CasualStereo: Casual Capture of Stereo Panoramas with Spherical Structure-from-Motion

Lewis Baker* University of Otago
Steven Mills† University of Otago
Stefanie Zollmann‡ University of Otago
Jonathan Ventura§ California Polytechnic State University

Figure 1: Example stereo panorama produced using our spherical structure-from-motion pipeline, rendered as a red-blue anaglyph.

ABSTRACT

Hand-held capture of stereo panoramas involves spinning the camera in a roughly circular path to acquire a dense set of views of the scene. However, most existing structure-from-motion pipelines fail when trying to reconstruct such trajectories, due to the small baseline between frames. In this work, we evaluate the use of spherical structure-from-motion for reconstructing handheld stereo panorama captures. The spherical motion constraint introduces a strong regularization on the structure-from-motion process which mitigates the small-baseline problem, making it well-suited to the use case of stereo panorama capture with a handheld camera. We demonstrate the effectiveness of spherical structure-from-motion for casual capture of high-resolution stereo panoramas and validate our results with a user study.

Index Terms: Human-centered computing—Interaction paradigms—Virtual reality—; Computer vision—Image and video acquisition—3D imaging

1 INTRODUCTION

With the continuous rise of immersive viewing devices such as head-mounted displays (HMDs), there is a growing need for consumer-grade methods of virtual reality (VR) content capture. The ideal VR content capture modality is inexpensive, lightweight and easy-to-use. For example, most smartphones and cameras today offer a panoramic capture mode, in which the user can stitch together a 360 degree field-of-view image by turning the camera in a circle. However, HMD viewing is greatly enhanced by stereoscopic viewing to give a sense of depth to the user, which a normal panorama cannot provide.

A viable alternative is the stereo panorama [22,29,33] which does provide stereoscopic, 360 degree viewing, and example of which is shown in Figure 1. To capture a stereo panorama, the camera is spun in a circle while facing outward. Columns from the left and right side of each image are concatenated to form the right- and left-eye panoramas, respectively. The disadvantages of this approach are the need for a perfectly circular trajectory and a high density of views. The angular resolution of the resulting panorama is determined by the number and density of the input images.

Recent work explores "casual" capture of stereo panoramas [3, 30, 40], where the camera can be handheld and spun in a roughly circular trajectory. To recover the pose of each camera, traditional structure-from-motion methods [31] are used. Once the camera poses have been determined, new views on a perfectly circular path are synthesized using flow-based blending. However, this particular case of camera motion is especially difficult for traditional structure-from-motion (SfM) methods, because of the small baseline between views and the relative lack of overlap between frames.

In this work, we evaluate the use of spherical structure-from-motion [37] as an effective alternative for reconstructing handheld stereo panorama sequences. Spherical SfM differs from general SfM in that the camera is assumed to move on the surface of an imaginary sphere; i.e., the camera is assumed to maintain a constant distance from the origin and to always be facing directly outward. Handheld stereo panorama capture matches these assumptions well, since the camera is held in an outstretched hand and spun roughly in a circle.

The spherical constraint removes disambiguity in two-view relationships and provides a strong, implicit regularization on the SfM result. In addition, because each camera pose is determined completely by the rotation of the camera, there is no need for scale propagation and incremental reconstruction as in traditional monocular SfM. In this paper we show that, using spherical structure-from-motion, we can produce a stereo panorama from an input video sequence in minutes rather than hours as required by previous work [30].

In this paper, we review the theory of spherical structure-from-motion and describe our spherical SfM and stitching pipeline for causal stereo panorama generation. We then evaluate the use of spherical SfM for processing handheld-captured videos and the resulting stereo panoramas produced using the camera pose estimates within a technical evaluation as well as within a user evaluation. Our evaluation on several handheld captured sequences shows that we can reconstruct casual stereo panorama capture trajectories and produce high-quality stereo panoramas, and that even untrained users can produce suitable input videos for our pipeline.

Specifically, our contributions are as follows:

- We describe a spherical structure-from-motion and stitching pipeline for reconstruction of stereo panoramas from casually-captured input videos.
- We compare the speed and reliability of our approach to COLMAP [31], a state-of-the-art general SfM method, on several input videos. We show that spherical SfM reconstructs

*e-mail: bakelew@gmail.com
†e-mail: steven@cs.otago.ac.nz
‡e-mail: stefanie@cs.otago.ac.nz
§e-mail: jventu09@calpoly.edu

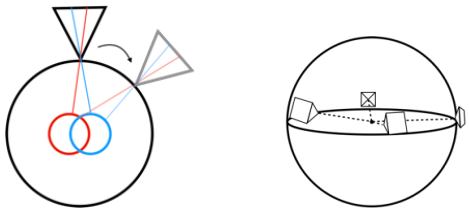

Figure 2: *Left:* Illustration of idealized stereo panorama capture. Images (black triangles) are taken in a dense set of locations along a circular path. Left- and right-eye panoramas (red and blue circles) are stitched together from different columns in the images. *Right:* Under the spherical motion constraint, the camera is allowed to move freely on the surface of the unit sphere. We show that a user causally capturing a stereo panorama with a camera held in an outstretched hand will produce a trajectory that agrees with this constraint well enough to produce a high-quality stereo panorama using our method.

the videos more reliably and quickly than general SfM.

