DISTORTION-ROBUST SPHERICAL CAMERA MOTION ESTIMATION VIA DENSE OPTICAL FLOW

Sarthak Pathak, Alessandro Moro, Hiromitsu Fujii, Atsushi Yamashita, Hajime Asama

Department of Precision Engineering, The University of Tokyo, Japan

ABSTRACT

Conventional techniques for frame-to-frame camera motion estimation rely on tracking a set of sparse feature points. However, images taken from spherical cameras have high distortion which can induce mistakes in feature point tracking, offsetting the advantage of their large fields-of-view. Hence, in this research, we attempt a novel approach of using dense optical flow for distortion-robust spherical camera motion estimation. Dense optical flow incorporates smoothing terms and is free of local outliers. It encodes the camera motion as well as dense 3D information. Our approach decomposes dense optical flow into epipolar geometry and the dense disparity map, and reprojects this disparity map to estimate 6 DoF camera motion. The approach handles spherical image distortion in a natural way. We experimentally demonstrate its accuracy and robustness.

Index Terms— Spherical vision, Distortion, Optical flow

1. INTRODUCTION

Frame-to-frame camera motion estimation is very fundamental to visual odometry/structure from motion. Conventionally, many sparse feature points are tracked and camera motion is back-calculated based on its projective geometry. Recently developed full-view spherical cameras, such as the Ricoh Theta, are considerably advantageous over perspective cameras [1]. They have no directional bias of information, and image information never goes out of view. However, spherical images do not exist on a planar manifold. They can be expressed as planar images in the equirectangular projection, which induces severe distortion that is highly non-linear to camera motion, especially to rotation as shown in Fig. 1. This induces many mistakes in tracking sparse points.

Several approaches have tried to deal with such distortion. [2] tried to project spherical images to planar cubes. A-KAZE [3], spherical versions of SIFT [4], ORB [5], and SURF [6] tried to robustly match points across distorted views. [7] implemented SIFT [8] on an interpolation-based regularized spherical grid. However, these sparse approaches throw away vast, useful dense information and make for unstable estimation. A single mismatched outlier can drastically affect the result. In contrast, there are several dense optical flow algorithms which can track all image pixels at the cost of spatial smoothening [9]. Due to this smoothening, they suppress local mismatches and can provide a reliable consensus of camera motion, as was exploited in [10], and also in [11, 12, 13]. Similarly, [14] and [15] estimated motion by separating rotational and translational flows.

(a) Equirectangular Spherical Image (b) Image Rotated by 30 deg. (pitch)

Fig. 1. Spherical images undergo strong non-linear distortions on rotation, as shown in the zoomed-in red area.

In this research, we explore spherical camera motion estimation using dense optical flow, similar to [10]. The dense optical flow field is decomposed to epipolar geometry and the dense disparity, and then densely reprojected to estimate camera motion. We take advantage of the complete spherical field of view, i.e. the property that they can be rotated to any orientation without loss of information. Initially, 2 spherical images are rectified to an equirectangular stereo pair through multiple image rotations, via an iterative energy minimization based on equirectangular dense optical flow. This step simultaneously estimates the 5 DoF epipolar geometry and the dense disparity map. Next, this dense disparity map is utilized to reproject one of the images to a third image to give the full 6 DoF estimate of the third image. This is analogous to conventional motion estimation in which sparse points are tracked across 3 views, and the triangulated 3D points from the first two are used to find the pose of the 3rd, and so on. The purpose of this paper is to demonstrate a novel method, and to show the superiority of using dense optical flow. This results in a robust, accurate estimate, as demonstrated later by experiments. In our previous work [16], a method for dense

This work was in part supported by the Council for Science, Technology and Innovation, "Cross-ministerial Strategic Innovation Promotion Program (SIP), Infrastructure Maintenance, Renovation, and Management" (funding agency: NEDO). Contact: <last-name>@robot.t.u-tokyo.ac.jp

3D reconstruction based on 2 spherical images was detailed. In this work, we expand it to an iterative estimation and dense reprojection in order to estimate the frame-to-frame motion in a distortion resistant manner.

