Joint Calibration Method of Thermal Infrared-Visible Based on Cross Modal Feature Matching

Shan Su [1], Li Yan [1, 2], Yuquan Zhou [1], Changjun Chen [1,2,*]

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China - sushan@whu.edu.cn (S.S.); lyan@sgg.whu.edu.cn (L.Y.); zhouyuquan@sgg.whu.edu.cn (Y.Z.); chencj@whu.edu.cn (C.C.)
[2] Hubei Luojia Laboratory, Wuhan 430079, China

Keywords: Binocular Stereo Vision System, Thermal Infrared Camera, RGB Camera, Calibration, Cross Modal.

Abstract

Aiming at the problem of limited imaging quality of monomodal optical cameras in low-light environments, this paper constructs a thermal infrared-RGB binocular stereo vision system and proposes a joint calibration framework for infrared and RGB cameras to provide a high-precision geometric alignment basis for multimodal image fusion. First, a high-precision geometric calibration method is used to eliminate the internal distortion of the infrared camera and establish the mapping relationship between its pixel coordinate system and physical space. Second, a cross-modal extrinsic calibration strategy based on common view targets is designed. A specially designed heated and temperature-controlled chessboard calibration board for thermal infrared is used to enhance the feature contrast in the infrared image through temperature control. Combined with a cross-modal feature matching algorithm, the spatial pose transformation matrix between the infrared and RGB cameras is accurately solved to align multimodal images. Experimental results show that the proposed thermal infrared–RGB binocular calibration method can significantly improve calibration accuracy and robustness, providing effective technical support for visual perception and target recognition in low-light environments.

1. Introduction

In low-light environments, monomodal optical cameras, limited by illumination conditions, often struggle to capture high-quality images, thereby restricting visual perception capabilities. For instance, in nighttime or low-illumination scenarios, images captured by RGB cameras may suffer from insufficient brightness, low contrast, and loss of details, severely affecting the usability of the images. This issue is particularly prominent in mobile mapping tasks, as mobile mapping systems (such as autonomous vehicles, UAV surveying, and mobile robots) need to obtain accurate environmental information in real time under complex and variable lighting conditions to achieve safe navigation, target recognition, and precise mapping. However, the insufficient performance of traditional monomodal optical imaging systems in low-light environments limits the application of mobile mapping technologies in nighttime or low-illumination scenarios.

To enhance visual perception capabilities and overcome the limitations of low-light environments, multimodal image fusion technology has emerged. Among these technologies, the fusion of infrared and RGB imaging has become a highly promising solution. Infrared imaging (Su et al., 2024), by capturing thermal radiation information, effectively addresses the perceptual failure of RGB in no-light, smoke/haze scenarios. It provides complementary details to RGB images, thereby significantly improving target recognizability and the accuracy of environmental perception. Therefore, this paper constructs a thermal infrared–RGB binocular system based on a thermal infrared camera and simultaneously implements a cross-modal alignment method for thermal infrared and RGB images. The fusion of thermal infrared and RGB binocular vision not only enhances image quality but also provides mobile mapping systems with more comprehensive environmental perception capabilities, thereby improving their robustness and reliability in complex environments.

The construction of a thermal infrared-RGB binocular system hinges on solving the geometric alignment issue between the two modalities, which relies on precise camera calibration techniques. However, during the calibration process, RGB imaging depends on corner/edge features, while infrared imaging relies on temperature gradient features. This results in limited feature visibility of traditional checkerboard calibration boards in infrared imaging, thereby restricting the accuracy and applicability of existing calibration methods. Moreover, when dealing with cross-modal extrinsic calibration, existing methods often overlook the differences in imaging characteristics between infrared and RGB, leading to insufficient calibration accuracy and robustness.

