Supplementary Material for Paper 1640

Anonymous Author(s)
Submission Id: 1640

1 PRELIMINARIES OF THE SURROUND-VIEW SYSTEM

This section describes how to generate a surround-view from images captured by the cameras in the surround-view system.

Given the ground coordinate system O_G and a surround-view system consisting of multiple cameras, the pose of camera C_i is denoted by $\boldsymbol{T}_{C_iG}$. For a point $\boldsymbol{P}_G = [X_G, Y_G, Z_G, 1]^T$ in O_G, its corresponding pixel coordinate $\boldsymbol{p}_{C_i}$ in the camera coordinate system of C_i is given by,

$$\boldsymbol{p}_{C_i} = \frac{1}{Z_{C_i}} \boldsymbol{K}_{C_i} \boldsymbol{T}_{C_iG} \boldsymbol{P}_G \tag{1}$$

where Z_{C_i} is the depth of $\boldsymbol{P}_G$ in C_i's coordinate system, and $\boldsymbol{K}_{C_i}$ is the intrinsic matrix of camera C_i, which can be estimated by Zhang's salient work [3] and some subsequent work of others [1, 4]. It's worth mentioning that the poses of cameras in the surround-view system are usually determined by offline calibration. In our solution, the scheme proposed by Shao *et al.* in [2] is adopted.

Consider a point $\boldsymbol{p}_G = [u_G, v_G, 1]^T$ on the bird's-eye-view image. Its corresponding point on the ground plane is denotaed by $\boldsymbol{P}_G = [X_G, Y_G, Z_G = 0]^T$ with respect to the ground coordinate system. The relationship between $\boldsymbol{p}_G$ and $\boldsymbol{P}_G$ can be represented as,

$$\boldsymbol{p}_G = \boldsymbol{K}_G \boldsymbol{P}_G \tag{2}$$

and the transformation matrix $\boldsymbol{K}_G$ is defined as,

$$\boldsymbol{K}_G = \begin{bmatrix} \frac{1}{d_{X_G}} & 0 & \frac{W}{2d_{X_G}} \\ 0 & -\frac{1}{d_{Y_G}} & \frac{H}{2d_{Y_G}} \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where d_{X_G} and d_{Y_G} define the physical size of each pixel[1], and W and H are the width and height of the synthesized surround-view, respectively. It is worth mentioning that since $Z_G = 0$, it is ignored implicitly here. By combining Eq. 1 and Eq. 2, we can get,

$$\boldsymbol{p}_{C_i} = \frac{1}{Z_{C_i}} \boldsymbol{K}_{C_i} \boldsymbol{T}_{C_iG} \boldsymbol{K}_G^{-1} \boldsymbol{p}_G \tag{4}$$

Eq. 4 actually depicts the relationship of a point $\boldsymbol{p}_{C_i}$ on the image plane of camera C_i and its projection $\boldsymbol{p}_G$ on the surround-view. From Eq. 4, we can generate a bird's-eye-view image by projecting the undistorted image of camera C_i onto the ground,

$$\boldsymbol{I}_{GC_i}(\boldsymbol{p}_G) = \boldsymbol{I}_{C_i}(\boldsymbol{p}_{C_i}) \tag{5}$$

where $\boldsymbol{I}_{C_i}$ is the undistorted fisheye image captured by camera C_i, and $\boldsymbol{I}_{GC_i}$ is the ground projection of $\boldsymbol{I}_{C_i}$, namely the bird's-eye-view image. Then, the surround-view image can be synthesized with appropriate stitching seams.

[1]More accurately, each pixel in the surround-view image corresponds to a $d_{X_G} \times d_{Y_G}$ physical area on the ground plane.

2 JACOBIANS OF THE BI-CAMERA ERROR

As mentioned in the manuscript, the bi-camera error term $\varepsilon^{bi}_{\boldsymbol{p}_G}$ of a point $\boldsymbol{p}_G$ on the surround-view is defined as,

$$\begin{aligned} \varepsilon^{bi}_{\boldsymbol{p}_G} = \frac{1}{|\boldsymbol{\mathcal{P}}|} \sum_{\boldsymbol{p}_s \in \boldsymbol{\mathcal{P}}} & \boldsymbol{I}_{C_i}\left(\lambda^{C_i}_{\boldsymbol{p}_G} \boldsymbol{K}_{C_i} \exp\left(\xi^{\wedge}_{C_iG}\right) \boldsymbol{K}_G^{-1} \boldsymbol{p}_G + \boldsymbol{p}_s\right) \\ & - \gamma_{ij} \boldsymbol{I}_{C_j}\left(\lambda^{C_j}_{\boldsymbol{p}_G} \boldsymbol{K}_{C_j} \exp\left(\xi^{\wedge}_{C_jG}\right) \boldsymbol{K}_G^{-1} \boldsymbol{p}_G\right) \end{aligned} \tag{6}$$

