Dense 3D Reconstruction from Two Spherical Images via Optical Flow-based Equirectangular Epipolar Rectification

Sarthak Pathak, Alessandro Moro, Atsushi Yamashita, and Hajime Asama
Department of Precision Engineering, The University of Tokyo
{pathak,moro,yamashita,asama}@robot.t.u-tokyo.ac.jp

***Abstract*—In this research, a technique for dense 3D reconstruction from two spherical images clicked at displaced positions near a large structure is formulated. The technique is based on the use of global variational information i.e. the dense optical flow field in a unique rectification based refinement of the epipolar geometry between the two images in a vertically displaced orientation. A non-linear minimization is used to globally align the 2D equirectangular optical flow field (i.e. pixel displacements). The magnitude component of the resultant optical flow field can directly be converted to a dense 3D reconstruction. Thus, the epipolar geometry as well as the 3D structure can be estimated in a single minimization. This method could be useful in measurement and reconstruction of large structures such as bridges, etc. using a robot equipped with a spherical camera, thus helping in their inspection and maintenance.**

(a) 2D Equirectangular (b) Spherical

Fig. 1. Spherical Images encode information from all directions and can be represented in the (a) 2D Equirectangular and (b) Spherical Projection.

I. INTRODUCTION

Spherical cameras that can capture information from all directions in real-time (Figure 1) have many applications in robotics. One popular application of these cameras is omnidirectional 3D reconstruction of various structures. Particularly, high quality 3D surface reconstruction of large infrastructures such as bridges using robots can help in creating 'digital copies' of them for easier, offline inspection [1], [2], which can be a very tedious task if done manually. For such purposes, the surface needs to be mapped from up-close in order to locate cracks and other surface defects. Hence, a spherical camera which can see the entire structure at once from up-close is particularly useful and advantageous for localization and reconstruction. Keeping these objectives in mind, a technique for dense reconstruction of large structures from spherical images is formulated. This research focuses on dense reconstruction from two images clicked near a structure.

Spherical cameras have several important advantages over perspective cameras. One obvious advantage being that they can see, and hence measure the entire structure at once. Moreover, due to this, translational displacement in spherical cameras causes a significantly lesser change in image content, as opposed to perspective cameras. This allows the use of global, dense information to estimate the properties of motion, which may not be possible otherwise. Considering these advantages, a technique for dense reconstruction using two spherical images is proposed.

In order to do so, two steps are required. The first involves a precise estimation of the epipolar geometry, or the 5 degree of freedom motion parameters between the two camera positions (3 for rotation, 2 for direction of translation). Following this, the dense pixel-to-pixel disparities need to be estimated along epipolar lines and converted to 3D point locations. Usually, the first step can be done by matching sparse interest points in both images. However, this involves noise and outliers and hence is not globally consistent to produce a dense reconstruction of good quality. Hence, non-linear refinement steps are typically used to align the epipolar lines. However, this too is a sparse procedure affected by remaining outliers.

Instead, [3] showed how variational techniques like dense optical flow can help in accurate epipolar estimation for perspective cameras, if the images have a high consistency of content. It is mentioned that the regularization steps of dense optical flow algorithms prevent any local outliers and hence can give a very accurate epipolar geometry estimate. The concept is further extended to jointly optimize epipolar geometry and dense optical flow. Hence, in this research, a similar technique is applied to spherical images. Two spherical images are rectified in a vertically displaced orientation using a novel dense refinement strategy in the 2D equirectangular projection allowing the use of planar image processing techniques, completely bypassing the high distortion.

In our other work [4], the use of dense optical flow to refine epipolar geometry was also shown. However, there could be a drop in accuracy because the optical flow vectors are estimated on a 2D equirectangular image and projected to a 3D unit spherical surface. Instead, in this research, it is ensured that the optimization remains in a 2D equirectangular format and uses quantities that are directly computed on the 2D image.

The other advantage of doing this is that the resulting optical flow field also forms the dense disparity map and can directly be converted to a 3D reconstruction. Thus, epipolar geometry refinement and the 3D reconstruction can be obtained in a single optimization step.