To address the aforementioned issues, this study proposes a joint calibration framework for infrared and RGB cameras, which is dedicated to providing a high-precision geometric alignment basis for multimodal image fusion. The core ideas of the research include two key steps:

Firstly, a high-precision geometric calibration method is employed to correct the internal distortion of the infrared camera. This process precisely establishes the mapping relationship between its pixel coordinate system and the physical space, thereby providing an accurate geometric foundation for subsequent multimodal image fusion. Secondly, for the extrinsic calibration between infrared and RGB cameras, a cross-modal calibration strategy based on common view targets is designed. A specially designed heated and temperature-controlled chessboard calibration board for thermal infrared imaging is utilized. This board presents significant temperature contrast in infrared images, enhancing the feature contrast in thermal infrared images while retaining the corner features of the chessboard in the RGB band. It effectively solves the problem of difficult feature point

extraction in infrared imaging due to insufficient temperature differences. Subsequently, a cross-modal feature matching method is used to accurately solve the spatial pose transformation matrix between the infrared and RGB cameras, ensuring the robustness of the alignment between thermal infrared and RGB images. The final calibration error is 0.295 pixels.

Therefore, the calibration method proposed in this paper is capable of solving the high-precision calibration problem between thermal infrared and RGB cameras. It further enhances the visual perception capabilities of mobile mapping systems in complex environments and provides a solid foundation for the application of mobile mapping technologies in a wider range of scenarios.

2. Related Works

2.1 Monocular camera calibration

2.1.1 Traditional calibration methods based on passive targets

Traditional single-camera calibration methods usually rely on calibration targets with checkerboard patterns or circular markers. These methods extract the coordinates of feature points (such as corners or centers of circles) and use their geometric mapping relationship with the camera's spatial coordinates for calibration, thereby solving for the camera parameters. The mathematical derivation of traditional calibration methods is rigorous, and the calculation process is efficient and concise, making them easy to implement and apply. The most classic calibration methods include the Tsai two-step method and the Zhang Zhengyou calibration method (Zhang et al., 2000).

However, the calibration accuracy of these traditional methods is highly dependent on the extraction accuracy of the feature points. In practical applications, they are easily affected by environmental factors, such as lighting conditions, the surface reflectivity of the calibration target, and the degree of camera distortion. Moreover, the placement posture of the calibration target and the camera's viewing angle must also be considered during the calibration process. If the angle between the calibration target and the camera's optical axis is too small, or if the calibration target does not fully cover the camera's field of view, it may lead to deviations in the calibration results.

Despite these limitations, traditional calibration methods have the advantages of simple operation and low cost. They can provide high calibration accuracy under ideal conditions.

2.1.2 Calibration method based on phase target

In recent years, calibration methods based on phase targets have gradually attracted attention. These methods use the phase information of fringe patterns as feature points and obtain the unwrapped phase map through algorithms such as phase-shifting or Fourier transform, thereby achieving high-precision calibration. For example, Ma (Ma et al., 2014) proposed a feature extraction method using fringe pattern sets as phase target features, and solved the wrapped phase through a three-step phase-shifting method.

However, the "defocused" images were obtained by applying Gaussian filtering to clear images, rather than real experimentally captured images. Wang (Wang et al., 2019) demonstrated the robustness and accuracy of the camera calibration method based on orthogonal fringes through simulation and experiments, but did not provide comparative experiments with other calibration methods under the same conditions in either simulation or experimental scenarios. Liu (Liu et al., 2024) proposed a calibration method based on encoded phase-shifting fringe patterns (Phase-Shifting Fringe, PSF), establishing a mapping relationship between the virtual phase plane and the original phase points through arbitrary quadrilateral interpolation, thereby achieving high-precision camera calibration. However, phase unwrapping is a key step in phase target calibration, but it is easily affected by noise and ambient light in practice, leading to discontinuities and errors in the phase map.

2.1.3 Other calibration methods

Genovese (Genovese et al., 2024) proposed a camera calibration method based on a single image. This method uses a random speckle pattern that covers the entire sensor and combines it with Digital Image Correlation (DIC) technology to achieve model-free distortion correction. Zhu (Zhu et al., 2024) proposed a calibration method based on monocular 3D priors, which is capable of recovering the complete 4-DOF (degrees of freedom) intrinsic parameters from monocular images without relying on specific 3D objects or strong geometric assumptions.

2.2 Multi camera calibration

Yang (Yang et al., 2024) addressed the alignment issue between thermal imagers and other sensors, proposing an autonomous targetless extrinsic calibration framework for thermal imagers, RGB cameras, and LiDAR sensors in mobile robots. By analyzing the characteristics of thermal imaging, they utilized thermal bridges and the PnL algorithm based on line features to achieve autonomous targetless calibration between LiDAR and RGB cameras as well as between LiDAR and thermal imagers.

