Scale Estimation of Monocular SfM for a Multi-modal Stereo Camera

Shinya Sumikura[1], Ken Sakurada[2],
Nobuo Kawaguchi[1], and Ryosuke Nakamura[2]

[1] Nagoya University, Japan
sumikura@ucl.nuee.nagoya-u.ac.jp, kawaguti@nagoya-u.jp
[2] National Institute of Advanced Industrial Science and Technology, Japan
{k.sakurada,r.nakamura}@aist.go.jp

Abstract. This paper proposes a novel method of estimating the absolute scale of monocular SfM for a multi-modal stereo camera. In the fields of computer vision and robotics, scale estimation for monocular SfM has been widely investigated in order to simplify systems. This paper addresses the scale estimation problem for a stereo camera system in which two cameras capture different spectral images (e.g., RGB and FIR), whose feature points are difficult to directly match using descriptors. Furthermore, the number of matching points between FIR images can be comparatively small, owing to the low resolution and lack of thermal scene texture. To cope with these difficulties, the proposed method estimates the scale parameter using batch optimization, based on the epipolar constraint of a small number of feature correspondences between the invisible light images. The accuracy and numerical stability of the proposed method are verified by synthetic and real image experiments.

1 Introduction

This paper addresses the problem of estimating the scale parameter of monocular Structure from Motion (SfM) for a multi-modal stereo camera system (Fig. 1). There has been growing interest in scene modeling with the development of mobile digital devices. In particular, researchers in the field of computer vision and robotics have exhaustively investigated scale estimation methods for monocular SfM to benefit from the simplicity of the camera system [5,14]. There are several ways to estimate the scale parameter — for example, integration with other sensors such as inertial measurement units (IMUs) [19] or navigation satellite systems (NSSs), such as the Global Positioning System (GPS). Also, some methods utilize the prior knowledge of the sensor setups [13,23]. In this paper, the scale parameter of monocular SfM is estimated by integrating the information of different spectral images, such as those taken by RGB and far-infrared (FIR) cameras in a stereo camera setup, whose feature points are difficult to directly match by using descriptors (e.g., SIFT [15], SURF [2], and ORB [22]).

With the development of the production techniques of FIR cameras, they have been widely utilized for deriving the benefits of thermal information in

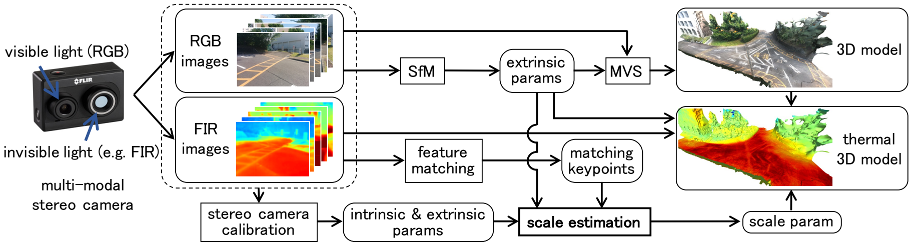


Fig. 1. Flowchart of the proposed scale estimation and the application example: thermal 3D reconstruction.

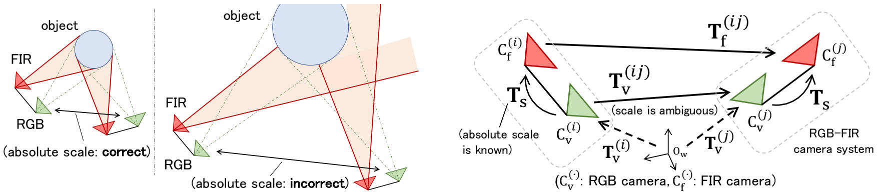
the form of infrared radiation emitted by objects, such as infrastructure inspection [8,11,16,29,30], pedestrian detection in the dark [3], and monitoring volcanic activity [27]. Especially for unmanned aerial vehicles (UAVs), a stereo pair of RGB and FIR cameras, which we call a *multi-modal stereo camera*, is often mounted on the UAV for such inspection and monitoring. Although the multi-modal stereo camera can capture different spectral images simultaneously, for example, in the case of structural inspection, it is labor-intensive to compare a large number of image pairs. To improve the efficiency of the inspection, SfM [1,24] and Multi-View Stereo (MVS) [7,12,25] can be used for *thermal 3D reconstruction* (Fig. 1). The estimation of the absolute scale of the monocular SfM is needed in order to project FIR image information to the 3D model (Fig. 2a). However, it is difficult to match feature points between RGB and FIR images directly. Moreover, the number of matching points between FIR images is comparatively small due to the low resolution and the lack of thermal texture in a scene. Although machine learning methods, such as deep neural networks (DNNs) [6,9,31], can be used to match feature points between different types of images, the cost of dataset creation for every camera and scene is quite expensive.

To estimate the scale parameter from only the information of the multi-modal camera system, we leverage the stereo setup with a constant extrinsic parameter and a small number of feature correspondences between the same modal images other than the visible ones (Fig. 1). More concretely, the proposed method is based on a least-squares method of residuals by the epipolar constraint between the same modal images. The main contribution of this paper is threefold: first, the formulation of the scale estimation for a multi-modal stereo camera system; second, the verification of the effectiveness of the formulation through synthetic and real image experiments; and third, experimental thermal 3D mappings as one of the applications of the proposed method.

2 Related work

2.1 Thermal 3D reconstruction

The FIR camera is utilized with other types of sensors for thermal 3D reconstruction because the texture of FIR images is poorer than that of visible ones,


object
FIR
RGB
(absolute scale: **correct**)
object
FIR
RGB
(absolute scale: **incorrect**)
$\mathrm{C}_\mathrm{f}^{(i)}$
$\mathbf{T}_\mathrm{f}^{(ij)}$
$\mathrm{C}_\mathrm{f}^{(j)}$
$\mathbf{T}_\mathrm{s}$
$\mathbf{T}_\mathrm{v}^{(ij)}$
(scale is ambiguous)
$\mathbf{T}_\mathrm{s}$
(absolute scale is known)
$\mathrm{C}_\mathrm{v}^{(i)}$
$\mathbf{T}_\mathrm{v}^{(i)}$
o_w
$\mathbf{T}_\mathrm{v}^{(j)}$
$\mathrm{C}_\mathrm{v}^{(j)}$
RGB–FIR camera system
($\mathrm{C}_\mathrm{v}^{(\cdot)}$: RGB camera, $\mathrm{C}_\mathrm{f}^{(\cdot)}$: FIR camera)

(a) Diagrams of thermal projection

(b) Definition of camera poses

Fig. 2. **(a)** Examples of projection when the absolute scale of RGB camera poses is correct (*left*) or incorrect (*right*). Green and red lines indicate the projection of the object in the RGB and FIR images, respectively. When the scale is incorrect, the reprojection of the FIR images is misaligned with the object. **(b)** Definition of camera poses for the i^{th} and j^{th} viewpoints. $\mathbf{T}_\mathrm{v}^{(\cdot)}$, $\mathbf{T}_\mathrm{f}^{(\cdot)}$ and $\mathbf{T}_\mathrm{s}$ represent the global poses of the RGB camera $\mathrm{C}_\mathrm{v}^{(\cdot)}$, FIR camera $\mathrm{C}_\mathrm{f}^{(\cdot)}$, and the relative pose between them, respectively. $\mathbf{T}_{(\cdot)}^{(ij)}$ represents the relative pose between the same type of cameras, $\mathrm{C}_{(\cdot)}^{(i)}$ and $\mathrm{C}_{(\cdot)}^{(j)}$.

especially for indoor scenes. Oreifej et al. [20] developed a fully automatic 3D thermal mapping system for building interiors using light detection and ranging (LiDAR) sensors to directly measure the depth of a scene. Additionally, depth image sensors are utilized to estimate the dense 3D model of a scene based on the Kinect Fusion algorithm [17] in the works of [16,29].

