IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 9, 2020, accepted April 14, 2020, date of publication April 20, 2020, date of current version May 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988664

Large Field of View Cooperative Infrared and Visible Spectral Sensor for Visual Odometry

YUBO NI[1,2], YUE WANG[1,2], SHOUCHANG YANG[2], RUIXIANG CHEN[2], AND XIANGJUN WANG[1,2]

[1]State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China
[2]MOEMS Education Ministry Key Laboratory, Tianjin University, Tianjin 300072, China

Corresponding author: Xiangjun Wang (tjuxjw@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51575388, and in part by the Beijing Key Laboratory of Urban Spatial Information Engineering under Grant 2019207.

ABSTRACT Thermal images produced by the long-wavelength infrared (LWIR) camera are robust and independent from environmental illumination change. They can help the standard visible light camera working under the complicated environmental condition. Breaking through the traditional stereo multi-spectral sensor consisting of a visible-light camera and a LWIR camera, a novel architecture of large field of view (FOV) cooperated infrared and visible spectral sensor for visual odometry is proposed. The novel sensor is equipped with two visible cameras, four infrared cameras covering 120 degrees FOV in horizontal under both bands. Distribution of cameras and related peripheral devices are specifically designed which makes the sensor's volume less than 100 cm (length) $\times$ 10 cm (height) $\times$ 10 cm (width). The sensor's cameras calibration, distortion correction and measurement principle are elaborated. Feature-based method for visible and multi-windowed optimization-based image alignment for infrared is designed for the visual odometer based on the different imaging mechanism and distribution of cameras in the sensor. The frames and estimated poses management from both bands are proposed. Moreover, all proposed methodologies can be implemented in the sensor's embedded processor. The electrical power consumption is only 12W. Experiments of the sensor's evaluation are performed, experimental results show that large FOV cooperated multi-spectral cameras can efficiently improve the robustness of visual odometry. The real-time performance of the sensor is higher than 10fps with disparity map construction under both bands.

INDEX TERMS Infrared, visible, stereo vision, visual odometer, calibration, direct method.

I. INTRODUCTION

Localizing and estimating its ego-motion in 3D space are crucial tasks for autonomous vehicles, mobile robots and Unmanned Aerial Vehicles (UAVs) [1]. Currently, these tasks are achieved by using LiDAR's, monocular and stereo imagery, *etc*. LiDAR has already played an important role in this researching area. The main drawbacks of LiDAR are: The price of LiDAR is expensive; The weight, power consumption is not affordable for some platforms; As an active sensor, the signal noise of environments and others LiDAR's emission can affect LiDAR's performance [2]. Camera, as a passive sensor, has its own advantage in information acquirement, which can play an important part in 3D data capturing. Vision-based navigation is an important area of research in robotics' sensing and mapping, especially for simultaneous localization and mapping (SLAM). Stereo vision-based vision odometry is widely used and developed for 3D reconstruction, indoor localization and mapping *etc*. For constructed SLAM system, it's mainly divided into two parts: the front end and the back end [3]. The front end is the visual odometer (VO), which roughly estimates the motion of the camera based on the information of adjacent images and provides a good initial value for the back end. The back end is the optimization procedure for long time localization and mapping. The implementation methods of VO can be divided into two categories according to whether features are extracted or not: feature point-based methods and direct methods without feature points. For the complicated environment, only visible cameras or other spectral cameras are not suitable. The multi-spectral cameras based rig is developed and evaluated.

The associate editor coordinating the review of this manuscript and approving it for publication was Kai Li.

For surveillance systems, multi-spectral cameras have their advantage, such as working under low visibility or lighting conditions, adding a richer set of information based on the different reflection properties of the object. For environment information acquirement, such as some surveillance and driving assistance system, multi-spectral cameras rig composed a thermal (infrared) and an optical (visible) sensors (such as telescopic sight) are widely discussed and developed. Visible and infrared cameras are worked as complementary for 2D imagery data, the research is focusing on image fusion and registration [4], [5], feature matching and coexistence of visible and infrared bands' information [6], [7]. For visual SLAM, visible camera working in low or complicated illumination scene remain considerably challenging, thermal information based SLAM framework appeared. For vision odometer, the image fusion based method has no advantage in processing speed and efficiency. 3D data extraction and analyzing under these two bands, for example [8] use a bumblebee stereo vision camera cooperated with a Near-InfraRed (NIR) camera, constructed a sparse disparity map. [9] combine RGB information and thermal feature together, use back-end optimization with loop closure, update map and location. [10] cooperated thermal-infrared camera with LiDAR for density map construction. All these works use the thermal-infrared camera as complementary of other sensors. But these works' has limited FOV sensing in thermal.

Multi-camera rig constructed by [11], [12], uses four gray/color visible cameras recorded related rich textural visual feature under traffic scenarios. They use the wide-angle lens to cover related large FOV. For large FOV cameras' sensors, [13] use a fish-eye camera cooperated with stereo sensors, achieved a large FOV based dense mapping strategy. [14] use 16 NIR camera and fish-eye visible camera, covering 360-degree in the horizontal, constructed system for self-driving vehicle localization and 3D scene perception. [15] constructs a stereo embedded system for underwater imaging, with the sensor's captured information and Bundle Adjustment (BA), they mapped the underwater scene. For common visual odometry, all these instruments and related methods can achieve good results relying on back-end optimization such as photometric bundle adjustment. Their strategy of keyframe tracking and organization are implemented with back-end optimization together. Their instruments only captured information.

We constructed a multi-stereo visual odometer, which equipped long-wavelength infrared (LWIR) and visible color cameras in a limited space. The prototype is shown in figure 1. Our sensor is a typical front-end visual odometer for SLAM. As a visual odometer, our system not only captures images but also produce disparity maps, estimates the system's initial pose while the system is moving.

In this paper, Section II gives a overview of basic hardware construction of sensor. Section III details the proposed methodology and implementation strategy, including a comprehensive calibration strategy for this instrument and a distortion control for visible cameras *etc*. Some related performance evaluation are also demonstrated in section III. Section IV presents the experiments set-up, the key performance about our infrared based visual odometer strategy. Section V gives a overall assessment, section VI concludes this work and gives a preview on future work.

FIGURE 1. Prototype of cooperative infrared and visible spectral sensor.

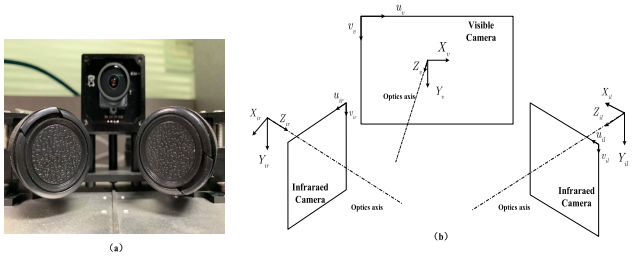


FIGURE 2. Coordinate defined of cameras in each group of the sensor; (a) Composition of one group with a visible camera and two infrared cameras; (b) Definition of each camera's coordination.

II. SET UP OF COOPERATIVE INFRARED AND VISIBLE SPECTRAL SENSOR

In this section, we briefly introduce our cooperative infrared and visible spectral sensor. It's designed as a passive sensor for autonomous mobile robots. To cover 5m to 30m depth of the scene, the distance of baseline (in u direction of cameras) between the same modal cameras is roughly 0.8m and 0.55m separately. The main features are:

1) Equipped imagers in system are only passive sensors: visible and infrared cameras, without any active lighting sources.
2) Relative large field of view(FOV): 120 degrees in horizontal for each visible camera and 60 degrees in horizontal for each infrared camera, which can sense wide area.
3) Embedded computing devices are equipped in sensor. It has imaging processing and 3D data analyzing processor.
4) All cameras and embedded equipment are arranged in a limited space;

A. HARDWARE SETUP

We equipped a rack with two visible color camera (25Hz, 1920 × 1080 pixels), four long-wavelength infrared (LWIR/FIR) detectors (50Hz, 384 × 288 pixels, detects radiations in the range 7.5∼ 14 μm). Advanced RISC Machines (ARM) based embedded processor and related peripheral devices (*i.e.* Nvidia TX2 and related PCB boards) are developed and deployed in our system. We place one visible and two infrared cameras as a group on the left side of the rack, the other group on the right side. The basic coordinate definition and architecture of one group are shown in figure 2.