Li and Cai (Li et al., 2023) proposed a calibration and real-time target matching method for a heterogeneous multi-camera system composed of thermal infrared cameras and visible spectrum (VS) cameras. This method enables better perception of surrounding information in complex environments and has been widely applied in many intelligent unmanned devices, such as drones and patrol robots. Edlinger (Edlinger et al., 2023) designed a calibration method for thermal imaging cameras. This method involves placing a calibration-patterned board on a heated background, solving the problem of traditional methods being unable to achieve calibration in the infrared spectrum.

3. Thermal Infrared -Visible Binocular System

3.1 Theoretical basis of binocular stereo vision

3.1.1 Pinhole Camera Model

The monocular pinhole camera model is the geometric foundation of optical imaging systems. Through this geometric model, the camera can map the coordinate points in a three-dimensional scene to a two-dimensional image plane. The mapping process is shown in figure 1. Here, $P_w(x_w, y_w, z_w)$ represents the coordinates of a 3D point in the world coordinate system, $P_{uv}(u, v)$ represents the pixel coordinates, O-x-y-z represents the world coordinate system, O'-x'-y'-z' represents the camera coordinate system, and f represents the focal length. The core idea is that light travels in straight lines, and after passing through the camera's optical center O, it is

projected onto the imaging plane. The pinhole camera model describes the transformation between the world coordinate system and the camera coordinate system using a rotation matrix R_w^c and a translation vector t_w^c, which represent the camera's pose. It then maps the 3D coordinates to normalized image coordinates through the intrinsic parameter matrix. The pinhole camera model provides the correspondence between the spatial location of objects and image pixels for monocular vision, and it is the theoretical foundation for understanding disparity calculation, depth recovery, and 3D reconstruction in binocular stereo vision.

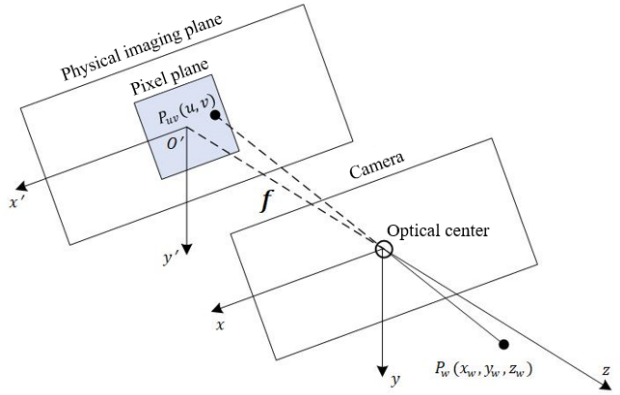

Figure 1. Mapping process of pinhole camera model

Project point A into the camera coordinate system based on the current image pose:

$$P_C = R_w^c P_w + t_w^c \tag{1}$$

Then, normalize P_C to obtain the projection X on the normalized image plane $(x_c / z, y_c / z, 1)$.

Project the normalized coordinates onto the pixel coordinate system based on the camera's focal length.

$$\begin{cases} u = \alpha f \dfrac{y_c}{z_c} + c_x \\ v = \beta f \dfrac{x_c}{z_c} + c_y \end{cases} \tag{2}$$

Here, α and β are the scaling factors in the directions of x and y, respectively, and c_x, c_y are the translations in the directions of x and y, respectively. After combining the scaling factors and translations, the transformation from the world coordinate system to the camera pixel coordinate system is ultimately given by:

$$P_{uv} = \frac{1}{z_c} K(R_w^c P_w + t_w^c) \tag{3}$$

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

K is the intrinsic parameter matrix, which is obtained through camera calibration.

