HMD-Guided Image-Based Modeling and Rendering of Indoor Scenes

Daniel Andersen[1[0000−0002−5640−1845]] and Voicu Popescu[1[0000−0002−8767−8724]]

Purdue University, West Lafayette IN 47907, USA
{andersed,popescu}@purdue.edu

Abstract. We present a system that enables a novice user to acquire a large indoor scene in minutes as a collection of images sufficient for five degrees-of-freedom virtual navigation by image morphing. The user walks through the scene wearing an augmented reality head-mounted display (AR HMD) enhanced with a panoramic video camera. The AR HMD shows a 2D grid of a dynamically generated floor plan, which guides the user to acquire a panorama from each grid cell. After acquisition, panoramas are preliminarily registered using the AR HMD tracking data, corresponding features are detected in pairs of neighboring panoramas, and the correspondences are used to refine panorama registration. The registered panoramas and their correspondences support rendering the scene interactively with any view direction and from any viewpoint on the acquisition plane. An HMD VR interface guides the user who optimizes visualization fidelity interactively, by aligning the viewpoint with one of the hundreds of acquisition locations evenly sampling the floor plane.

Keywords: Augmented reality · 3D acquisition · image-based rendering.

1 Introduction

Applications such as virtual tourism, real estate advertisement, or cultural heritage preservation require rendering real world scenes convincingly at interactive rates. However, efficient photorealistic acquisition of real world scenes is a challenging problem. Traditional texture mapped geometric models are difficult to acquire to a level of completeness necessary for high-fidelity rendering. The alternative approach of image-based modeling and rendering (IBMR) has been proposed over twenty years ago. The scene is captured with a database of rays, which is queried at run time to show the scene from the desired view. Assembling the output image is fast, and good results are obtained as long as the image-based model covers densely the entire viewing volume.

However, efficient image-based modeling of a large indoor space remains an open problem. Practical image-based modeling approaches acquire 2D ray databases, i.e., panoramas, which confine the user to the acquisition location. Image-based modeling that enables virtual scene navigation with more degrees

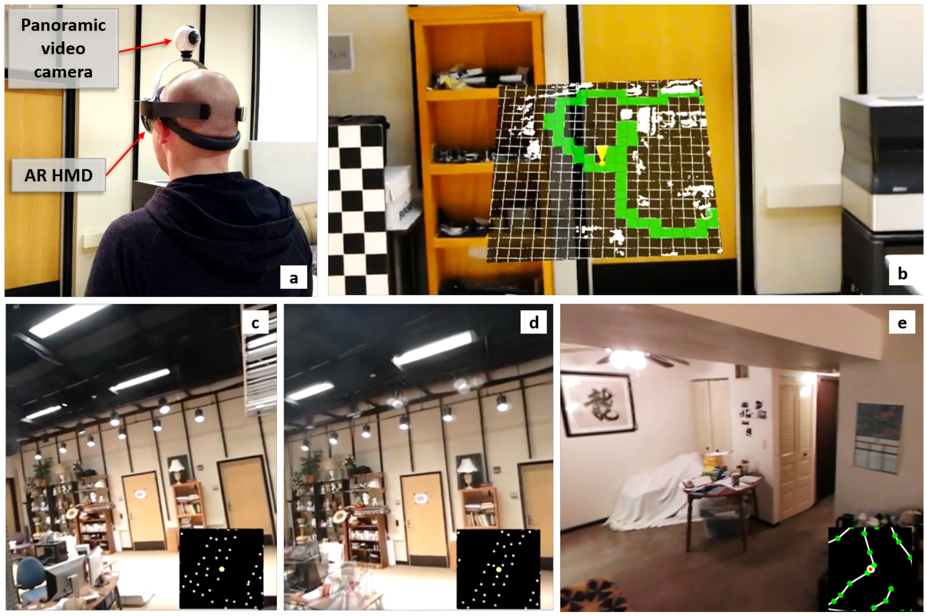

Fig. 1. AR-guided acquisition (a), acquisition map visualization (b), guided VR visualization frame from viewpoint at acquisition location (c) at barycenter of acquisition location triplet (d), and at midpoint of acquisition path segment (e).

of freedom have the disadvantages of expensive acquisition devices, of long acquisition times, and of reliance on operator expertise. Another challenge is the lack of immediate feedback during acquisition, which makes it difficult for the user to capture the scene reliably from all necessary viewpoints. Returning to the scene long after initial acquisition to acquire additional viewpoints is impractical.

After acquisition, the panoramas are processed offline with a pipeline that preregisters the panoramas using the AR HMD tracking data, triangulates acquisition locations, refines panorama registration using corresponding features in panorama through a RANSAC approach, enriches the set of correspondences using the scene geometry proxy acquired by the AR HMD, and builds panorama 3D morphing meshes by triangulating correspondences.

The resulting 3D morphing meshes supports interactive five degree of freedom (5-DOF) visualization through a virtual reality (VR) HMD, with correct depth perception. The desired image is rendered by morphing and blending the three panoramas that define the acquisition location triangle that contains the viewpoint. We do not focus on the difficult task of reconstructing a high-fidelity 6-DOF 3D mesh; instead, we recognize that most reasonable views of the scene will be on the 2D plane near head height and five degrees of freedom sufficiently describes the scene.

Like any image-based model, our model caters to the set of all possible output views with non-uniform fidelity. We allow the user to take advantage of the highest fidelity provided by our model by displaying through the VR HMD a visualization map with the current viewpoint position and the nearby panorama acquisition locations. Therefore, the VR HMD does not only show the scene to the user, but also guides the user who can optimize visualization quality interactively by easily aligning their viewpoint with the acquisition viewpoints.

The frame in Fig. 1c has a viewpoint that is near one of the acquisition locations, where the morph converges to the identity function, and artifact-free frames are rendered by resampling the acquisition panorama. The frame in Fig. 1d has a viewpoint (yellow dot at visualization map center) that is between its neighboring acquisition locations (white dots on map). Parts of the scene rich with correspondences, such as the book shelves, doors, walls, and floors, are visualized with high quality. Parts of the scene where correspondences are sparse, such as the nearby geometry in the bottom left corner of the frame and the near row of ceiling spotlights, exhibit ghosting artifacts. The visualization map can also show the acquisition path (gray line segments in Fig. 1e). When the user translates the viewpoint along a segment of the acquisition path, a high-fidelity visualization is provided by morphing between two consecutive panoramas saved with higher density along the acquisition path (green circles in Fig. 1e). Leveraging the high density of the acquisition and the visualization of the acquisition locations, the user achieves a high-quality interactive visualization of the scene, with brief transitions between acquisition locations, and with photorealistic pan-tilt sequences from viewpoints aligned with acquisition locations.

