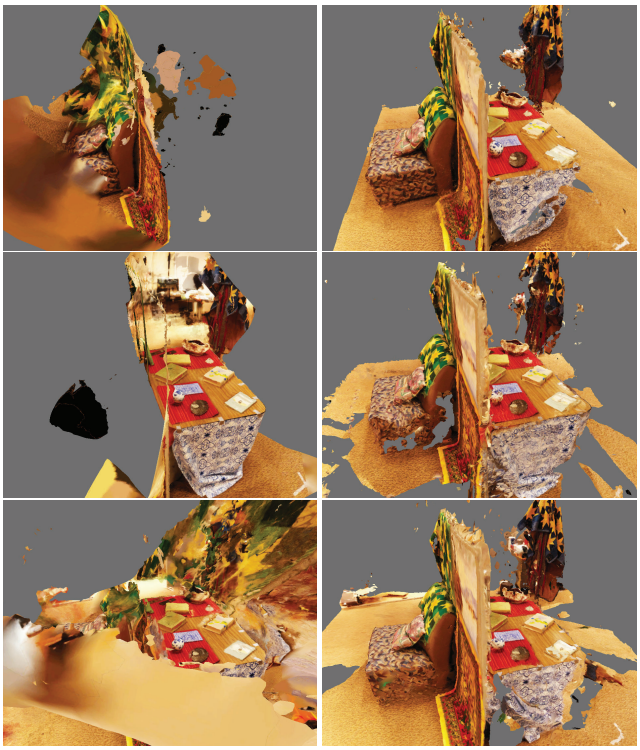

Figure 7: Example scene reconstructions generated from datasets acquired by study participants in both Control (left column) and Experimental (right column) conditions. Each row corresponds to an individual participant.

participants achieved a complete reconstruction under the Control condition, it was much more likely that participants without our AR HMD guidance would acquire images that led to only a partial (or highly distorted) reconstruction. Only 3 out of 10 reconstructed models from the Control condition achieved over 50% of the ground truth scene reconstructed to within 1cm, while all 10 of the 10 reconstructed models in the Experimental condition did. One cause of the difference is the wall divider that bisects the target scene; participants in the Control condition would adequately capture each individual side but would neglect the transition from one side to another, which is needed by the reconstruction software to automatically register the scene into a single frame of reference. The Experimental condition's ability to use the camera rig's estimated position as an initial guess in the reconstruction software also provides a great advantage.

Participants' responses to the post-session NASA-TLX questionnaire are detailed in Fig. 9. The Experimental condition was found to significantly increase the amount of Physical Demand (TLX-2) on the participant. We attribute this largely to the weight and discomfort of the AR HMD and the fatigue from holding the camera rig to precisely align with the suggested view. The NASA-TLX metric of Performance (TLX-4), where a lower score indicates a higher self-appraisal of success in accomplishing the task, showed significant improvement. No other values in the NASA-TLX questionnaire (Mental Demand, Temporal Demand, Effort, Frustration) were found to have statistically significant differences.

Fig. 10 shows the results of our 5-point Likert scale post-session questionnaire. While participants reported significantly greater discomfort (Q2) using the Experimental approach, participants found the Experimental approach significantly more enjoyable (Q1) than the Control approach. While the mere novelty of using new technology may be a contributing factor, we hypothesize that our AR guidance acts as an example of gamification for what would typically