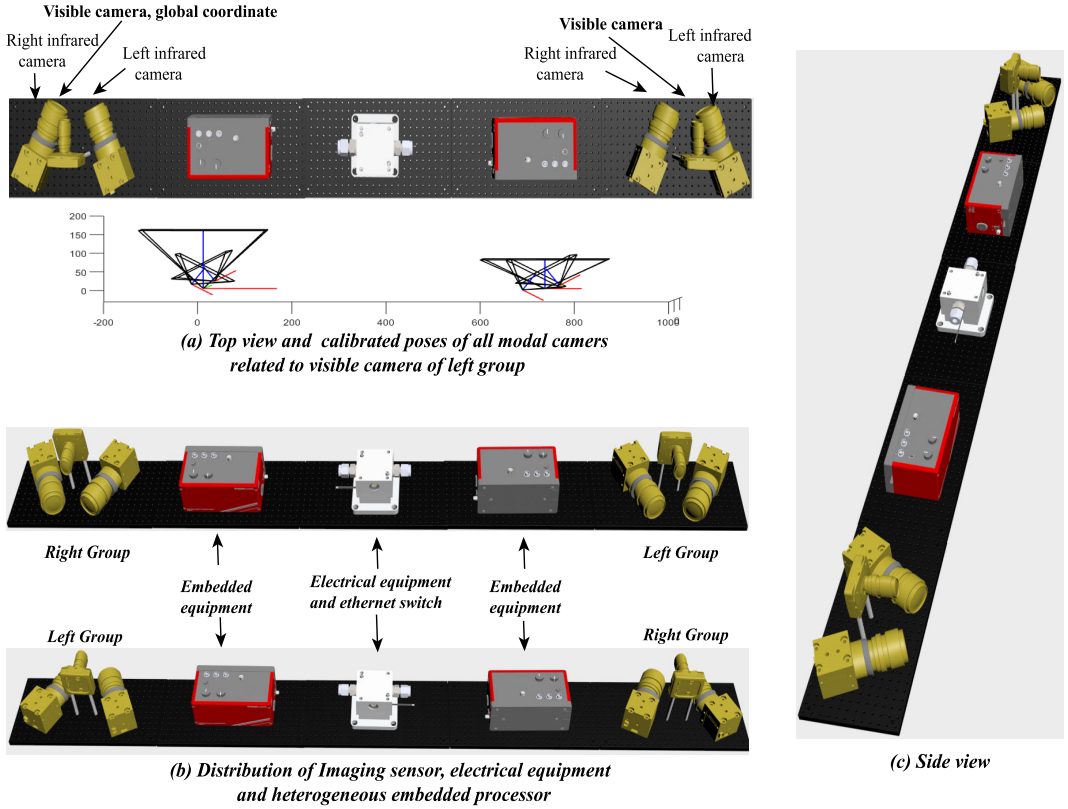


FIGURE 3. (a) Estimated poses of all modal cameras related to visible camera of left group; (b) Front view of the sensor;(c) Side view of te sensor.

The volume of our system is less than 100cm(length) × 10cm(height) × 10cm(width). The uncooled LWIR detector is adopted for the sensor can efficiently reduce electric consummation and volume of the system. The global view of our system is shown in the figure 3.

B. ATTRIBUTES

Our sensor is treated as three cooperated binocular stereo vision *i.e.* visible cameras from both groups construct a visible stereo vision unit. Disparity map calculated by visible cameras is shown in figure 4. We combine two statics stereo infrared cameras. Left infrared cameras (captured left part related to visible camera) from both groups construct a left infrared stereo vision unit, right infrared cameras from both groups construct a right infrared stereo vision unit. Disparity map calculated by both infrared stereo cameras is shown in the figure 5.

We believe this is a good set up since color images captured by visible cameras can provide rich details for object detection, segmentation, and the bag of visual word construction. Meanwhile, infrared detectors can provide higher contrast and sensitivity based on the thermal signatures. The reason why we use four infrared detectors is the resolution and FOV of common LWIR images is less than commercial optical sensors. Using the uncooled LWIR detector instead of NIR gives us some benefits:

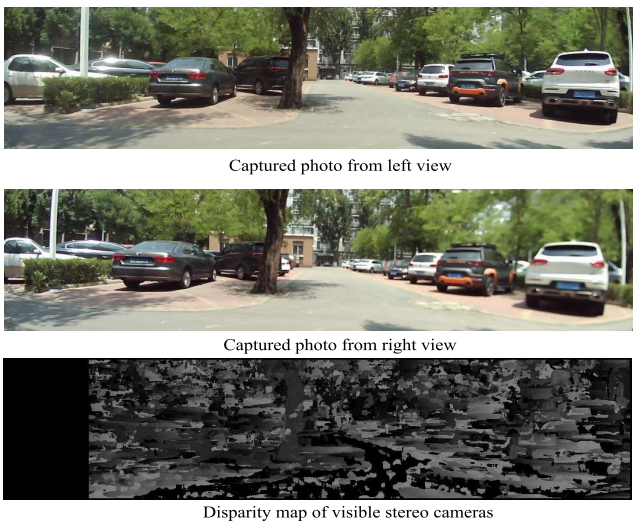


FIGURE 4. Color image captured by visible cameras and related disparity map, where we set minimum number of disparity is 32 pixel.

1) For potential threat from distances, especially some small moving object like the vehicle, infrared detector based methods have its own advantage comparing to optical imaging;
2) NIR detector not only captures thermal signal, also capture some optical reflection. NIR's details are not rich, comparing to visible camera's color image. NIR has

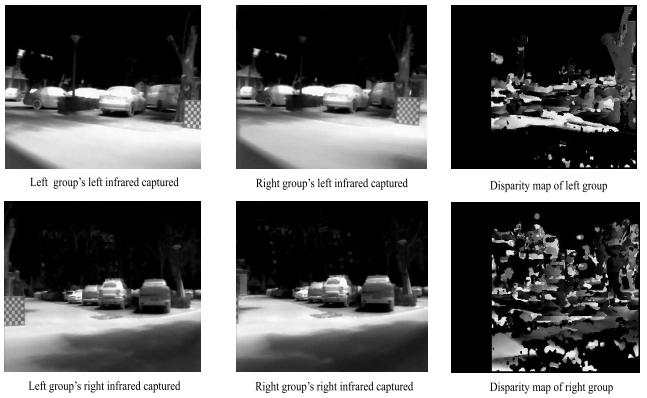


FIGURE 5. Gray image captured by infrared cameras and related disparity map, where we set minimum number of disparity is 16 pixel.

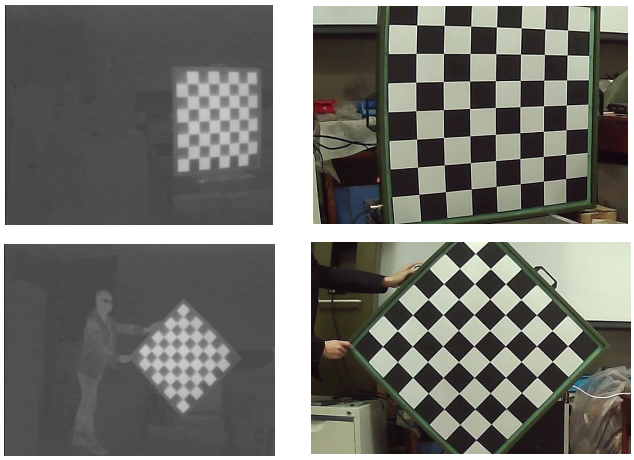


FIGURE 6. Self-heating chessboard captured by undistorted detectors.

no advantage in contrast and sensitivity for surrounding heat profile(from $-73°$C to $+349°$), comparing to uncooled LWIR [16].

