On Depth Error from Spherical Camera Calibration within Omnidirectional Stereo Vision

Michael Groom*, Toby P. Breckon*†
Department of {Engineering* | Computer Science†} Durham University, Durham, UK

***Abstract*—As a depth sensing approach, whilst stereo vision provides a good compromise between accuracy and cost, a key limitation is the limited field of view of the conventional cameras that are used within most stereo configurations. By contrast, the use of spherical cameras within a stereo configuration offers omnidirectional stereo sensing. However, despite the presence of significant image distortion in spherical camera images, only very limited attempts have been made to study and quantify omnidirectional stereo depth accuracy.**

In this paper we construct such an omnidirectional stereo system that is capable of real-time 360° disparity map reconstruction as the basis for such a study. We first investigate the accuracy of using a standard spherical camera model for calibration combined with a longitude-latitude projection for omnidirectional stereo, and show that the depth error increases significantly as the angle from the camera optical axis approaches the limits of the camera field of view.

In contrast, we then consider an alternative calibration approach via the use of perspective undistortion with a conventional pinhole camera model allowing omnidirectional cameras to be mapped to a conventional rectilinear stereo formulation. We find that conversely this proposed approach exhibits improved depth accuracy at large angles from the camera optical axis when compared to omnidirectional stereo depth based on a spherical camera model calibration.

I. Introduction

Ever increasing advances in low-cost camera technology have seen the rise of widely-available consumer-grade spherical cameras, offering full omnidirectional spherical panoramic video via dual-lens 180°× 180° image capture at the cost of a few hundred dollars (Ricoh Theta V; Samsung Gear 360; Garmin Virb 360; Insta360 One X - ∼2020+).

The ready availability of such devices naturally facilitates the consideration of 360° depth recovery by using two within a conventional stereo configuration. Whilst prior work [1] has shown such an approach can offer all-round depth recovery for a range of generalised video surveillance [2] or vehicle autonomy [3] sensing scenarios, the accuracy of the depth obtained in relation to the challenges of spherical camera calibration remains largely unquantified.

Whilst the cost of alternative technologies for 3D scene depth recovery, such as LiDAR [4], remain prohibitively high for many applications, stereo vision based sensing offers a good compromise between accuracy and cost and is used widely in robotics [5], [6], [7], [8], object recognition [9], [10], and 3D scene reconstruction [11], [12]. However, a key limitation of common stereo vision solutions is the limited Field of View (FoV) afforded by the use of conventional cameras. Whilst the use of spherical cameras within a stereo configuration can overcome this issue to offer omnidirectional stereo sensing [1], [13], [14], such an approach is not without additional challenges in terms of effective camera placement, calibration and synchronisation.

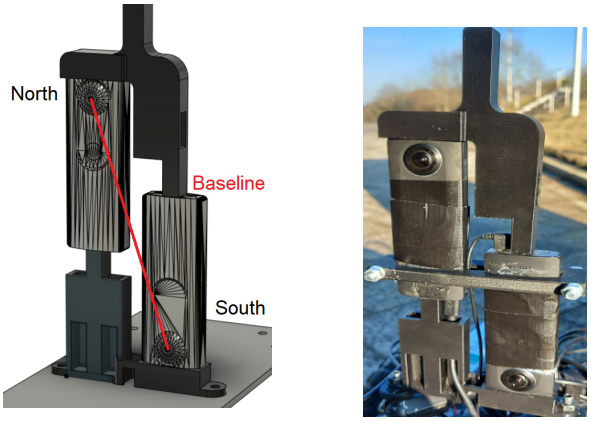

Fig. 1: Bi-polar spherical camera stereo configuration

Whilst recent work [1], [15], [16] has addressed a number of these issues, the choice and impact of camera calibration strategy remains largely unexplored [17], [18]. Following convention, the earlier work of Lin et al. [1] makes use of the spherical camera model proposed by Mei and Rives [19] to enable camera calibration as a means to the recovery of stereo disparity between two spherical cameras mounted in a bi-polar configuration. However, whilst this work focuses on effective camera mounting for use in autonomous vehicle sensing and the recovery of Cartesian from angular depth [1], it makes no attempt to evaluate the resulting disparity obtained via such a stereo setup in terms of quantifiable depth error.

This paper makes a number of contributions to address the limitations of prior work on this topic [1], [15], [16], and specifically gives further insight into the impact of alternative calibration approaches, as follows:

- a variant of the bi-polar omnidirectional stereo camera configuration of [1] is proposed to offer improved camera accessibility and stability in practical terms (Sec. III-A).
- the spherical camera calibration model proposed in [19] combined with a longitude-latitude projection is evaluated in the context of this omnidirectional stereo camera configuration and is shown to suffer from significant depth error at large angles from the optical axis (Sec. IV, with reference dataset release Sec. III-D).