Our proposed approach consists of 2 steps. The first is a 5 DoF epipolar estimation between two equirectangular images I_1 and I_2 displaced by a rotation $\mathbf{R_{1,2}}$ and translation $\mathbf{t_{1,2}}$. The spherical image pair is converted into a rectified equirectangular stereo pair via energy minimization, based on the dense optical flow between them. The final resultant optical flow field in the equirectangular projection automatically forms the dense disparity map of the two images. In the second step, this dense disparity map is used for a pixel-to-pixel reprojection of I_2 on a third image I_3, which is displaced from I_1 by $\mathbf{R_{1,3}}$ and $\mathbf{t_{1,3}}$. A dense, photometric reprojection error is minimized in order to obtain a full 6 DoF estimate of the third image. The translation scale $|\mathbf{t_{1,2}}|$ between the first two images is set to one without loss of generality. From the 3rd image onwards, the estimation is 6 DoF.

2. DENSE OPTICAL FLOW-BASED MOTION ESTIMATION

The first step estimates two-frame motion i.e. epipolar geometry and the dense disparity map, simultaneously. Our method can naturally handle the distortion of spherical images on a 2D equirectangular grid. If two spherical images are vertically displaced without any rotation between them, all corresponding pixels lie on the same vertical line, unlike complex curves as done in [17]. This is shown below in Fig. 2. Similar to [10], we use dense optical flow in an iterative energy minimization to rectify the two images to this arrangement.

First the rotation between the two images can be corrected by aligning them to the same orientation. Following this, they can both be rotated in order to align the translation vector in a vertical direction. In this arrangement, the equirectangular images should have a vertically oriented dense optical flow field. Thus, we estimate the epipolar geometry that best fits the desired vertically oriented dense optical flow field. The

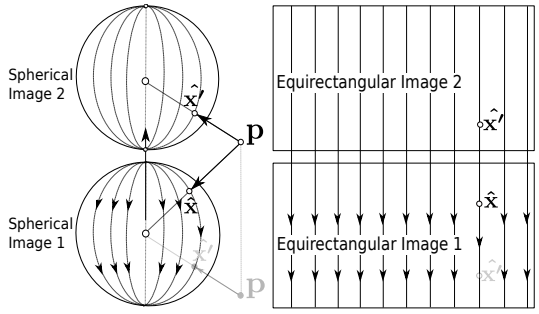


Fig. 2. If the cameras are displaced perfectly in the vertical direction, all equirectangular pixel movements vertical.

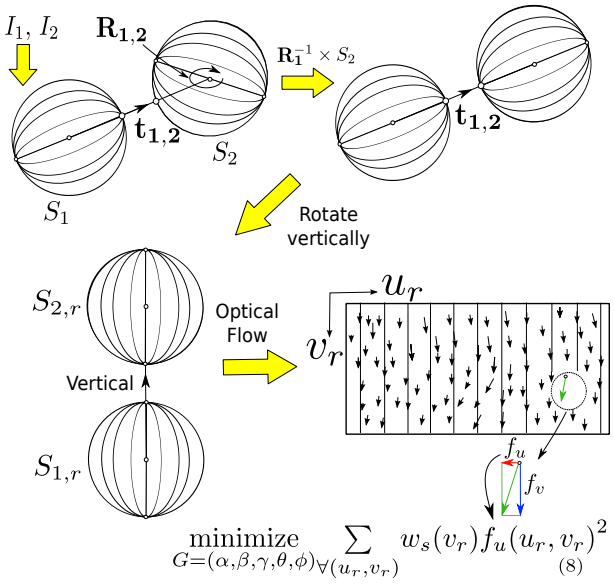


Fig. 3. Rectifying an arbitrary pair of spherical images.

pipeline is shown in Fig. 3.