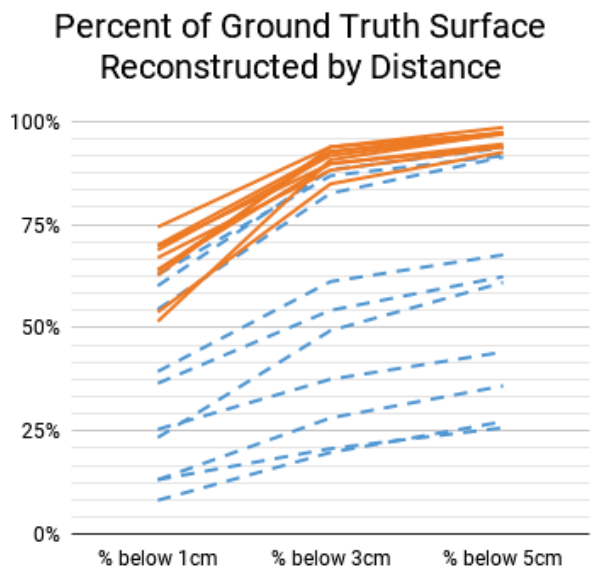

Figure 8: Result of comparison between reconstructed models generated from images acquired by user study participants, using threshold distance between ground truth model of target scene and nearest point on participant's reconstruction. Dashed blue lines: control condition. Solid orange lines: experimental condition.

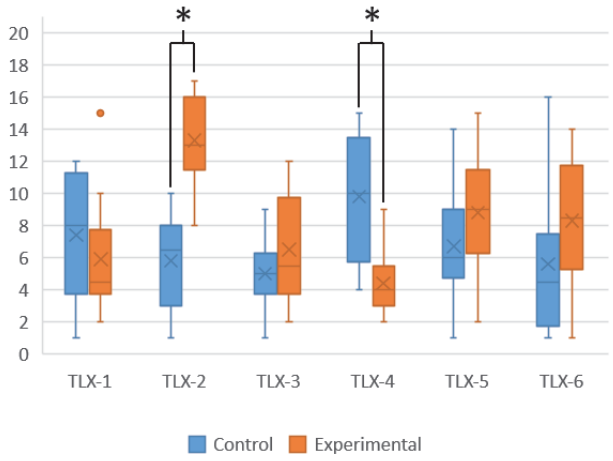

Figure 9: Box-and-whisker plot for each of the six NASA-TLX subscale scores, for both Control and Experimental conditions. All scores are on a 21-point scale. A star (*) indicates that a statistically significant ($p < 0.05$) difference was found between the two conditions.

be a tedious task for an operator. Participants using the Experimental also reported significantly increased ease at knowing which areas should be captured (Q6) and which areas had already been captured (Q7), suggesting that our approach helps offload the cognitively demanding task of maintaining a mental map of the scene. Our approach also seems to help participants learn how to do 3D capture (Q4) significantly more than the Control condition, making AR assistance useful for training purposes. No statistically significant difference ($p = 0.11$) was found between conditions in the level of interest of participants in doing 3D capture.

For our analysis of the blurriness of acquired images in Control and Experimental conditions, we computed a paired two-sided T-test, comparing the mean image quality between the population of Control sessions and the population of Experimental sessions. Using the aforementioned image quality metric of De and Masilamani, we found no statistically significant difference in image quality between Control (0.001287 ± 0.00017) and Experimental (0.001290 ± 0.00012) conditions ($P = 0.95$). We conclude that our feature of automatic image capture sufficiently avoids motion blur.

Additional findings: Several participants mentioned that when using the AR HMD based guidance, they did not pay any attention to the camera view on the smartphone screen; instead, they only focused on the AR overlay rendered by the headset. While further research is needed, we believe that this suggests a decreased cognitive load, as matching a pre-defined 6-DOF pose requires less mental analysis than the higher-dimensional analysis of evaluating a detailed on-screen image for overlap of salient features.

The FOV of the AR HMD's onboard camera is greater than the display's FOV, which means that the fiducial marker does not need to be within the display's FOV to be tracked. However, several participants found themselves attempting to keep the entire camera rig within the display's FOV, leading to stretching the arm straight out and resulting in greater fatigue and slowness during acquisition. It is possible that future AR HMDs with larger FOVs will mitigate this issue; however, a more limited AR overlay that is local only to the region immediately surrounding the rig's camera may also solve this issue.