where $\boldsymbol{I}_{C_i}$ and $\boldsymbol{I}_{C_j}$ are undistorted images captured by C_i and C_j, respectively. $\boldsymbol{K}_{C_i}$ and $\boldsymbol{K}_{C_j}$ are intrinsics of C_i and C_j, respectively. ξ_{C_iG} and ξ_{C_jG} are poses of C_i and C_j in Lie algebra form, respectively. $\boldsymbol{K}_G$ stands for the transformation matrix from the surround-view coordinate system to the ground coordinate system. $\boldsymbol{\mathcal{P}}$ is a set that contains the relative pixel coordinates of all the utilized points to $\boldsymbol{p}_{C_i}$, and is defined as,

$$\boldsymbol{\mathcal{P}} = \{[i, j]^T \,|\, i, j = -2, 0, 2\}. \tag{7}$$

Then, in this section, Jacobians of the bi-camera error term to both the camera pose and the inverse depth will be deduced in detail.

Jacobian to the pose. The Jacobian J_p of the bi-camera error term $\varepsilon^{bi}_{\boldsymbol{p}_G}$ to camera C_i's pose ξ_{C_iG} can be expressed as,

$$J_p = \frac{\partial \varepsilon^{bi}_{\boldsymbol{p}_G}}{\partial \xi^T_{GC_i}}. \tag{8}$$

It can be decomposed to 4 parts with the chain rule,

$$J_p = \frac{\partial \varepsilon^{bi}_{\boldsymbol{p}_G}}{\partial \boldsymbol{I}_{C_i}} \cdot \frac{\partial \boldsymbol{I}_{C_i}}{\partial \boldsymbol{p}^T_{C_i}} \cdot \frac{\partial \boldsymbol{p}_{C_i}}{\partial \boldsymbol{P}^T_{C_i}} \cdot \frac{\partial \boldsymbol{P}_{C_i}}{\partial \xi^T_{C_iG}}. \tag{9}$$

Next, we discuss these 4 parts one by one.

(1) $\partial \varepsilon^{bi}_{\boldsymbol{p}_G} / \partial \boldsymbol{I}_{C_i}$ is the derivative of the error $\varepsilon^{bi}_{\boldsymbol{p}_G}$ to pixel intensities of image $\boldsymbol{I}_{C_i}$. Actually, from Eq. 6, it's easy to know that this term is equal to one,

$$\frac{\partial \varepsilon^{bi}_{\boldsymbol{p}_G}}{\partial \boldsymbol{I}_{C_i}} = 1. \tag{10}$$

(2) $\partial \boldsymbol{I}_{C_i} / \partial \boldsymbol{p}^T_{C_i}$ is the average intensity gradient, which is generally computed by the Sobel operator, of image $\boldsymbol{I}_{C_i}$ at all the pixels in the local window $\boldsymbol{\mathcal{P}}$ whose center is $\boldsymbol{p}_{C_i}$. Actually, this term can also be approximated just by the intensity gradient at $\boldsymbol{p}_{C_i}$ (the window of the Sobel operator needs to be enlarged accordingly). Thus, $\partial \boldsymbol{I}_{C_i} / \partial \boldsymbol{p}^T_{C_i}$ can be given as,

$$\frac{\partial \boldsymbol{I}_{C_i}}{\partial \boldsymbol{p}^T_{C_i}} = \begin{bmatrix} \frac{\partial I_{C_i}}{\partial u_{C_i}} & \frac{\partial I_{C_i}}{\partial v_{C_i}} \end{bmatrix} \triangleq \begin{bmatrix} \nabla \boldsymbol{I}^{u_{C_i}}_{C_i} & \nabla \boldsymbol{I}^{v_{C_i}}_{C_i} \end{bmatrix} \tag{11}$$

where u_{C_i} and v_{C_i} are both coordinate values of $\boldsymbol{p}_{C_i}$.