II. Related Research and Our Approach

Many 3D reconstruction methods involving spherical cameras assume known camera positions, achieved either by calibration, or mechanical alignment of the two camera positions [5], [6], [7]. Specifically, [6] and [7] assume the camera is displaced in the vertical direction. Following this, they expand their spherical images to the 2D equirectangular projection along the vertical axis. It follows that all pixel displacements on the spherical image will follow vertical lines in the equirectangular image. Thus, disparity estimation is made much easier by searching within the same vertical line on the equirectangular images, as shown in Figure 2. This is a very convenient as it allows direct use of 2D image processing techniques on a spherical image, unaffected by the distortion. However, this cannot be applied directly to arbitrary robot motion. Hence, in this research, a method that can automatically determine the precise camera motion and rectify the images to such a vertical orientation is proposed. This is possible with spherical images because they can be rotated to any orientation without loss of information.

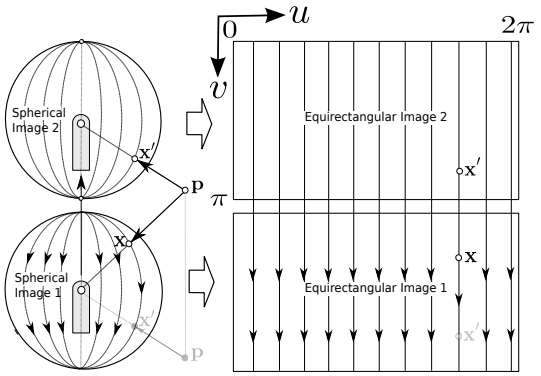


Fig. 2. If the cameras are displaced perfectly in the vertical direction, all pixel movements are in the vertical direction in the equirectangular projection.

In a similar spirit, to correct for small errors of vertical alignment (which can create huge errors in the disparity estimation), [8] used a point-matching based technique to refine the matched points to the same vertical line. [9] also introduced a generalized method of rectifying spherical image epipolar geometry. They conducted simulations and found the errors of the refinement method for different levels of noise and number of matched points. However, even after RANSAC-based filtering [10], point matching can involve outliers that can induce noise in the epipolar geometry, making disparity estimation suffer greatly. Thus, typical structure from motion approaches like [11] use computationally intensive approaches like bundle adjustment which globally optimize multiple camera positions. Following this, a multi-view stereo algorithm like [12] is used to estimate the dense cloud. Instead, with sufficient translation, it should be possible to estimate the dense structure using only two images as done by [7] and [6].

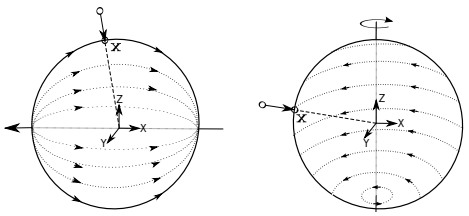
In place of sparse point-based methods, dense motion estimation methods like [3], [13], [14] and [15] are highly applicable for spherical images. Specifically, [3] explained how dense optical flow can help in precise epipolar estimation of closely displaced perspective images. They argued that the regularization property of dense optical flow prevents local outliers and leads to a more precise epipolar geometry. In this research, the same concept is applied to spherical images. Combined with the vertical rectification method of [8], [9], epipolar geometry is refined in a non-linear minimization of the dense optical flow field on the 2D equirectangular projection. The final state of optical flow also forms the dense disparity map and can directly be converted to a dense 3D model.

III. Spherical Camera Motion and Equirectangular Rectification

A. Overview

Considering the objectives mentioned in the previous section, a method for 3D reconstruction using dense optical flow is presented. The camera positions are assumed to be arbitrary, but with sufficient translation between them. Since the dense optical flow cannot be computed if the orientations of the spherical images differ a lot, a feature point-based approach is used for initial rectification. First, the rotation between both images is removed by derotation, and then they are both rotated to bring the epipoles in the vertical direction as shown in Figure 2. Following this, the dense optical flow is estimated using a recent, accurate algorithm like DeepFlow [16]. The deviation from the perfectly rectified state is defined as an error measure and minimized over the epipolar geometry in a non-linear least squares approach. Finally, the magnitude component of the resulting optical flow field is directly converted to the dense 3D map. The Ricoh Theta S spherical camera is used this research [17].

As the name implies, all pixels are formed on the surface of a sphere, where each pixel $\mathbf{x}$ is a unit vector. In pure translational camera motion, pixels move in circular arcs diverging away from a pole and converging to a diametrically opposite pole (epipoles). For pure rotation, the pixels move in loops around the rotation axis. Both are shown in Figure 3.