A combination of SfM and MVS is an alternative method for the 3D scene reconstruction. Ham et al. [8] developed a method to directly match feature points between RGB and FIR images, which works only in rich thermal-texture environments. Under similar conditions, the method proposed by Truong et al. [21] performs SfM using each of RGB and FIR images independently, aligning the two sparse point clouds.

Whereas the measurement range of the LiDAR sensor is longer than that of the depth image sensor, it has disadvantages in sensor size and weight, and is more expensive compared to RGB and depth cameras. Additionally, the depth image sensor can directly obtain dense 3D point clouds of a scene; however, it is unsuitable for wide-area measurement tasks because the measurement range is comparatively short. As mentioned, this study assumes thermal 3D reconstruction of wide areas for structural inspection by UAVs as an application. Thus, this paper proposes a scale estimation method of monocular SfM for a multi-modal stereo camera with the aim of thermal 3D reconstruction using an RGB–FIR camera system.

2.2 Scale estimation for monocular SfM

There are several types of scale estimation methods for monocular SfM based on other sensors and prior knowledge.

To estimate the absolute scale parameter of monocular SfM, an IMU is utilized as an internal sensor to integrate the information of the accelerations and angular velocities with vision-based estimation using the extended Kalman filter

(EKF) [19]. As an external sensor, location information from NSSs (e.g., GPS) can be used to estimate the similarity transformation between the trajectories of monocular SfM and the GPS information based on a least-squares method.

Otherwise, prior knowledge of the sensor setups is utilized for scale estimation. Scaramuzza et al. [23] exploit the nonholonomic constraints of a vehicle on which a camera is mounted. The work by Kitt et al. [13] utilizes ground planar detection and the height from the ground of a camera.

The objective of this study is to estimate the scale parameter of monocular SfM from only multi-modal stereo camera images without other sensor information, for versatility. For example, in the case of structural inspection using UAVs, IMUs mounted on the drones suffer from vibration noise, and the GPS signal cannot be received owing to the structure. Additionally, assumptions of sensor setups restrain the application of scale estimation. Therefore, the proposed method utilizes only input image information and pre-calibration parameters.

As one of the scale estimation methods for a multi-modal stereo camera, which uses the information only from such a camera system, Truong et al. [21] proposed a method based on an alignment of RGB and FIR point clouds. This method requires the point cloud created only from FIR images. Thus, it is not applicable to scenes with non-rich thermal texture, such as indoor scenes. Otherwise, considering a multi-modal stereo camera as a multi-camera cluster with non-overlapping fields of view, we can theoretically apply scale estimation methods of monocular SfM for such a multi-camera cluster to a multi-modal stereo camera. The work by Clipp et al. [4] estimates the absolute scale of monocular SfM for a multi-camera cluster with non-overlapping fields of view by minimizing the residual based on the epipolar constraint between two viewpoints. This method does not perform the batch optimization, which utilizes multiple image pairs, and does not take the scale parameter into account when performing the bundle adjustment (BA) [28].

Thus, in this paper, we compare the proposed scale estimation method with the ones of Truong et al. [21] and by Clipp et al. [4].

3 Scale estimation

3.1 Problem formulation

In this section, we describe a novel method of estimating a scale parameter of reconstruction results from monocular SfM. Here we use a stereo system of RGB and FIR cameras (i.e., RGB–FIR) as an example of a multi-modal stereo camera system. Fig. 2b expresses the global and relative transformation matrices of a system composed of two viewpoints with an RGB–FIR camera system.

We start with a given set of RGB images $\left\{\mathrm{I}_\mathrm{v}^{(1)}, \mathrm{I}_\mathrm{v}^{(2)}, \cdots, \mathrm{I}_\mathrm{v}^{(n)}\right\}$, and FIR images $\left\{\mathrm{I}_\mathrm{f}^{(1)}, \mathrm{I}_\mathrm{f}^{(2)}, \cdots, \mathrm{I}_\mathrm{f}^{(n)}\right\}$, whose k^th images, $\mathrm{I}_\mathrm{v}^{(k)}$ and $\mathrm{I}_\mathrm{f}^{(k)}$, are taken simultaneously using an RGB–FIR camera system whose constant extrinsic parameter is

$$\mathbf{T}_\mathrm{s} = \begin{bmatrix} \mathbf{R}_\mathrm{s} & \mathbf{t}_\mathrm{s} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix}. \tag{1}$$

$\mathbf{R}_{\mathrm{s}}$ and $\mathbf{t}_{\mathrm{s}}$ represent the rotation matrix and the translation vector between the two cameras of the camera system, respectively. Those matrix and vector are estimated via calibration in advance. Additionally, we assume that the k^{th} images, $\mathrm{I}_{\mathrm{v}}^{(k)}$ and $\mathrm{I}_{\mathrm{f}}^{(k)}$, are taken by the k^{th} cameras, $\mathrm{C}_{\mathrm{v}}^{(k)}$(RGB) and $\mathrm{C}_{\mathrm{f}}^{(k)}$(FIR), with the global extrinsic parameters, $\mathbf{T}_{\mathrm{v}}^{(k)}$ and $\mathbf{T}_{\mathrm{f}}^{(k)}$, respectively. Note that $\mathrm{C}_{\mathrm{v}}^{(k)}$ and $\mathrm{C}_{\mathrm{f}}^{(k)}$ comprise the pair of cameras in the RGB–FIR camera system. $\{\mathbf{T}_{\mathrm{v}}^{(k)}\}$ can be estimated except for its absolute scale by monocular SfM of the RGB images.

Using $\mathbf{T}_{\mathrm{v}}^{(i)}$ and $\mathbf{T}_{\mathrm{v}}^{(j)}$, the relative transformation between $\mathrm{C}_{\mathrm{v}}^{(i)}$ and $\mathrm{C}_{\mathrm{v}}^{(j)}$ is computed by $\mathbf{T}_{\mathrm{v}}^{(j)}\mathbf{T}_{\mathrm{v}}^{(i)^{-1}}$. To solve the scale ambiguity, a scale parameter $s \in \mathbb{R}$ is introduced. Then, the relative transformation $\mathbf{T}_{\mathrm{v}}^{(ij)}$ between $\mathrm{C}_{\mathrm{v}}^{(i)}$ and $\mathrm{C}_{\mathrm{v}}^{(j)}$ including the scale parameter s is expressed by

$$\mathbf{T}_{\mathrm{v}}^{(ij)} = \begin{bmatrix} \mathbf{R}_{\mathrm{v}}^{(ij)} & s \cdot \mathbf{t}_{\mathrm{v}}^{(ij)} \\ \mathbf{0}^{\mathsf{T}} & 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{R}_{\mathrm{v}}^{(ij)}$ and $\mathbf{t}_{\mathrm{v}}^{(ij)}$ are the rotation matrix block and the translation vector block of $\mathbf{T}_{\mathrm{v}}^{(j)}\mathbf{T}_{\mathrm{v}}^{(i)^{-1}}$, respectively. The goal is to estimate the correct $s \in \mathbb{R}$.