All rotations are done pixel-wise between equirectangular images after projecting each equirectangular pixel (u, v) to its spherical unit vector $\hat{\mathbf{x}} = [x, y, z]^T$, with bilinear interpolation to fill the gaps due to discretization of pixels. Thus, I_1 and I_2 are converted to two rectified equirectangular images $I_{1,r}$ and $I_{2,r}$ via multiple rotations. In this state, the dense optical flow $\mathbf{f_v}(u_r, v_r)$ between the two rectified images is estimated and an energy function that depends on the horizontal component of the optical flow field $f_u(u_r, v_r)$ is defined. In order to densely estimate the energy function over the whole image, a weighting scheme $w_s(v_r)$ is used to counter the stretching of pixels in the equirectangular image. Here, the epipolar geometry is represented in 5 DoF by $G = (\alpha, \beta, \gamma, \theta, \phi)$, the first three being euler rotation angles, and the last two being translation parameters expressed in spherical coordinates. A minimization of this energy over the epipolar geometry gives the final rectified state:

$$\underset{G=(\alpha,\beta,\gamma,\theta,\phi)}{\text{minimize}} \sum_{\forall(u_r,v_r)} w_s(v_r) f_u(u_r, v_r)^2. \quad (1)$$

This is iteratively solved by optimizing the rotation and translation components separately till convergence, via the Levenberg-Marquardt routine. We used the Deepflow [9] algorithm to compute the dense optical flow between $I_{1,r}$ and $I_{2,r}$. In order to avoid recomputing the optical flow field in every iteration, we simply reproject the dense optical flow state followed by rotational transformations.

At the end of the minimization, we obtain a vertically oriented optical flow field. Its magnitude component $|\mathbf{f_v}(u_r, v_r)|$ directly forms the dense disparity map. The final estimate of

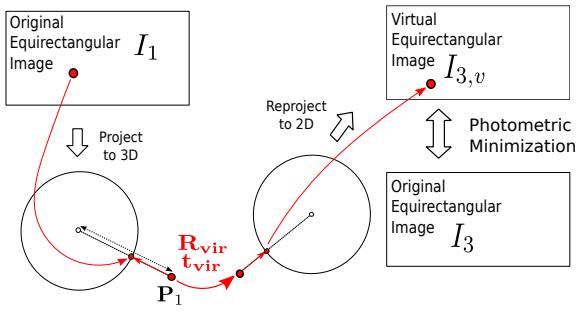


Fig. 4. 3D Reprojection of I_1 to a virtual image $I_{3,v}$ followed by dense photometric minimization

$G = (\alpha, \beta, \gamma, \theta, \phi)$ is converted to $\mathbf{R_{1,2}}$ and $\mathbf{t_{1,2}}$. The magnitude of $\mathbf{t_{1,2}}$ is set as 1, without loss of generality.

Once the disparity is obtained, the radius $r(u_r, v_r)$ of each pixel can be calculated from its disparity $|\mathbf{f_v}|$ in the same manner as given in [18] to obtain $\mathbf{P_1}(u, v)$, its 3D coordinates. These 3D coordinates can be reprojected to the expected position and orientation of I_3. We perform a pixel-wise transformation of every point $\mathbf{P_1}(u, v)$ in I_1 to form a virtual image $I_{3,v}$ which should be the same as image I_3. This pixel-wise transformation also incorporates bilateral interpolation in order to fill the gaps. Finally, a photometric reprojection error between $I_{3,v}$ and I_3 is minimized over $\mathbf{t_{1,3}}$ and $\mathbf{R_{1,3}}$. The same weighting scheme $w_s(v)$ is used.

$$\underset{\left[\mathbf{R}_{1,3}|\mathbf{t}_{1,3}\right]}{\text{minimize}} \sum_{\forall(u,v)} w_s(v)\big(I_3(u,v) - I_{3,v}(u,v)\big)^2. \qquad (2)$$

This problem can also be solved by the same Levenberg-Marquardt scheme as used in the previous step. For both steps, iterative rectification and reprojection error minimization, a coarse-to-fine scheme was used by downsampling the images three times. Thus, if there is a sequence of n images, the position of the first image I_1 can be set as the origin. The position of the second image I_2 can be estimated via the iterative epipolar estimation alone and the scale of translation set as 1. From I_3 onwards, the estimation becomes 6 DoF and can proceed via the dense reprojection alone.