In summary, our paper contributes a complete image based modeling and rendering system, with an AR HMD interface that guides a novice user to achieve a complete inside-looking-out acquisition of a large indoor space in minutes, and with a VR HMD interface that guides the user to optimize visualization fidelity.

2 Prior Work

We first discuss prior image-based modeling and rendering techniques relevant to our approach, and then we discuss prior work on guided scene acquisition.

2.1 Image-Based Modeling and Rendering

Aliaga and Carlbom presented a plenoptic stitching method that acquired image based models of indoor environments by moving an omnidirectional camera through a room in a regular pattern [2]. Synthetic views were generated by interpolating images captured from surrounding views. However, their approach did not provide interactive guidance on how to efficiently traverse the interior of the environment, so pre-planning was required. The visualization method also required a virtual camera location to be surrounded by a closed loop of acquired images, which leads to tedious acquisition paths when the environment is cluttered. In contrast, our approach gives guidance during acquisition, and the user only needs to visit each acquisition location a single time.

Bradley et al. presented a system for virtual navigation through a real-world scene by switching the view between densely sampled panoramas along a series of corridors [3]. Zhang and Zhu similarly built virtual tours from spherical panoramas acquired at regular intervals on a tabletop [30]. The sampling was dense but lacked morphing, so the user perceives discontinuities during translation. Such acquisition was also limited to a pre-defined path planned by consulting an existing map of the scene.

Some prior work examines the question of interpolation between panoramas for image-based navigation. Chiang et al. presented a method for image-based interpolation between cylindrical panoramas, but the method required manual input for determining adjacency between neighboring images [4]. Several methods achieve panorama interpolation, but only in one dimension along the path of acquisition [17, 14, 31], which results in a 4-DOF visualization. Kawai et al. extended such works to support navigation in two dimensions by bilinear interpolation of panoramas, but without automatically finding correspondences between panoramas [16]. Kawai later reformulated the problem as a sparse light field supporting transitions between panoramas but at the cost of highly noticeable artifacts [15]. Xiao and Shah's tri-view morphing method allows for novel viewpoints of a scene by grouping triples of neighboring cameras; however, it only makes use of conventional images and not panoramas, and correspondences must be enriched manually [28].

Shi presented a method for interpolation of cubemap panoramas for image-based navigation [25]. However, without a method to guide the user to capture such imagery at a minimum density, there is no guarantee of coverage or of minimum quality as a user navigates through the scene. Davis et al.'s work on unstructured light fields offers some visual feedback during image-based acquisition of a target object, but their work focuses on outside-looking-in object acquisition, while we focus on inside-looking-out scene acquisition [6].

RGB-D depth maps have been used to achieve visually impressive results for scene capture and reconstruction. Hedman et al. presented a recent work for high-quality image-based modeling and rendering of indoor scenes by combining RGB color images from traditional cameras with RGB-D depth-enhanced images [11]. Dai et al. created a method for globally consistent 3D reconstruction using a hand-held depth sensor [5]. Compared to their work, our models are much farther to the image end of the geometry-image continuum, with the benefits of a simpler acquisition device and of a simpler acquisition procedure.

Recent work by Huang et al. uses panoramic video camera footage to create 6-DOF VR videos, where a user can view the captured environment with depth cues and head orientation and translation [13]. However, this work targets a single fixed viewing location with some ability to move the head within a small viewing volume, whereas we capture floor spaces of hundreds of square meters. This prior work focuses on leveraging existing video footage to build an image-based model, as opposed to our goal of guidance during acquisition to capture the best set of images.

2.2 Guided Scene Acquisition

Several prior acquisition systems attempt to guide the user to acquire imagery of scenes from as-yet-uncaptured viewpoints. Tuite et al. illustrated sparsely-sampled regions in a 3D reconstruction of a building facade as markers on a smartphone map, prompting users to take pictures from the necessary viewpoints [26, 27]. However, their approach targets reconstruction of outdoor building facades, while our approach focuses on acquiring indoor environments. The indoor environments we target lack precise GPS tracking, which requires more active tracking such as in an AR HMD. Also, because many indoor environments are more likely to rapidly change appearance than outdoor buildings, the prior work's emphasis on multi-user capture over long periods of time is less suitable. Instead, we focus on providing guidance for a single user to rapidly capture an indoor environment, all in a single scanning session.

Rusinkiewicz et al. introduced an interactive method for capturing 3D models of hand-held objects while showing the in-progress model to the user [23]. Such approaches provide implicit guidance from a single viewpoint, but further manipulation of the object is required to uncover missing regions of the model. The equivalent action in our use case (acquiring large indoor environments) would be to physically traverse the environment, and so we provide additional guidance in the form of a top-down map that reduces the redundant physical traversal needed by the user. Diverdi et al. presented a method for interactively constructing an environment map; however, the output of a single environment map only provides a rough approximation of scene geometry [7]. Ahn et al. created a method to plan the placement of 3D scanning equipment in the context of digital heritage [1]. Given a top-down map and user-selected regions of interest, they automatically determined locations to place a 3D scanner to achieve a high-quality scan with sufficient coverage; however, it is not suited for casual scanning or acquisition of areas without a prior map or manual direction. Pan et al. presented an AR interface for acquiring texture imagery of a hand-held object by indicating rotations for the user to perform to reveal unscanned areas to a camera [21]. However, this particular interface is suitable only for small manipulable objects, rather than inside-out capture of a room. Pankratz and Klinker used a video pass-through AR HMD to visualize room-scale marker calibration during iterative refinement, but did not focus on capturing scenes [22].

2.3 Additional Prior Work

Recently, guided acquisition of room-sized scenes has been explored in fully-autonomous contexts, where a robot utilizes next-best-view (NBV) analysis of scene geometry to determine efficient trajectories for geometric capture [9, 29]. In contrast, our work focuses on providing guidance to a human user in a casual context where robotic acquisition is infeasible, such as in cluttered environments that are difficult for robots to navigate but easy for humans to walk through.