3.2 Derivation of scale parameter s

With $\mathbf{T}_{\mathrm{v}}^{(ij)}$ and $\mathbf{T}_{\mathrm{s}}$, the relative transformation $\mathbf{T}_{\mathrm{f}}^{(ij)} = \mathbf{T}_{\mathrm{s}}\mathbf{T}_{\mathrm{v}}^{(ij)}\mathbf{T}_{\mathrm{s}}^{-1}$ between the two FIR cameras, $\mathrm{C}_{\mathrm{f}}^{(i)}$ and $\mathrm{C}_{\mathrm{f}}^{(j)}$, can be computed as

$$\mathbf{T}_{\mathrm{f}}^{(ij)} = \begin{bmatrix} \mathbf{R}_{\mathrm{s}}\mathbf{R}_{\mathrm{v}}^{(ij)}\mathbf{R}_{\mathrm{s}}^{-1} & s \cdot \mathbf{R}_{\mathrm{s}}\mathbf{t}_{\mathrm{v}}^{(ij)} + (\mathbf{I} - \mathbf{R}_{\mathrm{s}}\mathbf{R}_{\mathrm{v}}^{(ij)}\mathbf{R}_{\mathrm{s}}^{-1})\mathbf{t}_{\mathrm{s}} \\ \mathbf{0}^{\mathsf{T}} & 1 \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} \mathbf{A}^{(ij)} & s \cdot \mathbf{b}^{(ij)} + \mathbf{c}^{(ij)} \\ \mathbf{0}^{\mathsf{T}} & 1 \end{bmatrix}, \tag{4}$$

where $\mathbf{A}^{(ij)} = \left[\mathbf{a}_1^{(ij)} \middle| \mathbf{a}_2^{(ij)} \middle| \mathbf{a}_3^{(ij)}\right] = \mathbf{R}_{\mathrm{s}}\mathbf{R}_{\mathrm{v}}^{(ij)}\mathbf{R}_{\mathrm{s}}^{-1}$, $\mathbf{b}^{(ij)} = \mathbf{R}_{\mathrm{s}}\mathbf{t}_{\mathrm{v}}^{(ij)}$ and $\mathbf{c}^{(ij)} = (\mathbf{I} - \mathbf{R}_{\mathrm{s}}\mathbf{R}_{\mathrm{v}}^{(ij)}\mathbf{R}_{\mathrm{s}}^{-1})\mathbf{t}_{\mathrm{s}}$. An essential matrix $\mathbf{E}^{(ij)}$ between $\mathrm{C}_{\mathrm{f}}^{(i)}$ and $\mathrm{C}_{\mathrm{f}}^{(j)}$ can be derived from $\mathbf{T}_{\mathrm{f}}^{(ij)}$ and expressed as

$$\mathbf{E}^{(ij)} = \left[s\mathbf{b}^{(ij)} + \mathbf{c}^{(ij)}\right]_{\times}\mathbf{A}^{(ij)} \tag{5}$$

$$\begin{aligned} &= s \cdot \left[\mathbf{b}^{(ij)} \times \mathbf{a}_1^{(ij)} \middle| \mathbf{b}^{(ij)} \times \mathbf{a}_2^{(ij)} \middle| \mathbf{b}^{(ij)} \times \mathbf{a}_3^{(ij)}\right] \\ &\quad + \left[\mathbf{c}^{(ij)} \times \mathbf{a}_1^{(ij)} \middle| \mathbf{c}^{(ij)} \times \mathbf{a}_2^{(ij)} \middle| \mathbf{c}^{(ij)} \times \mathbf{a}_3^{(ij)}\right]. \end{aligned} \tag{6}$$

The epipolar constraint between the two FIR images, $\mathrm{I}_{\mathrm{f}}^{(i)}$ and $\mathrm{I}_{\mathrm{f}}^{(j)}$, corresponding to the FIR cameras, $\mathrm{C}_{\mathrm{f}}^{(i)}$ and $\mathrm{C}_{\mathrm{f}}^{(j)}$, is formulated as

$$\mathbf{p}_k^{(j)^{\mathsf{T}}}\mathbf{E}^{(ij)}\mathbf{p}_k^{(i)} = 0, \tag{7}$$

where $\mathbf{p}_k^{(i)} = \left[x_k^{(i)},\, y_k^{(i)},\, 1\right]^\mathsf{T}$ and $\mathbf{p}_k^{(j)} = \left[x_k^{(j)},\, y_k^{(j)},\, 1\right]^\mathsf{T}$ are the k^th corresponding feature points between $\mathrm{I}_\mathrm{f}^{(i)}$ and $\mathrm{I}_\mathrm{f}^{(j)}$, in the form of normalized image coordinates [10]. A normalized image point $\mathbf{p}_k^{(i)}$ is defined as

$$\mathbf{p}_k^{(i)} = \mathbf{K}_\mathrm{f}^{-1} \left[u_k^{(i)},\, v_k^{(i)},\, 1\right]^\mathsf{T} \quad \text{with} \quad \mathbf{K}_\mathrm{f} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{8}$$

where $\mathbf{K}_\mathrm{f}$ is the intrinsic parameter matrix of the FIR camera. $\left[u_k^{(i)}, v_k^{(i)}\right]$ is the feature point in pixels in $\mathrm{I}_\mathrm{f}^{(i)}$ and is the k^th corresponding feature point with $\left[u_k^{(j)}, v_k^{(j)}\right]$ in $\mathrm{I}_\mathrm{f}^{(j)}$. Additionally, the normalized image point is also defined as

$$\mathbf{p}_k^{(i)} = \mathbf{X}_l^{(i)} \Big/ Z_l^{(i)}, \tag{9}$$

where $\mathbf{X}_l^{(i)} = [X_l^{(i)}, Y_l^{(i)}, Z_l^{(i)}]^\mathsf{T}$ is the l^th 3D point in the coordinate system of the i^th FIR camera $\mathrm{C}_\mathrm{f}^{(i)}$. Here, $\mathbf{X}_l^{(i)}$ corresponds to the feature point $\mathbf{p}_k^{(i)}$ on $\mathrm{I}_\mathrm{f}^{(i)}$.