3. EXPERIMENTAL EVALUATION

In order to test the robustness of our proposed approach, we evaluated it against a recent sparse feature descriptor that provides high resilience to non-linear distortions via a non-linear scale space - A-KAZE [3]. The Ricoh Theta S spherical camera was used for all experiments. It provides completely stitched equirectangular images without the need for additional calibration. Initially, a set of two textured cardboards were used to provide an environment easy for disparity estimation - the 'board' sequence. A set of 3 images I_1, I_2, and I_3 of equirectangular image resolution 800×400 pixels were captured from left to right, as shown in Fig. 5. Following this, I_1 was rotated with a known pitch angle to a new image $I_{1,p}$ to create a fourth image in the sequence. This induces large distortions in the equirectangular image, as shown in Fig. 6. Thus, an image sequence I_1, I_2, I_3, and $I_{1,p}$ is created, with I_1 and $I_{1,p}$ in the same position, with a known rotation angle.

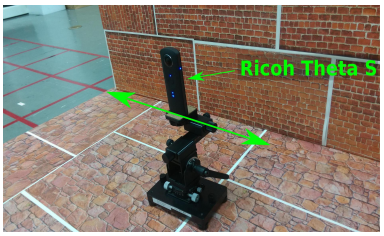
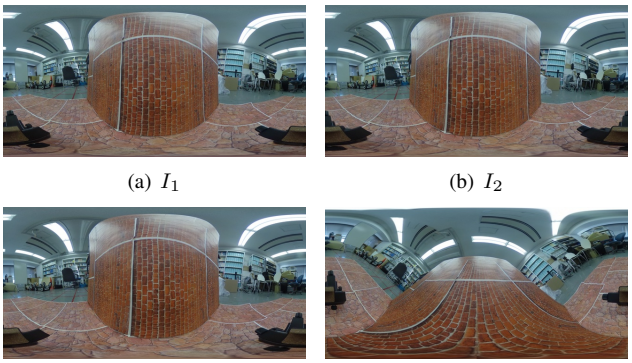


Fig. 5. Experimental Setup consisting of two sheets of cardboard forming a perpendicular structure. 3 images I_1, I_2, and I_3 were captured at arbitrary positions from right-to-left.



(a) I_1 (b) I_2

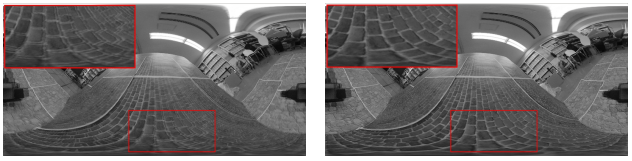
(c) I_3 (d) $I_{1,p}$: I_1 rotated by 50 deg. pitch

Fig. 6. Image sequence captured for evaluation using two cardboard sheets - 'board' sequence. Large distortions are induced in I_1 by rotating it with a pre-decided pitch angle.

Estimation was performed linearly as $I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow I_{1,p}$. The error of estimation of $I_{1,p}$ w.r.t. I_1 was used as the estimation error. Essentially, we checked the ability to resist the distortion induced by the pitch rotation and 'close the loop'. This estimation was done in two ways and compared. The first was by sparse features tracking with A-KAZE [3] using the same method proposed in [19]. The second way was using our dense optical flow-based method. Further, in order to provide an environment where disparity estimation is difficult, a similar experiment was repeated by arbitrarily capturing three images in a cluttered room, as shown in Fig. 7 - the 'room' sequence (please see footnote for link to dataset)[1]. An example of dense photometric minimization in the 'board' sequence is shown in Fig. 8.