- We show that spherical SfM reconstruction is accurate enough to produce high-quality stereo panoramas through qualitative evaluation and a quantitative user study.

2 Related Work

Our method builds on recent structure-from-motion methods that constrain the solution to motion on a sphere [37], and applies them to stereo panorama reconstruction. This section provides an overview of these areas.

2.1 Structure-from-Motion

Structure-from-motion is the process of determining the camera poses and 3D structure of a scene from a video or collection of images. Most systems take an incremental approach, where the reconstruction is grown from an initial pair of images (Cf. [31, 34]). An alternative strategy is to attempt to first determine the camera rotations independent of their translations [38] and then estimate the translations after.

These techniques form the basis of structure-from-motion techniques that can model scenes from large collections of images [31]. Such techniques rely on general camera motion and, as we shall see, can fail in the limited motion made in panorama capture scenarios. Most small-motion SfM methods [12, 19, 39] use the inverse depth parameterization and depth map regularization to achieve accurate triangulation, but are not designed to reconstruct a 360 stereo panorama. SfM with a spherical panoramic camera (Cf. [28, 36]) can be used to construct a stereo panorama [20] but requires a special panoramic camera, whereas we use a normal perspective camera, as would be found on a smartphone or other consumer device.

In this work we explore the use of a spherical motion constraint [37] to regularise the reconstruction and overcome the small-baseline problem inherent in stereo panorama capture. Introducing a spherical constraint reduces the number of point correspondences needed between image pairs, as described in Sect. 3, and improves robustness to the limited motion, as shown in Sect. 5.

2.2 Stereo Panorama Reconstruction

The reconstruction of stereo panoramas is a well researched field. Early works used input from cameras that where moved on a circular trajectories [22, 29, 33] in order to capture the imagery necessary to create separate left- and right-eye panoramas.The idea is that when using the centre strip of each image a normal panoramic image is generated. When using strips from the left side, a right-eye panorama is generated and vice-versa for the left-eye panorama. These approaches come with the disadvantage of visual artifacts such as seams and discontinuities when the input camera sequence is not perfectly acquired.

The Megastereo [30] and Megaparallax [3] systems compensate for these artifacts by adopting a view synthesis approach. Both methods use Structure-From-Motion (SfM) in combination with flow-based ray blending in order to compute more accurate synthetic views. Zhang and Liu [40] use purely stitching-based methods instead of 3D reconstruction in order to stitch the left and right panoramas from casual input video, but require a stereo camera.

Stereo panoramas only provide horizontal parallax from a fixed viewing position. Other techniques such as plenoptic modeling [26], view-dependent image-based rendering [8], unstructured lumigraph or light field rendering [5, 7] and more recent methods [14, 18, 25] go beyond stereo panoramas to provide free viewpoint rendering using view synthesis. In recent years, methods based on deep neural networks [10, 11, 21, 41, 42] have also shown impressive results for view synthesis. However, these approaches require a density of input views which make them generally unsuitable for casual capture of an entire 360 degree scene.

An alternative approach is to use a stereo or multi-camera rig to facilitate capturing the large set of views needed to stitch a stereo panorama. Companies such as Facebook and Google provide camera setups that are specifically designed for 360 degree capture and consist of a ring of outward facing cameras, such as the Google Jump [2] or Facebook Surround360 [9]. These camera rigs provide a sparse set of views with known extrinsic and intrinsic calibration; view interpolation is then used to produce smooth stereo panoramic output. Similarly, Schroers et al. presented an approach for computing omnistereo videos from a sparse set of 16 cameras that also allow for virtual head motion [32]. While these rigs are ideal for professional panoramic video capture, they are less than ideal for "casual" capture given the cost, size, and weight of the camera rig. Hedman and Kopf [15] enabled fast reconstruction of a scene for free-viewpoint rendering from casually captured smartphone video, but require a stereo camera and integrated inertial measurement unit (IMU) for initial depth and motion estimates.

For casual capture of stereo panoramas with a handheld perspective camera, Bertel and Richardt [3] noted that one of the main challenges is that "structure-from-motion is also known to not be robust for our desired narrow-baseline inside-out capturing scenario." In our work, we address this problem by proposing the use of spherical SfM [37] for the generation of stereo panoramas.

3 Spherical Structure-from-Motion

Casual stereo panorama involves moving a hand-held camera in a roughly circular trajectory. This trajectory makes it the an ideal candidate for being represented by the geometry of spherical camera motions. In this section, we will discuss the geometry of spherical camera motions. We will give expressions for the absolute and relative pose matrices induced by spherical motion and derive the form of the essential matrix.

The assumption behind spherical SfM is that the camera lies on an imaginary sphere with its optical axis parallel with the sphere normal [37]. This allows for a simplified mathematical representation. For an outward-facing camera, the 3×4 camera extrinsics matrix P can be expressed using a 3×3 rotation matrix R and a 3×1 vector $\mathbf{z}$:

$$\mathsf{P} = [\mathsf{R} \mid -\mathbf{z}]. \tag{1}$$

where $\mathbf{z} = [0\ 0\ 1]^\top$. This is in contrast to general SfM where the vector $\mathbf{z}$ is replaced by a general translation vector, $\mathbf{t}$. Fixing $\mathbf{z}$ in this way represents the fact that the camera stays a fixed distance (which we take to be unit) from the origin and always points away from the origin. The camera pose is determined completely by the rotation of the camera, which only has three degrees of freedom, instead of six as in the unconstrained case.