A recent work presented a method for acquiring textured 3D models of indoor scenes using an AR HMD [8]. However, the purpose of that work was to create a

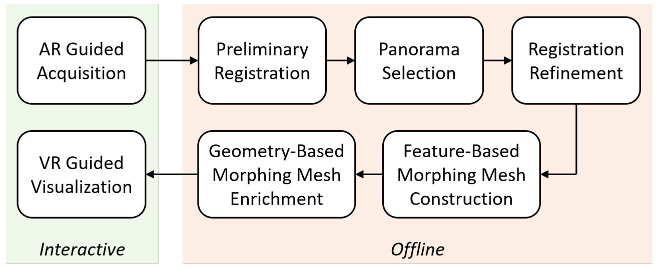

Fig. 2. System pipeline overview.

fixed-memory texture atlas during acquisition to color the AR HMD's on-board 3D geometry capture. The resulting mesh typically contains holes that cannot be filled in after acquisition due to discarding of unused color data. Guidance is not provided to the user during acquisition.

3 System overview

Fig. 2 gives an overview of the architecture of our system. The user acquires scene panoramas with interactive AR guidance (Section 4), the panoramas are prepared offline for morphing (Section 5), and the panoramas are then morphed in a guided VR interactive visualization of the scene (Section 6).

4 AR Guided Acquisition

The goal of the acquisition stage is to capture a complete and dense set of scene images as quickly as possible, without the prerequisite of user expertise. We achieve this goal with an acquisition device that not only captures images of the scene, but also guides the user for efficient and reliable acquisition.

4.1 Acquisition Device

Our acquisition device consists of an AR HMD enhanced with a panoramic video camera (Fig. 1a). The panoramic camera's pose is calibrated with respect to the AR HMD. Our approach relies on and assumes that the AR HMD provides inside-out tracking and some generation of rough geometric data. As the user walks through the scene, the AR HMD tracks the user's position and orientation, it builds a map of the scene, and it overlays onto the user's field of view the locations from where the scene is yet to be acquired.

The AR HMD does have a built in video camera, which, in principle, could be used to acquire the images needed to build the image-based model of the scene. However, the AR HMD camera has a small field of view, which makes it unsuitable for our purpose. To support free view rotation during the interactive visualization, the scene must be acquired in all view directions from each acquisition location. Covering all view directions with a small field of view camera leads to long acquisition times and blurry images from head rotation. Also, when the user pans and tilts their head to cover all view directions from a given acquisition location, it is difficult to enforce a single viewpoint constraint, which reduces the quality of the interactive visualization that has to cover the residual translation by morphing.

To improve acquisition efficiency, we capture the scene with a 360° panoramic video camera that is rigidly attached to the AR HMD. Each frame is a complete spherical pinhole panorama, which can be trivially resampled to a high-quality output image when the desired viewpoint matches an acquisition location. The resulting image-based model captures the scene with very high fidelity from the hundreds of viewpoints from where panoramas are acquired. The camera records continuously during acquisition as a panoramic video stored to the camera's flash memory. After acquisition, the panoramic video is processed offline into an image-based model of the scene as described in Section 5.

4.2 AR Interface

Our system relies on an AR interface to guide the user towards a fast, dense, and complete acquisition. To support 5-DOF interactive virtual navigation, i.e., two translations and three rotations, the user acquires panoramas on a horizontal plane at the user's head height, which will also be the height from where the scene is rendered during visualization. Acquisition density and coverage is controlled by partitioning the acquisition plane with a uniform 2D grid (e.g., with 0.5m x 0.5m cells for Fig. 1b).

The 2D grid is shown as a 2D map floating in front of the user (Fig. 1b). The map rotates as the user changes direction to maintain an intuitive user perspective orientation. The map shows the parts of the grid that are yet to be discovered (empty dark cells), the parts that are inaccessible due to floor obstacles such as furniture (white), the parts that have already been traversed during acquisition (green), as well as the user's current position and orientation (yellow, at the center of the map). The floor obstacles are computed from the coarse geometric scene model acquired by the AR HMD through active depth sensing. The scene does not have to be acquired from inaccessible cells since during a typical virtual navigation, the user does not want, and is prevented from, assuming inaccessible positions (e.g., inside a book shelf or above a desk). Acquisition is complete once all accessible grid cells are traversed (Fig. 3).

As we were designing the interface, we found that relying on a first person visualization is inefficient. We initially rendered markers in the scene to illustrate target locations. This was difficult for the user to align their head with the marker, because the marker grows larger as the user approaches it. We had also tried rendering the 2D grid over the scene floor; acquisition suffered from excessive downwards head tilting needed to consult the grid visualization. In both cases, the AR HMD's low field of view meant that indicators anchored to world locations were often hidden to the user and thus could not communicate guidance. Instead we found that adding an additional perspective in the form

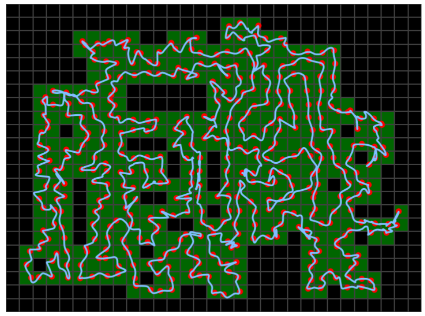

Fig. 3. Acquisition result. The acquisition path (blue line) has traversed all accessible grid cells (green). A panorama is selected for each grid cell (red dots).

of a virtual map was most suitable for our use case and has proven to be an efficient way of guiding the user, who can intuitively turn left and right to guide the yellow dot through the grid cells that are yet to be traversed.

5 Panorama Morphing Setup

In this section, we describe our offline processing of the acquired data to transform it into a set of 3D morphing meshes suitable for interactive visualization.

5.1 Preliminary Registration

Once scanning is complete (usually about 5 to 10 minutes), the AR HMD has acquired (1) a video sequence of spherical panoramas along a path that intersects all empty grid cells on the floor plan, (2) a video sequence acquired by the AR HMD on-board video camera, (3) pose tracking data for the frames of the on-board video camera, and (4) a coarse geometric of the scene captured by the active depth camera built into the AR HMD. At the beginning of the scanning session we synchronize the frame sequences of the panoramic and of the on-board cameras by flashing a light. After synchronization, the pose tracking data for the on-board camera frames is transferred to the panoramic frames.