(a) (a) Translation (b) (b) Rotation

Fig. 3. Pixel motion on the spherical image for the camera undergoing (a) Pure Translation (b) Pure Rotation. (Spherical projection)

Any general camera motion is a combination of translation and rotation. Since spherical images contain information from all directions, they can be rotated without loss of information. Hence, the camera motion can effectively de-rotated to bring it to the same state as Figure 3 (a). Further, it can also be rotated to bring the epipoles in a vertical alignment, similar to Figure 2. In effect, by knowing the precise epipole position, and the rotation between two spherical images, they can be perfectly rectified to bring all corresponding pixels on the same vertical line. In this rectified state, if the two images are expanded to an equirectangular projection, the resultant dense optical flow field must have no horizontal component (Figure 2). Thus, a non-linear least squares minimization is defined in order to optimize the epipole and the rotation to achieve this state.

B. Notations

The two spherical images are denoted by $S_1(\mathbf{x})$ and $S_2(\mathbf{x})$, where $\mathbf{x}(x, y, z) \in S_1, S_2$ denotes a pixel (a unit vector). The rotation between the two images is denoted by the rotation matrix $\mathbf{R}$, constructed from the three XYZ euler angles (α, β, γ). $\mathbf{Rv}$ represents the rotation matrix that rectifies each image (after derotation) to the vertically rectified orientation of Figure 2. The epipole is represented by $\mathbf{q}$, the pole towards the direction of motion. $\mathbf{Rv}$ can be determined by knowing $\mathbf{q}$. Since $\mathbf{q}$ only represents a direction, it is represented in spherical coordinates (θ, ϕ). Thus, $G(\alpha, \beta, \gamma, \theta, \phi)$ forms the parameter vector that defines the relative pose or epipolar geometry between two images.

C. Rectification

The vertical rectification follows the following format. First, $S_2(\mathbf{x})$ is derotated to the same orientation as $S_1(\mathbf{x})$:

$$S_{2,1}(\mathbf{x}) = \mathbf{R}^{-1} \times S_2(\mathbf{x}) = \Big(\mathbf{R_x}(\alpha)\mathbf{R_y}(\beta)\mathbf{R_z}(\gamma)\Big)^{-1} \times S_2(\mathbf{x}) \tag{1}$$

where $S_{2,1}(\mathbf{x})$ indicates $S_2(\mathbf{x})$ in the same orientation as $S_1(\mathbf{x})$, and $\mathbf{R_x}(\alpha)$, $\mathbf{R_y}(\beta)$, and $\mathbf{R_z}(\gamma)$ denote the individual rotation matrices in the x, y, and z axes. Following this, $\mathbf{Rv}$ is determined as follows. The angle ω between the epipole $\mathbf{q}$ and the vector $\mathbf{n}(0, 0, 1)$ is:

$$\omega = \arccos\left(\frac{\mathbf{q} \cdot \mathbf{n}}{|\mathbf{q}||\mathbf{n}|}\right) \tag{2}$$

The axis of rotation $\mathbf{a}$ is the cross product of $\mathbf{q}$ and $\mathbf{n}$:

$$\mathbf{a} = \mathbf{q} \times \mathbf{n} \tag{3}$$

Thus, $\mathbf{Rv}$ can be written as a rotation matrix of angle ω around the axis $\mathbf{a}$:

$$\mathbf{Rv} = \mathbf{R_a}(\omega) \tag{4}$$

Finally, both images are rotated by $\mathbf{Rv}$ to the rectified orientations, $S_{1,r}$ and $S_{2,r}$ and thereafter expanded to the equirectangular projection for refinement.

$$S_{1,r}(\mathbf{x}) = \mathbf{Rv} \times S_1(\mathbf{x}) \tag{5}$$

$$S_{2,r}(\mathbf{x}) = \mathbf{Rv} \times S_{2,1}(\mathbf{x}) \tag{6}$$

D. Non-Linear Minimization

In this rectified state, the images should take the form of Figure 2. When expanded to the equirectangular projection, all pixel movements are expected to be in the vertical direction. Thus, if the dense optical flow field between the equirectangular projections of the two images is estimated, it is expected to have no horizontal component. The recently proposed, accurate DeepFlow algorithm [16] is employed for this purpose. Using this approach, the horizontal component of the dense optical flow can be minimized at every pixel with respect to the parameter vector $G(\alpha, \beta, \gamma, \theta, \phi)$. If $\mathbf{f}$ is the two-dimensional optical flow vector in the equirectangular projection at the equirectangular pixel $\mathbf{u}(u, v)$, define the following least-squares minimization is defined:

$$\underset{G(\alpha,\beta,\gamma,\theta,\phi)}{\text{minimize}} \sum_{\forall \mathbf{u}(u,v)} \mathbf{f}_u^2 \tag{7}$$

where $\mathbf{f}_u$ is the horizontal component of optical flow in the equirectangular image. The complete rectification and refinement pipeline is shown in Figure 4. Such problems are easily solvable by the popular Levenberg-Marquardt approach [18]. However, there are still two unsolved problems. In a general case where the images could have a very large difference of orientation, optical flow cannot be computed. Further, the Levenberg-Marquardt approach also requires a good initial value, close to the optimum. To tackle these, the feature-point based, 8-point RANSAC is modified for use on spherical images.

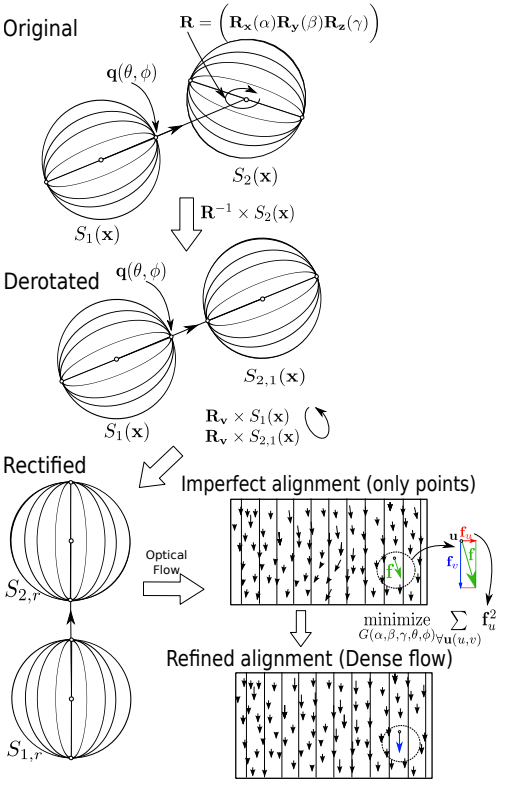


Fig. 4. Rectification-based refinement

E. Initialization

An initial estimation using the regular, 8-point RANSAC approach. In the same manner as perspective images, the essential matrix $\mathbf{E}$ defines the mapping between corresponding points in two images:

$$\mathbf{x'}^t \mathbf{E} \mathbf{x} = 0 \tag{8}$$

where $\mathbf{x}$ and $\mathbf{x'}$ are the corresponding spherical image points, written as $\begin{bmatrix}\mathbf{xyz}\end{bmatrix}^\top$ unit vectors. The original 8-point approach [19] suggests many normalizations, however the points in spherical images are already unit vectors and hence no further normalizations are necessary. They are filtered in a RANSAC [10] algorithm to obtain the set of inliers. The essential matrix $\mathbf{E}$ is decomposed using singular value decomposition to give us the rotation matrix and translation vector, that are converted to the initial values of the parameter vector $G(\alpha, \beta, \gamma, \theta, \phi)$.

At this stage, most typical structure from motion approaches would filter the point matches in a non-linear approach similar to [11], [8], [9]. However, as mentioned earlier, basing it on point matches which can contain outliers even after RANSAC filtering can induce errors in the final epipolar geometry estimate. Such errors make the disparity estimation very noisy (which can be particularly sensitive when searching along epipolar lines). Instead, the dense optical flow is used for this purpose. Apart from the advantage of not having local outliers due to regularization, the final rectified optical flow field in the vertical direction also directly forms the disparity image and can directly be converted to the 3D reconstruction, as shown in the next section.