- the alternative fish-eye camera calibration approach proposed in [20], using a perspective undistortion with a conventional pinhole camera model, is evaluated in the context of the same omnidirectional stereo camera configuration and is shown to offer significantly improved depth accuracy at large angles from the optical axis and hence more stable depth recovery across an omnidirectional FoV (Sec. IV, with reference dataset release Sec. III-D).
- the impact of alternative spherical camera calibration strategies [19], [20] on depth accuracy is verified over multiple stereo disparity estimation approaches [21], [22] achieving real-time performance from low-cost consumer-grade spherical camera hardware (maximal 14.4 fps performance, Sec. IV-B).

Overall, these contributions show that an appropriate choice of calibration approach, in terms of the equirectangular projection used for conversion to an rectilinear image composition, is crucial to the realisation of effective omnidirectional stereo with spherical cameras. Futhermore, they show that the most obvious choice for the practitioner, via the use of a spherical camera model such as Mei et al. [19], is not necessarily the most suitable choice for stereo accuracy.

II. Related Work

We present an overview of related work in terms of: Spherical Camera Models (Section II-A), Stereo Disparity Estimation (Section II-B) and Omnidirectional Stereo (Section II-C).

A. Spherical Camera Models

With reference to spherical imaging, Geyer and Daniilidis [23] introduce a basic projection model for panoramic images whilst Barreto and Araujo [24] establish a general model for image formation in central catadioptric images. The subsequent work of Mei and Rives [19] builds upon these models and formed a unified spherical camera projection model. A new parameter, ξ, is proposed that compensates for any difference between the idealised spherical camera model centre and the actual camera centre which is obtained via a calibration process similar to the seminal conventional stereo approach of [25]. Li [26] propose a similar model to [19] for use with multiple camera configurations offering a combined 360° FoV from a set of conventional image projections. As such, [26] did not include a term to model differences in ideal and actual spherical camera centres similar to the ξ parameter outlined in [19] but did reformulate the conventional rectilinear stereo problem for spherical cameras by defining disparity and depth for such spherical stereo systems.

Contemporary low-cost spherical cameras (such as the Ricoh Theta S cameras [27] used in this study, Section III-A) use a dual fish-eye lens configuration to achieve an 360° omnidirectional FoV. Scaramuzza et al. [20] present a camera calibration technique for omnidirectional cameras, assuming that the images are captured under a fish-eye distortion and proposed a novel image projection function to undistort them. This image projection function is a Taylor series expansion, whose coefficients are found during a calibration process, similar to the seminal work of Zhang [25]. The work of Scaramuzza et al. [20] is subsequently refined by Urban et al. [28] to achieve notably more stable, robust and accurate camera calibration (up to a factor of 7). Such fish-eye undistortion approaches essentially enable images produced by spherical cameras to be treated as conventional rectilinear images.

B. Stereo Correspondence

In many stereo vision applications, such as automotive depth sensing [7], a stereo approach must be chosen that is a trade-off between accuracy and real-time performance. Of late, the KITTI stereo benchmark (2015) [29] provides a leaderboard for contemporary state of the art in stereo correspondence approaches with testing performed on challenging outdoor scene environments. In recent years, deep learning based stereo correspondence algorithms, [30], [22], have dominated the this benchmark.

Within this study, PSMNet [22] is identified as a representative deep learning based approach due to a favourable trade-off between accuracy and real-time performance within the KITTI benchmark [29]. In addition, the seminal Semi-Global Block Matching (SGBM) [21] approach is selected in order to facilitate direct comparison to the previous omnidirectional stereo work of Lin et al. [1] and due to its widespread use in real-time automotive sensing applications [7], [10], [31].

C. Omnidirectional Stereo

Contemporary work addressing calibration explicitly within an overall omnidirectional stereo approach is generally sparse within the literature [15], [1], [13], [14].

Earlier work by Ma et al. [15] uses the spherical camera of [26], resulting in stereo disparity maps that *"look a little messy"*[15]. This work is subsequently built on by Lin et al. [1] using the general spherical camera model of Mei et al. [19] to again present stereo disparity of variable qualitative appearance.

Goa and Shen [13] use a dual fish-eye lens stereo configuration similar to that used here but use a lens-specific fish-eye camera calibration method that is reliant on manufacturer data for the optical lens characteristics in order to assist with the intrinsic calibration and hence develop an online self-calibration approach for extrinsic parameter estimation.

Won et al. [14] propose an end-to-end deep neural network approach for omnidirectional stereo depth estimation that itself implicitly encompasses the intrinsic and extrinsic camera calibration steps. This approach extracts features from four orthogonal fish-eye camera views and uses spherical sweeping and cost volume aggregation to produce omnidirectional disparity maps. Related end-to-end deep-learning enabled approaches have also been proposed by [32], [33], [34], [35] but again do not explicitly address calibration in the traditional sense.

III. Methodology

By contrast to prior work, here we present a methodology that explicitly addresses the issue of calibration within omnidirectional stereo and its impact on stereo depth accuracy.