3) For object localization, rigid or non-rigid motion extraction and analyzing, the color image can provide visual detail. Meanwhile, the thermal images can provide surroundings heating information. This can help our system working well in an extreme brightness change environment.

We combine visible and LWIR advantage to construct the sensor. LWIR detectors are working as complementary of visible, in a cooperation fashion, comparing to pixel registration of NIR and visible [17]. With our accurately estimated cameras' relative pose, calibrated intrinsic and distortion coefficients, a cooperative multi-modal stereo visual odometer is established.

III. IMPLEMENTATION

In this section, we will introduce the key component and performance of our system: (1) Related poses estimation of all cameras with different bands, a specified self-heating chessboard for global calibrate our system are introduced. (2) Two-stage distortion control for large FOV visible stereo; (3) Methodology of ego-motion estimation by two LWIR stereos units.

A. CAMERAS' CALIBRATION

Accurate sensors' calibration is the key to obtaining reliable 3D data. The challenges of the system's calibration are: (1) Common features extracted from both LWIR and visible cameras are limited or hard to be found, especially from natural scene; (2) To cover relative large FOV, non-overlapping or limited common filed of views cameras are equipped in our system, especially uncooled LWIR cameras in the same group have very limit FOV in common.

All cameras from different modals are following the pinhole camera model. All cameras' intrinsic, distortion coefficients are calibrated separately in advance. The main challenge of our system's calibration can be regarded as a relative camera pose estimation problem. The global coordinate is defined at the visible camera of the left group.

We develop a self-heating chessboard for visible and LWIR bands calibration. Common features *i.e.* corner for visible and LWIR bands are constructed by heating resister in each cell of the chessboard. The main advantage of this design are: (1) The heating distribution and parameters of the chessboard can be adjusted based on the LWIR detector; (2) Designed chessboard can be used off the premises with a portable electronic generator. The effect of our self-heating chessboard are shown in the figure 6.

To calibrate related pose of left visible camera with camera j (right visible camera or other infrared camera) in sensor, we define 4×4 translation matrix T_i^j as the pose of left visible camera related to its chessboard. Meanwhile the pose of camera j related to its chessboard (same or related fixed to left visible camera's chessboard) is defined as K_i^j, translation matrix of camera j to left visible camera is defined as C_{lv}^j. The error function of camera j and left visible camera is defined as following:

$$R^j = \sum_{i=0}^{i<n_{lv}^j} \left\| \xi \left(C_{lv}^j T_i^j B_{lv}^j K_i^j \right) \right\| \tag{1}$$

With B_{lv}^j is the relative pose of chessboards. In some cases, B_{lv}^j can be treated as identity. n_{lv}^j is the number of the calculated poses between left visible camera and camera j. ξ (.) is $\mathfrak{se}(3)$ lie algebra representation. To reduce error conduction, we put all the errors in the system together, set up a global optimization method for these six cameras related to the left visible camera. We use [18] set up optimization procedure. The calibration in sensor can be treated as:

$$\underset{\{C_{lv}^1 \ldots C_{lv}^5\}}{minimize} \sum_{j \in sensor} R^j \tag{2}$$

The calibrated poses of each camera related to left visible camera are shown in the figure 7, related flow chart are also demonstrated in the figure 7.

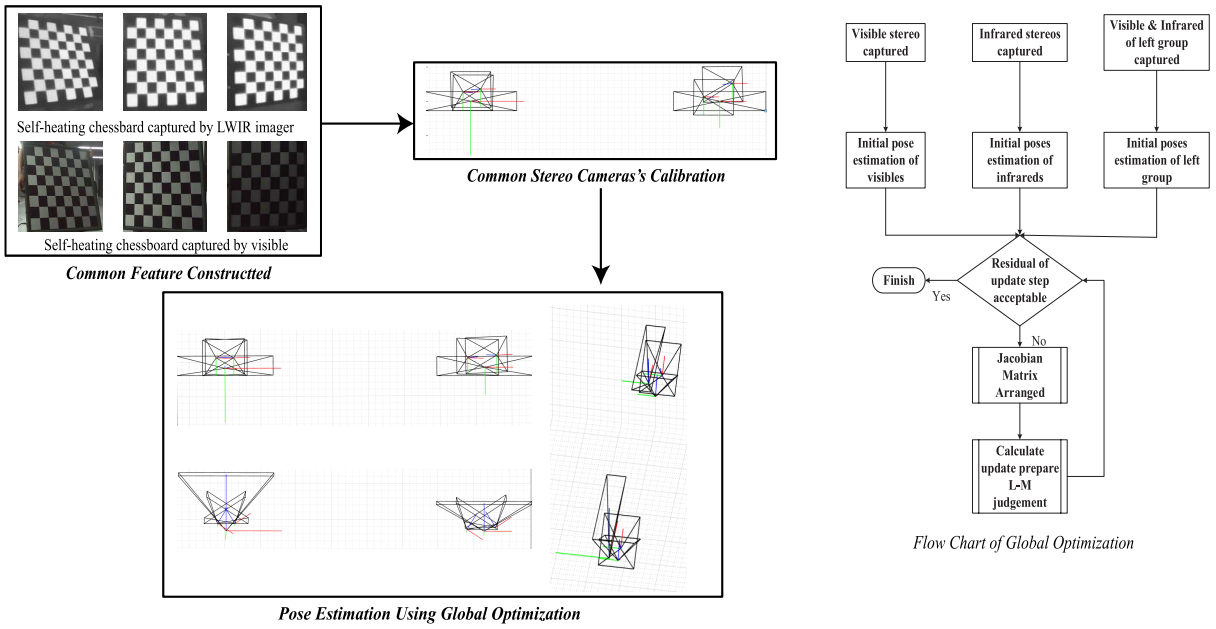


FIGURE 7. System's calibration result.

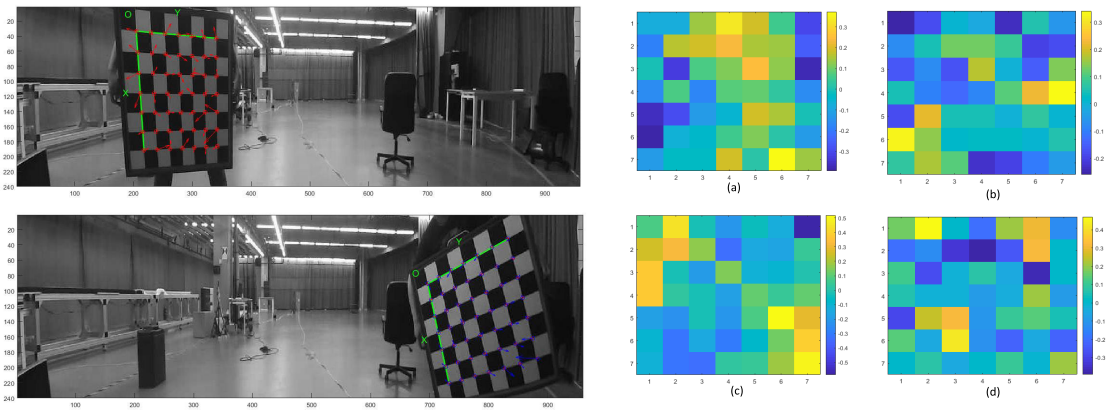


FIGURE 8. Cropped images calibration after full size distortion correction, (a) *u* direction reprojected error of each chessboard corner; (b)*v* direction reprojected error of each chessboard corner; (c) *u* direction reprojected error of each chessboard corner; (d)*v* direction reprojected error of each chessboard corner;

B. LARGE FOV VISIBLE STEREO ODOMETER

For the visible band part, the initial resolution of our visible cameras is 1920 × 1080, which brings related high computational burden. For the large FOV sensor, distortion is much heavier at the edge of images. So we cut images vertically, reduce to 400 (Considering the processing ability of the embedded system and sensing area in vertical). We pre-process visible images by two stages. Before image cropped, images are calibrated by the fish-eye model [19], and remap centrally. Then we use cropped images (1920 × 400) to do the rest calibration. The undistorted result of our visible cameras is shown in figure 8. After full-size correction and remapping, the cropped visible image reprojection error is less than 0.8 pixels.