Thus, the overall minimization is as follows:

1) Initialize $G(\alpha, \beta, \gamma, \theta, \phi)$ with 8-point RANSAC
2) Derotate S_2 with $\mathbf{R}^{-1} = \left(\mathbf{R_x}(\alpha)\mathbf{R_y}(\beta)\mathbf{R_z}(\gamma)\right)^{-1}$ to form $S_{2,1}$, in the same orientation as S_1
3) Calculate $\mathbf{Rv}$ from $\mathbf{q}$ and rotate both S_1 and $S_{2,1}$ to form vertically rectified images $S_{1,r}$ and $S_{2,r}$
4) Expand $S_{1,r}$ and $S_{2,r}$ to the 2D equirectangular projection and estimate optical flow using DeepFlow [16]
5) Calculate least squares error $\sum_{\forall \mathbf{u}(u,v)} \mathbf{f}_u^2$
6) Proceed with the Levenberg-Marquardt [18] approach from step 2 onwards to optimize $G(\alpha, \beta, \gamma, \theta, \phi)$

IV. 3D Reconstruction

In the final rectified state, the optical flow field should be aligned vertically in the equirectangular projection. Hence, taking its magnitude component directly gives us the disparity of each pixel in the equirectangular projection. A simple calculation as shown in Figure 5 is enough to convert it to the pixel-wise 3D structure.

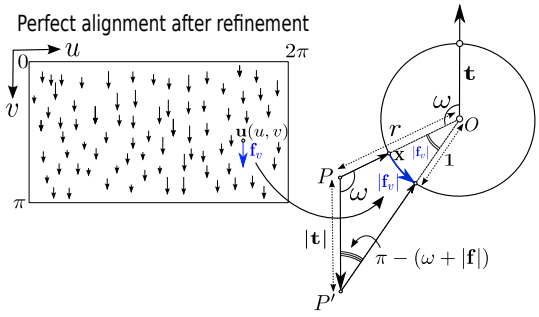


Fig. 5. 3D reconstruction from the final optical flow state

For a point $\mathbf{x}$ on the sphere, ω denotes the angular distance from the topmost point of the sphere, i.e. the epipolar point. The vertical magnitude of $\mathbf{f}_v$ on the equirectangular image is a difference of latitudes on the sphere, and thus forms the angular disparity on the sphere. Thus, the law of sines in triangle POP' gives the radius r of 3D point P:

$$r = |\mathbf{t}| \times \frac{\sin(\omega + |\mathbf{f}_v|)}{\sin(|\mathbf{f}_v|)} \tag{9}$$

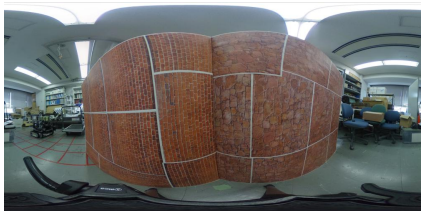
The magnitude of the translation vector is set as $|\mathbf{t}| = 1$ without loss of generality to give the final structure.

V. Experiments

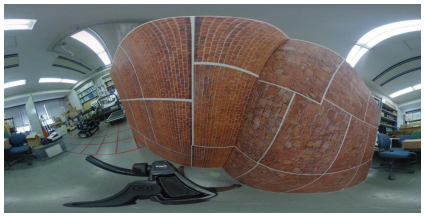
An experiment was conducted in an artificial setup in order to check for the quality of 3D reconstruction with the propose method. Two cardboard slabs of 1m $\times$ 1m with textures pasted on them boards were placed perpendicular to each other as shown in Figure 6. Two spherical images were captured at arbitrary orientations, displaced by a distance of 5 cm in an approximate vertical direction. The camera movement (5 cm) is very small with respect to the scale of the structure (1m). Thus, it can be assumed to be a large structure. The image pair is shown in Figure 7.

Fig. 6. Experimental Setup for 3D Reconstruction

The proposed algorithm was run on these images. Figure 8 shows the initialization step using 8-point RANSAC. Figure 10 shows the resultant images. Figure 9 shows the second image deroted, before vertical rectification. Figure 10 (a) shows the initial vertically rectified state, estimated only from point matching. It can be seen that the optical flow vectors (i.e. pixel movements) are not vertically aligned because of outliers. Meanwhile, after the proposed refinement, in Figure 10 (b),



(a) Image 1



(b) Image 2

Fig. 7. Two spherical images clicked at arbitrary orientations at displaced positions

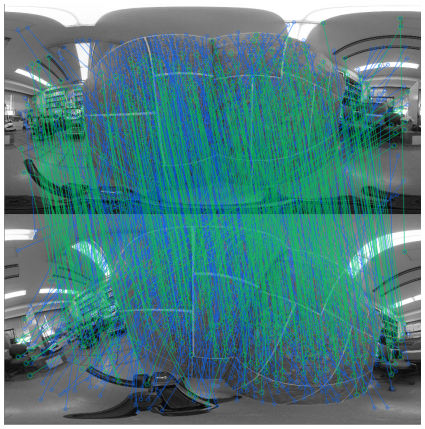


Fig. 8. RANSAC with the 8-point algorithm for initial estimation. All point matches are in blue, while the inliers are in green.

they are now aligned in the vertical direction, as expected. Figure 10 (c) shows their magnitude, i.e. the disparity map.