5.2 Panorama Selection

We select two sets of panoramas from the spherical panoramic video sequence to be incorporated into the image-based model of the scene: *grid cell panoramas*, for scene visualization from anywhere on the acquisition plane; and *acquisition path panoramas*, for quality visualization from anywhere on the acquisition path.

The grid cell panoramas are selected by finding the best panorama for each accessible 2D grid cell (Fig. 3). The best panorama is the one that has an acquisition location closest to the cell center, based on the tracked pose data. We exclude panoramas with high panning angular velocity since they are blurry. The

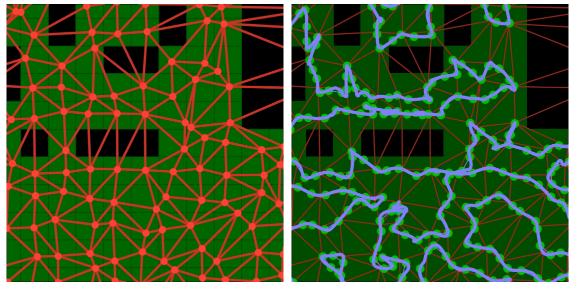

Fig. 4. Detail view of acquisition area. Left: Grid cell panoramas (red dots) triangulated in panorama triplets (red lines). Right: Acquisition path (blue line), and acquisition path panoramas (green dots).

angular velocity is estimated based on the tracked pose data. The acquisition viewpoints of the panoramas in the grid cell set are 2D Delaunay triangulated on the acquisition plane. This triangulation defines a *panorama triplet* for every output visualization viewpoint (Fig. 4, left). Panorama triplets may occasionally cross unvisited grid cells. This is not a problem when cells are unvisited due to containing low obstacles (desks, chairs) below head height. However, if an unvisited cell contains a wall, some morphing results could incorporate views from both sites of the wall, leading to unwanted artifacts. In practice, we cast rays against the AR HMD's coarse geometry along the triangulation edges, and discard triplets that would cross head-height obstacles.

The acquisition path panoramas are chosen from the panoramic video sequence at equal distance intervals, leveraging the tracked posed data again. The distance is smaller than the grid cell size (e.g., 0.25m vs 0.5m) to provide a higher quality visualization when the output viewpoint is on the acquisition path, as compared to when it is in the middle of a panorama triplet. As before, frames with high panning rotational velocity are avoided. Fig. 4, right, shows in detail the triangulated grid cell panorama locations and the path panorama locations.

5.3 Registration Refinement

The poses provided by our AR HMD are only accurate to about 2cm in translation and about 2 degrees in rotation. Also, drift can accumulate in the AR HMD's pose estimation; while the HMD can internally correct itself using loop closure techniques, the raw pose data we capture during acquisition is not automatically corrected. Using estimated poses directly in the construction of our image-based model would result in reduced visualization quality from image instability along smooth visualization paths, and from ghosting when transitioning between panoramas. We refine panorama registration (1) by detecting panorama features, (2) by estimating feature correspondences between neighboring panoramas, (3) by removing outlier correspondences, and (4) by performing a global pose graph optimization of all panorama poses. Throughout our registration refinement pipeline we work with the original spherical panorama images gathered during acquisition, without resampling to a cube map.

(1) For each panorama, we compute image features using SPHORB [32], an ORB feature generalization that operates directly in the spherical domain.

(2) We define correspondences between features of panorama pairs, based on the SPHORB feature distance. Rather than computing a full pairwise set of correspondences between all panoramas, which would be computationally expensive, we only find correspondences between adjacent panoramas based on AR HMD provided poses. First, we find correspondences between panoramas that are connected by an edge of a panorama triplet (determined by the AR HMD's estimated poses). Second, we find correspondences between consecutive acquisition path panoramas. Third, we find correspondences between each acquisition path panorama p_a and the grid cell panorama p_b with the closest acquisition location to that of p_a. These additional correspondences ensure that the two sets of panoramas are correctly registered together.

(3) The resulting set of correspondences contains many outliers unsuitable for registration refinement, so we find inliers with an iterative RANSAC approach [10] on each pair of neighboring panoramas (p_1, p_2) for which correspondences were found. Each iteration selects a subset of correspondences, uses the subset to estimate the essential matrix $\mathbf{E}$ of (p_1, p_2), and computes the subset's reprojection error. Correspondences are marked as inliers or outliers by comparing reprojection error against a threshold. We work directly in the spherical domain and follow the approach of Pagani and Stricker [20] in approximating the geodesic reprojection error ϵ_p with the projected distance of a ray to the epipolar plane, according to Equation 1, where (f_1, f_2) are a pair of corresponding features mapped onto the unit sphere that belong to two neighboring panoramas.

$$\epsilon_p = \frac{|f_2^T \mathbf{E} f_1|}{\|f_2\| \|\mathbf{E} f_1\|} \tag{1}$$

(4) The final step of panorama registration refinement performs a global non-linear least squares optimization of panorama poses based on the inlier correspondences validated by the previous step. This optimization determines poses for all panoramas that are globally consistent. Although pairwise relative poses have been implicitly computed in the previous outlier removal step, the essential matrix determined between a pair of panoramas is sensitive to noise and to the RANSAC parameters. To ensure that relative poses are consistent when moving between multiple panorama triplets, it is important to do a global optimization. Each panorama pose is represented with 6 parameters, i.e., 3 translations and 3 rotations. To limit the change in panorama acquisition location, we constrain translation to be within 0.1m of the estimated translation. The rotations are unconstrained. The error targeted by the optimization is the sum of squares of correspondence pair errors. The error for a pair of corresponding features (f_1, f_2) is computed based on an approach described by Pagani et al. [19]. The essential matrix $\mathbf{E}$ of the panoramas (p_1, p_2) is computed based on current pose estimates of the two panoramas. f_1 and f_2 are mapped to 3D points on the unit sphere