With proper illumination condition, for visible cameras, the feature-based indirect method [20], [21], can be implemented. Since visible based visual odometry is widely researched, we adjust the baseline and related propriety of our cameras and realize stereo's tracking thread part of ORB_SLAM2 [22].

C. LARGE FOV LWIR BASED VISUAL ODOMETER

Infrared sensors' resolution is related small (384 × 288), compared to visible cameras. Therefore, amounts of keypoints extracted under infrared sensing are limited. Related low-textured is common for infrared sensors, matching keypoints from stereo are hardly be found. Four infrared cameras are equipped in system, feature-based (indirect) method for infrared images will bring unnecessarily high computational burden. To overcome these issues, we adopt direct methods' theory, combine LWIR detectors in our system, propose and implement a novel infrared stereos direct visual odometer.

Direct methods, in contrast to feature matching based indirect methods, estimate geometry directly from images, *i.e.* the raw sensor's images [23]–[25]. Our instrument brings some advantages for direct method adaptation:

1) Two static stereo infrared cameras not only cover a related large field of view but also calculate directly absolute scale basing on each known baseline stereo vision. This setup can avoid mono camera's scale problems;
2) Direct methods use all information from the images including corners, edges, weakly textured and repetitive image regions. This is suitable for a low textural LWIR detector;
3) For the visible based direct method, the photometric error is calculated directly on pixel intensities, which is sensitive to sudden illumination changes between consecutive frames. Meanwhile, pixel intensities from images captured by LWIR are only depended on the thermal distribution of the scene. From the system's captured images, we found that thermal distribution is more stable comparing to visible while illumination varied [30];

Camera's motion (*i.e.* its ego-motion) related to observed static objects in the scene is tracked by image alignment to the reference keyframe. All keyframes and its keypoints are arranged by a sliding window. Suppose a set of keypoint ρ_i in a reference frame I_i, which are observed in frame I_j. Image alignment of these two frames *i,j* by the direct method can be formulated as:

$$E_{ji} = \omega_p \left\| I_j \left[p'\right] - I_i \left[p\right] \right\|_\gamma \tag{3}$$

With $\|.\|_\gamma$ is the Huber norm. The intensity-based loss function should be more robust since we only use intensity error of pixel. Huber loss function is not sensitive to the outlier, which is widely used in classification. Using Huber penalties, a gradient-dependent weighting ω_p, which down-weights pixels with high gradient, given by:

$$\omega_p = \frac{c^2}{c^2 + \left\| \nabla I_i(p) \right\|_2^2} \tag{4}$$

Further, p' stands for the projected point position of p with inverse depth d_p, *i.e.*:

$$p' = \Pi_K \left(T_{ji} \Pi_K^{-1} \left(p, d_p \right) \right) \tag{5}$$

With Π_K is denoted as the intrinsic matrix of cameras, T_{ji} is denoted as the transformation of a point from frame i to frame j, with $T_{ji} = T_j T_i^{-1}$. For visible camera based direct method, a brightness transfer function is proposed in [26], *i.e.* $e^{-a_i}(I_i - b_i)$. This affine brightness transfer is a inverse function of linear response function. [27], [28] use this formula to adjust the exposure time (controlled by a_i) and brightness (controlled by b_i). For LWIR, we also use this formula to adjust the captured infrared heating distribution. Image alignment of equation 3 is then modified to:

$$E_{ji} = \omega_p \left\| I_j \left[p'\right] - b_j - \frac{e^{a_j}}{e^{a_i}} \left(I_i \left[p\right] - b_i\right) \right\|_\gamma \tag{6}$$

With a set of keyframes $\mathcal{F}$, all points in frame i are denoted as ρ_i. The other frame j can observe point p in ρ_i overall frame is denoted as $obj(p)$. Since two static stereo infrared cameras are equipped, each static stereo infrared unit has its own residual. Modify equation 6, static stereo's residual can be represented as:

$$E_i^{(L,R)} = \omega_p \left\| I_i^R \left[p^R\right] - b_i^R - \frac{e^{a_j^R}}{e^{a_i^L}} \left(I_i^L \left[p^L\right] - b_i^L \right) \right\|_\gamma \tag{7}$$

The total energy function of one stereo infrared unit can be established by combining the multi-view stereo geometry part (equation 6) and static stereo part (equation 7). We use λ weights the constraints of static stereo, the full photometric error over all frames and points can be represented as:

$$E_{total} = \sum_{i \in \mathcal{F}} \left(\sum_{p \in \rho_i} \sum_{j \in obj(p)} E_{ji} + \lambda E_i^{(L,R)} \right) \tag{8}$$

All infrared cameras are equipped and installed rigidity on the rack of the system, which means that transformation T_{ji} of all cameras in the system is the same while sensor is moving. Naturally, we can adjust Equ. 8, make all the residual of two infrared stereo units together. But this idea is not good. The reasons are: (1) The public FOV of these two units are nearly zero degrees, the common objects can not be observed by two infrared stereo units simultaneously. Residuals of different units are hard to be established; (2) Two units' residual in one energy loss function will nearly double the dimension of the Hessian matrix comparing to one unit, which will bring computation burden to the system. Keyframe management in RAM will be more difficult. To avoid these problems, we use multi-threading technology to implement the direct method in sensor's embedded system. Each infrared stereo unit's residual function is established and solved by its own thread. Each unit has its own keyframes management and marginalization. A control thread use frame captured id and the system's time to identify each unit's captured frame, transform each unit's pose to the system's global coordinate, adjust these two units' pose. The two LWIR stereos based factor graph of the energy function for infrared stereo units are shown in the figure 9.

In this example shown in figure 9, 5 points are observed by each unit's four keyframes. Observed points from the different units have no strict relationship. Each infrared stereo unit's thread has its own keyframe management and marginalization, control thread adjust these two units' keyframes, construct a system's keyframe management. For common frame id's keyframe, system pose is calculated by coordinate transformation to the system's coordinate with the bias of two units' estimated poses is acceptable (for KF1, KF4, KF5). For no common frame id's keyframe, the system's pose

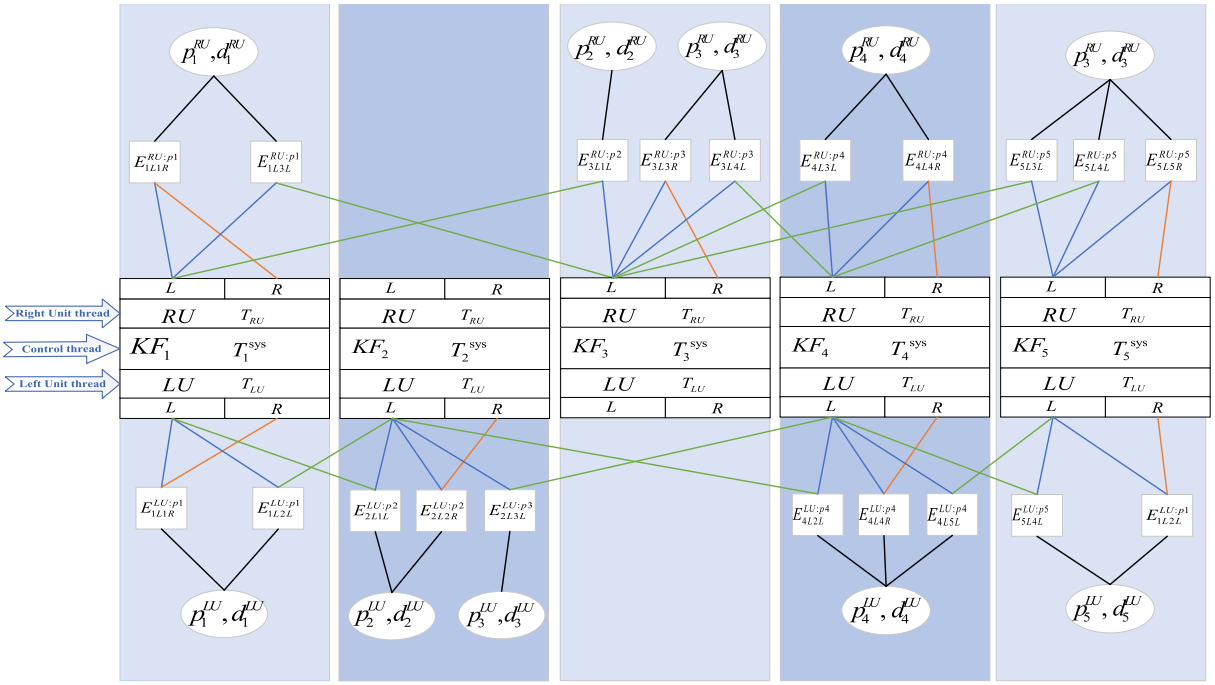


FIGURE 9. Factor graph for two LWIR stereos direct method model.

accepts related unit's estimated pose(for KF2, we use left unit's estimated pose, for KF3, we use the right unit's estimated pose). For each unit, the energy factor is constructed by related points and observed cameras. Finally, all poses are transformed into visible camera coordinate of the left unit.