[1]Dataset available at: www.robot.t.u-tokyo.ac.jp/%7Epathak/research

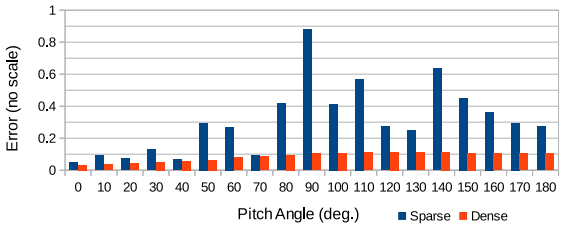
Fig. 7. Sample experimental image from a cluttered room



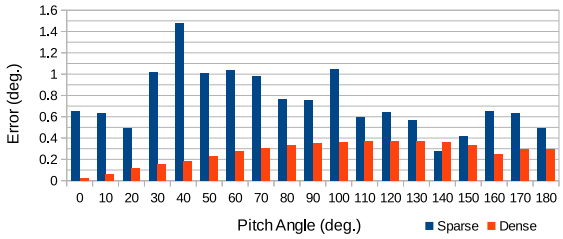
(a) Initial state. (b) Final, optimized state

Fig. 8. Minimizing the dense photometric reprojection error between $I_{3,v}$ and I_3 (red area zoomed-in for visibility).

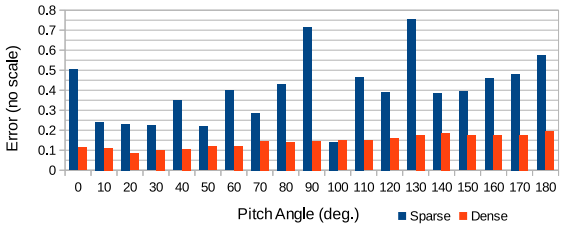
The pitch angle was varied from 0 to 180 degrees and errors between the true and estimated positions and orientations of $I_{1,p}$ were plotted with respect to each angle, as shown in Fig. 9(a). The position error is without real-world scale and is expressed as a factor of the translation distance and the orientation error is expressed in the quaternion angle. The results in Fig. 9(a) and (b) for the 'board' sequence show that the estimation derived from the proposed dense optical flow approach is much more stable and robust to the strong distortions induced by the pitch rotations whereas the estimation obtained by sparse A-KAZE [3] is unstable and affected quite strongly. From Figs. 9(c) and (d), which show the errors of the 'room' sequence, it can be seen that even in conditions that make 3D reconstruction or disparity estimation difficult (which increased the average position errors), the use of dense optical flow showed more robustness and lower errors.



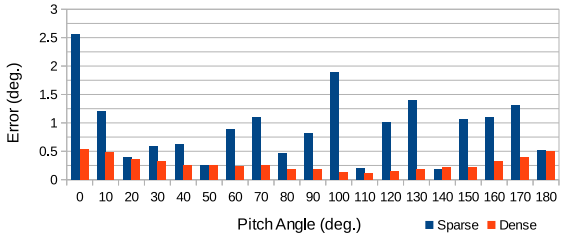
(a) 'board' sequence: Position Errors vs. Angle of Pitch (without scale, fraction of $I_1 \rightarrow I_2$ distance)



(b) 'board' sequence: Orientation Errors vs Angle of Pitch (degrees)



(c) 'room' sequence: Position Errors vs. Angle of Pitch (without scale, fraction of $I_1 \rightarrow I_2$ distance)



(d) 'room' sequence: Orientation Errors vs Angle of Pitch (degrees)

Fig. 9. Estimation errors. Estimation using dense optical flow shows lower errors, and is robust to distortion.

4. DISCUSSION AND CONCLUSION

In this research, we built a frame-to-frame motion estimation framework for spherical cameras using dense optical flow. Instead of conventional sparse feature point matching and 3D triangulation, we estimated 2-frame epipolar geometry and dense disparity in an iterative, distortion resistant manner using dense optical flow. Then, instead of reprojecting sparse 3D points and minimizing a geometric error, we reprojected the recovered disparity map and minimized a dense photometric error to obtain a 6 DoF estimate.