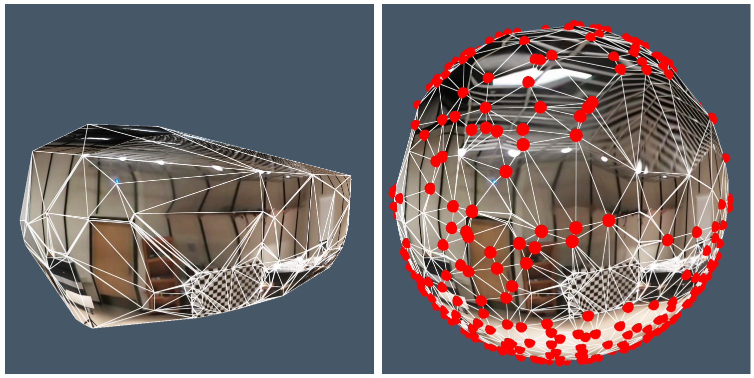

Fig. 5. Morphing mesh computed as 3D convex hull of feature point projections on unit sphere (left). Morphing mesh enriched with points (red dots) from the coarse 3D scene model acquired by the AR HMD (right).

and the error is computed as a locally projected measure of the geodesic distance from the epipolar line. Minimizing $f_2^T \mathbf{E} f_1$, which would be appropriate for Euclidean space correspondences, would minimize the sine of the geodesic distance in our context of spherical panorama correspondences. Instead, we minimize the value d_{et} defined in Equation 2.

$$d_{et} = \tan\left(\arcsin(f_2^T \mathbf{E} f_1)\right) = \frac{f_2^T \mathbf{E} f_1}{\sqrt{1 - (f_2^T \mathbf{E} f_1)^2}} \tag{2}$$

Both ϵ_p and d_{et} are valid approximations of the geodesic reprojection error, with ϵ_p yielding slightly better registration refinement given noisy features [20]. However, we found that optimizing over ϵ_p during registration refinement was much slower to converge, and so we favor ϵ_p during our inlier selection to ensure only high-quality matches are selected, and use d_{et} during the computationally expensive registration refinement step.

After registration refinement, we re-triangulate the grid cell panorama viewpoints to account for any shift in these viewpoints during optimization. Since the translation degrees of freedom are constrained, the viewpoints shift little, which preserves the uniform acquisition property of the grid cell panorama set.

5.4 Morphing Mesh Construction

When the viewpoint is located at one of the panorama acquisition locations, the scene can be visualized with high quality in any view direction by resampling the panorama to the output image. To support translations between the panorama acquisition locations, we triangulate panorama correspondences to construct a *morphing mesh*. We build morphing meshes for each grid cell panorama triplet, and for each acquisition path panorama pair.

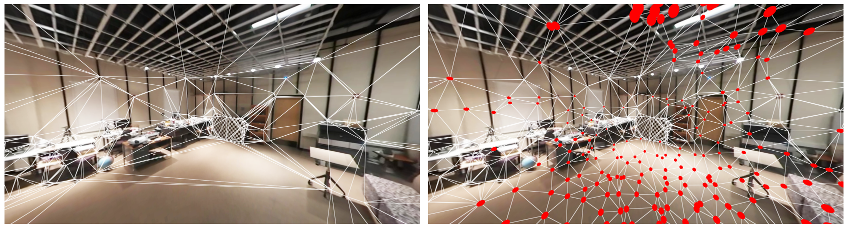

Fig. 6. Morphing mesh computed from features (left) and enriched with geometry points, shown as red dots (right).

A morphing mesh is built in three stages: (1) the mesh is triangulated from image features, (2) the mesh is enhanced using the coarse 3D geometry acquired by AR HMD's active sensors, and (3) the mesh is modified to ensure topological consistency over the panorama triplet or the panorama path segment pair.

(1) We define mesh vertices from correspondences between neighboring panoramas detected during the registration refinement stage. Given a panorama triplet (p_1, p_2, p_3), we compute the transitive closure S of correspondence pairs. A three-way correspondence (f_1, f_2, f_3) is included in S iff the two-way correspondences (f_1, f_2), (f_2, f_3), and (f_3, f_1) exist between panoramas (p_1, p_2), (p_2, p_3), and (p_3, p_1), respectively. For each three-way correspondence (f_1, f_2, f_3), we define a 3D scene feature point as the point closest to the three panorama rays through f_1, f_2, and f_3. The feature 3D points are projected on a unit sphere centered at the panorama triplet barycenter, and their projections are triangulated by computing their 3D convex hull. An example morphing mesh constructed from feature points is visualized in Fig. 5, left, on the unit sphere, and in Fig. 6, left.

(2) The feature-based morphing mesh is sparse in featureless regions (Fig. 5, left, Fig. 6, left). We increase the fidelity of the morphing mesh in a second stage by adding vertices based on the coarse geometric model acquired by the AR HMD with its built-in active depth camera (Fig. 5, right, Fig. 6, bottom). This geometric model is neither complete nor precise, so it could not be used alone. However, when combined with the panorama features, the coarse geometry increases the fidelity of the transitions between panorama viewpoints. 3D points from the coarse geometry are iteratively added to the morphing mesh only in regions where the morphing mesh is sparse. Morphing mesh triangles that are above a threshold size are subdivided by adding a vertex at the triangle center. The 3D position of the new vertex is defined by intersecting the ray from the barycenter of the panorama triplet with the coarse geometry. If the ray fails to intersect the incomplete coarse geometric model, a random point is selected inside the triangle to be subdivided. The process continues until all triangles in the morphing mesh are sufficiently small, or until no further intersections with the coarse geometry are possible.

(3) Even though we compute connectivity by computing the 3D convex hull, triangulation is performed in the 2D domain of the surface of the unit sphere. This requires topological consistency anywhere within the panorama triplet; no

mesh triangle should flip orientation as the viewpoint translates away from the barycenter. We enforce morphing mesh topological consistency with an iterative approach [31], which eliminates triangles that yield inconsistent orientation.

We have described the construction of the morphing mesh in the context of a panorama triplet. For panorama pairs defined by acquisition path segments, the process is similar and simpler. No three-way correspondences are needed during the feature-based morphing construction, and the two-way correspondences between the two panoramas are used directly. The barycenter of the panorama triplet is replaced with the midpoint of the path segment. Finally, topological consistency is achieved by checking triangle orientation at the midpoint and two endpoints of the path segment.