3.1.2 Distortion Model

The discrepancy between the idealized assumptions of the pinhole camera model and actual optical systems is the fundamental cause of image distortion. Real lenses, due to non-ideal optical characteristics and off-axis aberrations, prevent light rays from strictly following the straight-line projection path. Additionally, material inhomogeneities (such as local refractive index variations caused by temperature gradients or manufacturing errors) further disrupt the uniform propagation of light rays. Mechanical assembly deviations can cause misalignment between the projection coordinate system and the ideal coordinate system. Thermal effects and environmental interferences, especially detector sensitivity drift or lens deformation caused by temperature changes in uncooled infrared cameras, introduce nonlinear radiometric distortions. The combined effect of these factors causes the actual imaging process to deviate from the ideal pinhole model. Therefore, precise distortion mathematical modeling and calibration techniques must be employed for correction, thereby enhancing the accuracy and reliability of applications such as photogrammetry and computer vision.

The ideal imaging formula for a pinhole camera is:

$$\begin{cases} x = \dfrac{X}{Z} \cdot f_x + c_x \\ y = \dfrac{Y}{Z} \cdot f_y + c_y \end{cases} \tag{5}$$

(X,Y,Z) represents the coordinates of a spatial point, f_x and f_y are the focal lengths, and (C_x,C_y) is the coordinates of the optical center O. In actual imaging, distortion terms need to be introduced to correct the above deviations.

The symmetric distortion caused by lens curvature has the following mathematical form:

$$\begin{cases} x' = x\left(1 + k_1 r^2 + k_2 r^4 + \cdots\right) \\ y' = y\left(1 + k_1 r^2 + k_2 r^4 + \cdots\right) \end{cases} \tag{6}$$

Here, $r^2 = x^2 + y^2$, and k_1, k_2 are the radial distortion coefficients, and higher-order terms can be neglected or fitted with a more complex polynomial. The distortion caused by the optical axis not being perpendicular to the lens plane or by assembly tilt has the form:

$$\begin{cases} x' = x + 2p_1 xy + p_2(r^2 + 2x^2) \\ y' = y + 2p_2 xy + p_1(r^2 + 2y^2) \end{cases} \tag{7}$$

p_1, p_2 are the tangential distortion coefficient, which is related to the tilt angle of the lens.

3.2 Thermal Infrared-Visible Calibration Board

To address the geometric distortion correction and radiometric consistency requirements of the uncooled thermal infrared-RGB binocular system, this paper designs a multimodal calibration board. This calibration board integrates high-precision geometric structures, dynamic temperature control modules, and multi-band radiometric characteristics. Geometric calibration is used to correct the spatial resolution degradation and nonlinear distortions of the thermal infrared camera caused by the thermal diffusion effect. Radiometric correction can establish a unified radiometric model between infrared and RGB images, thereby improving the depth estimation accuracy of binocular stereo vision. Additionally, the calibration board can simulate real-world temperature distributions (30–150°C) to verify the system's stability under extreme conditions, demonstrating its environmental adaptability. The core parameters of the calibration board are shown in Table 1.

This calibration board, featuring a 12×9 black-and-white checkerboard structure (with each cell measuring 25mm×25mm), achieves high-precision geometric distortion correction and integrates a closed-loop PID temperature control

system to support multi-physics coupling analysis. In infrared mode, the black cells have a high emissivity of up to 0.9 (in the 8–12μm wavelength band). Combined with the sub-pixel corner extraction algorithm (OpenCV cornerSubPix), it can enhance the corner positioning accuracy to ≤0.1 pixels. This effectively corrects the radial/tangential distortions and nonlinear responses of uncooled thermal infrared cameras caused by the thermal diffusion effect. Meanwhile, the temperature control module precisely regulates the target temperature (30–150℃, with an accuracy of ±1.5℃) through power-off/on cycles, simulating the radiative characteristics of high-temperature targets to verify sensor dark current drift and detector temperature sensitivity. In RGB mode, the low reflectivity of the white cells (<5%) significantly reduces stray light interference. This supports the disparity calculation and spatial consistency verification in the joint calibration of binocular cameras. Moreover, the dynamic characteristics of the temperature control module can further provide experimental evidence for the temperature drift correction of multimodal imaging systems.