For comparison, an inaccurate dense reconstruction attempted without using the proposed refinement is shown in Figure 11. Finally, in Figure 12, the dense reconstruction using the proposed method is shown.

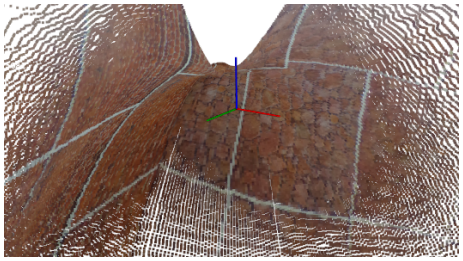
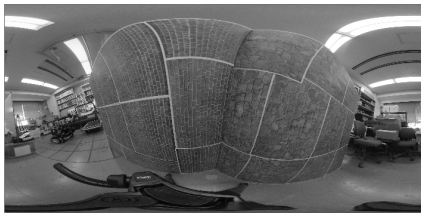
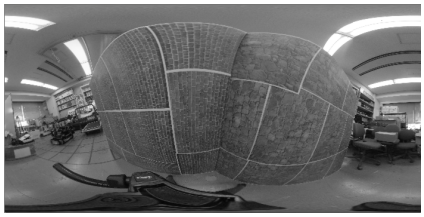


Fig. 11. Dense reconstruction using vertical rectification without using the proposed refinement (inaccurate)

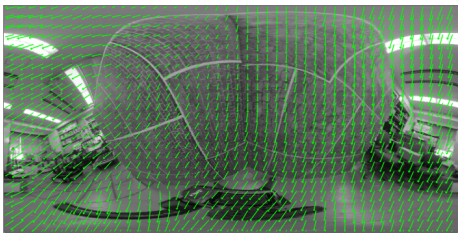


(a) Image 1

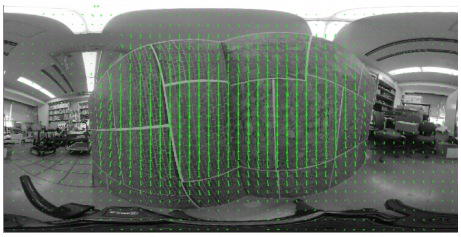


(b) Image 2

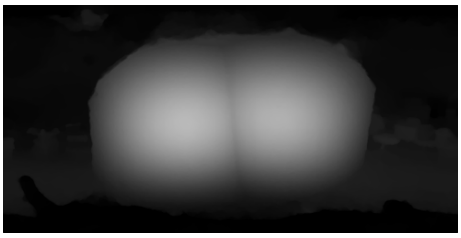
Fig. 9. Derotating image 2 to the same orientation as image 1 prior to vertical rectification



(a) Vertically rectified state before proposed refinement: The pixel movements are not aligned in a perfectly vertical direction.



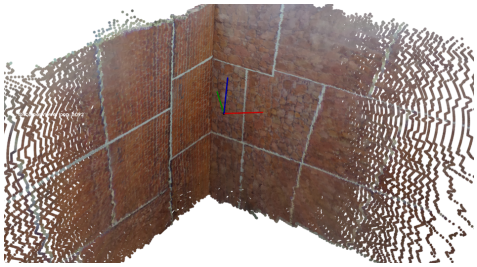
(b) Vertically rectified state after proposed refinement: Pixel movements aligned along the vertical direction, as expected



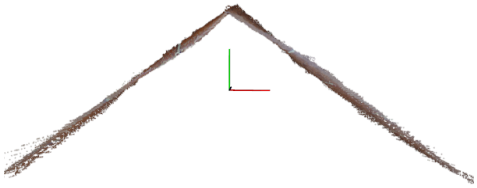
(c) Magnitude component of optical flow field: Disparity Image

Fig. 10. Output of the algorithm applied to the image pair

(a) Experimental Setup with camera position



(b) Dense reconstruction using rectification with proposed refinement, shown along with camera position



(c) Top view of the reconstruction

Fig. 12. 3D Reconstructed Views and camera positions after the proposed refinement. These reconstructions have been made by inserting the *RGB* color of each pixel in the corresponding 3D location, in order to show the quality of the resulting dense reconstruction. No meshing or texture mapping has been done.