(3) $\partial \boldsymbol{p}_{C_i} / \partial \boldsymbol{P}^T_{C_i}$ is the derivative of a pixel's 2D coordinate to its 3D position in the camera coordinate system. From the pin-hole

camera model, we have

$$\frac{\partial \boldsymbol{p}_{C_i}}{\partial \boldsymbol{P}_{C_i}^T} = \begin{bmatrix} \frac{f_x^i}{Z_{C_i}} & 0 & -\frac{f_x^i X_{C_i}}{Z_{C_i}^2} \\ 0 & \frac{f_y^i}{Z_{C_i}} & -\frac{f_y^i Y_{C_i}}{Z_{C_i}^2} \end{bmatrix} \tag{12}$$

where f_x^i and f_y^i are focal lengths of C_i, and X_{C_i}, Y_{C_i} and Z_{C_i} are coordinate values in three axes of $\boldsymbol{P}_{C_i}$ in C_i's coordinate system.

(4) $\partial \boldsymbol{P}_{C_i}/\partial \boldsymbol{\xi}_{C_iG}^T$ is the derivative of the 3D point $\boldsymbol{P}_{C_i}$ to the camera pose $\boldsymbol{\xi}_{C_iG}$,

$$\frac{\partial \boldsymbol{P}_{C_i}}{\partial \boldsymbol{\xi}_{C_iG}^T} = \begin{bmatrix} \boldsymbol{I}_{3\times3} & -\boldsymbol{P}_{C_i}^{\wedge} \end{bmatrix} \tag{13}$$

where $\boldsymbol{I}$ is a 3×3 identity matrix and $\boldsymbol{P}_{C_i}^{\wedge}$ is the 3×3 anti-symmetric matrix generated from $\boldsymbol{P}_{C_i}$. By merging the four terms in Eqs. 10∼13, we can get the final form of the Jacobian $\boldsymbol{J}_p$,

$$\boldsymbol{J}_p = \begin{bmatrix} \nabla \boldsymbol{I}_{C_i}^{u_{C_i}} & \nabla \boldsymbol{I}_{C_i}^{v_{C_i}} \end{bmatrix} \begin{bmatrix} \frac{f_x^i}{Z_{C_i}} & 0 & -\frac{f_x^i X_{C_i}}{Z_{C_i}^2} \\ 0 & \frac{f_y^i}{Z_{C_i}} & -\frac{f_y^i Y_{C_i}}{Z_{C_i}^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{3\times3} & -\boldsymbol{P}_{C_i}^{\wedge} \end{bmatrix}. \tag{14}$$

Jacobian to the inverse depth. The Jacobian $\boldsymbol{J}_d$ of the bi-camera error term $\varepsilon_{\boldsymbol{p}_G}^{bi}$ to point $\boldsymbol{p}_{C_j}$'s inverse depth $\lambda_{\boldsymbol{p}_G}^{C_j}$ can be expressed as,

$$J_d = \frac{\partial \varepsilon_{\boldsymbol{p}_G}^{bi}}{\partial \lambda_{\boldsymbol{p}_G}^{C_j}}. \tag{15}$$

With the chain rule, it can also be decomposed as,

$$J_d = \frac{\partial \varepsilon_{\boldsymbol{p}_G}^{bi}}{\boldsymbol{P}_{C_i}^T} \cdot \frac{\partial \boldsymbol{P}_{C_i}}{\partial \boldsymbol{P}_{C_j}^T} \cdot \frac{\partial \boldsymbol{P}_{C_j}}{\partial \lambda_{\boldsymbol{p}_G}^{C_j}}. \tag{16}$$

Next, these three simpler parts are discussed one by one.

(1) $\partial \varepsilon_{\boldsymbol{p}_G}^{bi}/\partial \boldsymbol{P}_{C_i}^T$ is the derivative of the error $\varepsilon_{\boldsymbol{p}_G}^{bi}$ to $\boldsymbol{p}_G$'s corresponding 3D position $\boldsymbol{P}_{C_i}$ in C_i's camera coordinate system. This term can be obtained by combining Eqs. 10 ∼ 12, which is given as,

$$\frac{\partial \varepsilon_{\boldsymbol{p}_G}^{bi}}{\boldsymbol{P}_{C_i}^T} = \begin{bmatrix} \nabla \boldsymbol{I}_{C_i}^{u_{C_i}} & \nabla \boldsymbol{I}_{C_i}^{v_{C_i}} \end{bmatrix} \begin{bmatrix} \frac{f_x^i}{Z_{C_i}} & 0 & -\frac{f_x^i X_{C_i}}{Z_{C_i}^2} \\ 0 & \frac{f_y^i}{Z_{C_i}} & -\frac{f_y^i Y_{C_i}}{Z_{C_i}^2} \end{bmatrix}. \tag{17}$$

