Computers, Materials & Continua
DOI:10.32604/cmc.2022.022053
Article

Tech Science Press

Binocular Vision Positioning Method for Safety Monitoring of Solitary Elderly

Lihua Zhu[1], Yan Zhang[1], Yu Wang[1,*] and Cheire Cheng[2]

[1]School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, 210000, China
[2]Department of Electrical and Electronic Engineering, Colorado State University, Colorado, United States
*Corresponding Author: Yu Wang. Email: wangyu78@njust.edu.cn

Received: 26 July 2021; Accepted: 13 September 2021

Abstract: In nowadays society, the safety of the elderly population is becoming a pressing concern, especially for those who live alone. There might be daily risks such as accidental falling or treatment attack on them. Aiming at these problems, indoor positioning could be a critical way to monitor their states. With the rapidly development of the imaging techniques, wearable and portable cameras are very popular, which could be set on human individual. And in view of the advantages of the visual positioning, the authors propose a binocular visual positioning algorithm to real-timely locate the elderly indoor. In this paper, the imaging model has been established with the corrected image data from the binocular camera; then feature extraction has been completed to provide reference to adjacent image matching based on the binary robust independent elementary feature (BRIEF) descriptor, finally the camera movement and the states of the elderly have been estimated to distinguish their falling risk. In the experiments, the real-sense D435i sensors were adopted as the binocular cameras to obtain indoor images, and three experimental scenarios have been carried out to test the proposed method. The results show that the proposed algorithm can effectively locate the elderly indoor and improve the real-time monitoring capability.

Keywords: Indoor positioning; binocular vision; feature matching; solitary elderly; safety monitoring

1 Introduction

Due to the decrease of the physical function of the elderly and the influence of various chronic diseases, it is of great social significance to improve the safety monitoring of their daily behaviors, especially for those who live alone. Indoor positioning technology with high accuracy and reliability can help to determine the individual position in real time. Various technologies for indoor positioning, such as, wireless Wi-Fi, Bluetooth, ultrasonic positioning, radio frequency identification (RFID) and ultra-wideband (UWB), have been proposed in the past decades. Wi-Fi [1] and Bluetooth [2] mainly rely on the signal strength, which are easily get interfered. Ultrasonic positioning is greatly affected by multipath effects and non-line-of-sight propagation. RFID positioning system [3] has strong anti-interference, but it requires a large amount of hardware, such as the deployment of recognizers and antennas, which is too complex to be applied. UWB technology [4,5] has the advantages of high positioning accuracy and strong penetration, while the delay module needs to be accurately calibrated. The vision-based technology generally employs the camera to collect visual information, and the image processing algorithms are applied to achieve the indoor positioning. It is merit in low-cost, high accuracy, and robustness in complex environment, makings it popular in various fields.

Throughout the development of the visual positioning, Sattler et al. [6] proposed a visual positioning method based on image database, which is capable of fast recovering the indoor coordinates through image characteristics, but it requires a number of environmental images in advance, so that it is not adaptive to unknown environment. Davision et al. [7] proposed a real-time monocular vision slam system, which is able to perform real-time positioning in unknown environment, but it requires external information to construct the spatial scale, while the error estimation of depth is still rather large. Binocular vision could obtain the depth information through stereo matching. The RGB-D detection calculates the distance between the potential object and camera by sending infrared light or pulse to the object, which is capable of reducing the uncertainty of the monocular scale, but it is too expensive to be widely used. By contrast, the low-cost binocular scheme can achieve the same function as long as two cameras are combined.