7.4 Limitations

Our approach relies on an environment in which the AR HMD can (1) track itself relative to the environment, (2) generate a reasonable rough geometric model of the scene, and (3) display content clearly to the operator. Poorly-lit scenes, moving objects, or reflective materials can disrupt the AR HMD's tracking. However, our approach does work robustly in outdoor environments provided that the scene is in shadow or the weather is overcast, as evidenced in Fig. 5; the display of the AR HMD is faint but still usable. Such overcast weather is in fact preferred for outdoor photogrammetry as the ambient light prevents strong shadows from being baked into the model.

Our method of view generation is entirely geometry-based in that we do not analyze image features at runtime, and implicitly assume that all parts of the scene are equally feature-rich. Future work could analyze frames from AR HMD's camera at runtime to identify feature-rich or feature-poor areas and adapt the view generation accordingly.

Our acquisition path between suggested views is computed by only considering Euclidean 3D distance between views. However, this can lead to inefficient movement because physically stepping around the scene is more laborious than standing still while moving the camera rig. We hypothesize that an acquisition path that takes into account the human factors of movement would reduce acquisition time and fatigue.

8 Conclusions and Future Work

In this paper, we have presented a novel HMD-based approach for AR guidance during 3D acquisition that helps ensure sufficient images are acquired for reconstruction coverage and quality. We have demonstrated several quality reconstructions generated from our acquisition approach, as well as a user study that reveals that novice participants using our method can achieve more complete reconstructions.

We are interested in the potential for expanding our guided acquisition to a multi-user approach, where multiple operators can simultaneously acquire a large environment in parallel [24]. By shar-

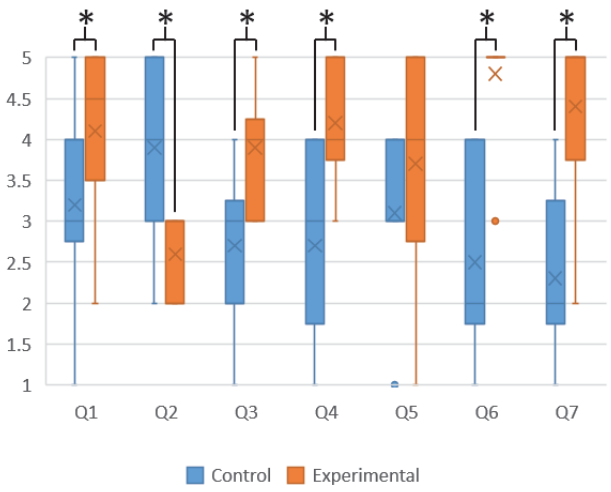

Figure 10: Box-and-whisker plot for the seven questions in the post-session questionnaire. All scores are on a 5-point Likert scale. A star (*) indicates that a statistically significant ($p < 0.05$) difference was found between the two conditions.

ing a common coordinate system and tracking the position of other users, the workload of acquisition could be greatly parallelized.

Additionally, our work has revealed an important future research direction of AR interfaces that guide users to precisely assume a handheld 6-DOF pose in open space, as opposed to the simpler problem of annotating a surface location on a physical object. We hope our work both justifies and encourages future research into such interface design.

Combining the flexibility of a handheld acquisition device with the world-aligned visualization and tracking of an AR HMD is a pairing that achieves high quality results and has direct, practical application in the use case of photogrammetric acquisition. As AR interfaces become integrated into day-to-day life we anticipate that such multimodal combinations of devices will become increasingly beneficial.

References

[1] A. Adán, B. Quintana, and S. A. Prieto. Autonomous mobile scanning systems for the digitization of buildings: A review. *Remote Sensing*, 11(3):306, 2019.

[2] Agisoft. Agisoft Metashape, 2019.

[3] J. Ahn and K. Wohn. Interactive scan planning for heritage recording. *Multimedia Tools and Applications*, 75(7):3655–3675, 2016.