The spherical assumption also leads to a simplified form for the relative pose between two cameras. Given two outward-facing cameras with extrinsics $\mathsf{P}_1 = [\mathsf{R}_1 \mid -\mathbf{z}]$ and $\mathsf{P}_2 = [\mathsf{R}_2 \mid -\mathbf{z}]$, we can now derive the relative pose $[\mathsf{R}_{1\to2} \mid \mathbf{t}_{1\to2}]$ between them. The relative rotation is

$$\mathsf{R}_{1\to2} = \mathsf{R}_2\mathsf{R}_1^\top. \tag{2}$$

and the relative translation is

$$\mathbf{t}_{1\to2} = \mathbf{r}_3 - \mathbf{z}. \tag{3}$$

where $\mathbf{r}_3$ denotes the third column of $\mathsf{R}_{1\to2}$.

The essential matrix $\mathsf{E}_{1\to2}$ relates corresponding camera normalised (i.e. calibrated) homogeneous points $\mathbf{x}_1$ and $\mathbf{x}_2$ in two images such that

$$\mathbf{x}_2^\top \mathsf{E}_{1\to2}\mathbf{x}_1 = 0. \tag{4}$$

If the two images have relative pose $[\mathsf{R}_{1\to2}|\mathbf{t}_{1\to2}]$ then

$$\mathsf{E}_{1\to2} = [\mathbf{t}_{1\to2}]_\times \mathsf{R}_{1\to2}, \tag{5}$$

where $[\mathbf{a}]_\times$ is the skew-symmetric matrix such that $[\mathbf{a}]_\times\mathbf{b} = \mathbf{a}\times\mathbf{b}\ \forall\ \mathbf{b}$.

Plugging in the Equations 2 and 3, the essential matrix relating outward-facing cameras is

$$\mathsf{E}_{1\to2} = [\mathbf{r}_3 - \mathbf{z}]_\times \mathsf{R}_2\mathsf{R}_1^T. \tag{6}$$

When the cameras are inward-facing, meaning that they point toward the center of the sphere instead of away from it, the translation vector and the essential matrix are both negated. However, since the essential matrix is only defined up to scale, the essential matrix for inward- and outward-facing cameras undergoing the same relative rotation is equivalent.

Ventura [37] developed minimal-case solvers for the essential matrix between two cameras under the spherical motion assumption, requiring only three correspondences. Furthermore, they allow the essential matrix to be decomposed into a unique relative pose solution, instead of the four possible relative poses normally present when dealing with unconstrained camera motion.

In general, an essential matrix E can be decomposed into two rotations R_a and R_b and a translation $\hat{\mathbf{t}}$ with unknown scale. This gives four possible solutions for the relative pose, namely, $[\mathsf{R}_a \mid \hat{\mathbf{t}}]$, $[\mathsf{R}_a \mid -\hat{\mathbf{t}}]$, $[\mathsf{R}_b \mid \hat{\mathbf{t}}]$, or $[\mathsf{R}_b \mid -\hat{\mathbf{t}}]$. However, only one of these relative poses is consistent with spherical motion with an outward facing camera. Let $\mathbf{t}_a$ and $\mathbf{t}_b$ be corresponding translation vectors for rotation solutions R_a and R_b, respectively, determined by Equation 3. We then compute a score for each rotation representing how close the rotation's corresponding translation is to the translation solution $\hat{\mathbf{t}}$. We compute the scores as

$$s_a = \frac{|\mathbf{t}_a \cdot \hat{\mathbf{t}}|}{||\mathbf{t}_a||}, s_b = \frac{|\mathbf{t}_b \cdot \hat{\mathbf{t}}|}{||\mathbf{t}_b||}. \tag{7}$$

The solution with higher score is chosen as the correct relative pose:

$$[\mathsf{R}|\ \mathbf{t}] = \begin{cases} [\mathsf{R}_a|\ \mathbf{t}_a] & \text{if } s_a > s_b, \\ [\mathsf{R}_b|\ \mathbf{t}_b] & \text{otherwise.} \end{cases} \tag{8}$$

This unique decomposition is especially helpful when dealing with the small-baseline problem. We noticed in our experiments (Sect. 5) that sometimes the partial reconstructions produced by COLMAP [31] are inverted so that the cameras point inward instead of outward. This suggests that COLMAP mistakenly selects the wrong decomposition of the essential matrix when initializing the reconstruction, leading to a failure of the entire process. By enforcing the constraint of outward-facing cameras, our system avoids this problem entirely.

4 CasualStereo: Constrained Panorama Capture

To capture a stereo panorama the user turns on the spot with their device held out, creating a roughly circular motion path, as illustrated in Figure 3. The preferred capture orientation is to hold the camera in "portrait" orientation to maximize the vertical field-of-view. Using portrait orientation causes a small visual overlap between views; in other words, feature points are not in the field of view for very long. In addition, because the circle of motion has a relatively small radius (the length of a human arm) there is not a large baseline between frames for point triangulation. A second issue is that the camera moves in a complete circle before ever reaching a loop closure, so accumulation of drift is unavoidable with general visual tracking.

This capture setup is especially difficult for incremental structure-from-motion systems [31,34] because there is usually no good initial pair of keyframes to start the SfM process. General essential matrix estimation with a small baseline is unstable and doesn't give good results, and scale propagation is especially difficult. In addition, reconstruction of 3D points from an initial pair of keyframes gives poor depth estimates due to the small baseline, which in turn make it difficult to register new views to the reconstruction.