For binocular indoor monitoring of the elderly, the cameras are mounted on the individual, the most critical part is the positioning of the carrier in successive images. The popular implementations of binocular positioning mainly include the optical flow method and feature point method. The optical flow method is under the illumination invariance assumption [8], it uses the optical information of the image pixels to calculate the speed between two frames, and then estimates the camera motion with polar geometry. But the practice cannot completely agree with the illumination invariance assumption, resulting in low accuracy in use. Feature-based methods mainly refer to the point feature, the edge feature and the block feature. Among them, the point feature method outperforms the other two methods in identification and anti-noise ability [9]. It usually goes through feature extraction and feature matching to achieve the image association [10–12]. And the feature points of good quality are repeatable and unique, which are inevitable in popular feature-point method such as the scale-invariant feature transform (SIFT) algorithm, speeded-up robust features (SURF) algorithm, oriented features from accelerated segment test (FAST) and rotated brief (ORB) algorithm [13]. The SIFT [14] algorithm fully considers the illumination, rotation impact, scale invariance, view angle changing, affine transformation, and noise stability, but its computation is too complex to perform real-time application. SURF [15] is a kind of an upgrade of the SIFT, which not only has SIFT's advantage of high accuracy and robustness, but also has lower computational complexity as the descriptor dimension of SURF is lower than that of the SIFT. However, SURF is still not fast enough for real-time operation. In comparison, the use of ORB [16] can guarantee the real-time binocular positioning, it uses improved FAST angle [17–19] extraction algorithm to increase the rapidity with enhancement of the illumination invariance, scale invariance and rotation invariance. In view of its good robustness and real-time capability, the ORB algorithm has been adopted in this paper.

The feature matching of the binocular positioning can be divided into two aspects, namely the matching of the successive frames and the matching between the left and right images. For the matching of the successive frame matching, the widely-used direct matching method has large computation burden, so that this paper adopts the approximate nearest-neighbor algorithm to simplify the process. The depth information of the feature point can be obtained from the left and right images, also known as the stereo matching. Which can be divided into the global stereo matching algorithm and local stereo matching algorithm [20]. The global stereo matching algorithm that uses the global information of the image has a heavy computation burden, which is not suitable for the real-time application. The classic area-based algorithm based on regional gray levels detection [21,22] has good matching accuracy and fast speed but it is weak in anti-interference, which is adopted to perform binocular matching in this paper. With the feature matching and association of successive images, the relative positions and the motion status of carrier can be estimated by projecting the three dimensional (3D) perspective-n-point (PnP) to two dimensional (2D) plane.

Given the needs to the accurate and real-time stereo matching of the indoor positioning for the elderly living alone, this paper contributes to propose a binocular positioning based on the feature point extraction algorithm that unconventionally uses the BRIEF descriptor to construct the cost function to obtain accurate depth information. The proposed algorithm is dedicated to find the feature correspondence between two images by tracking the ORB feature points, and the PnP 3D-2D model has been built to perform the real-time motion estimation of the individual. After all, the position and movement tracking of the elderly could be timely monitored. Thus, it is possible to send out warning messages for potential risks. Through a series of experiments, the feasibility and effectiveness of the proposed indoor positioning technique have been evaluated.

The organization of this paper is as follows. Section 2 presents the general principles of the binocular positioning and the schematic process of the proposed design. And it describes the feature matching, the stereo matching method, and the implementation details of the BRIEF descriptor in Section 3. Section 4 verifies the proposed algorithm through a set of indoor tests. Section 5 concludes the paper.

2 Formulation of the Binocular Positioning Problem

The main flow diagram of the binocular positioning system is shown in Fig. 1. Firstly, the image sequence is acquired through the binocular camera, and stereo correction is then performed on the acquired left and right images. Secondly, the local stereo matching algorithm based on the BRIEF descriptor is employed to complete the estimation of depth information, and the ORB feature points in the image are extracted. The camera coordinate system of the camera is used as the world coordinate system of the system to calculate the world coordinates of the key points; then the approximate nearest-neighbor algorithm is used to achieve the matching between two adjacent frames of images, and the random sample consensus (RANSAC) algorithm is applied to eliminate the approximate nearest neighbor algorithm to a certain extent. Finally, the PnP 3D-2D model is used to estimate the pose of the carrier [23,24].