The epipolar constraint of Equation (7) can be expanded to

$$\mathbf{u}_k^{(ij)} \left(s \cdot \mathbf{f}^{(ij)} + \mathbf{g}^{(ij)}\right) = 0 \tag{10}$$

with

$$\mathbf{u}_k^{(ij)} = [x_k^{(i)} x_k^{(j)},\, x_k^{(i)} y_k^{(j)},\, x_k^{(i)},\, y_k^{(i)} x_k^{(j)},\, y_k^{(i)} y_k^{(j)},\, y_k^{(i)},\, x_k^{(j)},\, y_k^{(j)},\, 1], \tag{11}$$

$$\mathbf{f}^{(ij)} = \left[\left[\mathbf{b}^{(ij)} \times \mathbf{a}_1^{(ij)}\right]_1, \left[\mathbf{b}^{(ij)} \times \mathbf{a}_1^{(ij)}\right]_2, \cdots, \left[\mathbf{b}^{(ij)} \times \mathbf{a}_3^{(ij)}\right]_2, \left[\mathbf{b}^{(ij)} \times \mathbf{a}_3^{(ij)}\right]_3\right]^\mathsf{T}, \tag{12}$$

$$\mathbf{g}^{(ij)} = \left[\left[\mathbf{c}^{(ij)} \times \mathbf{a}_1^{(ij)}\right]_1, \left[\mathbf{c}^{(ij)} \times \mathbf{a}_1^{(ij)}\right]_2, \cdots, \left[\mathbf{c}^{(ij)} \times \mathbf{a}_3^{(ij)}\right]_2, \left[\mathbf{c}^{(ij)} \times \mathbf{a}_3^{(ij)}\right]_3\right]^\mathsf{T}. \tag{13}$$

If the coordinates of the feature points have no error, Equation (10) is completely satisfied. However, in reality, the equation is not completely satisfied because coordinates of feature points usually have some error and the scale s is unknown. In such a case, the scalar residual $e_k^{(ij)}$ is defined as

$$e_k^{(ij)} = \mathbf{u}_k^{(ij)} \left(s \cdot \mathbf{f}^{(ij)} + \mathbf{g}^{(ij)}\right). \tag{14}$$

Likewise, the residual vector $\mathbf{e}^{(ij)}$ can be defined by

$$\mathbf{e}^{(ij)} = \mathbf{U}^{(ij)} \left(s \cdot \mathbf{f}^{(ij)} + \mathbf{g}^{(ij)}\right) \;\text{ with }\; \mathbf{U}^{(ij)} = \left[\; {\mathbf{u}_1^{(ij)}}^\mathsf{T} \middle| \; {\mathbf{u}_2^{(ij)}}^\mathsf{T} \middle| \cdots \middle| \; {\mathbf{u}_n^{(ij)}}^\mathsf{T} \;\right]^\mathsf{T}, \tag{15}$$

where n is the number of corresponding feature points between $\mathrm{I}_\mathrm{f}^{(i)}$ and $\mathrm{I}_\mathrm{f}^{(j)}$. Using a least-squares method, the scale parameter s can be estimated by

$$s = \operatorname*{arg\,min}_{s \,\in\, \mathbb{R}} \frac{1}{2} \sum_{i,j,i \neq j} \left|\left|\mathbf{e}^{(ij)}\right|\right|^2. \tag{16}$$

Collectively, the scale estimation problem comes down to determining s, such that the error function,

$$J(s) = \frac{1}{2} \sum_{i,j,i \neq j} \left|\left| \mathbf{U}^{(ij)} \left(s \cdot \mathbf{f}^{(ij)} + \mathbf{g}^{(ij)} \right) \right|\right|^2 \tag{17}$$

is minimized. Thus, the scale parameter s is determined by solving the equation $\mathrm{d}J(s)/\mathrm{d}s = 0$ in terms of s. Therefore, the scale s is computed by

$$s = - \sum_{i,j,i \neq j} \left(\mathbf{f}^{(ij)^\mathsf{T}} \mathbf{U}^{(ij)^\mathsf{T}} \mathbf{U}^{(ij)} \mathbf{g}^{(ij)} \right) \Big/ \sum_{i,j,i \neq j} \left(\mathbf{f}^{(ij)^\mathsf{T}} \mathbf{U}^{(ij)^\mathsf{T}} \mathbf{U}^{(ij)} \mathbf{f}^{(ij)} \right). \tag{18}$$

3.3 Alternative derivation

In Equation (2), the scale parameter s and the relative translation vector $\mathbf{t}_\mathrm{v}^{(ij)}$ between the two RGB cameras, $\mathrm{C}_\mathrm{f}^{(i)}$ and $\mathrm{C}_\mathrm{f}^{(j)}$, are multiplied. The scale parameter s can be alternatively applied to the translation vector $\mathbf{t}_\mathrm{s}$ in $\mathbf{T}_\mathrm{s}$, in contrast to Equations (1) and (2). This introduction of s is reasonable because multiplying $\mathbf{t}_\mathrm{s}$ by s is geometrically equivalent to multiplying $\mathbf{t}_\mathrm{v}^{(ij)}$ by $1/s$. Therefore, we can also estimate the scale parameter of monocular SfM, which has scale ambiguity, from

$$\mathbf{T}_\mathrm{v}^{(ij)} = \begin{bmatrix} \mathbf{R}_\mathrm{v}^{(ij)} & \mathbf{t}_\mathrm{v}^{(ij)} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{T}_\mathrm{s} = \begin{bmatrix} \mathbf{R}_\mathrm{s} & s \cdot \mathbf{t}_\mathrm{s} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix}. \tag{19}$$

When using Equation (19) for scale estimation, the $\mathbf{A}^{(ij)}$, $\mathbf{b}^{(ij)}$ and $\mathbf{c}^{(ij)}$ in Equation (4) are

$$\mathbf{A}^{(ij)} = \mathbf{R}_\mathrm{s} \mathbf{R}_\mathrm{v}^{(ij)} \mathbf{R}_\mathrm{s}^{-1}, \; \mathbf{b}^{(ij)} = (\mathbf{I} - \mathbf{R}_\mathrm{s} \mathbf{R}_\mathrm{v}^{(ij)} \mathbf{R}_\mathrm{s}^{-1}) \mathbf{t}_\mathrm{s} \; \text{ and } \; \mathbf{c}^{(ij)} = \mathbf{R}_\mathrm{s} \mathbf{t}_\mathrm{v}^{(ij)}. \tag{20}$$

The rest of the derivation procedure remains the same.

Hereinafter, the formula for the scale estimation based on Equations (1) and (2) is called Algorithm (1), whereas the formula based on Equation (19) is called Algorithm (2).

3.4 Scale-oriented bundle adjustment

After an initial estimation of the scale parameter by Equation (18) of Algorithm (1) or (2), we perform the bundle adjustment (BA) [28]. Before the scale estimation, the camera poses of the RGB cameras are precisely estimated via monocular SfM, except for its absolute scale. Thus, our BA optimizes the scale parameter s rather than the translation vectors of the RGB cameras.