Spherical SfM is a natural fit for casual stereo panorama capture, because the camera moves on an approximately circular path. In spherical SfM, it is assumed that camera center maintains a constant radius from the origin, so that the camera moves along the surface of an imaginary sphere. Furthermore, the camera viewing axis is assumed to always be parallel to the normal of the sphere, i.e. the viewing axis is coincident with ray from from center of the sphere to camera center. For handheld capture, the center of rotation would be approximately at the shoulder of the arm holding the camera, and the viewing axis pointing out along the user's outstretched arm.

4.1 Spherical SfM Reconstruction Pipeline

In this section we describe our spherical SfM reconstruction pipeline for casual stereo panorama capture. The major modifications to the spherical SfM pipeline introduced in previous work [37] are the use of SIFT features [23], automatic sequence sub-sampling, constrained loop closure search, L2 rotation averaging [6], and full 3D bundle adjustment to improve speed, accuracy and robustness. The input to the pipeline is the video captured by the user and the intrinsic parameters of the camera. The output is the estimated pose of an evenly distributed sub-sequence of frames from the video. Each step of the pipeline is described in the following.

4.1.1 Frame-to-frame tracking

We process each frame in the video sequentially. In the first frame detect SIFT feature points and extract their descriptors [23]. Then we track these features into the next frame using pyramidal Lucas-Kanade (LK) tracking [24, 35]. We found that using LK tracking, instead of detecting and matching SIFT features by their descriptors, provides higher accuracy, sub-pixel feature tracks which are critical for 3D reconstruction with small-baseline image pairs. When we track a feature with LK we also copy its new descriptor to the tracked feature point in the next frame instead of re-computing the descriptor. After LK tracking we also detect new SIFT features that are reasonably far from existing features in the image.

We then use these matches to estimate the spherical essential matrix [37] and separate inliers and outliers in a Preemptive RANSAC loop [27]. Outlier matches are removed from further consideration. The result of this process is an estimated rotation $\mathsf{R}_{i\to j}$ and a set of inlier feature matches between each pair of consecutive frames i, j in the sequence. If the amount of rotation between the two frames is less than one degree, we drop frame j from the sequence and restart the matching process between frames i and frame $j+1$. This loop continues until we have found the next frame in the sequence with a large enough rotation to the current frame, at which point the image pair is added to the reconstruction.

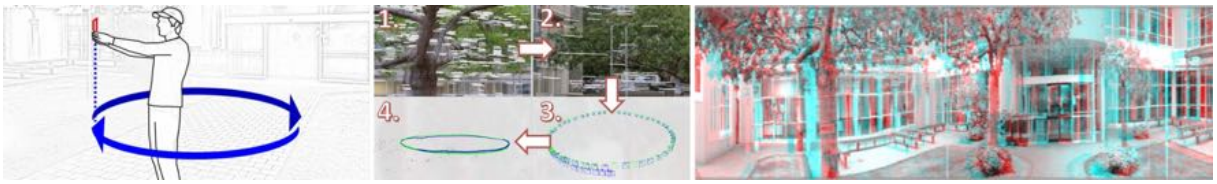

Figure 3: Overview of our approach. Left: The user spins around their axis to capture the stereo panorama. Centre: Tracking, loop detection, rotation averaging, and bundle adjustment are performed. Right: Stereo panorama is stitched, and stereo anaglyphs can be rendered.

This process of adaptively selecting a subset of reasonably spaced frames from the video helps to avoid sub-sequences with very small or no motion in the video, as these cause problems in the later triangulation and bundle adjustment steps. The adaptive sub-sampling also establishes a rough cap on the maximum time required for the reconstruction process. Note that, without the spherical constraint, it is difficult to automatically prune the video in this way, because in general the essential matrix only defines the translational part of the relative pose up to scale. For example, Bertel et al. [3] manually subsample their input sequences before processing them in COLMAP [31], since COLMAP is unable to perform this subsampling automatically.

To form an initial set of poses for the cameras in the reconstruction, we integrate the estimated relative pose between image pairs, i.e., $R_j = R_{i\to j}R_i$ for each consecutive pair of images i, j in the reconstruction. The first camera is fixed to have the identity rotation.

4.1.2 Loop closure

We search for the loop closures using the first thirty and last thirty frames of the sequence, after sub-sampling the sequence to one-degree increments as described above. For each frame in the first thirty frames, we attempt to calculate the relative pose to each of the last thirty frames. Specifically, we match nearest neighbor SIFT features, making use of the ratio test to reject ambiguous matches. For each candidate loop closure pair we run the Preemptive RANSAC matching procedure to estimate the spherical relative poses between the two images. Each loop closure with greater than 100 inliers is accepted. The loop closures are integrated into the initial pose estimate using rotation averaging as described next.

4.1.3 Rotation averaging

Pose drift accumulates during the frame-to-frame tracking and relative pose integration process. To reduce drift, we perform L2 rotation averaging [6] over the graph of relative rotation estimates produced by the frame-to-frame tracking and loop closure processes. The result of this process is a corrected initial pose estimate for every frame in the sequence.