A. Camera Configuration

Our camera configuration is a variant of that used by Lin et al. [1] and similarly uses a top-to-bottom (North-South) vertical camera mounting solution. However, to afford improved camera stability (and also accessibility in practical terms) the vertical baseline is shortened compared to that of Lin et al. [1] and a horizontal offset component between the cameras is introduced. The arrangement maintains the camera blind spots in areas of no interest whilst improving stability of the resultant shorter height rig for on-vehicle mounting (Fig. 1). To compensate for this horizontal offset, the camera images are rotated to create a single vertical stereo system along a diagonal baseline between the camera optical centres (Fig. 1).

Our configuration uses two Ricoh Theta S cameras [27] interfaced via USB 3.0 to a Intel Core i7-6700HQ 2.60GHz CPU / NVidia GeForce GTX 970M GPU. Each camera consists of two fish-eye lenses with a ($\approx 190^\circ \times \approx 190^\circ$) FoV that combine to provide 360° omnidirectional scene coverage per camera via one 640×640 image per camera lens. Following the notation of [1], cameras will be denoted as North-Front (N_{front}), North-Back (N_{back}), South-Front (S_{front}), and South-Back (S_{back}) (Fig. 1).

B. Spherical Camera Calibration

Conventional stereo configurations use two cameras, $\{L, R\}$ modelled using the pinhole camera model [36]. A point $\mathbf{X} = (X, Y, Z)^T$ in the world coordinate system is projected to a point $\mathbf{x}$ on the image plane. Stereo correspondence approaches (Section II-B) then identify feature matches between corresponding points in the two image planes, from which subsequent triangulation enables the estimation of scene depth, Z. Within this formulation, epipolar geometry is used to constrain corresponding feature locations (i.e. pixels) to the same vertical position (i.e. row) in the image, hence reducing the correspondence search space significantly and is expressed mathematically as:

$$x_L^T \mathbf{F} x_R = 0\,, \tag{1}$$

where x_L and x_R are projections of the same world point, $\mathbf{X}$, onto the left and right image planes whilst the fundamental matrix, $\mathbf{F}$, denotes the geometric relationship between these two image planes. The fundamental matrix encapsulates both the intrinsic (camera projection matrix, $\mathbf{K}$) and extrinsic (essential matrix, $\mathbf{E}$) parameters of a given stereo configuration [36] in addition to the radial and tangential lens distortion coefficients, $\mathbf{D}$, which are recovered during camera calibration [25].

Contemporary dense stereo correspondence approaches (Section II-B) rely upon this epipolar constraint to facilitate image rectification under the pinhole camera model such that feature correspondences will occur along corresponding image rows in rectified stereo images.

However, within spherical images the epipolar lines instead appear as conics around the image centre and limit the use of epipolar geometry to similarly constrain feature correspondence in this way. If unaddressed, this in turn precludes the use of contemporary stereo correspondence approaches on spherical images.

To overcome this issue, spherical images must first be transformed such that their epipolar lines no longer appear as conics and regular image rectification can be performed to provide row-wise feature correspondences. We investigate two such image calibration methods to address this (Sections B.1 / B.2), as a conduit to onward stereo depth recovery (Section III-C).

B.1 Longitude-Latitude Projection (Mei-Rives):

Within the spherical camera model proposed by Mei and Rives [19], points are projected onto a spherical surface centred around the camera centre. Firstly, a point $\mathbf{X} = (X, Y, Z)^T$ is projected onto a unit sphere by:

$$X_s = \frac{X}{\|X\|}\,. \tag{2}$$

This point is then transformed to a new reference frame centred at $Cp = (0, 0, \xi)^T$ as $\mathbf{X} = (X, Y, Z + \xi)^T$. The parameter ξ is used to model the difference between the model and real camera centres. This point is then projected to the point m on the normalised plane with coordinates $m = (\frac{X}{Z+\xi}, \frac{Y}{Z+\xi}, 1)^T$. Finally, a generalised camera matrix $\mathbf{K}$ is used to project the point m from the normalised plane to the image plane. $\mathbf{K}$ is defined as:

$$\mathbf{K} = \begin{bmatrix} f_1\eta & f_1\eta\alpha & u_0 \\ 0 & f_2\eta & v_0 \\ 0 & 0 & 1 \end{bmatrix}\,, \tag{3}$$

where (u_0, v_0) are the coordinates of the principal point P, f_1 and f_2 are the focal lengths in pixels, α is the skew and η is dependant on lens geometry [19]. Using the approach of [19], radial and tangential lens distortions are corrected by applying a final undistortion, $\mathbf{D} = (k_1, k_2, p_1, p_2)$, where k and p are radial and tangential distortion coefficients respectively.