Using the scale parameter s, the reprojection error $\boldsymbol{\delta}_{k,l}^{(i)}$ of the l^th FIR 3D point $\mathbf{X}_l = [X_l, \, Y_l, \, Z_l]^\mathsf{T}$ (in the world coordinate system) in the FIR image $\mathrm{I}_\mathrm{f}^{(i)}$ is defined as

$$\boldsymbol{\delta}_{k,l}^{(i)} = \mathbf{x}_k^{(i)} - \pi^{(i)} \left(s, \mathbf{X}_l \right), \tag{21}$$

where $\mathbf{x}_k^{(i)}$ represents the k^{th} feature point in the i^{th} FIR image $\mathrm{I}_{\mathrm{f}}^{(i)}$ and corresponds to $\mathbf{X}_l$. The projection function $\pi^{(i)}(\cdot)$ for the i^{th} FIR camera is

$$\pi^{(i)}\left(s, \mathbf{X}_l\right) = \left[f_x X_l^{(i)} / Z_l^{(i)} + c_x,\ f_y Y_l^{(i)} / Z_l^{(i)} + c_y\right]^{\mathsf{T}}, \tag{22}$$

where $\mathbf{X}_l^{(i)} = [X_l^{(i)},\ Y_l^{(i)},\ Z_l^{(i)}]^{\mathsf{T}}$ is computed by

$$\mathbf{X}_l^{(i)} = \mathbf{R}_{\mathrm{s}} \mathbf{R}_{\mathrm{v}}^{(i)} \mathbf{X}_l + s \cdot \mathbf{R}_{\mathrm{s}} \mathbf{t}_{\mathrm{v}}^{(i)} + \mathbf{t}_{\mathrm{s}} \quad \big(\text{when using Algorithm (1)}\big), \tag{23}$$

$$\mathbf{X}_l^{(i)} = \mathbf{R}_{\mathrm{s}} \mathbf{R}_{\mathrm{v}}^{(i)} \mathbf{X}_l + \mathbf{R}_{\mathrm{s}} \mathbf{t}_{\mathrm{v}}^{(i)} + s \cdot \mathbf{t}_{\mathrm{s}} \quad \big(\text{when using Algorithm (2)}\big). \tag{24}$$

The cost function $\mathcal{L}(\cdot)$ composed of the reprojection errors is defined by

$$\mathcal{L}\left(s, \left\{\mathbf{X}_l\right\}, \mathbf{K}_{\mathrm{f}}; \left\{\mathbf{T}_{\mathrm{v}}^{(i)}\right\}, \mathbf{T}_{\mathrm{s}}\right) = \sum_{i,k,l} \rho_{\mathrm{h}} \left(\left|\left| \boldsymbol{\delta}_{k,l}^{(i)} \right|\right|^2 \Big/ \sigma_{\mathrm{r}}^2 \right), \tag{25}$$

where $\rho_{\mathrm{h}}(\cdot)$ is the Huber loss function and σ_{r} is the standard deviation of the reprojection errors. The optimized scale parameter s is estimated as follows:

$$s = \underset{s \in \mathbb{R}, \{\mathbf{X}_l\}, \mathbf{K}_{\mathrm{f}}}{\arg\min}\ \mathcal{L}\left(s, \left\{\mathbf{X}_l\right\}, \mathbf{K}_{\mathrm{f}}; \left\{\mathbf{T}_{\mathrm{v}}^{(i)}\right\}, \mathbf{T}_{\mathrm{s}}\right). \tag{26}$$

Equation (26) is a non-convex optimization problem. Thus, it should be solved using iterative methods such as the Levenberg–Marquardt algorithm, for which an initial value is acquired by Equation (18) of Algorithm (1) or (2). See the details of the derivation above in Section 1 of the supplementary material paper.

4 Synthetic image experiments

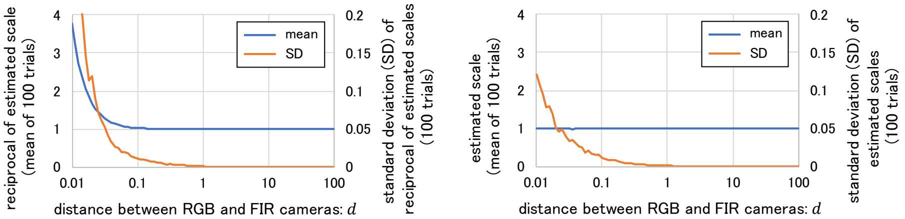
In Section 3, we described the two approaches of resolving scale ambiguity, with differences in the placement of the scale parameter s. In this section, we investigate, via simulation, the effect of noise given to feature points on scale estimation accuracy when varying the baseline length between the two cameras of the multi-modal stereo camera system.

The scale parameter is estimated in the synthetic environment with noise in both Algorithms (1) and (2). Preliminary experiments in the synthetic environment show that scale parameters can be estimated correctly using the proposed method when no noise is added to the feature points. See the details under the noise-free settings in Section 2 of the supplementary material paper.

4.1 Experimental settings

The procedure for the synthetic image experiments is as follows:

1. Scatter 3D points $\mathbf{X}_i \in \mathbb{R}^3$ $(i = 1, 2, \cdots, n_{\mathrm{p}})$ randomly in a cubic space with a side length of D.



(a) Algorithm (1)

(b) Algorithm (2)

Fig. 3. Mean (left vertical axes) and standard deviation (right vertical axes) of a hundred estimated scales under feature point noise of $\sigma_{\mathrm{n}} = 0.001$. Both horizontal axes represent the baseline length d between the two cameras of the RGB–FIR camera system, which is varied in the range of $[10^{-2}, 10^{2}]$. Note that the true value of the scale $s_{\mathrm{true}} = 1.0$ here. The accuracy and stability of the estimated scales are different between (a) and (b), especially for $d < 0.1$.

2. Arrange n_{c} RGB–FIR camera systems in the 3D space randomly. More concretely, a constant relative transformation of an RGB–FIR camera system $\mathbf{T}_{\mathrm{s}}$ is given, and the absolute camera poses of the RGB cameras $\mathbf{T}_{\mathrm{v}}^{(k)}$ ($k = 1, 2, \cdots, n_{\mathrm{c}}$) are set randomly. Then, the absolute camera poses of the FIR cameras $\mathbf{T}_{\mathrm{f}}^{(k)}$ ($k = 1, 2, \cdots, n_{\mathrm{c}}$) are computed by $\mathbf{T}_{\mathrm{f}}^{(k)} = \mathbf{T}_{\mathrm{s}}\mathbf{T}_{\mathrm{v}}^{(k)}$.
3. For all $k = 1, 2, \cdots, n_{\mathrm{c}}$, reproject the 3D points $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$, $\mathbf{X}_{n_{\mathrm{p}}}$ to the k^{th} FIR camera using $\mathbf{T}_{\mathrm{f}}^{(k)}$. Then, determine the normalized image points $\mathbf{p}_i^{(k)}$ ($i = 1, 2, \cdots, n_{\mathrm{p}}$) using Equation (9). Gaussian noise with a standard deviation $\sigma_{\mathrm{n}} \geq 0$ can be added to all of the reprojected points.
4. Estimate the scale parameter s, using both Algorithms (1) and (2) with outlier rejection based on Equation (14). Note that the true value of the scale parameter is 1.0 because the RGB camera positions are not scaled.