4.1.4 Structure initialization

Feature tracks are assembled from the inlier feature matches found during the frame-to-frame tracking and loop closure steps. We then find an initial 3D point estimate for each feature track by applying direct linear transform (DLT) triangulation [13] with all observations in each track.

We found that having all observations available for triangulation is critical for these sequences because of the small baseline between frame pairs. In contrast to an incremental bundle adjustment, where it is critical to find good image pairs for the initial point triangulation, we are able to use all available views in the sequence to initialize the bundle adjustment procedure.

4.1.5 Bundle adjustment

Finally, we optimize the re-projection error in a bundle adjustment procedure, updating the 3D point locations and the camera rotations in an iterative manner. We parameterize points as 3D vectors instead of applying an inverse depth parameterization as has been previously used for spherical SfM [37].

Specifically, we find the rotations $\mathsf{R}_1, \ldots, \mathsf{R}_m$ and 3D point positions $\mathbf{p}_1, \ldots, \mathbf{p}_n$ that minimise the total re-projection error:

$$\sum_{i,j} \mathsf{M}_{i,j}\sigma(||\mathbf{o}_{i,j} - \pi(\mathsf{K}(\mathsf{R}_i\mathbf{p}_j - \mathbf{z}))||^2) \tag{9}$$

where M is a Boolean matrix indicating the visiblity of point j in camera i, $\mathbf{o}_{i,j}$ is the observation of point j in camera i, K is the intrinsics matrix, R_i is the estimated rotation for camera i, and $\mathbf{p}_j$ is the estimated position of point j, and $\pi(\cdot)$ is the projection operator. We apply a robust cost function $\sigma(x) = \log(1+x)$ to reduce the effect of outliers, and Ceres Solver [1] to perform the optimization.

4.2 Stereo Panorama Stitching

Once the SfM pipeline has completed, we have the camera pose estimates necessary to stitch together stereo panoramas. For the ideal case of a perfect circular trajectory with equal angular spacing between frames, Shum and Szeliski describe the geometry of stereo (or multiperspective) panoramas [33]. Under ideal capture conditions, each input view lies on a circular camera trajectory with an angle of θ around the circle. The columns in an input image are indexed by ϕ, the horizontal angle of the column from the center of the image. Panoramas at different horizontal offsets can be produced by sampling column ϕ from each input view and stitching them into a single panoramic image (Figure 2).

With casual capture, the actual cameras deviate slightly from the circular trajectory in terms of pose, and view density. Megastereo [30] and Megaparallax [3] use the input views to synthesize a new set of views perfectly spaced on the ideal circular trajectory. We adopt their approach to produce the necessary image columns for the output stereo panorama. We review the method briefly here.

4.2.1 Plane estimation

We robustly fit a plane to the camera centers in a RANSAC procedure. Then all cameras are rotated so that the plane normal coincides with the global up vector $\mathbf{y} = [0\ 1\ 0]^T$. This correction step brings the camera poses closer to the ideal circular trajectory.

4.2.2 Organization of images

We now order the images by their signed angle around a unit circle on the X-Z plane. To compute the signed angle, we first project the camera centers to the plane and compute each camera's angle θ_i around the circle. Let $\mathbf{x} = [1\ 0\ 0]^T$ and $\mathbf{c}_i = [c_x\ c_y\ c_z]^T$ be the camera center for camera i. We compute the signed angle θ_i as [3]

$$\theta_i = \text{atan2}(-c_z, c_x) \tag{10}$$

4.2.3 Flow-based blending

Now that the cameras are aligned and ordered, we can synthesize a panorama for a given ϕ by synthesizing each column, indexed by θ. We set the radius of the circular trajectory of synthetic views to be 0.5 and set the focal length of the synthesized views to be $1.2f$ where f is the focal length of the input views, in order to reduce

Table 1: Spherical SfM processing times from several sequences, and devices. All sequences were captured at 30fps and successfully reconstructed with Spherical SfM.

Dataset	Camera	Resolution	Frames	Time (s)
street [37]	Sony a5100	1920x1080	435	256
mountain [30]	Canon S95	720x1280	230	177
shrine	iPhone X	1080x1920	602	557
nature path	iPhone X	1080x1920	697	610
campus	iPhone X	1080x1920	667	520
courtyard	iPhone X	1080x1920	690	505

missing pixels at the top and bottom of the output panoramas. At each synthetic viewpoint we essentially render a synthetic image using a planar geometric proxy at a constant depth; we used a depth of 10 in our experiments.

For each pixel to be synthesized we select the best pair of consecutive input views to use for view synthesis. Let $\mathbf{r}_D^*$ be the ray to be synthesized, projected to the X-Z plane. The best pair of consecutive input views is the pair whose camera centers, when projected to the X-Z plane, lie on either side of $\mathbf{r}_D^*$ and are in front of the view to be synthesized. Note that, since our synthetic views lie in a perfect circle on the X-Z plane, every pixel in a column of the panorama will be synthesized from the same image pair.

To synthesize each pixel we will sample a pixel from the left and right images and linearly blend them with blending weight α. Again working on the X-Z plane, let α_{LD} be the angle between the ray to be synthesized and the ray to the left camera center, and let α_{LR} be the angle between the rays to the left and right cameras. Then the blending weight is computed as $\alpha = \alpha_{LD}/\alpha_{LR}$.