Li [26] proposed that a longitude-latitude projection could be used to transform spherical images, transforming conic epipolar lines to straight lines as desired. A longitude-latitude projection is defined as:

$$\begin{aligned} u &= f_s\theta\,, \\ v &= f_s\phi\,, \end{aligned} \tag{4}$$

where θ and ϕ are the polar and azimuth angles if the two epipoles in the spherical image are defined as two poles of the coordinate system. As noted by Li [26] when first proposing the use of longitude-latitude projections for transforming spherical images, this approach produces regions near the epipoles where the error in estimated depth becomes large. The epipole is the point at which the baseline intersects the spherical image plane. After the longitude-latitude projection is performed on a spherical image, the epipoles are placed at the vertical edges of the image.

We use the same technique as [1] to implement the longitude-latitude projection such that the projection matrix is defined:

$$[P] = \begin{bmatrix} \frac{|u|}{\theta_u} & 0 & 0 \\ 0 & \frac{|v|}{\phi_v} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (5)$$

where $|u|$ is the image width, $|v|$ is the image height, θ_u is the horizontal FoV of the camera and ϕ_v is the vertical FoV. For our Ricoh Theta S cameras (Section III-A), $\theta_u = \phi_v = \frac{19}{18}\pi$ rad and $|u| = |v| = 640$ pixels, giving the radius of spherical camera model $f_s \approx 193$ px/rad in Eqn. (4) and Eqn. (9).

B.2 Perspective Undistortion (Scaramuzza-Urban):

The work of Scaramuzza et al. [20], with subsequent refinement by Urban et al. [28], allows spherical cameras to be treated as conventional pinhole cameras with rectilinear image planes by proposing that a Taylor series can be used to approximate a mapping function to project spherical to rectilinear images. A 2D image point $\mathbf{m} = [u, v]^T$ can be mapped to its corresponding scene point $\mathbf{X_c} = [X_c, Y_c, Z_c]^T$ in the camera coordinate system through an imaging function, $g()$:

$$\mathbf{X_c} = \lambda g(\mathbf{m}) = \lambda(u, v, f(\rho))^T, \quad (6)$$

with $\lambda > 0$ and $\rho = \sqrt{u^2 + v^2}$ as the radial Euclidean distance to the image centre. Instead of defining a specific mapping function $g()$, [20] instead approximated it with a Taylor series:

$$f(\rho) = a_0 + a_2\rho^2 + \cdots + a_n\rho^n. \quad (7)$$

This imaging function transforms the image to a rectilinear geometry such that the epipolar lines are now straight, and thus contemporary dense stereo correspondence approaches can be applied to the stereo image pair. However, to represent a hemispherical view using a perspective projection, an infinitely large image is required, which results in this imaging function limiting the FoV of the omnidirectional camera to less than $180°$.

The Taylor series coefficients from Eqn. (7) are used to create lookup tables for applying the resultant perspective undistortion. When performing the undistortion step, a scale factor must be chosen [20]. The scale factor alters the distance between the undistorted image plane and the camera centre. This causes the scale factor to act akin to a change in focal length (i.e. zoom) that alters the camera FoV slightly, as shown in Fig. 2. The largest possible FoV available after this transformation step is desirable to make the stereo configuration as effective as possible in terms of retaining near omnidirectional scene coverage. However, this introduces a significant loss of detail from the centre of the source image and practically makes stereo calibration with a conventional calibration target difficult (Fig. 2).

To partially address this issue, our output image resolution is increased by a factor of two (i.e. 1280×1280 from 640×640 resolution) in order to minimise the loss of detail suffered. The perspective undistortion is then applied to these padded images, which facilitate greater detail retention in the centre of the image via this increased pixel sampling to thus accommodate a wider FoV. In practice, a scale factor $= 5.0$ is empirically found to produce optimal results in terms of this trade-off, and offers only a marginal reduction from a full $180°$ FoV for a single camera view (Fig. 2).

Source Image SF = 2.5 SF = 5.0 SF = 7.5

Fig. 2: Exemplar impact of differing scale factors (SF) used during perspective undistortion of spherical images via Scaramuzza-Urban.

C. Stereo Depth Estimation

Post calibration (Section III-B), stereo disparity is first obtained using three variant estimation approaches: Semi-Global Block Matching (SGBM) [21]; SGBM with Weighted Least Squares filtering (SGBM-WLS); and PSMNet [22].

SGBM enables ease of comparison with the previous omnidirectional stereo work of Lin et al. [1] with our implementation using the Birchfield-Tomasi sub-pixel metric [37] and additional post-processing steps including uniqueness checks, speckle filtering and quadratic interpolation.

SGBM-WLS additionally filters disparity maps from SGBM using a Weighted Least Squares (WLS) filter in the form of a Fast Global Smoother [38] (parameters used: $\lambda = 800$, $\sigma = 1.2$). All additional SGBM and SGBM-WLS parameters match those used in [39] within the KITTI benchmark [29].

We use the original pre-trained PSMNet model [22], trained on the KITTI benchmark dataset [29] and crop images from 1280×1280 to 1280×400 for the Scaramuzza-Urban method results due to GPU RAM limitations.