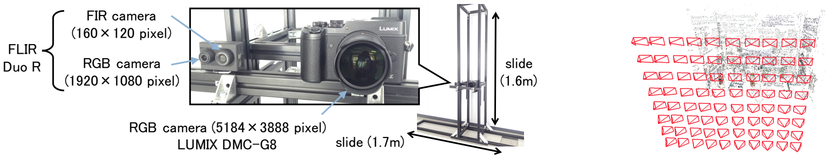
In this paper, we define $n_{\mathrm{p}} = 1000$, $D = 2000$ and $n_{\mathrm{c}} = 100$. In addition, the relative pose $\mathbf{T}_{\mathrm{s}}$ between the two cameras of the camera system is set as

$$\mathbf{T}_{\mathrm{s}} = \begin{bmatrix} \mathbf{R}_{\mathrm{s}} & \mathbf{t}_{\mathrm{s}} \\ \mathbf{0}^{\mathsf{T}} & 1 \end{bmatrix} \text{ with } \mathbf{R}_{\mathrm{s}} = \mathbf{I} \text{ and } \mathbf{t}_{\mathrm{s}} = \begin{bmatrix} d & 0 & 0 \end{bmatrix}^{\mathsf{T}}, \tag{27}$$

where $d > 0$ is the distance between the two cameras of the RGB–FIR camera system. d and σ_{n} are set depending on the simulation.

4.2 Effects of feature point detection error

We consider the effect of noise given to feature points on scale estimation accuracy when varying a baseline length of the stereo camera system. Setting $\sigma_{\mathrm{n}} = 0.001$, we estimate scale parameters s 100 times and compute a mean and a standard deviation (SD) of $1/s$ $\big($in Algorithm (1)$\big)$ or s $\big($in Algorithm (2)$\big)$, with respect to each of the various baseline lengths d between the two cameras



(a) Camera system and its mount **(b)** Grid-aligned cameras

Fig. 4. (a) Camera system and its mount used in our experiment. The camera mount is used to capture images along the grid-aligned viewpoints. The stage, to which the camera system is fixed, can be moved in both vertical and horizontal directions. **(b)** Grid-aligned camera poses estimated by SfM. The images are captured using (a).

of the camera system. Fig. 3 shows the relationship between d, the means and the SDs of the estimated scales for both Algorithms (1) and (2).

In Fig. 3b, the scale parameters are stably estimated in the region where d is relatively large ($0.1 < d$) because the means are $s = s_{\text{true}} = 1.0$ and the SDs converge to 0.0. On the contrary, in the region where d is relatively small ($d < 0.1$), the SD increases as d decreases but the means maintain the correct value of $s_{\text{true}} = 1.0$. Meanwhile, in Fig. 3a the means of the scale parameters are less accurate than the ones in Fig. 3b in the region where d is relatively small ($d < 0.1$). In addition, the SDs in Fig. 3a are larger than the ones in Fig. 3b.

Hence, it is concluded that the estimated scales obtained by Algorithm (2) are more accurate and stable than the ones obtained by Algorithm (1). Additionally, the baseline length between the two cameras of a multi-modal stereo camera system should be as long as possible for scale estimation.

5 Real image experiments

5.1 Evaluation method

We apply the proposed method to the experimental environment to verify that the method is capable of estimating the absolute scales of outputs from monocular SfM which uses a multi-modal stereo camera. For this verification, we need to prepare results of monocular SfM in which the actual distances between the cameras are already known. Therefore in this experiment, the multi-modal stereo camera system is fixed to the stage on the camera mount as shown in Fig. 4a, and we capture RGB and FIR images while moving the camera system on a grid of 100[mm] intervals. The stage of the camera mount, where the camera system is fixed, can be moved in both vertical and horizontal directions. Fig. 4b shows an example of grid-aligned camera poses estimated by SfM, whose images are captured using the camera mount shown in Fig. 4a.

Let $d^{(ij)}$ be the actual distance between the two RGB cameras $(\mathrm{C}_{\mathrm{v}}^{(i)}, \mathrm{C}_{\mathrm{v}}^{(j)})$ and $L^{(ij)}$ be the distance between $(\mathrm{C}_{\mathrm{v}}^{(i)}, \mathrm{C}_{\mathrm{v}}^{(j)})$ in the result of the monocular SfM, which has scale ambiguity. The estimated actual distance $\hat{d}^{(ij)}$ is computed by

$$\hat{d}^{(ij)} = s \cdot L^{(ij)} \big(\text{in Algorithm (1)}\big) \text{ and } \hat{d}^{(ij)} = L^{(ij)}/s \; \big(\text{in Algorithm (2)}\big), \quad (28)$$

where s is the scale parameter in Equation (2) and Equation (19), respectively. Additionally, the relative error $\epsilon^{(ij)}$ of $\hat{d}^{(ij)}$ can be defined as

$$\epsilon^{(ij)} = \frac{\hat{d}^{(ij)} - d^{(ij)}}{d^{(ij)}} \times 100[\%]. \quad (29)$$

The RGB–FIR camera system used in our experiment is shown in Fig. 4a. The RGB camera in the camera system is a LUMIX DMC–G8 (Panasonic Corp.) or the RGB camera part of a FLIR Duo R (FLIR Systems, Inc.), depending on the experimental setting of the baseline length. The FIR camera is the FIR camera part of the FLIR Duo R.

The procedure for the experiment is as follows:

1. Capture the RGB and FIR image pairs using the camera system and its mount shown in Fig. 4a. Additionally, some supplementary RGB and FIR images are added to stabilize the process of monocular SfM and scale estimation.
2. Perform a process of monocular SfM using the captured RGB images.
3. Compute feature point matches of the FIR images using SIFT [15] descriptor, whose outliers are rejected via RANdom SAmple Consensus (RANSAC) based on a five-point algorithm [18,26].
4. Estimate the scale parameter by Algorithms (1) and (2).
5. Compute a mean of $\epsilon^{(ij)}$ with all the combinations, which is defined as

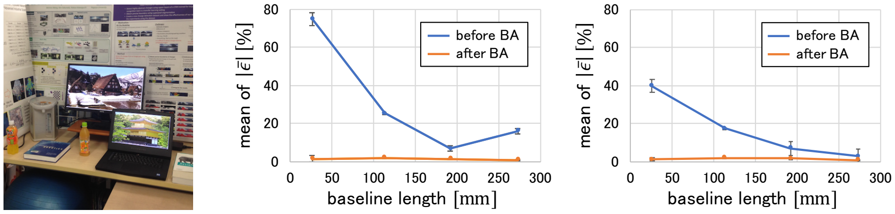
$$\bar{\epsilon} = \frac{1}{N(N-1)/2} \sum_{i<j} \epsilon^{(ij)}, \quad (30)$$

where N is the number of RGB images taken in a grid.

When detecting and describing feature points, FIR images are converted to gray-scaled images. FLIR Duo R outputs FIR images whose pixels contain values of radiation temperature. To convert them to gray-scaled images, a mean μ and a standard deviation σ_{p} of pixels for each image are computed, and then pixel values with a range of $[\mu - 2\sigma_{\mathrm{p}}, \mu + 2\sigma_{\mathrm{p}}]$ are mapped to $[0, 2^8 - 1]$.