6 VR-Guided Interactive Visualization

Like all image-based models, our model captures the scene with non-uniform fidelity. Based on the user desired viewpoint, our model has three levels of fidelity. The first level is when the user viewpoint is inside a panorama triplet. In this case, the output image is rendered by projectively texture mapping the morphing mesh with a blend of the three triplet panoramas. The blending weights are defined by the user's viewpoint barycentric coordinates inside the triplet triangle. The user can look in any direction from within the triangle triple, leveraging the complete 360° textures and meshes. In the case where the user is not within any panorama triplet, we can either display nothing to the user or we can use the most recently visited panorama triplet at the cost of additional distortion.

The second level of fidelity is when the user viewpoint is near an acquisition path segment. In this case, we switch to a morph between the two panoramas of the segment endpoints. The segment is shorter than the triplet triangle edge, and only two panoramas are blended, so the quality of the output image is higher than in the inside the triplet case. Again, the user can look in any direction from anywhere along the acquisition path segment.

The third and highest level of fidelity of our image-based model is when the user viewpoint is near one of the panorama acquisition locations. In this case the output image is rendered by texturing the morphing mesh with a single panorama, which approaches a resampling of the panoramic frame and therefore has high fidelity in any view direction. Note that we only clamp blending weights for texture color and not for 3D viewpoint position, so that tracked HMD navigation in VR is natural and does not "stick".

We have designed a VR visualization interface that allows the user to take advantage intuitively of the highest model fidelity available in the proximity of their viewpoint. The interface shows a map of panorama acquisition locations in the bottom right corner of the frame (Fig. 1, c and d). The map shows the acquisition path segments (Fig. 1, e) and the current user location. As during acquisition, the user can easily align their visualization location with one of the panorama acquisition locations. There are hundreds of acquisition locations and so there is always one nearby. The typical navigation pattern is to translate the viewpoint to an acquisition location, to pan and tilt the viewpoint while remaining at the acquisition location to take advantage of the highest fidelity of our image-based model, and then to move to the next acquisition location, either along an acquisition path segment, or through a panorama triplet triangle.

7 Results and Discussion

Our AR acquisition device uses a Microsoft HoloLens [18] AR HMD coupled with a Samsung Gear 360 panoramic camera [24]. The camera captures 3,840 x 1,920 panoramic frames at 30fps. We visualize our image-based models interactively and immersively using an HTC Vive HMD [12]. Our image-based models are also suitable for visualization on conventional displays.

7.1 AR Guided Acquisition

To demonstrate our system, we acquired several indoor environments (one environment at multiple resolutions), and we conducted a user study in which ten novice participants acquired a reference environment. The acquisition cell size is 0.5m x 0.5m, and the path segment length is 0.25m, unless otherwise specified.

Test scenes Table 1 gives acquisition details for our four test environments. All were acquired with hundreds of evenly distributed panoramas in 8min or less.

Table 1. Acquisition performance for four indoor environments.

Environment (Figures)	Floor space [m] x [m]	Path length [m]	Capture time [s]	Grid cell panoramas	Path segm. panoramas
Lab (1c, 1d, 7 top)	10x13	172	307	244	688
Home (1e, 7 btm)	7x10	105	284	157	420
Lobby (7 mid)	9x9	131	248	215	524
Office (8)	2.5x4	31	86	35	124

We acquired the *Office* scene at multiple spatial resolutions (Table 2). As the length of the grid cell is halved, the grid cell area is quartered, so the number of grid cell panoramas will quadruple, assuming the scene is an open floor area. For floor areas with obstacles, this factor varies: from Table 2, the number of grid cell panoramas grows by a factor of 2.82 and of 3.55 as grid cell length shrinks from 0.5m to 0.25m to 0.125m. Since the panoramic video camera records continuously as the user moves along the acquisition path, we expect *acquisition time* to double (not quadruple) when the grid cell's length is halved; direct movement between two locations requires the same path regardless of grid cell size. Once floor obstacles are included, this factor is affected by the user's ability to visit

Fig. 7. Results for *Lab* (top row), *Lobby* (middle row), and *Home* (bottom row) environments, with viewpoint between acquisition locations (left), and near an acquisition location (right).

all accessible parts of the floor non-redundantly. From Table 2, acquisition times grow by a factor of 1.29 and 2.18, respectively.

The actual acquisition paths corresponding to Table 2 are shown in Fig. 8. For grid cell length of 0.5m and 0.25m, the acquisition path tends to have long straight portions, but in the case of a grid cell length of 0.125m, the trajectory tends to be made up of small imprecise loops. Limitations in comfortable head and neck motion leads to gradual rather than sharp turns which would be needed to efficiently sample the space at such a high resolution.

Acquisition user study An important goal of our work is to develop an acquisition system that allows novice users to acquire a complex indoor environment in minutes. We have gathered initial evidence for reaching this goal in a user study with ten first-time users of our system who acquired the same large indoor environment (i.e., the *Lab*). The users, who had general experience with AR/VR HMDs, were asked to traverse every accessible grid cell in the room that they felt they could reasonably reach. The users were briefed in five minutes or less on how to use the acquisition system. The briefing did not include a suggested scanning strategy so as not to bias participants.

Table 2. Acquisition performance for *Office* environment as a function of grid cell size.

Grid size [m]x[m]	Path length [m]	Time [s]	Grid cell panoramas	Path segm. panoramas
.5 x .5	31	86	35	62
.25 x .25	39	111	99	154
.125 x .125	81	242	352	648

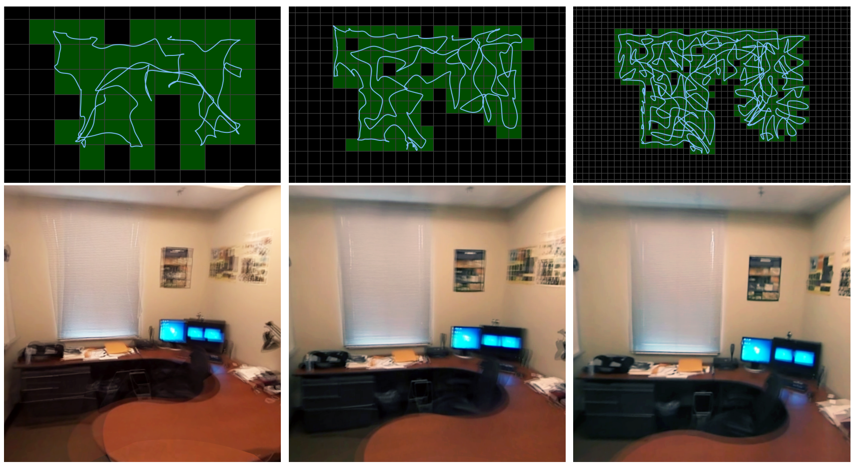

Fig. 8. Top row: Paths during acquisition of the *Office* scene at varying grid cell sizes: 0.5m (left), 0.25m (middle), and 0.125m (right). Bottom row: Corresponding visualization frames, rendered from a panorama triplet barycenter.