From stereo disparity, calculating scene depth, Z, for a conventional stereo configuration is performed via the simple relation:

$$Z = \frac{bf}{d}, \quad (8)$$

where Z is the depth, b is the baseline distance between the camera left/right image pair, f is the camera focal length, and d is the disparity, $d = u_l - u_r$, calculated via one of the aforementioned disparity estimation approaches from corresponding pixel positions, u_l and u_r, in the left and right images.

By contrast, calculating depth from disparity for a spherical camera stereo configuration is more complex with a full derivation presented in [1]. In summary, for our vertical spherical stereo camera configuration (Section III-A) the distance, ρ_N, can be calculated with reference to the centre of the North camera, N_i for $i \in \{Front, Back\}$, as:

$$\rho_N = b\frac{\sin(v_S/f_s)}{\sin(d/f_s)}, \quad (9)$$

where v_S is the vertical pixel coordinate in the South camera image and both d and f_s are defined as the vertical disparity and radius of the spherical camera model as set out in [1]. Subsequent conversion to depth, Z, is then performed via:

$$Z = \rho_N \sin(\phi_N) \cos(\theta_N) \tag{10}$$

where $\{\theta_N, \phi_N, \rho_N\}$ are the polar coordinates of our 3D point, $\mathbf{X}$, with respect to the North camera following [1].

D. Experimental Setup

Calibration via the Mei and Rives [19] approach (henceforth denoted as Mei-Rives) as well as the Scaramuzza et al. [20] calibration approach, inclusive of the refinements of Urban et al. [28], (henceforth denoted as Scaramuzza-Urban) is performed using a planar 8×6 chessboard (grid separation size: $80.8 \times 80.8mm$) to collect a total of 150 image pairs for each of the front and back camera pairs from which 50% are randomly selected for use in the calibration optimisation process[1]. Chessboard patterns are automatically detected using corner detection to sub-pixel accuracy [40], [41] as per [1].

Mei-Rives: is performed using the implementation of [40], which extends the work of Zhang [25] based upon [42], to obtain the spherical and extrinsic camera parameters.

Scaramuzza-Urban: is performed using the implementation of [43], to obtain the coefficients for the Taylor series that describes the imaging function of Eqn. (7). Stereo calibration is then performed on the rectilinear images obtained (see example in Fig. 3, lower centre column) using standard camera calibration [44], [25] to obtain intrinsic and extrinsic parameters.

For both approaches, due to numerical instability in the Levenberg–Marquardt optimisation, optimal stereo calibration is achieved via multiple optimisation runs with differing termination criteria (i.e. # iterations, ϵ change in parameters). The final calibration results are selected based on the lowest RMS error (Table I) with a secondary qualitative check of post-rectification vertical alignment.

Resultant stereo depth sensing accuracy is measured, for each of Mei-Rives [19] and Scaramuzza-Urban [20], [28], using a planar target at a known distance from the cameras. Two sets of images pairs with the target, spanning the camera FoV, at 3 metre and 2 metre distances are captured containing 67 and 75 images pairs respectively[1].

Post calibration, disparity is recovered using each of SGBM, SGBM-WLS, and PSMNet from which stereo depth is then calculated (Section III-C). The four corners of the planar target are then identified within the resulting stereo depth map and the mean depth across the interior region of the target calculated in addition to the centre position of the target relative to the image centre.

IV. Results & Discussion

We present our results in terms of the numerical calibration accuracy obtained for stereo calibration via each of the Mei-Rives and Scaramuzza-Urban approaches (Section IV-A) and subsequently the impact on stereo depth accuracy for each calibration approach (Section IV-B) with subsequent additional discussion in Section (Section IV-C).

[1]Supporting dataset release: http://doi.org/10.15128/r13197xm075.

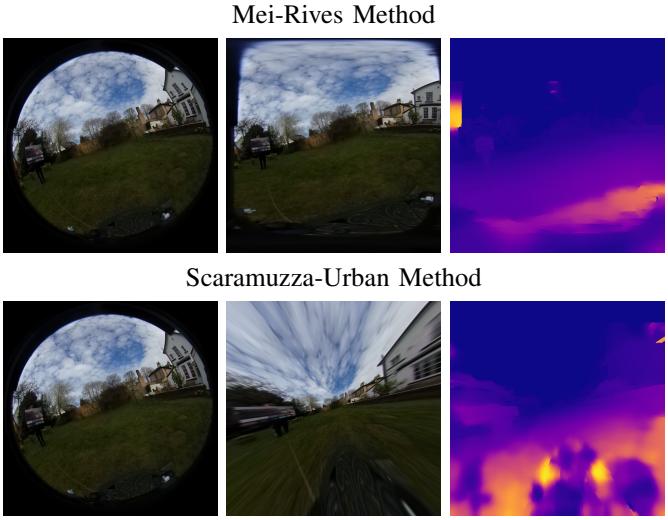

Fig. 3: Spherical camera images before (left) and after rectification with each of the approaches (middle) and resulting stereo disparity from SGBM-WLS (right).