To balance the accuracy and speed, Gauss-Newton based windowed optimization is used to solve Equ. 8 of each infrared units. For one stereo infrared cameras unit, all optimized variable including: camera poses (T_j, T_i), brightness of heating distribution (a_i, a_j, b_i, b_j) and depth of point p in keyframe i d_p. For keyframes i, j. we define $\xi_{ji} = ln(T_{ji})^\vee = ln(T_j T_i^{-1})^\vee$ replacing (T_j, T_i), $a_{ji} = \frac{exp(a_j)}{exp(a_i)}$, $b_{ji} = b_j - a_{ji} b_i$, and inverse depth $\rho_i = \frac{1}{d_p}$. Equation 6 can be modified by following formula:

$$E_{ji} = \omega_p \left\| I_j \left[\Pi_K \left(T_{ji} \Pi_K^{-1} (p, d_p) \right) \right] - I_i \left(a_{ji} I_i [p] - b_{ji} \right) \right\|_\gamma \tag{9}$$

To solve equation 9, Jacobian matrix should be discussed. The Jacobian matrix of equation 9 can be mainly divided by following formula:

$$\mathbf{J}_{ji} = \left[\frac{\partial E_{ji}}{\partial a_{ji}}, \frac{\partial E_{ji}}{\partial b_{ji}}, \frac{\partial E_{ji}}{\partial \xi_{ji}}, \frac{\partial E_{ji}}{\partial \rho_i} \right] \tag{10}$$

For a_{ji} and b_{ji}, the partial derivation is simple. For pose ξ_{ji} and inverse depth ρ_i, we use the derivative chain rule to construct a close formula of the Jacobian matrix block. Basing on basic camera model:

$$\begin{cases} \rho_i^{-1} x_i = \Pi_K X_i; \\ \rho_j^{-1} x_j = \Pi_K (\exp(\xi_{ji}^\wedge) X_i); \end{cases} \tag{11}$$

With x_i and x_j are point p projected in camera i and j coordinate. For pose ξ_{ji}, the Jacobian matrix block can be developed by equation 12.

$$\begin{aligned} \frac{\partial E_{ji}}{\partial \xi_{ji}} &= \omega_h \begin{bmatrix} g_x \\ g_y \\ 0 \end{bmatrix}^T \begin{bmatrix} fx & 0 & 0 \\ 0 & fy & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{\partial x_j}{\xi_{ji}} \\ &= \omega_h \begin{bmatrix} g_x f_x \rho_j \\ g_y f_y \rho_j \\ -g_x f_x \rho_j u_j' - g_y f_y \rho_j v_j' \\ -g_x f_x v_j' u_j' - g_y f_y (1 + v_j'^2) \\ g_y f_y v_j' u_j' + g_x f_x (1 + u_j'^2) \\ -g_x f_x v_2' + g_y f_y u_2' \end{bmatrix} \end{aligned} \tag{12}$$

With g_x, g_y are image gradient at point p in frame i, f_x, f_y are the focal length of camera, u_j', v_j' are camera normalized coordinates. For inverse ρ_i, the jacobian matrix block can be developed by equation 13.

$$\frac{\partial E_{ji}}{\partial \rho_i} = \omega_h \begin{bmatrix} g_x \\ g_y \\ 0 \end{bmatrix}^T \begin{bmatrix} f_x \rho_i^{-1} \rho_j \left(t_{ji}^x - u_j' t_{ji}^z \right) \\ f_y \rho_i^{-1} \rho_j \left(t_{ji}^y - v_j' t_{ji}^z \right) \\ 0 \end{bmatrix} \tag{13}$$

with $\left[t_{ji}^x, t_{ji}^y, t_{ji}^z \right]^T$ is the displacement of frame i,j. We use Gausse-Newton optimization with equation 12, 13, the pose

and related parameters can be estimated. The keyframes and keypoints marginalization strategy, we use Shur complement presented in [27]–[29].

D. FRAMES AND POSES MANAGEMENT

Visible cameras are connected and arranged by one Nvidia TX2 embedded equipment. Infrared cameras are implemented to another Nvidia TX2. Two embedded equipment is connected via Ethernet.

For some multi-camera system, frame synchronization is very important, especially for image fusion based method. Six cameras' frame synchronization with different resolution, different sensing area and different propriety of the camera (camera's frame rate, visible or LWIR) is unnecessary. Instead, we deploy frame synchronization for the same band cameras. Visible stereo has its own keyframe management under feature-based judgment. Two infrared stereos manage their own keyframe basing on their thermal distribution. We use network time protocol (NTP) for two embedded system time synchronization, transmit calculated poses with each system's synchronized time via Ethernet protocol and manage them.

With visible cameras' grabbing frame rate roughly 25Hz, infrared grabbing frame rate roughly 50Hz, frame synchronized of all cameras will loss information of infrared sensors, bring unnecessary electrical power load to our system. We use synchronized embedded systems' time and calibrated poses of multi-cameras to cooperate visible and infrared band camera. Both embedded equipment *i.e.* Nvidia TX2, will transform disparity map (under left visible camera's coordinate), estimated poses marked with keyframes via Ethernet protocol, under ROS middle-ware. We use two embedded Nvidia TX2 to balance the processing load of multi-cameras, implement efficient front end visual odometer strategies. The flow chart of related frame management and pose management is shown in the figure 10

For poses management, the example we presented in figure 9 will be detailed. For infrared stereos, frame synchronization is deployed. Image processing, windowed optimization and marginalization are processed in the same unit under its own thread control. Each infrared stereo estimates the system's pose under the camera's coordinate of the left unit. The pose judgment of system's pose ξ_{sys} can be represented as:

$$\xi_{sys} = \begin{cases} \ln(T_{lg-sys}T_{lg})^{\vee}; \; if \; T_{rg} \; is \; abscent \\ \ln(T_{rg-sys}T_{rg})^{\vee}; \; if \; T_{lg} \; is \; abscent \\ \ln(T_{lg-sys}T_{lg})^{\vee}; \; if \; \left\| \ln\left(R_{lg}R_{rg}^{-1}\right)^{\vee} \right\| \leq \theta \\ \qquad\qquad and \; \left\| t\left(T_{lg}T_{rg}^{-1}\right) \right\| \leq \epsilon \\ marginalized; \; if \; \left\| \ln\left(R_{lg}R_{rg}^{-1}\right)^{\vee} \right\| \geq \theta \\ \qquad\qquad or \; \left\| t\left(T_{lg}T_{rg}^{-1}\right) \right\| \geq \epsilon \end{cases} \tag{14}$$

where R_{lg}, R_{rg} is the rotation part of T_{lg}, T_{rg}. $t(.)$ represents the translation part of the transform matrix. The chosen of θ and ϵ is basing on the calibrated poses result of two infrared cameras in the same group.