Let $\mathbf{x}_L$ and $\mathbf{x}_R$ be projections of a point on the proxy geometry in the left and right images, respectively. To compensate for parallax in the images before blending, we calculate a correction to $\mathbf{x}_L$ and $\mathbf{x}_R$ based on the optical flow between the two images [3, 30].

We denote the left and right images as I_L and I_R, respectively. We compute bidirectional optical flow [4] between I_L and I_R to obtain flow maps F_{LR} and F_{RL}. Local flow displacements F_{LR}^* and F_{RL}^* are calculated as

$$F_{LR}^*(\mathbf{x}_L) = \mathbf{x}_R - \mathbf{x}_L - F_{LR}(\mathbf{x}_L) \tag{11}$$

$$F_{RL}^*(\mathbf{x}_R) = \mathbf{x}_L - \mathbf{x}_R - F_{RL}(\mathbf{x}_R). \tag{12}$$

Then the corrected sample locations are

$$\mathbf{x}_L^* = \mathbf{x}_L + \alpha \cdot F_{LR}^*(\mathbf{x}_L) \tag{13}$$

$$\mathbf{x}_R^* = \mathbf{x}_L + (1-\alpha) \cdot F_{RL}^*(\mathbf{x}_R). \tag{14}$$

Finally, each pixel is synthesized using linear blending by

$$I_D(\mathbf{x}_D) = (1-\alpha) \cdot I_L(\mathbf{x}_L^*) + \alpha \cdot I_R(\mathbf{x}_R^*). \tag{15}$$

5 Technical Evaluation

In our evaluation, we compared our spherical SfM pipeline to a typical incremental SfM pipeline. We choose COLMAP [31] for our comparison, as it is currently regarded as one of the best performing and most robust SfM systems. We ran our video sequences through COLMAP using known camera intrinsics, sequential matching, and loop-closure. In many cases, COLMAP would fail to complete the reconstruction. Some of these cases produced partial maps before terminating early, and other cases failed to initialize the reconstruction. Table 1 summarizes the sequences we tested.

In general, we found that COLMAP does not reliably reconstruct handheld circular motion sequences as needed for stereo panorama creation. Probably for this reason, Bertel et al. [3] recommend first subsampling the input video, reconstructing this sparse set of views with COLMAP, and then registering the remaining views followed by a final bundle adjustment. However, as discussed in Section 4.1, selecting this subset of views requires manual intervention.

Table 2: Survey of stereo panorama stitching methods.

Method	Images	Time	Hardware
Megastereo [30]	100 - 300	2-3 hours	Single camera
Jump [2]	16	231 s	16-camera rig
Schroers et al. [32]	16	20 min.	16-camera rig
Instant3D [15]	20 - 200	35 s	Stereo camera
Ours	200 - 700	3 - 10 min.	Single camera

In Table 2, we present a survey of methods that are designed specifically for reconstructing stereo panoramas from circular motion video. Our method is flexible in that it only requires a single handheld camera instead of a stereo camera [15] or multi-camera rig [2, 32]. In comparison to Megastereo [30], our reconstruction time is much faster (3-10 minutes compared to 2-3 hours).

5.1 Stereo panorama results

Figure 4 shows example stereo panoramas produced using our method, rendered as red-blue anaglyphs. We collected the videos with an iPhone, holding the phone in an outstretched hand while spinning roughly in a circle. Each video is about 20 to 30 seconds in length. Figure 5 shows a comparison of our method with Megastereo. Both methods produce effective stereo panoramas, although Megastereo uses negative disparities for distant parts of the scene, so we have larger disparities for objects close to the viewer.

Although we do not have ground truth camera poses to compare against, these panoramas demonstrate that our pipeline is able to reconstruct the camera trajectories accurately enough in order to produce clean, well-stitched images. In addition, inspection of the apparent parallax shows that nearby objects indeed have greater disparity than distant objects throughout the panoramas. Example disparity maps computed using semi-global block stereo matching [16, 17] are included in the supplemental material.

In some sequences, we noticed visual artifacts near the loop closure point. These are caused by the camera at the end of the sequence deviating too far from the circular trajectory. For example, if the camera moves too far inward or outward, or vertically or horizontally, then the spherical motion assumption breaks and the estimated relative pose between the first and last cameras in the sequence is inaccurate. The flow-based blending helps to hide this inaccuracy and smoothly blend through it.

6 User Evaluation

In addition to the technical evaluation, we performed a user evaluation. There were two main goals for the user evaluation. The first goal was to create a dataset of videos captured by casual users in order to further evaluate our approach. Here we were interested if there were any difference in results if we give the users no specific instructions (other than capture a panorama on a mobile phone) or if we give them detailed instructions to follow in the capturing process. Our intention here was to test whether untrained users capturing a panorama would naturally perform spherical motion suitable for reconstruction by our pipeline without us needing to instruct them about to the requirements of our system.

The second goal was to test whether users would perceive a difference when viewing the resulting stereo panoramas in a VR headset in contrast to a standard mono panorama. Here we wanted to test if the panoramas created by our pipeline would appear convincing and natural to the participants, and how the parallax effect in the stereo panoramas would be perceived by participants. The study received ethical approval by the University of Otago Ethics committee (D19/161).

Figure 4: Sample stereo panoramas produced using our pipeline, rendered as red-blue anaglyphs. Each input video was captured with a handheld iPhone, without the use of a tripod or any other device. From top to bottom: garden, grove, bay, courtyard, beach.