Calibration Method	Mean Reprojection Error	RMS Reprojection Error
Mei-Rives	0.135	0.182
Scaramuzza-Urban	0.557	0.580

TABLE I: Stereo calibration results using the Mei-Rives and Scaramuzza-Urban Methods. Reprojection errors and RMS values are averaged across the four cameras.

A. Calibration Accuracy

Table I presents the mean absolute reprojection error and Root Mean Squared (RMS) reprojection error results calculated from the stereo calibration performed on imagery with either of the Mei-Rives or Scaramuzza-Urban calibration methods applied. It can be observed that the errors for the Mei-Rives are significantly lower than those achieved for the Scaramuzza-Urban method (Table I). On this basis alone, it would appear that the Mei-Rives approach is likely to result in lower disparity estimation errors than Scaramuzza-Urban and hence offer superior stereo depth accuracy.

B. Stereo Depth Accuracy

Following the experimental setup of Section III-D, we present the absolute error in stereo depth, for results obtained on either of the Mei-Rives or Scaramuzza-Urban calibration methods, as a function of the distance from the image centre (Fig. 4) for the 2 and 3 metre target placement experiments respectively. To ensure a fair comparison, a total of 8 image pairs are removed from the 3m image set as the target was not visible in the reduced FoV of the Scaramuzza-Urban approach. It

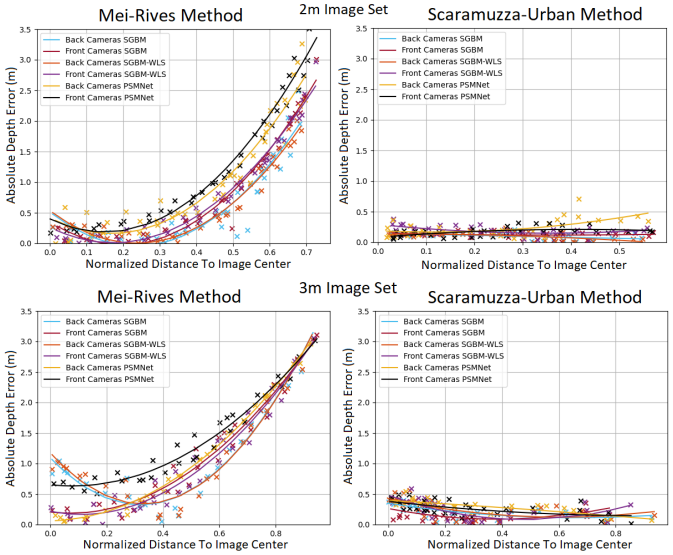

Fig. 4: Absolute Depth Error (metres, m) vs. Normalized Horizontal Distance to Image Centre on the 2 metre (upper) and 3 metre (lower) experimental image sets. Distance to image centre is normalized by the half the image width.

can be observed that whilst the resultant depth error for the Scaramuzza-Urban method remains stable (Fig. 4), the depth error for the Mei-Rives method increases significantly as a function of the distance from the centre of the image (towards the epipoles). This trend is repeated for all of the stereo disparity estimation approaches used (SGBM, SGBM-WLS, PSMNet; Fig. 4).

Further qualitative analysis is provided via Fig. 5 where the line of best fit obtained from depth error measurements obtained using each approach with SGBM-WLS in Fig. 4 (lower), is overlain onto the corresponding disparity map. This illustrates how depth accuracy varies significantly across the camera FoV when using a spherical camera model with a longitude-latitude projection (Mei-Rives, Fig. 5 left). The original spherical camera images, resultant rectification via equirectangular projection and resulting stereo disparity corresponding to the stereo error visualisation of Fig. 5 are additionally shown in Fig. 3.

As an ancillary result, with reference to our stereo configuration (Section III-A), processing throughput for the Mei-Rives/SGBM method combination can achieve 14.4 fps (outperforming [1]) whilst Scaramuzza-Urban/SGBM can only achieve 5.8 fps (matching [1]).

C. Discussion

Whilst the numerical calibration error for the Mei-Rives method is tolerable (Table I), the instability of depth error across the stereo FoV shown in Fig. 4 is too significant for practical omnidirectional stereo sensing based on this approach.

This phenomenon is entirely attributable to points appearing close to the epipoles within a stereo formulation of the Mei-Rives spherical camera model as previously identified in the original work of [26]. Such a high level of inaccuracy towards the edges of the images (i.e. $>$ 1 metre at 50-60% of distance from centre to image edge) means that although the use of a fish-eye lens and the Mei-Rives spherical camera model allows for stereo disparity information to be calculated for a large FoV in a omnidirectional manner, a significant portion of the generated disparity map is effectively useless.

By contrast, although similarly tolerable, the higher numerical calibration error for the Scaramuzza-Urban method (Table I) instead translates to a stable depth error across the stereo FoV shown in Fig. 4.

Although the Mei-Rives method offers both higher throughput and a greater FoV owing to its inherent projection model, it is clear that the Scaramuzza-Urban method results in a more viable omndirectional stereo sensing solution.