Table 1. Calibration board parameters and performance

Parameters	Technical Specifications	Function
Size	Effective dimensions: 300mm × 225mm; Single cell: 25mm × 25mm.	Supports calibration of wide field-of-view imaging systems. The cell size is matched with the detector pixel size (17μm), which can be used to verify the spatial resolution (approximately 1.47 times the pixel size).
Material & Properties	Emissivity of black cells: 0.9 ± 0.05 (8–12μm wavelength band); Reflectivity of white cells: <5% (visible light wavelength band).	Ensures high contrast in infrared imaging and reduces the impact of ambient light interference on visible light calibration.
Temperature Control Module	Operating temperature range: 30–150℃, temperature control accuracy ±1.5℃.	Simulates the temperature distribution of real-world scene targets and verifies the stability of the thermal imager under extreme conditions.

3.3 Calibration method for thermal infrared-visible binocular system

3.3.1 Intrinsic Parameter Calibration Based on the Zhang Zhengyou Method

Zhang Zhengyou calibration algorithm is a widely used monocular camera calibration method. Its core lies in the detection of corner points and geometric constraints of a checkerboard calibration plate. By capturing images of a planar calibration plate with known spatial coordinates from different viewpoints, a linear equation system is established based on the correspondence between image corner points and real three-dimensional coordinates. The method estimates the intrinsic and extrinsic parameter matrices of the camera by minimizing the reprojection error. Its mathematical foundation is the camera imaging model:

$$\mathbf{X}_c = \mathbf{R}[\mathbf{X}_w - \mathbf{t}]\mathbf{K}^{-1} + \mathbf{d} \tag{8}$$

$\mathbf{X}_c$ represents the image coordinates, $\mathbf{X}_w$ represents the world coordinates, $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector, respectively, and $\mathbf{K}$ is the intrinsic parameter matrix (including the focal length f_x, f_y, and the principal point (c_x, c_y). The distortion vector $\mathbf{d}$ includes radial distortions k_1, k_2, and tangential distortions p_1, p_2. The distortion vector includes radial distortions k_1, k_2, and tangential distortions p_1, p_2. Assuming the calibration board is located in the plane Z=0, and high-precision 2D corner coordinates are obtained through sub-pixel corner detection, a linearized equation system is constructed. Based on this, the 3D point $X_w = (X_i, Y_i, 0)^T$ in the world coordinate system is projected into the camera coordinate system to obtain the normalized coordinates $\mathbf{X}_{c/n} = \mathbf{R}(\mathbf{X}_w - t) / Z_c^{'}$. By using the intrinsic parameter matrix $\mathbf{K}$, it is reprojected onto the image plane, and its predicted value is:

$$\hat{\mathbf{x}} = \mathbf{K}\mathbf{X}_{c/n} = \mathbf{K}\mathbf{R}(\mathbf{X}_w - \mathbf{t})\mathbf{K}^{-1}\mathbf{X}_w^{'} \tag{9}$$

After introducing radial distortions k_1, k_2 and tangential distortions p_1, p_2, the complete observation model is:

$$\mathbf{x}_i = \hat{\mathbf{x}} + \begin{bmatrix} k_1 r^2 + k_2 r^4 + p_1(2xy) + p_2(r^2 + 2x^2) \\ k_1 r^2 + k_2 r^4 + p_2(2xy) + p_1(r^2 + 2y^2) \end{bmatrix} \tag{10}$$

Here, $r^2 = x^2 + y^2$, (x,y) are the normalized image coordinates

To linearize this model, we perform a Taylor expansion on A and retain the first-order terms:

$$\mathbf{x}_i \approx \mathbf{A}\mathbf{X}_w^{'} + \mathbf{b} + \mathbf{d}(\mathbf{x}_i) \tag{11}$$

$\mathbf{A}$ is the combination of the camera matrix and the extrinsic parameters, $\mathbf{b}$ is the translation term, and $\mathbf{d}(\mathbf{x}_i)$ is the distortion vector based on the current estimate. Substituting all corner point observations $\mathbf{x}_{i,j}$ into the model, we construct a nonlinear least squares problem and estimate the camera parameters by minimizing the reprojection error:

$$\min_{\mathbf{K},\mathbf{R},\mathbf{t},k_1,k_2,p_1,p_2} \sum_{i,j} \left\| \mathbf{x}_{i,j} - \left(\mathbf{A}\mathbf{X}_{w,j}^{'} + \mathbf{b} + \mathbf{d}(\mathbf{x}_{i,j}) \right) \right\|^2 \tag{12}$$