The users acquired the 10m x 13m scene in an average time of 7min5s (min: 4min57s, max: 11min1s), using a grid cell size of 0.5m x 0.5m. On average, 268 grid cell panoramas were acquired (min: 217, max: 365). Average distance traveled was 196m (min: 144m, max: 232m). Fig. 9 shows the acquisition paths for the ten users. There is great variability in the acquisition paths: some users cover the floor space with large cycles around the perimeter then fill in missing interior regions, while others cover the floor space progressively with paths reminiscent of space-filling curves. Users also had different completeness criteria; because the environment was a complex scene with many floor obstacles, users differed in their willingness to move into hard-to-reach areas to achieve greater coverage. In all cases, acquisition resulted in hundreds of evenly-spaced panoramas that allow for quality interactive visualization of the environment.

7.2 Panorama Morphing Setup

The acquired data is processed offline to prepare for interactive visualization. For our largest environment, i.e., the *Lab*, the entire offline processing took 227min.

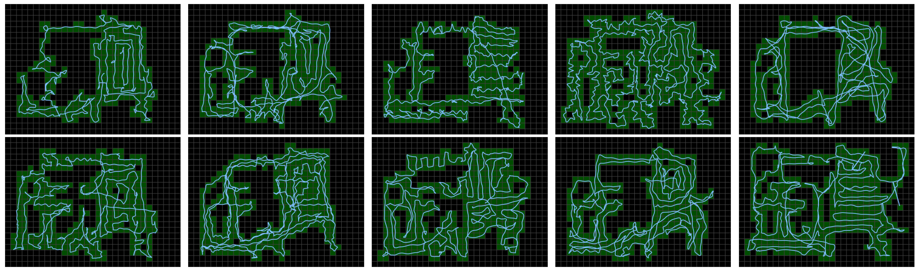

Fig. 9. Acquisition paths (blue lines) and explored grid cells (green) for first-time users of our system (grid cell of 0.5m x 0.5m).

Referring back to our system pipeline (Fig. 2), for the typical panorama shown in Fig. 5 and Fig. 6, the Registration Refinement stage finds 1,369 pairwise feature matches, 775 of which are inliers used during registration refinement. Registration refinement reduces the average feature reprojection error from 9.5 pixels to 1.3 pixels, which is a small error relative to the high panorama resolution of 3,840 x 1,920. For a typical panorama triplet, the Feature-Based Morphing Mesh Construction stage results in a morphing mesh with 1,126 points and 1,858 triangles, which is then refined in the Geometry-Based Morphing Mesh Enrichment stage to a final morphing mesh with 1,276 points and 2,158 triangles.

7.3 VR Guided Visualization

Our image-based models visualize the captured environments in the VR HMD at 90fps in stereo, by rendering the current low-polygonal-count 3D morphing mesh with projective texture mapping. High fidelity is achieved near one of the hundreds of panorama acquisition locations that sample the floor space uniformly, where the visualization converges to a resampling of the high-resolution spherical panorama acquired (e.g., Fig. 1c and Fig. 7, top right). Moderately high fidelity is achieved along the acquisition path, where the two segment endpoint panoramas are merged (e.g., Fig. 1e). The lowest level of fidelity is found at the center of the panorama triplet, when undersampled geometry can lead to ghosting artifacts (e.g., ceiling lights in Fig. 1d, vertical black lines on the far wall in Fig. 7, top left, or the far end of the long hallway in Fig. 7, middle right). Using the visualization map, the user can align their viewpoint with an acquisition location or segment where visualization fidelity is highest. Movement through the visualization is orientation-preserving and non-disorienting.

Fig. 8 shows visualization frames rendered from the center of panorama triplets for the three image-based models of the *Office* scene with decreasing acquisition grid cell size. As expected, the quality of the visualization increases, as a smaller grid cell size reduces the distance from the output viewpoint to the closest acquisition location, which reduces ghosting.

7.4 Limitations

The accuracy of geometry points in the morphing meshes is limited by the coarseness of the geometric model that our AR HMD acquires, and by the quality of the panorama registration. We also enforce topological consistency of the morphing mesh across the panorama triplet, which caps the maximum fidelity of the morphing mesh. Future work could explore a general, and not unit sphere-based, 3D triangulation of feature and geometry points, that allows for folds as the viewpoint translates within a panorama triplet.

Our selection of neighboring panoramas from which to find correspondences and feature points implies small baselines for optimization and triangulation. Another challenge is the registration of panoramas from adjacent rooms connected by a open door. Such panoramas do not share many common features as the panorama from room A sees only a small part of room B, and vice versa. Therefore, with the current implementation, multi room environments require assembling the overall image-based model from individual room models.

Our panoramic camera is mounted above the wearer's head, which results in a visualization that appears taller than the acquiring user's height. Since the morphing meshes are rendered in 3D, we do currently provide limited support for vertical viewpoint translation, as needed for example to cover the small range of vertical translation when walking with the VR HMD. Future prototypes could place the cameras at a lower height to better match a typical user height.

8 Conclusions and Future Work

We have presented a system for fast image-based modeling and rendering of indoor spaces, which guides the user with an AR interface towards complete and dense acquisition. A VR interface enables the user to optimize output image quality. The results of our pilot study are promising; however, we plan to conduct additional and more formal user studies to validate the effectiveness of the AR interface and the acceptability of the resulting VR visualization.