This novel approach took care of outliers due to the smoothing involved in dense optical flow and allowed the use of all the available information in an image. It provided epipolar geometry as well as the dense disparity, leading to a stable, robust estimate which was tested by inducing distortion in the equirectangular image. Experimental validation in different situations, including cluttered settings, demonstrated that the use of dense optical not only results in lower errors, but also vastly increases robustness to the distortion induced by rotation. A possible limitation is that it may not work across large translational displacements, which can make it difficult to compute dense optical flow, as also noted in [10]. Studying the effect of the baseline, and a full dense optical-flow based visual SLAM, remain as future work.

5. REFERENCES

[1] Zichao Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, "Benefit of large field-of-view cameras for visual odometry," in *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2016, pp. 801–808.

[2] Florian Kangni and Robert Laganiere, "Orientation and pose recovery from spherical panoramas," in *Proceedings of the IEEE International Conference on Computer Vision*, October 2007, pp. 1–8.

[3] Pablo Alcantarilla, Jesus Nuevo, and Adrien Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proceedings of the British Machine Vision Conference*, September 2013, pp. 13.1–13.11.

[4] Javier Cruz-Mota, Iva Bogdanova, Benoît Paquier, Michel Bierlaire, and Jean Philippe Thiran, "Scale invariant feature transform on the sphere: Theory and applications," *International Journal of Computer Vision*, vol. 98, no. 2, pp. 217–241, June 2012.

[5] Qiang Zhao, Wei Feng, Liang Wan, and Jiawan Zhang, "Sphorb: A fast and robust binary feature on the sphere," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 143–159, June 2015.

[6] Ana Cris Murillo, José Jesús Guerrero, and C Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Proceedings of the IEEE International Conference on Robotics and Automation*, April 2007, pp. 3901–3907.

[7] Juan David Adarve and Robert Mahony, "Spherepix: A data structure for spherical image processing," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 483–490, April 2017.

[8] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[9] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE International Conference on Computer Vision*, December 2013, pp. 1385 – 1392.

[10] Levi Valgaerts, Andrés Bruhn, Markus Mainberger, and Joachim Weickert, "Dense versus sparse approaches for estimating the fundamental matrix," *International Journal of Computer Vision*, vol. 96, pp. 212–234, January 2012.

[11] Florian Becker, Frank Lenzen, Jörg H Kappes, and Christoph Schnörr, "Variational recursive joint estimation of dense scene structure and camera motion from monocular high speed traffic sequences," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 269–297, December 2013.

[12] Timm Schneevoigt, Christopher Schroers, and Joachim Weickert, "A dense pipeline for 3d reconstruction from image sequences," in *Proceedings of the German Conference on Pattern Recognition*, September 2014, pp. 629–640.

[13] K Prazdny, "Egomotion and relative depth map from optical flow," *Biological cybernetics*, vol. 36, no. 2, pp. 87–102, February 1980.

[14] Randal C Nelson and John Aloimonos, "Finding motion parameters from spherical motion fields (or the advantages of having eyes in the back of your head)," *Biological Cybernetics*, vol. 58, no. 4, pp. 261–273, March 1988.

[15] Sarthak Pathak, Alessandro Moro, Atsushi Yamashita, and Hajime Asama, "A decoupled virtual camera using spherical optical flow," in *Proceedings of the IEEE International Conference on Image Processing*, September 2016, pp. 4488–4492.

[16] Sarthak Pathak, Alessandro Moro, Atsushi Yamashita, and Hajime Asama, "Dense 3d reconstruction from two spherical images via optical flow-based equirectangular epipolar rectification," in *Proceedings of the IEEE International Conference on Imaging Systems and Techniques*, October 2016, pp. 140–145.

[17] Akira Ohashi, Yuki Tanaka, Gakuto Masuyama, Kazunori Umeda, Daisuke Fukuda, Takehito Ogata, Tatsurou Narita, Shuzo Kaneko, Yoshitaka Uchida, and Kota Irie, "Fisheye stereo camera using equirectangular images," in *Proceedings of the 11th France-Japan congress on Mechatronics / the 9th Europe-Asia congress on Mechatronics / the 17th International Conference on Research and Education in Mechatronics (MECATRONICS-REM2016)*, June 2016, pp. 284–289.

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

[19] Alain Pagani and Didier 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.