3.3.2 Extrinsic Calibration Based on Cross-Modal Feature Matching

In the cross-modal extrinsic calibration task, this paper adopts the XoFTR (Cross-modal Feature Matching Transformer) algorithm proposed by Tuzcuoglu (Tuzcuoğlu et al., 2024) to achieve feature matching between RGB and infrared modalities. This method effectively addresses cross-modal differences through a two-stage training strategy: First, it employs Masked Image Modeling (MIM) pre-training to learn cross-modal features from visible-thermal image pairs. By randomly masking regions of the images and reconstructing the inter-modal correlation features, the model learns the common

expressions under heterogeneous radiative mechanisms. Subsequently, it applies a pseudo-thermal image augmentation strategy based on cosine transformation to non-linearly transform the intensity of RGB images, simulating the radiative property differences of thermal imaging and enhancing the model's adaptability to real cross-modal data. Its improved coarse-to-fine matching pipeline constructs many-to-many feature associations at 1/8 resolution through a multi-level Transformer architecture and then performs sub-pixel coordinate regression at 1/2 resolution using a custom decoder, significantly improving the matching accuracy of cross-modal features under view differences, scale changes, and low-texture scenes.

In this paper, we first utilize the feature matching capability of XoFTR to extract precise feature point correspondences from RGB and thermal infrared images. These correspondences not only have high accuracy at the pixel level but are also further refined by a sub-pixel refinement module to enhance the matching precision. To further improve the robustness of the calibration, we introduce a multi-view geometric constraint mechanism. By optimizing the extrinsic parameters across multiple viewpoints, we capture the geometric structure of the scene from different angles, thereby providing richer geometric information for the extrinsic calibration. We then employ the RANSAC algorithm to estimate the Essential Matrix between each pair of images, initially calculating the relative rotation and translation between cameras. Subsequently, using these preliminary estimated extrinsic parameters as initial values, we introduce a global optimization framework based on Bundle Adjustment. This framework minimizes the reprojection error and multi-view consistency error while optimizing the extrinsic parameters across all viewpoints, ensuring the stability and consistency of the calibration results across different scenes.

4. Experiments Results

In multi-camera systems, accurately obtaining the intrinsic and extrinsic parameters of cameras is crucial for subsequent image processing, 3D reconstruction, and visual measurement tasks. The Zhang Zhengyou calibration method, as a classic camera calibration technique, is widely used in various camera calibration scenarios due to its advantages in calibration accuracy and ease of operation. In this experimental section, the Zhang Zhengyou calibration method was employed to calibrate the intrinsic and extrinsic parameters of RGB and thermal infrared cameras. Additionally, an extrinsic calibration method based on cross-modal feature matching was used to verify and correct the relative pose of the binocular cameras, ensuring the accuracy of the extrinsic calibration. The RGB camera used is the RealSense D455, and the thermal infrared camera is the FLIR VUE PRO R. The binocular stereo vision system composed of these two cameras is shown in the figure, with 3D-printed structural components used for connection and fixation. Through careful design of the calibration process and multiple rounds of experimental data collection and analysis, the aim is to ensure the validity and reliability of the calibration method in this experimental environment and to provide an accurate parameter basis for subsequent visual applications based on these two cameras.

4.1 Calibration Results Using Zhang Zhengyou Method

4.1.1 Intrinsic Calibration

We employed the Zhang Zhengyou calibration method to calibrate the intrinsic parameters and perform image undistortion for both the infrared and RGB cameras. During this process, it is important to note that the calibration board should fill the entire image frame as much as possible to ensure a uniform distribution of feature points, thereby enhancing the correction accuracy. The undistortion results for RGB and thermal infrared images are shown in the figure 2. It can be observed that the RGB image exhibits little noticeable distortion, while the undistorted thermal infrared image shows more significant changes, with the checkerboard pattern being well-corrected.

4.1.2 Extrinsic Calibration

We used the undistorted RGB and thermal infrared images for extrinsic calibration. The specific experimental setup is shown in the figure 3.