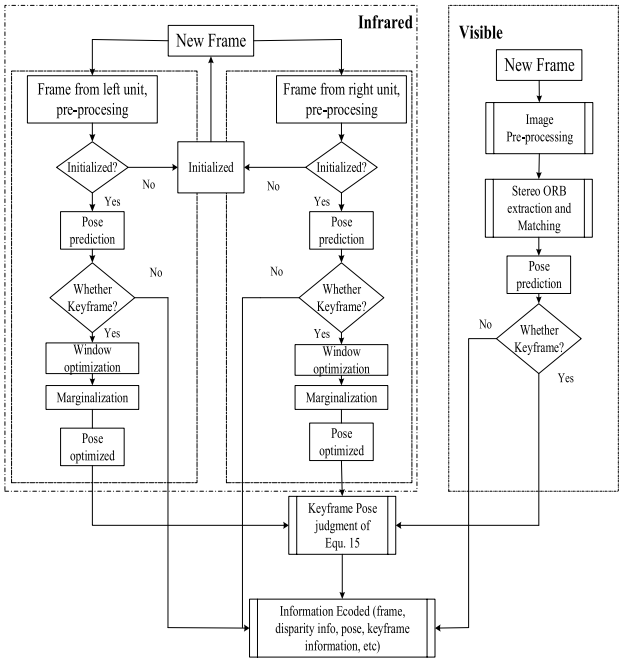


FIGURE 10. Flow chart of frame management and pose management.

IV. EXPERIMENTS AND RESULTS

In this section, we will show the key performance of our system which are not demonstrated in implementation part, *i.e.* (1) Infrared part's visual odometry: we will show the result of our direct method for infrared-based images, frame management of our proposed method; (2) Quantitative evaluation of the sensor's visual odometry. (3) Processing speed and efficient evaluation. To evaluate our system, we use our system recording and working indoor under different illumination, we also installed the system on top of the van, perform the outdoor test under natural heating distribution and random illumination. The outdoor setup test is shown in the figure 11.

A. VISUAL ODOMETER EVALUATION OF INFRARED

For visual odometer, the distribution of extracted keypoint candidates (either feature-based method or direct method) is important. Normally, the keypoints are distributed more uniform, the odometry is more robust. For example, ORB_SLMA2 uses grid-based Fast feature extraction, adjusts the threshold of each grid to extract more or fewer keypoints to make keypoints distributed more uniform compared to whole frame extraction. For low textural LWIR, grid-based feature extraction can't avoid extracted keypoints concentrate on some particular area. Comparison keypoints initial extraction from LWIR is shown in figure 12.

Figure 12 is a typical example of feature extraction for LWIR. There exist related rich textural features of heating distribution in this example. The main drawback of the indirect method comparing to the direct method are:

TABLE 1. Frame management strategy analyzing under absolutely stable scene.

Serial Number	Number of frames	Keyframes overlapped rate	$\|\theta_{max}\|$ in degree	$\|\epsilon_{max}\|$ in meter
sq-01	346	32.7%	0.30	0.09m
sq-02	175	22.3%	0.43	0.13m
sq-03	104	66.7%	0.85	0.39m
sq-04	406	16.3%	0.31	0.09m
sq-05	1052	9.5%	0.56	0.11m

FIGURE 11. Outdoor evaluation, (a) figure of traces where we recording and testing our system around Tianjin University, tracks in red is under heavy traffic, with moving vehicle and amounts of random movement pedestrian; tracks in orange is under median traffic, with moving vehicle, little pedestrian; tracks in blue is under related static area, little moving vehicle and pedestrian. (b) front view of our system mounted on the top of van. (c) side view of the our system installed on the van, sensor is marked under red ellipse.

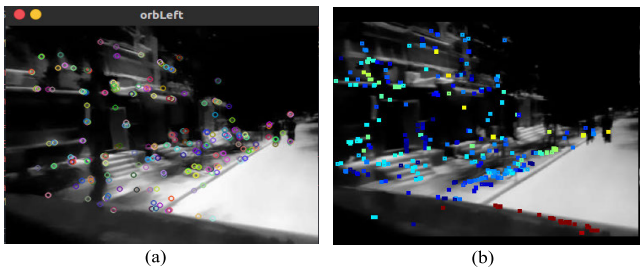


FIGURE 12. Comparison of keypoints extraction for LWIR, (a) ORB_SLAM2 based keypoint candidates extraction from LWIR images; (b) Our direct method's keypoints candidates.

1) Processing speed: keypoint candidates extracted in example (a) of figure 12 are used ORB_SLAM2 strategy. The processing speed for the rich textural scene is 8∼ 10ms (image resolution of LWIR is related small). For low textural condition, the processing time will obviously raise, tracking failed judgment will spend a related long time. Meanwhile, the direct method processing speed is roughly 2∼4ms for each frame; Processing time over 10ms is treated as loss of tracking, reinitialized.
2) Distribution of keypoints candidates: From the example we demonstrated in figure 12, keypoints candidates are distributed in both (a) of indirect method and (b) of direct method are related uniform. The problem of the indirect method is the matching strategy. For our multi-stereo system, the indirect method will use kyepoints matching strategy (Vector field consensus VFC, Euler distance *etc*), which will reduce amounts keypoints and lead an attenuation of keypoints distribution. In the example, about 35 percent of keypoints will be abandoned by feature matching. Direct method skips feature matching and PNP series solving steps, estimates cameras' poses basing on related uniform and rich distributed keypoints candidate.

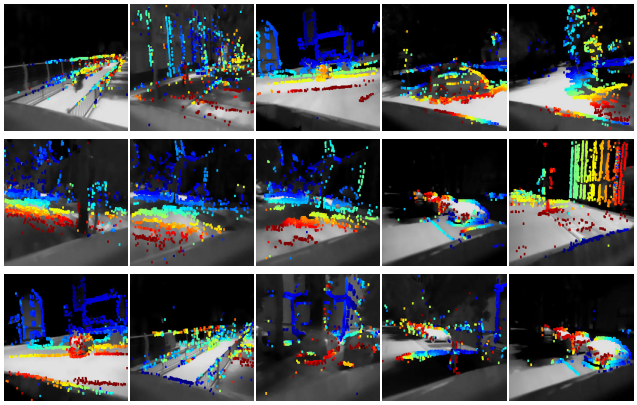


FIGURE 13. Under outdoor experiments, coarse depth maps for both left and right LWIR stereo unit, where feature tracking method lost occurred.

LWIR cameras' propriety is more suitable for the direct method because of "shutter control" and stable "illumination" comparing to visible. Failure of some direct strategy due to geometric distortion introduced by a rolling shutter (even for high frame rate camera). Sudden illumination change will lead to the image gradient failed. For the LWIR detector we used, the imaging mechanism is staring array. Detector absorbed infrared radiation from a scene simultaneously, adjusted each pixel intensity basing on the distribution of absorbed infrared radiation energy on staring array. This mechanism can avoid geometric distortion at each pixel induced by rolling shutter. It also provides a related stable distribution of pixel intensity with successive frames. We evaluated all recorded frames from our instrument. With un-distorted LWIR's frame, the direct method can run robustly. Under large FOV, some examples of keyframes managed by our proposed strategy are shown in figure 13, where ORB_SLAM2 tracking loss occurred.

From campus experiments, we choose five stable scenes. Overlapped means that keyframes extracted from different infrared units at the same time. The max number of gap θ and ϵ between two LWIR stereo group shown in the table 1.

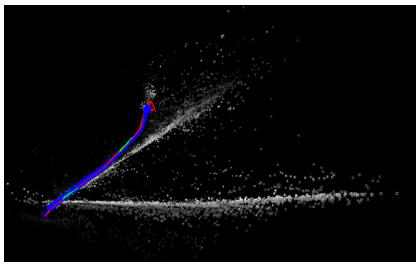


FIGURE 14. Example of our two LWIR stereos' direct method.