VI. Conclusion

In this research, a new method was proposed for dense reconstruction from two spherical images clicked at arbitrary positions using an optical flow-based refinement. Closely following the research of [6] and [7], it was extended it to be applicable for general camera motions. This can enable 3D reconstruction of large structures at once. For this purpose, the advantages of dense optical flow-based epipolar estimation were considered, as suggested by [3], and applied to spherical images. A dense reconstruction of a 1m $\times$ 1m structure using a movement of only 5 cm was estimated.

It is intended to extend this work to a full video based structure from motion pipeline using successive frames. The stereo disparity estimation can also be moved from an optical flow based approach to that of [7] and [6] for a more accurate reconstruction. It is expected that this research could help in the inspection and maintenance of large infrastructures by digitizing them and mapping their surface accurately.

Acknowledgement

This work was in part supported by the 'Cross-ministerial Strategic Innovation Promotion Program (SIP), Infrastructure Maintenance, Renovation, and Management', Council for Science Technology and Innovation (funding agency: NEDO).

References

[1] H. Fathi, F. Dai, and M. Lourakis, "Automated as-built 3d reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 149 – 161, April 2015.

[2] I. Brilakis, H. Fathi, and A. Rashidi, "Progressive 3d reconstruction of infrastructure with videogrammetry," *Automation in Construction*, vol. 20, no. 7, pp. 884 – 895, November 2011.

[3] L. Valgaerts, A. Bruhn, M. Mainberger, and J. Weickert, "Dense versus sparse approaches for estimating the fundamental matrix," *International Journal of Computer Vision*, vol. 96, pp. 212–234, January 2012.

[4] S. Pathak, A. Moro, A. Yamashita, and H. Asama, "3d reconstruction of structures using spherical cameras with small motion," in *Proceedings of the 16th International Conference on Control, Automation, and Systems*, October 2016.

[5] N. Kita, "Dense 3d measurement of the near surroundings by fisheye stereo," in *Proceedings of the IAPR Conference on Machine Vision Applications, 2011*, June 2011, pp. 148–151.

[6] H. Kim and A. Hilton, "Planar urban scene reconstruction from spherical images using facade alignment," in *Proceedings of the 11th IEEE Image Video and Multidimensional Signal Processing (IVMSP) workshop, 2013*, June 2013, pp. 1–4.

[7] H. Kim and A. Hilton, "3d scene reconstruction from multiple spherical stereo pairs," *International Journal of Computer Vision*, vol. 104, no. 1, pp. 94–116, August 2013.

[8] A. Banno and K. Ikeuchi, "Omnidirectional texturing based on robust 3d registration through euclidean reconstruction from two spherical images," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 491 – 499, April 2010.

[9] J. Fujiki, A. Torii, and S. Akaho, "Epipolar geometry via rectification of spherical images," in *Proceedings of the Third International Conference on Computer Vision/Computer Graphics Collaboration Techniques*, March 2007, pp. 461–471.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[11] A. Pagani and D. Stricker, "Structure from motion using full spherical panoramic cameras," in *Proceedings of the IEEE International Conference on Computer Vision (Workshops)*, November 2011, pp. 375–382.

[12] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 8, pp. 1362–1376, August 2010.

[13] A. Makadia and K. Daniilidis, "Rotation recovery from spherical images without correspondences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1170–1175, August 2006.

[14] S. Pathak, A. Moro, A. Yamashita, and H. Asama, "A decoupled virtual camera using spherical optical flow," in *Proceedings of the IEEE International Conference on Image Processing*, September 2016.

[15] A. Makadia and K. Daniilidis, "Direct 3d-rotation estimation from spherical images via a generalized shift theorem," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2003, pp. 217–224.

[16] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE International Conference on Computer Vision*, December 2013, pp. 1385 – 1392.

[17] "Ricoh theta," https://theta360.com/en/.

[18] M. Lourakis, "levmar: Levenberg-marquardt nonlinear least squares algorithms in c/c++," http://www.ics.forth.gr/~lourakis/levmar/, July 2004.

[19] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 580–593, June 1997.