To confirm the effect of the difference in baseline lengths between the RGB and FIR cameras, datasets of RGB and FIR images are taken with each of the four baseline lengths of the camera system: 273[mm], 192[mm], 113[mm] and 26[mm]. The systems with the first, second, and third baseline lengths use the LUMIX DMC–G8 as the RGB camera. The system with 26[mm] uses the RGB camera equipped on the FLIR Duo R. Considering the randomness of RANSAC, for each of the four baseline lengths, the scale estimation and computation of $\bar{\epsilon}$ are performed 100 times. Then, a mean and a standard deviation of $|\bar{\epsilon}|$ are calculated.

Also, pre-calibration of an RGB–FIR stereo camera system is needed to perform the proposed scale estimation procedure. Thus, we adopt the stereo calibration method in which a planar pattern such as a chessboard is used [32]. See the details in Section 3 of the supplementary material paper.



(a) Target scene **(b)** Algorithm (1) **(c)** Algorithm (2)

Fig. 5. **(a)** Target scene of evaluation. **(b)**, **(c)** The means of $|\bar{\epsilon}|$ with 100 trials under various baseline setups $(26, 113, 192$ and 273[mm]), in both Algorithms (1) and (2). The error bars indicate the range of $\pm 1\sigma$ of $|\bar{\epsilon}|$. σ is a standard deviation of $|\bar{\epsilon}|$ with 100 trials. It is found that the means of $|\bar{\epsilon}|$ before BA decrease as the baseline length becomes larger in both (b) and (c). Additionally, the means of $|\bar{\epsilon}|$ in (b) are larger the ones in (c). On the contrary, after BA, the means of $|\bar{\epsilon}|$ approach nearly zero in both (b) and (c).

5.2 Evaluation with a real scene

The experimental environment used in the evaluation is shown in Fig. 5a. The grid pattern along which the camera system is moved has 8 vertical × 10 horizontal grids. Thus, there are 80 RGB camera poses used for the evaluation. Additionally, 50 supplementary pairs of RGB and FIR images are included to stabilize the process of monocular SfM and scale estimation. Considering the randomness of RANSAC, we show the means and standard deviations of $|\bar{\epsilon}|$ with 100 trials of scale estimation. Figs. 5b and 5c show the results when using Algorithms (1) and (2), respectively.

In both Figs. 5b and 5c, the means of $|\bar{\epsilon}|$ before BA decrease as the baseline length becomes larger. Additionally, the mean values in Fig. 5b are larger than the ones in Fig. 5c across the whole range of baseline length. Those results denote the same pattern as the experiments in the synthetic environment in Section 4. Consequently, without BA, it is evident that the smaller error of scale estimation occurs when using the camera system with the longer baseline as indicated by the simulation in Section 4. In addition, the difference in numerical stability of the proposed method occurs in experiments with both synthetic and real images.

On the contrary, after BA, the means of $|\bar{\epsilon}|$ approach nearly zero in both Figs. 5b and 5c, even though large error occurred before BA. Especially, at the 26[mm] baseline length in Fig. 5b, the mean of $|\bar{\epsilon}|$ after BA is 1.64[%] whereas it is 74.9[%] before BA. Additionally, at the 273[mm] baseline length after BA, high accuracy of the scale estimation is achieved as the means of $|\bar{\epsilon}|$ are 0.832[%] under Algorithm (1) and 0.876[%] under Algorithm (2). The SDs also decrease after BA compared to the ones before BA. Summarizing the above, we conclude that scale parameters estimated by both Algorithms (1) and (2) are suitable for an initial value of BA as well as that our BA effectively refines the scale parameters with respect to the accuracy and variance.

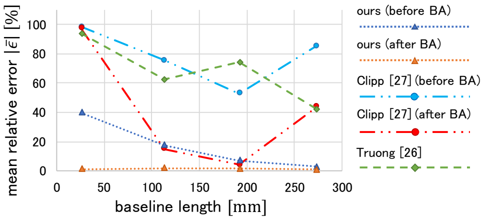
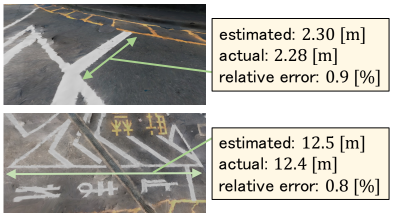


Fig. 6. The mean relative errors $|\bar{\epsilon}|$ acquired by Algorithm (2) of the proposed method (*ours*), [21] and [4]. *Ours* and [4] have randomness caused by RANSAC. Thus, we show the means of $|\bar{\epsilon}|$ with 100 trials for them.


estimated: 2.30 [m]
actual: 2.28 [m]
relative error: 0.9 [%]

estimated: 12.5 [m]
actual: 12.4 [m]
relative error: 0.8 [%]

Fig. 7. Evaluation of the practical result with road surface markings in the scene of Fig. 8a. High accuracy of the scale estimation is achieved as the relative errors are under 1.0[%].

5.3 Comparison with the existing method

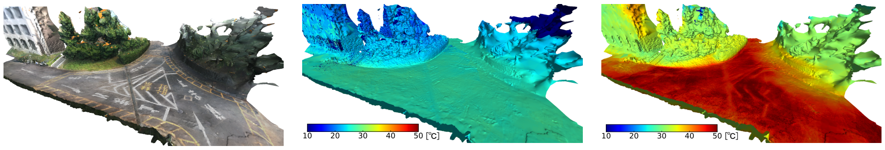
As mentioned in Section 2.2, we compare the proposed scale estimation method with the ones by Truong et al. [21] and by Clipp et al. [4]. We apply the two methods of [21] and [4] to the RGB–FIR image datasets used in Section 5.2, then evaluate the estimated scale parameter by calculating $\bar{\epsilon}$ accordingly. Fig. 6 shows the comparison of the accuracies of the scale parameters estimated by Algorithm (2) of the proposed method, [21] and [4]. The results of the proposed method and [4] present the means of $|\bar{\epsilon}|$ with 100 trials both before and after BA. In result by [21], we adopt s_t computed by Equation (6) in the paper of [21] as the scale parameter s.

As shown in Fig. 6, the $|\bar{\epsilon}|$ by [21] and [4] are much larger than the means of $|\bar{\epsilon}|$ by the proposed method throughout the whole range of baseline length. As for [21], the low accuracy mainly results from the erroneous 3D points reconstructed via SfM which uses only the FIR images. On the other hand, unlike our method, the method by [4] cannot deal with the epipolar residuals of multiple FIR image pairs. Thus, before BA, the means of $|\bar{\epsilon}|$ by [21] and [4] are much larger than the ones by the proposed method. Additionally, the BA in [4] does not optimize a scale parameter but rather rotations and translations. Thus, after BA, coupled with the poor initial estimation by [4], the BA is unstable as shown in Fig. 6.