For test serial number 3 in table 1, amounts of frames are related small, even though the keyframes overlapped rate is high (66.7%), number of overlapped frames is small (18 frames). A gap of right LWIR unit loss tracking and re-initialization occurred in this serial. The overlapped rate is low for test serial number 5 is that our van was running around the gymnastic of Tianjin University, one side with related rich textural building feature, another side is the low intensity of the whole image. In this case, the system's poses are relay on only one LWIR stereo unit temporarily. For a typical visual SLAM mapping task, our system's estimated poses' accuracy is not high. For front-end only, the calculated poses are robust enough.

From table 1, we can also see that the keyframes overlapped rate is normally under 50%. This demonstrates that separated optimization of our proposed method will efficiently reduce RAM cost and processing time.

For the front-end of the instruments, we designed a windowed optimization of keyframes and keypoints management for each LWIR stereo unit. It manages its estimated poses depends on its own optimized results. An example of estimated poses calculated and managed by both LWIR stereo units cooperated with its own windowed optimization is shown in figure 14. In the figure 14, trajectory of systems are represented in red while keyframes of both unit are represented in light blue, non keyframes are represented in dark blue. Two lines of keypoints in white are extracted by these two LWIR stereo units separately.

The main advantage of our proposed strategy is: using only one LWIR stereo, the tracking loss or initialization occurred more frequently. Large FOV's thermal sensing makes the system more robust. Two separated windowed optimizations make the system less complicated and more efficiency.

B. VISUAL ODOMETER EVALUATION OF SENSOR

In order to make a quantitative evaluation, the proposed sensor is evaluated under a series of indoor experiments. We mount our sensor on a 15m long track. A total station (SX-105T) is placed 3m behind the end of the track. Cooperated prism of the total station is installed with our sensor on the mechanical adapter. The essential geometry relationship of the prism and our sensor is calibrated after the installation. The mechanical adapter is connected with the track and driven by a motor. This controllable motor control the speed and acceleration of the sensor while it is reciprocating on the

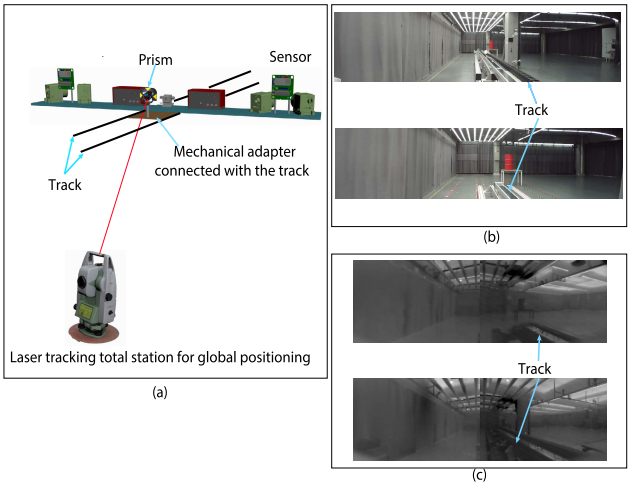


FIGURE 15. (a) Indoor experiments set up; (b) Visible images captured at both sides of the track while the sensor is moving; (c) Infrared images captured at both sides of the track while the sensor is moving.

TABLE 2. RMSE of absolute trajectory error under different condition.

Condition	Number of poses	RMSE [m]
normal	897	0.0593
strong light	471	0.0805
thermal interference	439	0.0736
illumination change	929	0.0984

track. The schematic of the experiments' setup is shown in the figure 15 (a). Captured images from both sides of the track are shown in 15 (b) and (c).

We use the total station capturing the motion of the sensor, measuring the position and the orientation while it's moving on the track. Meanwhile, we adjust the illumination and the heating distribution in the scene. The sensor is evaluated not only under the normal scene, but also under these three particular conditions:

1) Influence of the strong light: We place light source around the track. Since the FOV of sensor is large, the light source can be placed not far. Captured samples are shown in figure 16 (a);
2) Thermal interference: We provide heating source on the roof, which cause upper side of LWIR's image plane lose texture information. Captured samples are shown in figure 16 (b);
3) Change of the illumination condition: We adjust lighting system from 600 lux to 0.5 lux while the sensor is moving. The captured samples are shown in figure 16 (c).

Our sensor record 2736 keyframes and poses in total from these experiments. The total station provides ground-truth data for quantitative evaluation. The Euclidean distance is computed between the estimation results and the ground-truth with translational Root Mean Square Error (RMSE) of Absolute Trajectory Error in meters. The results are shown in table 2. From results, we can see that under these interferences, the dilution of accuracy is acceptable. The proposed

TABLE 3. Evaluation using our experiments data, where * is initial failed, x is tracking lost occurred, ✓ is working well.

Serial	Ours	ORB_SLAM2 RGB only	ORB_SLAM2 thermal only	DSO thermal only
sq-01	✓	x	x	*
sq-02	✓	✓	*	*
sq-03	✓	✓	*	*
sq-04	✓	*	*	x
sq-05	✓	x	x	x
strong light	✓	*	x	✓
thermal interference	✓	✓	*	*
illumination change	✓	x	x	✓

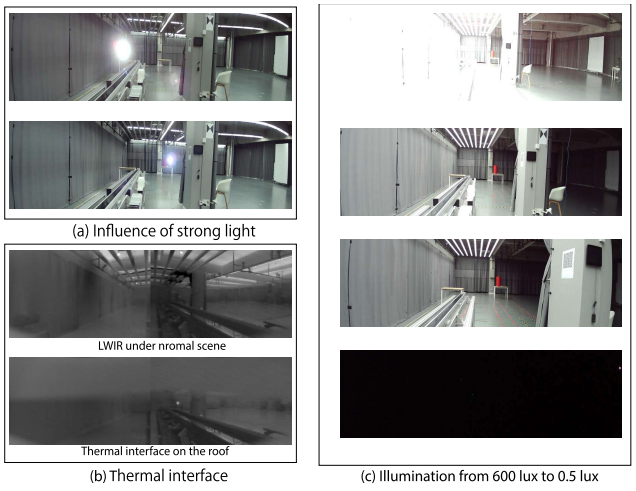


FIGURE 16. (a) Influence of strong light; (b) Thermal interface; (c) Illumination change.

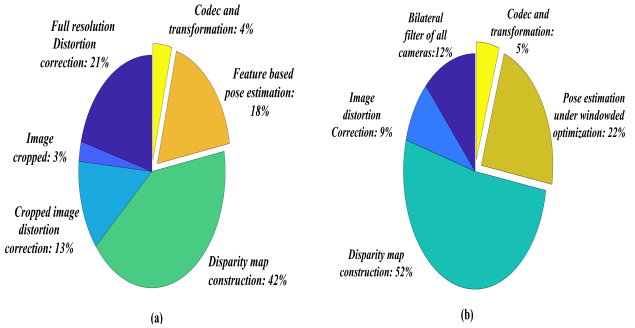


FIGURE 17. Runtime distribution of each procedure for each stereo unit with (a) visible stereos, (b) infrared stereos.

method combine advantage of thermal and visible spectral, provide more robust performance in motion estimation.

C. RUNTIME EVALUATION

We track the cost time of each procedure while we were doing our experiments. The run time distribution is presented as following:

For visible cameras' embedded processor, we implemented large FOV distortion correction and remapping central under full resolution(19∼21ms); Cropped image under full resolution undistorted image (2∼3ms); Corrected distortion and remapped image central under cropped image(12∼14ms); Constructed disparity map of stereo vision(40∼45ms); Estimated system's pose using ORB feature-based method(17∼19ms). All these procedures are implemented in one Nvidia TX2 using cooperated Arm kernel and CUDA acceleration.

For infrared cameras' embedded processor and each LWIR stereo unit, we implemented a bilateral filter for infrared cameras connector's noise adjustment; Corrected image distortion; Constructed two directions disparity under left visible camera's coordinate; Estimated system's pose using our proposed direct method strategy(14∼16ms). All these functions are implemented in another Nvidia TX2 with CUDA acceleration and related peripheral devices.

The time distribution of each procedure for visible stereo and one infrared stereo unit are shown in figure 17. When the system transformed information coded with full resolution disparity map, the frame rate of visible is 11∼12 fps, whereas infrared is 14∼15fps. For information coded only with pose and keyframe information, the frame rate of visible is 18∼19 fps, whereas infrared is 22∼23fps.