(2) $\partial \boldsymbol{P}_{C_i}/\partial \boldsymbol{P}_{C_j}^T$ is the derivative of $\boldsymbol{P}_{C_i}$'s 3D coordinate in C_i's coordinate system to its corresponding 3D point in C_j's coordinate system. This term is given as,

$$\frac{\partial \boldsymbol{P}_{C_i}}{\partial \boldsymbol{P}_{C_j}^T} = \boldsymbol{T}_{C_iG}\boldsymbol{T}_{C_jG}^{-1} = \boldsymbol{T}_{C_iC_j}. \tag{18}$$

where $\boldsymbol{T}_{C_iC_j}$ is the relative pose of C_j to C_i.

(3) $\partial \boldsymbol{P}_{C_j}/\partial \lambda_{\boldsymbol{p}_G}^{C_j}$ is the derivative of a 3D point $\boldsymbol{P}_{C_j}$ to its inverse depth. It can be expressed as,

$$\frac{\partial \boldsymbol{P}_{C_j}}{\partial \lambda_{\boldsymbol{p}_G}^{C_j}} = -\frac{1}{(\lambda_{\boldsymbol{p}_G}^{C_j})^2}\boldsymbol{K}_{C_j}^{-1}\boldsymbol{p}_{C_j} = -\frac{1}{(\lambda_{\boldsymbol{p}_G}^{C_j})}\boldsymbol{P}_{C_j}. \tag{19}$$

Thus, the final form of the Jacobian $\boldsymbol{J}_d$ is given by,

$$\begin{aligned} \boldsymbol{J}_d &= -\frac{1}{(\lambda_{\boldsymbol{p}_G}^{C_j})} \begin{bmatrix} \nabla \boldsymbol{I}_{C_i}^{u_{C_i}} & \nabla \boldsymbol{I}_{C_i}^{v_{C_i}} \end{bmatrix} \begin{bmatrix} \frac{f_x^i}{Z_{C_i}} & 0 & -\frac{f_x^i X_{C_i}}{Z_{C_i}^2} \\ 0 & \frac{f_y^i}{Z_{C_i}} & -\frac{f_y^i Y_{C_i}}{Z_{C_i}^2} \end{bmatrix} \boldsymbol{T}_{C_iC_j}\boldsymbol{P}_{C_j} \\ &= -\frac{1}{(\lambda_{\boldsymbol{p}_G}^{C_j})} \begin{bmatrix} \nabla \boldsymbol{I}_{C_i}^{u_{C_i}} & \nabla \boldsymbol{I}_{C_i}^{v_{C_i}} \end{bmatrix} \begin{bmatrix} \frac{f_x^i}{Z_{C_i}} & 0 & -\frac{f_x^i X_{C_i}}{Z_{C_i}^2} \\ 0 & \frac{f_y^i}{Z_{C_i}} & -\frac{f_y^i Y_{C_i}}{Z_{C_i}^2} \end{bmatrix} \boldsymbol{P}_{C_i}. \end{aligned} \tag{20}$$

As all derivative relationships between the bi-camera error term and the optimized variables have been deduced, the objective function of ROECS can then be minimized with any non-linear optimization scheme and thereby we can get accurate extrinsics of the SVS.

REFERENCES

[1] Fenglei Du and Michael Brady. 1993. Self-calibration of the Intrinsic Parameters of Cameras for Active Vision Systems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'93)*. IEEE, New York, USA, 477–482. https://doi.org/10.1109/CVPR.1993.341087

[2] Xuan Shao, Xiao Liu, Lin Zhang, Shengjie Zhao, Ying Shen, and Yukai Yang. 2019. Revisit Surround-view Camera System Calibration. In *International Conference on Multimedia and Expo (ICME'19)*. IEEE, Shanghai, China, 1486–1491. https://doi.org/10.1109/ICME.2019.00257

[3] Zhengyou Zhang. 1999. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In *IEEE International Conference on Computer Vision (ICCV'99)*. IEEE, Kerkyra, Greece, 666–673. https://doi.org/10.1109/ICCV.1999.791289

[4] Haijiang Zhu, Jinfu Yang, and Zhongtian Liu. 2009. Fisheye Camera Calibration with Two Pairs of Vanishing Points. In *International Conference on Information Technology and Computer Science (ITCS'09)*. IEEE, Kiev, Ukraine, 321–324. https://doi.org/10.1109/ITCS.2009.72