5.4 Practical examples

Fig. 8 presents temporal thermal 3D mappings as a practical example of thermal 3D reconstruction. RGB and FIR images are captured by a smartphone-based RGB–FIR camera system, composed of a FLIR One (FLIR Systems, Inc.) and a smartphone. The baseline length of the camera system is 154[mm].

A 3D mesh model shown in Fig. 8a is reconstructed from the RGB images using monocular SfM and MVS, and is then resized to the absolute scale estimated by the proposed method. The thermal 3D models shown in Figs. 8b and 8c are built by reprojecting FIR images to the 3D mesh model on a sunny day and on a rainy day, respectively. The thermal information is reprojected well as shown



(a) 3D mesh model **(b)** on a rainy day **(c)** on a sunny day

Fig. 8. Examples of temporal thermal 3D modeling. The 3D mesh model reconstructed from RGB images is shown in (a). (b) and (c) show the thermal 3D reconstructions on a rainy day and on a sunny day, respectively.

in Figs. 8b and 8c. In addition, as shown in Fig. 7, we measure the size of road surface markings in the 3D model (*estimated*) and in the real world (*actual*), as an evaluation of the estimated scales in practical scenes. The relative errors of the estimated size are approximately 0.8[%] in the scene in Fig. 8. See the additional results in Section 4 of the supplementary material paper.

6 Conclusion

In this paper, we have shown a novel method of estimating the scale parameter of monocular SfM for a multi-modal stereo camera system, which is composed of different spectral cameras (e.g., RGB and FIR) in a stereo camera setup. Owing to the difficulty of matching feature points directly between RGB and FIR images, we have leveraged a constant extrinsic parameter of the stereo setup and a small number of feature correspondences between the same modal images. Two types of formulae for scale parameter estimation, both of which are based on the epipolar constraint, were proposed in this paper. We have also verified the difference in scale estimation accuracy and stability between the two formulae in the synthetic and real image experiments. The cause for the difference in scale estimation stability requires further investigation.

Additionally, we have demonstrated a scale estimation of monocular SfM under the experimental environment using an RGB–FIR stereo camera, and we have verified its accuracy both before and after BA. The consequence shows that the proposed method can estimate an appropriate scale parameter and its accuracy depends on the baseline length between RGB and FIR cameras of a stereo camera system. Moreover, we have presented the thermal 3D modeling as an application of the proposed scale estimation method.

These results suggest that the proposed method is applicable to the construction of thermal 3D mappings using payload-limited vehicles, such as UAVs, on which an RGB–FIR camera system is mounted. Therefore, we conclude that the proposed method is suitable for scale estimation of monocular SfM.

Acknowledgements. This research is supported by the Hori Sciences & Arts Foundation, the New Energy and Industrial Technology Development Organization (NEDO) and JSPS KAKENHI Grant Number 18K18071.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building Rome in a day. In: International Conference on Computer Vision (ICCV). pp. 72–79 (2009)
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European Conference on Computer Vision (ECCV). pp. 404–417 (2006)
3. Bertozzi, M., Broggi, A., Caraffi, C., Rose, M.D., Felisa, M., Vezzoni, G.: Pedestrian detection by means of far-infrared stereo vision. Computer Vision and Image Understanding **106**(2), 194–204 (2007)
4. Clipp, B., Kim, J.H., Frahm, J.M., Pollefeys, M., Hartley, R.: Robust 6dof motion estimation for non-overlapping, multi-camera systems. In: IEEE Workshop on Applications of Computer Vision (WACV) (2008)
5. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: Monoslam: Real-time single camera slam. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **29**(6), 1052–1067 (2007)
6. DeTone, D., Malisiewicz, T., Rabinovich, A.: Toward geometric deep slam. arXiv preprint arXiv:1707.07410 (2017)
7. Furukawa, Y., Ponce, J.: Accurate, Dense, and Robust Multi-View Stereopsis. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **32**(8), 1362–1376 (2010)
8. Ham, Y., Golparvar-Fard, M.: An automated vision-based method for rapid 3d energy performance modeling of existing buildings using thermal and digital imagery. Advanced Engineering Informatics **27**(3), 395–409 (2013)
9. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3279–3286 (2015)
10. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
11. Iwaszczuk, D., Stilla, U.: Camera pose refinement by matching uncertain 3d building models with thermal infrared image sequences for high quality texture extraction. ISPRS Journal of Photogrammetry and Remote Sensing **132**, 33–47 (2017)
12. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3121–3128 (2011)
13. Kitt, B.M., Rehder, J., Chambers, A.D., Schonbein, M., Lategahn, H., Singh, S.: Monocular visual odometry using a planar road model to solve scale ambiguity. In: European Conference on Mobile Robots (2011)
14. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: International Symposium on Mixed and Augmented Reality (ISMAR). pp. 225–234 (2007)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60**(2), 91–110 (2004)
16. Mller, A.O., Kroll, A.: Generating high fidelity 3-d thermograms with a handheld real-time thermal imaging system. IEEE Sensors Journal **17**(3), 774–783 (2017)
17. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: International symposium on Mixed and augmented reality (ISMAR). pp. 127–136 (2011)
18. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE transactions on pattern analysis and machine intelligence **26**(6), 756–770 (2004)

19. Nützi, G., Weiss, S., Scaramuzza, D., Siegwart, R.: Fusion of imu and vision for absolute scale estimation in monocular slam. Journal of Intelligent & Robotic Systems **61**(1), 287–299 (2011)
20. Oreifej, O., Cramer, J., Zakhor, A.: Automatic generation of 3d thermal maps of building interiors. ASHRAE transactions **120**, C1 (2014)
21. Phuc Truong, T., Yamaguchi, M., Mori, S., Nozick, V., Saito, H.: Registration of rgb and thermal point clouds generated by structure from motion. In: International Conference on Computer Vision Workshop (ICCVW) (2017)
22. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: International Conference on Computer Vision (ICCV). pp. 2564–2571 (2011)
23. Scaramuzza, D., Fraundorfer, F., Pollefeys, M., Siegwart, R.: Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In: International Conference on Computer Vision (ICCV). pp. 1413–1419 (2009)
24. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016)
25. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV). pp. 501–518 (2016)
26. Stewénius, H., Engels, C., Nistér, D.: Recent developments on direct relative orientation. ISPRS Journal of Photogrammetry and Remote Sensing **60**, 284–294 (2006)
27. Thiele, S.T., Varley, N., James, M.R.: Thermal photogrammetric imaging: A new technique for monitoring dome eruptions. Journal of Volcanology and Geothermal Research **337**(Supplement C), 140–145 (2017)
28. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Vision Algorithms: Theory and Practice. pp. 298–372 (1999)
29. Vidas, S., Moghadam, P., Bosse, M.: 3d thermal mapping of building interiors using an rgb-d and thermal camera. In: International Conference on Robotics and Automation (ICRA). pp. 2311–2318 (2013)
30. Weinmann, M., Leitloff, J., Hoegner, L., Jutzi, B., Stilla, U., Hinz, S.: Thermal 3d mapping for object detection in dynamic scenes. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **2**(1), 53 (2014)
31. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4353–4361 (2015)
32. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **22**, 1330–1334 (2000)