V. DISCUSSION

From an engineering perspective, our instrument has its own advantage such as (1) Instrument's size and weight is related small; (2) Electrical power consumption is limited (12W total); (3) System not only captures images but also produces pose and keyframe information.

Comparing to other instruments in the context of large FOV visible and LWIR cooperated system or strategy [30], we use the multi-LWIR detector to cover the same FOV as visible cameras. LWIR captured images are not partial complementary of the visible camera. It's working as an independent component. Multi LWIR detector can bring us benefits such as larger detecting areas, but it also gives us challenges. Our proposed methodology makes them working as front-end of the visual odometer.

We use collected RGB and thermal images from indoor and outdoor experiments to evaluate ORB_SLAM2 performance, we also use thermal images to evaluate DSO performance (since RGB visible camera are using rolling shutter, not suitable for DSO series). The result are shown in table 3. Our instrument with implemented methodology is more robust compared to traditional visual odometry approach *i.e.* indirect method: ORB_SLAM2 and direct method: DSO.

For accuracy assessment: since we combine ORB_SLAM2 of RGB and direct method of thermal, the accuracy can be guaranteed for short term system's localization. For long term localization or mapping, since our system has not enough spare computational ability, local bundle adjustment can not be implemented in our front-end instruments. Without BA optimization, the accuracy of our instruments for long term activity will lose.

VI. CONCLUSION

In this work, we introduced and realized several improvements basing on the state of the art visual odometry approach to serve in the context of large FOV visible and LWIR cooperated instruments.

The developed hardware platforms represent efficient, low-cost solutions for two bands' visual odometer. In addition, to lighten the burden of the back-end system, our sensor not only captured the appearance and thermal information but also implemented front-end algorithm *i.e.* keypoints tracking, system's pose estimation *etc*. Our proposed direct method of LWIR has shown remarkable influence in the system's robust keyframes management and keypoints controlled. The outdoor test demonstrates that our system can work under complicated illumination conditions, and processing speed can maintain at least 10fps (with full resolution disparity map construction).

Our future perspectives are mainly centered on back-end designed for loop closure and mapping based on our constructed front end sensor. Furthermore, a more complicated motion segmentation strategy, research on vision odometry under heavy traffic and pedestrian basing on our recording and constructed date-set is working on.

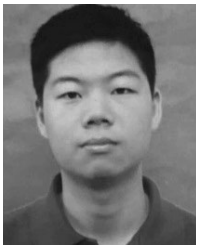
REFERENCES

[1] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: A survey from 2010 to 2016," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, p. 16, Jun. 2017.
[2] Y.-S. Shin, Y. S. Park, and A. Kim, "Direct visual SLAM using sparse depth for camera-LiDAR system," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–8.
[3] H. Liu, M. Chen, G. Zhang, H. Bao, and Y. Bao, "ICE-BA: Incremental, consistent and efficient bundle adjustment for visual-inertial SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1974–1982.
[4] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
[5] M. Yaman and S. Kalkan, "An iterative adaptive multi-modal stereo-vision method using mutual information," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 115–131, Jan. 2015.
[6] T. Mouats and N. Aouf, "Multimodal stereo correspondence based on phase congruency and edge histogram descriptor," in *Proc. 16th Int. Conf. Inf. Fusion*, Jul. 2013, pp. 1981–1987.
[7] F. Bonardi, S. Ainouz, R. Boutteau, Y. Dupuis, X. Savatier, and P. Vasseur, "PHROG: A multimodal feature for place recognition," *Sensors*, vol. 17, no. 5, p. 1167, 2017.
[8] F. Barrera Campo, F. Lumbreras Ruiz, and A. D. Sappa, "Multimodal stereo vision system: 3D data extraction and algorithm evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 437–446, Sep. 2012.
[9] L. Chen, L. Sun, T. Yang, L. Fan, K. Huang, and Z. Xuanyuan, "RGB-T SLAM: A flexible SLAM framework by combining appearance and thermal information," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5682–5687.
[10] Y.-S. Shin and A. Kim, "Sparse depth enhanced direct thermal-infrared SLAM beyond the visible spectrum," 2019, *arXiv:1902.10892*. [Online]. Available: http://arxiv.org/abs/1902.10892
[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[13] Z. Cui, L. Heng, Y. C. Yeo, A. Geiger, M. Pollefeys, and T. Sattler, "Real-time dense mapping for self-driving vehicles using fisheye cameras," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6087–6093.
[14] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, "Towards robust visual odometry with a multi-camera system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1154–1161.
[15] M. M. Nawaf, D. Merad, J. P. Royer, J. M. Boï, M. Saccone, M. Ben Ellefi, and P. Drap, "Fast visual odometry for a low-cost underwater embedded stereo system," *Sensors*, vol. 18, no. 7, p. 2313, 2018.
[16] M. Vollmer and K. P. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Hoboken, NJ, USA: Wiley, 2017.
[17] J. Han, E. J. Pauwels, and P. de Zeeuw, "Visible and infrared image registration in man-made environments employing hybrid visual features," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 42–51, Jan. 2013.
[18] Y. Ni, X. Wang, and L. Yin, "Relative pose estimation for multiple cameras using lie algebra optimization," *Appl. Opt.*, vol. 58, no. 11, p. 2963, Apr. 2019.
[19] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2006, pp. 5695–5701.
[20] R. Gomez-Ojeda, F.-A. Moreno, D. Zuniga-Noel, D. Scaramuzza, and J. Gonzalez-Jimenez, "PL-SLAM: A stereo SLAM system through the combination of points and line segments," *IEEE Trans. Robot.*, vol. 35, no. 3, pp. 734–746, Jun. 2019.
[21] R. Gomez-Ojeda and J. Gonzalez-Jimenez, "Robust stereo visual odometry through a probabilistic combination of points and line segments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 2521–2526.
[22] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.
[23] L. Heng and B. Choi, "Semi-direct visual odometry for a fisheye-stereo camera," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4077–4084.
[24] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct SLAM for omnidirectional cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 141–148.
[25] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. ECCV*, 2014, pp. 834–849.
[26] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," 2016, *arXiv:1607.02555*. [Online]. Available: http://arxiv.org/abs/1607.02555
[27] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.
[28] R. Wang, M. Schworer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3903–3911.
[29] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.
[30] W. Dai, Y. Zhang, D. Sun, N. Hovakimyan, and P. Li, "Multi-spectral visual odometry without explicit stereo matching," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 443–452.

YUBO NI received the B.S. and M.S. degrees from the Sino-European Institute of Aviation Engineering, Civil Aviation University of China, Tianjin, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in instrumentation science and technology with Tianjin University. His research interests include precision measurement technology and instruments, vision measurement technology, and smart sensors.

YUE WANG received the B.S. degree in instrumentation science and technology from the Chongqing University of Technology, Chongqing, China, in 2012, and the M.S degree in instrumentation science and technology from the Hefei University of Technology, Hefei, China, in 2015. He is currently pursuing the Ph.D. degree in instrumentation science and technology with Tianjin University. His research interests include vision measurement technology and 3D reconstruction.

SHOUCHANG YANG received the B.S. degree in measurement and control technology and instrument from Tianjin University, Tianjin, China, in 2018. He is currently pursuing the M.S degree in instrumentation science and technology. His research interests include infrared sensor and image processing.



RUIXIANG CHEN received the B.S. degree in measurement and control technology and instrument from Tianjin University, Tianjin, China, in 2018. He is currently pursuing the M.S degree in instrumentation science and technology. His research interests include deep learning and vision measurement technology.

XIANGJUN WANG received the B.S., M.S., and Ph.D. degrees in precision measurement technology and instruments from Tianjin University, Tianjin, China, in 1980, 1985, and 1990, respectively. He is currently a Professor and the Director of the Precision Measurement System Research Group, Tianjin University. His research interests include photoelectric sensors and testing, computer vision, image analysis, MOEMS, and MEMS.