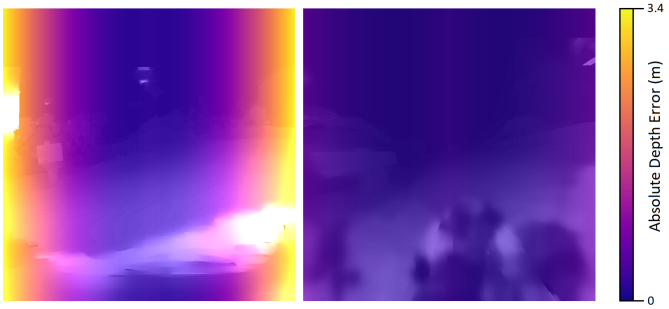

Fig. 5: Absolute Depth Error (m) overlain onto disparity maps generated using Mei-Rives (left) and Scaramuzza-Urban (right) with SGBM-WLS stereo disparity estimation .

V. Conclusion

Our work constructs a bi-polar omnidirectional stereo camera configuration using consumer-grade hardware, extending the work of [1] in terms of practicality and processing performance, as the basis for studying the impact of calibration on the resultant stereo depth accuracy.

Two calibration approaches are investigated, that vary in terms of the camera model used for equirectangular projection to a rectilinear image composition. We investigate the work of Mei and Rives [19] (Mei-Rives), that uses a spherical camera calibration model combined with a longitude-latitude projection, and compare this to the work of Scaramuzza et al. [20] / Urban et al. [28] (Scaramuzza-Urban), that uses a perspective undistortion with a conventional pinhole camera model.

Over multiple experiments and with multiple stereo disparity estimation approaches, we show that the Mei-Rives [19] approach results in significant stereo depth error at large angles from the optical axis whilst the Scaramuzza-Urban [20], [28] approach offer significantly improved depth accuracy at large angles from the optical axis and hence more stable depth recovery across an omnidirectional FoV.

Future work will investigate the use of auxiliary sensors within an extended omnidirectional stereo accuracy evaluation in addition to the use of targetless calibration approaches.

REFERENCES

[1] K. Lin and T. Breckon, "Real-time low-cost omni-directional stereo vision via bi-polar spherical cameras," in *Proc. Int. Conf. Image Anal. Recognit.* Springer, June 2018, pp. 315–325.

[2] M. Breszcz and T. Breckon, "Real-time construction and visualization of drift-free video mosaics from unconstrained camera motion," *IET J. Engineering*, vol. 2015, no. 16, pp. 1–12, August 2015.

[3] G. Payen de La Garanderie, A. Atapour-Abarghouei, and T. Breckon, "Eliminating the dreaded blind spot: Adapting 3d object detection and monocular depth estimation to 360° panoramic imagery," in *Proc. Euro. Conf. on Comput. Vis.* Springer, September 2018, pp. 812–830.

[4] L. Li, K. Ismail, H. Shum, and T. Breckon, "Durlar: A high-fidelity 128-channel lidar dataset with panoramic ambient and reflectivity imagery for multi-modal autonomous driving applications," in *Proc. Int. Conf. on 3D Vision*. IEEE, December 2021, pp. 1227–1237.

[5] M. J. Schuster, K. Schmid, C. Brand, and M. Beetz, "Distributed stereo vision-based 6d localization and mapping for multi-robot teams," *J. Field Robotics*, vol. 36, no. 2, pp. 305–332, 2019.

[6] C. Holder and T. Breckon, "Encoding stereoscopic depth features for scene understanding in off-road environments," in *Proc. Int. Conf. Image Anal. Recognit.* Springer, June 2018, pp. 427–434.

[7] F. Mroz and T. Breckon, "An empirical comparison of real-time dense stereo approaches for use in the automotive environment," *EURASIP J. Image Video Process.*, vol. 2012, no. 13, pp. 1–19, 2012.

[8] T. Kriechbaumer, K. Blackburn, T. Breckon, O. Hamilton, and M. Riva-Casado, "Quantitative evaluation of stereo visual odometry for autonomous vessel localisation in inland waterway sensing applications," *Sensors*, vol. 15, no. 12, pp. 31 869–31 887, December 2015.

[9] Y.-C. Du, M. Muslikhin, T.-H. Hsieh, and M.-S. Wang, "Stereo vision-based object recognition and manipulation by regions with convolutional neural network," *Electronics*, vol. 9, no. 2, p. 210, 2020.

[10] O. Hamilton, T. Breckon, X. Bai, and S. Kamata, "A foreground object based quantitative assessment of dense stereo approaches for use in automotive environments," in *Proc. Int. Conf. on Image Process.* IEEE, September 2013, pp. 418–422.

[11] R. Fan, Y. Liu, X. Yang, M. J. Bocus, N. Dahnoun, and S. Tancock, "Real-time stereo vision for road surface 3-D reconstruction," in *2018 IEEE Int. Conf. Imaging Syst. Tech. Proc.*, 2018, pp. 1–6.