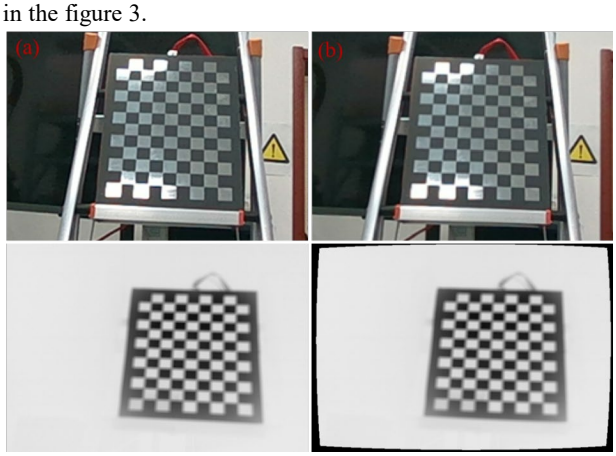

Figure 2. Image Undistortion

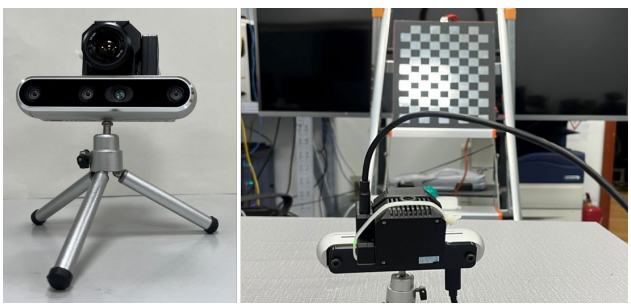

Figure 3. Experimental scene

However, due to the lower resolution and higher image noise of thermal infrared images, after using the Zhang Zhengyou method to obtain the relative pose between the RGB camera and the thermal infrared camera, we further employed a cross-modal feature matching method to perform sub-pixel registration between the two cameras. This step ensures and enhances the accuracy of the extrinsic calibration.

4.2 Extrinsic Calibration Based on Cross-Modal Feature Matching

This paper utilizes XoFTR to complete cross-modal feature matching and further calculates the relative pose between RGB and thermal infrared cameras. XoFTR is a cross-modal method for matching RGB and thermal infrared images. It addresses the matching challenges brought by modality differences

through pre-training with masked image modeling and fine-tuning with pseudo-thermal image augmentation. Meanwhile, XoFTR employs a coarse-to-fine matching process, performing coarse matching at the 1/8 scale and fine matching at the 1/2 scale, and enhancing precision through sub-pixel refinement.

To ensure the accuracy of cross-modal feature matching between RGB and thermal infrared images, this paper sets the confidence level for coarse matching feature points to 0.8, that is, only features with a confidence greater than 0.8 are retained. Similarly, the confidence level for fine matching feature points is set to 0.6. The final cross-modal feature matching results based on the XoFTR method are shown in Figure 4.

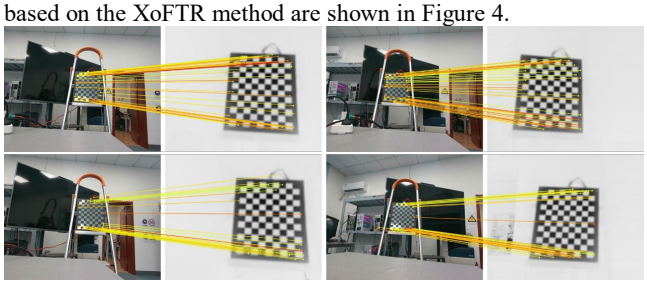

Figure 4. Cross-modal feature matching

We use the reprojection error as a quantitative metric to evaluate the accuracy of the extrinsic camera calibration. Specifically, we first use XoFTR to extract matching feature point pairs from RGB and thermal infrared images, compute the essential matrix using the eight-point algorithm, and simultaneously employ the RANSAC algorithm to eliminate mismatches, thereby enhancing the estimation accuracy of the essential matrix. Subsequently, we perform Singular Value Decomposition (SVD) on the essential matrix to obtain the relative rotation matrix and translation vector between the two cameras. We then reconstruct the feature points in 3D space using triangulation and reproject these 3D points onto the image plane using the estimated extrinsic parameters. The reprojection error is quantified by calculating the Euclidean distance between the reprojected points and the actually detected feature points. In this paper, we choose the mean value as the quantitative metric for calibration accuracy.