We currently only support static scenes; future work could support dynamic scenes by injecting moving geometry captured with RGB-D sensors into a captured scene. Automatic detection and removal of dynamic regions of the image-based model would also help deal with accidental intruders that interfere with acquisition, opening up the possibility of acquiring busy, in-use spaces. We are also interested in real-time use of image-based features for saliency or view-dependency during acquisition, which could allow our system to prioritize regions that would be of greater complexity or of greater interest to a viewer.

We believe our work demonstrates that image-based modeling and rendering of inside-looking-out indoor spaces can efficiently produce quality models that are ready to be integrated into applications.

9 Acknowledgments

We thank the ART research group at Purdue University for their feedback during development. This work was supported in part by the United States National Science Foundation under Grant DGE-1333468.

References

1. Ahn, J., Wohn, K.: Interactive scan planning for heritage recording. Multimedia Tools and Applications **75**(7), 3655–3675 (Apr 2016)
2. Aliaga, D.G., Carlbom, I.: Plenoptic stitching: a scalable method for reconstructing 3d interactive walk throughs. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 443–450. ACM (2001)
3. Bradley, D., Brunton, A., Fiala, M., Roth, G.: Image-based navigation in real environments using panoramas. In: IEEE International Workshop on Haptic Audio Visual Environments and their Applications. pp. 3 pp.– (Oct 2005)
4. Chiang, C.C., Way, D.L., Shieh, J.W., Shen, L.S.: A new image morphing technique for smooth vista transitions in panoramic image-based virtual environment. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology. pp. 81–90. VRST '98 (1998)
5. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM Transactions on Graphics (TOG) **36**(3), 24 (2017)
6. Davis, A., Levoy, M., Durand, F.: Unstructured light fields. In: Computer Graphics Forum. vol. 31, pp. 305–314. Wiley Online Library (2012)
7. DiVerdi, S., Wither, J., Höllerer, T.: All around the map: Online spherical panorama construction. Computers & Graphics **33**(1), 73–84 (2009)
8. Dong, S., Höllerer, T.: Real-time re-textured geometry modeling using microsoft hololens (2018)
9. Fan, X., Zhang, L., Brown, B., Rusinkiewicz, S.: Automated view and path planning for scalable multi-object 3D scanning. ACM Trans. Graph. **35**(6), 239:1–239:13 (Nov 2016)
10. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In: Readings in computer vision, pp. 726–740. Elsevier (1987)
11. Hedman, P., Ritschel, T., Drettakis, G., Brostow, G.: Scalable inside-out image-based rendering. ACM Transactions on Graphics (TOG) **35**(6), 231 (2016)
12. HTC: VIVE (2017), www.vive.com/us
13. Huang, J., Chen, Z., Ceylan, D., Jin, H.: 6-dof vr videos with a single 360-camera. In: Virtual Reality (VR), 2017 IEEE. pp. 37–44. IEEE (2017)
14. Jung, J.H., Kang, H.B.: An Efficient Arbitrary View Generation Method Using Panoramic-Based Image Morphing, pp. 1207–1212 (2006)
15. Kawai, N.: A simple method for light field resampling. In: ACM SIGGRAPH 2017 Posters. pp. 15:1–15:2. SIGGRAPH '17 (2017)
16. Kawai, N., Audras, C., Tabata, S., Matsubara, T.: Panorama image interpolation for real-time walkthrough. In: ACM SIGGRAPH Posters. pp. 33:1–33:2 (2016)
17. Kolhatkar, S., Laganire, R.: Real-time virtual viewpoint generation on the GPU for scene navigation. In: 2010 Canadian Conference on Computer and Robot Vision. pp. 55–62 (May 2010)

18. Microsoft: Microsoft HoloLens (2017), www.microsoft.com/en-us/hololens
19. Pagani, A., Gava, C.C., Cui, Y., Krolla, B., Hengen, J.M., Stricker, D.: Dense 3d point cloud generation from multiple high-resolution spherical images. In: VAST. pp. 17–24 (2011)
20. Pagani, A., Stricker, D.: Structure from motion using full spherical panoramic cameras. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 375–382. IEEE (2011)
21. Pan, Q., Reitmayr, G., Drummond, T.W.: Interactive model reconstruction with user guidance. In: 2009 8th IEEE International Symposium on Mixed and Augmented Reality. pp. 209–210 (Oct 2009)
22. Pankratz, F., Klinker, G.: [poster] ar4ar: Using augmented reality for guidance in augmented reality systems setup. In: Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on. pp. 140–143. IEEE (2015)
23. Rusinkiewicz, S., Hall-Holt, O., Levoy, M.: Real-time 3d model acquisition. ACM Transactions on Graphics (TOG) **21**(3), 438–446 (2002)
24. Samsung: Gear 360 Camera (2017), www.samsung.com/us/explore/gear-360
25. Shi, F.: Panorama Interpolation for Image-based Navigation. Master's thesis, University of Ottawa (12 2007)
26. Tuite, K., Snavely, N., Hsiao, D.Y., Smith, A.M., Popović, Z.: Reconstructing the world in 3d: bringing games with a purpose outdoors. In: Proceedings of the Fifth International Conference on the Foundations of Digital Games. pp. 232–239. ACM (2010)
27. Tuite, K., Snavely, N., Hsiao, D.y., Tabing, N., Popovic, Z.: Photocity: training experts at large-scale image acquisition through a competitive game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1383–1392. ACM (2011)
28. Xiao, J., Shah, M.: Tri-view morphing. Comput. Vis. Image Underst. **96**(3), 345–366 (Dec 2004)
29. Xu, K., Shi, Y., Zheng, L., Zhang, J., Liu, M., Huang, H., Su, H., Cohen-Or, D., Chen, B.: 3D attention-driven depth acquisition for object identification. ACM Trans. Graph. **35**(6), 238:1–238:14 (Nov 2016)
30. Zhang, Y., Zhu, Z.: Walk-able and Stereo Virtual Tour based on Spherical Panorama Matrix, pp. 50–58. Springer International Publishing, Cham (2017)
31. Zhao, Q., Wan, L., Feng, W., Zhang, J., Wong, T.T.: Cube2Video: Navigate between cubic panoramas in real-time. IEEE Transactions on Multimedia **15**(8), 1745–1754 (Dec 2013)
32. Zhao, Q., Feng, W., Wan, L., Zhang, J.: SPHORB: A fast and robust binary feature on the sphere. International Journal of Computer Vision **113**(2), 143–159 (Jun 2015)