[12] P. Cavestany, A. Rodriguez, H. Martinez-Barbera, and T. Breckon, "Improved 3D sparse maps for high-performance structure from motion with low-cost omnidirectional robots," in *Proc. Int. Conf. on Image Process.* IEEE, September 2015, pp. 4927–4931.

[13] W. Gao and S. Shen, "Dual-fisheye omnidirectional stereo," in *IEEE Int. Conf. Intell. Robots Syst.*, 2017, pp. 6715–6722.

[14] C. Won, J. Ryu, and J. Lim, "OmniMVS: End-to-end learning for omnidirectional stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8987–8996.

[15] C. Ma, L. Shi, H. Huang, and M. Yan, "3D Reconstruction from Full-view Fisheye Camera," *arXiv preprint arXiv:1506.06273*, 2015.

[16] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski, "Low-cost 360 stereo photography and video capture," *ACM Trans on Graphics*, vol. 36, no. 4, pp. 1–12, jul 2017.

[17] J. Heller and T. Pajdla, "Stereographic rectification of omnidirectional stereo pairs," in *Proc. Conf. Comput. Vis. Pattern Recognit.* IEEE, 2009, pp. 1414–1421.

[18] Z. Arican and P. Frossard, "Dense disparity estimation from omnidirectional images," in *Proc. Conf. on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 399–404.

[19] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *Proc. IEEE Int. Conf. Robot. and Automat.*, 2007, pp. 3945–3950.

[20] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in *Proc. IEEE Int. Conf. Comput. Vision Syst.*, 2006, pp. 45–45.

[21] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2008.

[22] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[23] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical implications," in *Eur. Conf. Comput. Vis.* Springer, 2000, pp. 445–461.

[24] J. P. Barreto and H. Araujo, "Issues on the geometry of central catadioptric image formation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2001, pp. II–II.

[25] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[26] S. Li, "Real-time spherical stereo," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 1046–1049.

[27] Ricoh, "Ricoh Theta S - User Guide," 2015, Accessed on: March. 22, 2021. [Online]. Available: https://support.theta360.com/uk/manual/s/

[28] S. Urban, J. Leitloff, and S. Hinz, "Improved wide-angle, fisheye and omnidirectional camera calibration," *ISPRS J. Photogrammetry Remote Sens.*, vol. 108, pp. 72–79, 2015.

[29] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.

[30] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, T. Drummond, H. Li, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *arXiv preprint arXiv:2010.13501*, 2020.

[31] O. Hamilton and T. Breckon, "Generalized dynamic object removal for dense stereo vision based scene mapping using synthesised optical flow," in *Proc. Int. Conf. on Image Process.* IEEE, September 2016, pp. 3439–3443.

[32] P. K. Lai, S. Xie, J. Lang, and R. Laganière, "Real-time panoramic depth maps from omni-directional stereo images for 6 dof videos in virtual reality," in *IEEE Conf. Virtual Real. 3D User Interfaces*. IEEE, 2019, pp. 405–412.

[33] N. Zioulis, A. Karakottas, D. Zarpalas, F. Alvarez, and P. Daras, "Spherical view synthesis for self-supervised 360 depth estimation," in *Int. Conf. on 3D Vision*. IEEE, 2019, pp. 690–699.

[34] X. Cheng, P. Wang, Y. Zhou, C. Guan, and R. Yang, "Omnidirectional depth extension networks," in *Int. Conf. Robot. Automat.* IEEE, 2020, pp. 589–595.

[35] Z. Lai, D. Chen, and K. Su, "OLANET: self-supervised 360° depth estimation with effective distortion-aware view synthesis and L1 smooth regularization," in *Int. Conf. on Multimedia and Expo*. IEEE, 2021, pp. 1–6.

[36] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[37] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 401–406, 1998.

[38] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5638–5653, 2014.

[39] OpenCV, M. Menze, and A. Geiger, "OpenCV Semi-Global Block Matching [OCV-SGBM] - KITTI," Accessed on: Dec. 26, 2020. [Online]. Available: http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php

[40] "OpenCV Omnidirectional Camara Calibration Tutorial, 4.5.1 Edition." [Online]. Available: https://docs.opencv.org/4.5.1/dd/d12/tutorial_omnidir_calib_main.html

[41] W. Förstner and E. Gülch, "A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features," in *ISPRS Intercommission Workshop*, 1987, pp. 281–305.

[42] B. Li, L. Heng, K. Koser, and M. Pollefeys, "A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern," in *2013 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1301–1307.

[43] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *IEEE Int. Conf. Intell. Robots Syst.*, 2006, pp. 5695–5701.

[44] "OpenCV Camera Calibration and 3D Reconstruction Tutorial, 4.5.1 Edition." [Online]. Available: https://docs.opencv.org/4.5.1/d6/d55/tutorial_table_of_content_calib3d.html