Figure 5: Anaglyph results compared to Megastereo [30] with closeups for better visibility. We achieve visually similar results compared to Megastereo. Please note that the disparities are computed differently (Megastereo's disparities are adjusted in a way that close objects are roughly at the depth of the image plane and have a near-zero disparity. Far objects are set to be behind the image plane and have a positive disparity). To be viewed with red/cyan glasses.

6.1 Study design

Task and Procedure Participants were asked to perform two tasks, a) capturing a stereo panorama and b) exploring a stereo panorama in a virtual reality headset.

Capture a 360 degree video The first task was to capture a 360 degree video in portrait mode. There were two conditions for this task. In condition A "Standard Panorama", the participants were asked to use a mobile phone to capture a video in portrait mode rotating around their own axis as if they would capture a panoramic image. In condition B: "Stereo Panorama", the participants were asked to use the mobile phone in portrait mode with one hand with the following instructions:

- Hold the phone away from their body with arm outstretched,
- Press the capture button with the other hand,
- Slowly turn around, trying to stay in the same position,
- Close the circle at 360 degrees.

After each condition we asked the participants to answer questions with regards to the capturing process. The order of the conditions for this task have not been randomised as we assume that most users have experience of capturing panoramas on a mobile phone (as confirmed in our demographics questionnaire).

360 degree video experience In the second task of the user evaluation, we let the participants experience a 360 degree rendering using a VR headset. There were two conditions used: condition A "Stereo" using the results of our casual stereo panorama computation and condition B: "Mono" simply showing the same panoramic image for both eyes. After each condition we asked the participants to fill in a questionnaire about the VR experience.

In addition, the participants were asked to fill a paper-based demographics questionnaire. Overall, the study took around 25 minutes to complete. Participants were able to decide not to take part in the project without any disadvantage to themselves.

The demographic data collected includes age, gender, ethnicity, and vision impairments, as well as familiarity with similar systems and technologies. The remaining experimental data includes answers regarding the capturing process including Likert-scale number responses to questions regarding the usability of the capturing procedure. No personally identifiable data was collected beyond those included in the demographic questionnaire, and every effort is made to ensure that no data can be linked to any individual participant.

For the experiment we used an iPhone X for the stereo panorama capture. For the VR experience, we used WebVR displayed on a mobile phone that was attached to an *Zeiss VR one* headset.

Hypotheses We aimed to evaluate whether users would find the casual stereo panorama capture process similar to a standard panorama capture as well as whether the visual quality of our results is sufficient to perceive a stereoscopic effect when viewing them on a VR headset. Thus, we had the following hypotheses:

- H1: Users will find the capturing process for capturing stereo panoramas similar in usability to standard panorama captures.
- H2: The quality of our casually captured stereo panoramas is good enough to convey a stereoscopic effect: Users will be more likely to perceive a stereoscopic effect when watching casual stereo panoramas in a VR headset compared to a standard mono panorama.

Participants We invited 12 participants that were recruited from undergraduate and graduate students (3 female, 9 male), with ages ranging from 21 to 31. All participants but one had either experience with the capture of panoramas (5) or knew about it (6).

6.2 Results

In order to compare both capturing methods, we used an excerpt of the SUS usability scale to capture usability aspects. The results show that the participants judged both capturing options very similarly. For both methods participants showed a slight tendency towards agreeing with "use frequently" and "learn quickly", disagreeing with "unnecessarily complex" and being "cumbersome". We used the Wilcoxon signed-rank test in order to measure statistical differences and did not find any statistically significant differences ($p > 0.05$). As a result we can confirm H1. In addition, we were interested if our method is robust enough to work with *casual* captures of stereo panoramas without specific instructions. We used the captured data from both conditions to test our method and achieve a similar success rate of over 90% for both conditions.

A Wilcoxon signed-rank test showed that there is a significant effect on the stereoscopic effect (p-value = 0.031) and descriptive statistics show that the participants are more likely to perceive a stereoscopic effect when being presented with the stereo panorama compared to the mono panorama. This confirms hypothesis H2.

6.3 Discussion

The results of the user evaluation confirmed our hypotheses.

For the capturing with and without instructions, we were able to confirm H1 and can assume that there is no difference in the usability of panorama capture for the users of our approach. It also worth to note that the participants judge both capturing task rather positively, with leaning towards using "use frequently" and "learn quickly" and rather disagreeing on "unnecessarily complex" and "cumbersome". The second aspect we were interested in with regards to the user capturing was if this data can be used to successfully create stereo panoramas with our method. 10 out of 12 people were able to capture videos for both conditions (with and without specific instructions) that could be used to successfully construct a stereo panorama. One sequence captured with condition A and one sequence captured with condition B could not be successfully used for creating a stereo panorama case. One failure case was caused by poor focus, and the other by failing to complete a full circle during the capture process. In summary, we experienced similar outputs for both capture conditions.

We were also able to confirm H2 showing that there is a significant difference in terms of the perceived stereoscopic effect. Participants were likely to experience a stereoscopic effect within the stereo panorama condition. However is worth to note that participants even had light stereoscopic effect feeling for the mono condition. Also, there was a higher variance in answers for the mono condition. Possibly participants were more undecided for the mono condition.