To further verify the reliability of the calibration results, we also conducted geometric consistency checks. Specifically, we selected multiple sets of matching feature points across several views and used these feature points to estimate the extrinsic parameters separately. By verifying whether the estimated extrinsic parameters satisfy geometric constraints such as coplanarity, parallelism, and perpendicularity, we assessed their geometric consistency. The final extrinsic parameter estimation error obtained in this paper was 0.295 pixels.

Therefore, the extrinsic calibration method proposed in this paper demonstrates high accuracy and stability in matching RGB and thermal infrared images. The binocular stereo vision system constructed based on this calibration method can effectively be applied to mobile measurement tasks in real-world scenarios.

5. Discussion and Conclusion

The calibration method proposed in this paper demonstrates its advantages in several aspects. First, the high-precision geometric calibration method is used to correct the internal distortions of the infrared camera, significantly improving the imaging quality of infrared images and providing an accurate geometric basis for subsequent multimodal image fusion. Second, the cross-modal extrinsic calibration strategy, using a heatable checkerboard calibration board, effectively addresses the issue of difficult feature point extraction in infrared imaging due to insufficient temperature differences, further enhancing the accuracy and robustness of the calibration. In addition, the cross-modal feature matching capability of the XoFTR algorithm further ensures the reliability of the calibration results.

However, our method also has some limitations. First, the calibration process requires the use of a specially designed heatable checkerboard calibration board, which increases the cost and complexity of the calibration equipment. Second, although the method demonstrates high calibration accuracy in the laboratory environment, environmental factors (such as temperature changes and lighting conditions) may have some impact on the calibration results in practical applications. Future work can further optimize the design of the calibration board to make it more adaptable to different environments. In addition, more efficient cross-modal feature matching algorithms can be explored to further improve the accuracy and efficiency of calibration.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 42394061).

References

Su, S., Yan, L., Xie, H., et al., 2024. Multi-Level Hazard Detection Using a UAV-Mounted Multi-Sensor for Levee Inspection. Drones, 8(3), 90. doi.org/10.3392/drones8030090.

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), 1330-1334. doi.org/10.1109/34.2000.888718.

Ma, M.C., Chen, X.C., Wang, K.Y., 2014. Camera calibration by using fringe patterns and 2D phase-difference pulse detection. Optik, 125(2), 671-674. doi.org/10.1016/j.ijleo.2012.06.045.

Ma, M.C., Chen, X.C., Wang, K.Y., 2014. Camera calibration by using fringe patterns and 2D phase-difference pulse detection. Optik, 125(2), 671-674. doi.org/10.1016/j.ijleo.2012.06.045.

Liu, C., Zhang, Q., Liang, F., et al., 2024. Effective camera calibration by using phase-shifting fringe patterns. Optics & Laser Technology, 169, 110084. doi.org/10.1016/j.optlastec.2024.110084.

Genovese, K., 2024. Single-image camera calibration with model-free distortion correction. Optics and Lasers in Engineering, 181, 108348. doi.org/10.1016/j.optlaseng.2024.05.038.

Zhu, S., Kumar, A., Hu, M., et al., 2024. Tame a wild camera: in-the-wild monocular camera calibration. Advances in Neural Information Processing Systems, 36.

Yang, M., 2024. Research on depth measurement calibration of light field camera based on Gaussian fitting. Scientific Reports, 14(1), 8774. doi.org/10.1038/s41598-024-62511-1.

Li, C., Cai, C., 2023. A calibration and real-time object matching method for heterogeneous multi-camera system. IEEE Transactions on Instrumentation and Measurement, 72, 1-12. doi.org/10.1109/TIM.2023.3260263.

Edlinger R, Himmelbauer G, Zauner G, et al. Visual odometry and mapping under poor visibility conditions using a stereo infrared thermal imaging system[J]. Electronic Imaging, 2023, 35: 1-7.

Tuzcuoğlu, Ö., Köksal, A., Sofu, B., Kalkan, S., Alatan, A.A., 2024. XoFTR: Cross-modal Feature Matching Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 4275-4286.