[4] D. Andersen and V. Popescu. An AR-guided system for fast image-based modeling of indoor scenes. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 501–502. IEEE, 2018.

[5] D. Andersen and V. Popescu. HMD-guided image-based modeling and rendering of indoor scenes. In *International Conference on Virtual Reality and Augmented Reality*, pp. 73–93. Springer, 2018.

[6] E. Azimi, L. Qian, N. Navab, and P. Kazanzides. Alignment of the virtual scene to the 3D display space of a mixed reality head-mounted display. *arXiv preprint arXiv:1703.05834*, 2018.

[7] M. Chen, E. Koc, Z. Shi, and L. Soibelman. Proactive 2D model-based scan planning for existing buildings. *Automation in Construction*, 93:165–177, 2018.

[8] K. De and V. Masilamani. Image sharpness measure for blurred images in frequency domain. *Procedia Engineering*, 64:149–158, 2013.

[9] S. Dong and T. Höllerer. Real-time re-textured geometry modeling using Microsoft HoloLens. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 231–237. IEEE, 2018.

[10] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 209–216. ACM Press/Addison-Wesley Publishing Co., 1997.

[11] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[12] Google. Google Pixel 3, 2018.

[13] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[14] C. Hoppe, M. Klopschitz, M. Rumpler, A. Wendel, S. Kluckner, H. Bischof, and G. Reitmayr. Online feedback for structure-from-motion image acquisition. In *BMVC*, vol. 2, p. 6, 2012.

[15] R. Huang, D. Zou, R. Vaughan, and P. Tan. Active image-based modeling with a toy drone. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2018.

[16] F. Langguth and M. Goesele. Guided capturing of multi-view stereo datasets. In *Eurographics (Short Papers)*, pp. 93–96, 2013.

[17] L. Liu, X. Xia, H. Sun, Q. Shen, J. Xu, B. Chen, H. Huang, and K. Xu. Object-aware guidance for autonomous scene reconstruction. *arXiv preprint arXiv:1805.07794*, 2018.

[18] A. Locher, M. Perdoch, H. Riemenschneider, and L. Van Gool. Mobile phone and clouda dream team for 3D reconstruction. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pp. 1–8. IEEE, 2016.

[19] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *ACM siggraph computer graphics*, vol. 21, pp. 163–169. ACM, 1987.

[20] Microsoft. Microsoft HoloLens, 2016.

[21] Q. Pan, G. Reitmayr, and T. W. Drummond. Interactive model reconstruction with user guidance. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pp. 209–210. IEEE, 2009.

[22] C. Peng and V. Isler. View selection with geometric uncertainty modeling. *arXiv preprint arXiv:1704.00085*, 2017.

[23] C. Peng and V. Isler. Adaptive view planning for aerial 3D reconstruction of complex scenes. *arXiv preprint arXiv:1805.00506*, 2018.

[24] F. Poiesi, A. Locher, P. Chippendale, E. Nocerino, F. Remondino, and L. Van Gool. Cloud-based collaborative 3D reconstruction using smartphones. In *Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017)*, p. 1. ACM, 2017.

[25] L. Qian, E. Azimi, P. Kazanzides, and N. Navab. Comprehensive tracker based display calibration for holographic optical see-through head-mounted display. *arXiv preprint arXiv:1703.05834*, 2017.

[26] M. Roberts, D. Dey, A. Truong, S. Sinha, S. Shah, A. Kapoor, P. Hanrahan, and N. Joshi. Submodular trajectory optimization for aerial 3D scanning. In *International Conference on Computer Vision*, 2017.

[27] N. Smith, N. Moehrle, M. Goesele, and W. Heidrich. Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark. In *SIGGRAPH Asia 2018 Technical Papers*, p. 183. ACM, 2018.

[28] E. Wakisaka, S. Kanai, and H. Date. Model-based next-best-view planning of terrestrial laser scanner for HVAC facility renovation. *Computer-Aided Design and Applications*, 15(3):353–366, 2018.