7 Conclusions and Future Work

While immersive viewing devices are becoming readily available, content creation for them remains a challenge. We have shown that spherical structure-from-motion can be used to enable robust, effective stereo panorama creation from casually captured video sequences. Our method does not require any specialised hardware – a single-camera mobile phone is all that is required. Compared to state-of-the-art methods for creating stereo panoramas from monocular video sequences, our method is substantially faster and the spherical constraint in the SfM component gives reliable reconstructions despite limited stereo baselines.

Future technical development could investigate relaxations of the spherical constraint. While this constraint is core to creating reliable stereo panoramas, it is not exactly met in real capture scenarios. In future work, the bundle adjustment phase (Sect. 4.1.5) could be extended to allow small deviations from this model.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation through Award #1464420 and the New Zealand Marsden Council through Grant UOO1724.

REFERENCES

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org.

[2] R. Anderson, D. Gallup, J. T. Barron, J. Kontkanen, N. Snavely, C. H. Esteban, S. Agarwal, and S. M. Seitz. Jump: Virtual reality video. *SIGGRAPH Asia*, 2016.

[3] T. Bertel and C. Richardt. MegaParallax. In *ACM SIGGRAPH 2018 Posters on - SIGGRAPH '18*, pp. 1–2. ACM Press, New York, New York, USA, 2018. doi: 10.1145/3230744.3230793

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, pp. 25–36. Springer, 2004.

[5] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 425–432. ACM, 2001.

[6] A. Chatterjee and V. Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 521–528, 2013.

[7] A. Davis, M. Levoy, and F. Durand. Unstructured light fields. In *Computer Graphics Forum*, vol. 31, pp. 305–314. Wiley Online Library, 2012.

[8] P. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. In *Rendering Techniques' 98*, pp. 105–116. Springer, 1998.

[9] Facebook. Surround360. http://github.com/facebook/Surround360.

[10] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker. Deepview: View synthesis with learned gradient descent. *arXiv preprint arXiv:1906.07316*, 2019.

[11] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deep Stereo: Learning to Predict New Views from the World's Imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5515–5524, jun 2016. doi: 10.1109/CVPR.2016.595

[12] H. Ha, S. Im, J. Park, H.-G. Jeon, and I. So Kweon. High-quality depth from uncalibrated small motion clip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5413–5421, 2016.

[13] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[14] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf. Casual 3d photography. *ACM Transactions on Graphics (TOG)*, 36(6):234, 2017.

[15] P. Hedman and J. Kopf. Instant 3D Photography. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 37(4):101:1–101:12, 2018.

[16] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 807–814. IEEE, 2005.

[17] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[18] J. Huang, Z. Chen, D. Ceylan, and H. Jin. 6-dof vr videos with a single 360-camera. In *2017 IEEE Virtual Reality (VR)*, pp. 37–44. IEEE, 2017.

[19] S. Im, H. Ha, G. Choe, H.-G. Jeon, K. Joo, and I. So Kweon. High quality structure from small motion for rolling shutter cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 837–845, 2015.

[20] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. S. Kweon. All-around depth from small motion with a spherical panoramic camera. In *European Conference on Computer Vision*, pp. 156–172. Springer, 2016.

[21] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):193, 2016.

[22] Y. Li, H.-Y. Shum, C.-K. Tang, and R. Szeliski. Stereo reconstruction from multiperspective panoramas. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):45–62, 2004.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[24] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[25] B. Luo, F. Xu, C. Richardt, and J. H. Yong. Parallax360: Stereoscopic 360Â° scene representation for head-motion parallax. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2018.2794071

[26] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 39–46, 1995.

[27] D. Nistér. Preemptive RANSAC for live structure and motion estimation. *Machine Vision and Applications*, 16(5):321–329, 2005.

[28] A. Pagani and D. Stricker. Structure from motion using full spherical panoramic cameras. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 375–382. IEEE, 2011.

[29] S. Peleg, M. Ben-Ezra, and Y. Pritch. Omnistereo: Panoramic stereo imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001. doi: 10.1109/34.910880

[30] C. Richardt and H. Zimmer. Megastereo : Constructing high-resolution stereo panoramas. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. doi: 10.1109/CVPR.2013.166

[31] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] C. Schroers, J.-C. Bazin, and A. Sorkine-Hornung. An omnistereoscopic video pipeline for capture and display of real-world vr. *ACM Transactions on Graphics*, 37(3):37:1–37:13, Aug. 2018. doi: 10.1145/3225150

[33] H.-Y. Shum and R. Szeliski. Stereo reconstruction from multiperspective panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 14–21. IEEE, 1999.

[34] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics (TOG)*, 25(3):835–846, 2006.

[35] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, 1991.

[36] A. Torii, A. Imiya, and N. Ohnishi. Two-and three-view geometry for spherical cameras. In *Proceedings of the sixth workshop on omnidirectional vision, camera networks and non-classical cameras*, pp. 81–88. Citeseer, 2005.

[37] J. Ventura. Structure from motion on a sphere. In *European Conference on Computer Vision*. Amsterdam, the Netherlands, 2016.

[38] K. Wilson and N. Snavely. Robust global translations with 1DSfM. In *European Conference on Computer Vision*, pp. 61–75. Springer, 2014.

[39] F. Yu and D. Gallup. 3d reconstruction from accidental motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3986–3993, 2014.

[40] F. Zhang and F. Liu. Casual stereoscopic panorama stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002–2010, 2015.

[41] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[42] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European conference on computer vision*, pp. 286–301. Springer, 2016.