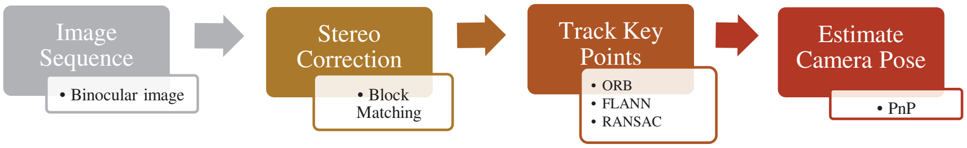


Figure 1: A block diagram showing the main components of a binocular positioning system

The detailed algorithm diagram is shown in Fig. 2. The image is stereo corrected according to the calibrated parameters of the binocular camera. And then the key points of the corrected left-side images would be extracted. If the current frame is the first image since the system started, the key points in the image are directly extracted, and the world coordinates frame of the left-side camera is used to calculate the coordinates of the key points, the current position and the posture matrix. Otherwise, the current left-side image is tracked by key points to find the correspondence of the pixels to further estimate the real-time pose. With the obtained pose, new key points could be matched to evaluate the depth information and restore the camera coordinates and world coordinates.

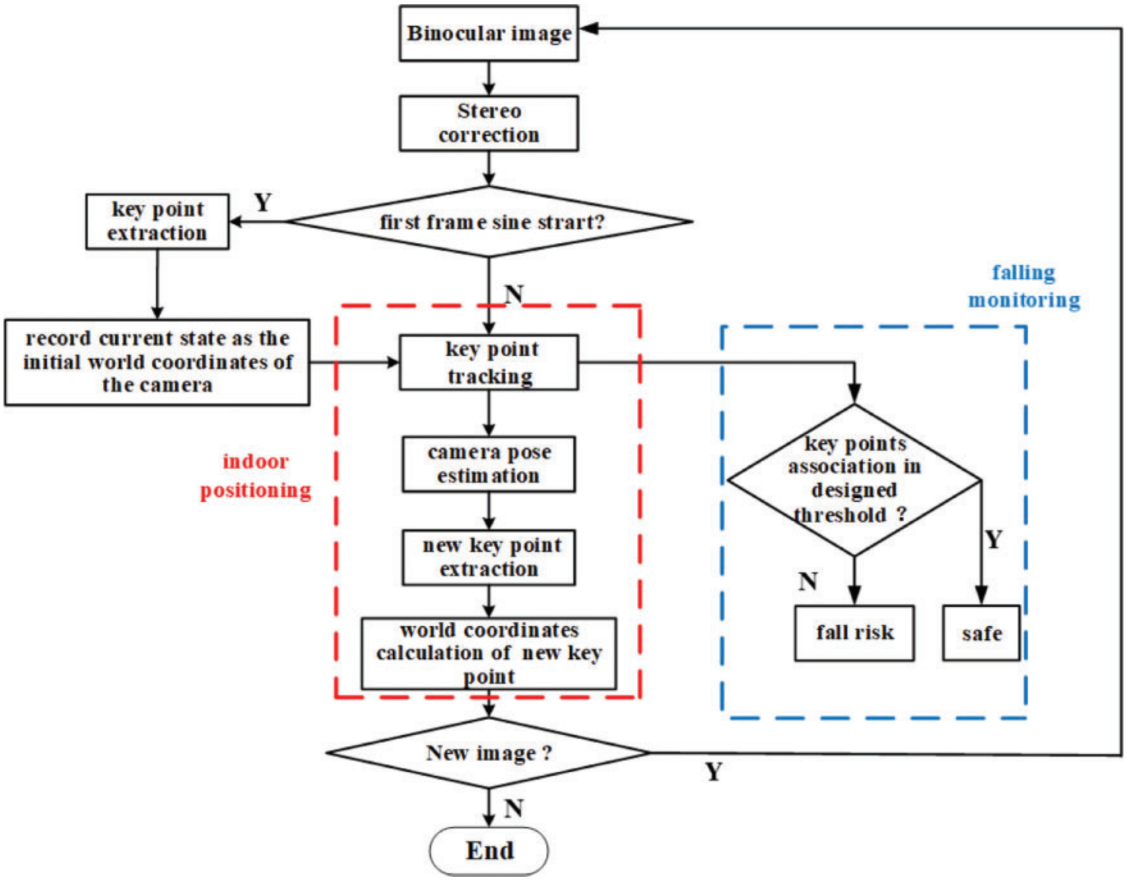


Figure 2: Stereo visual scheme of indoor positioning and safety evaluation

In the key point tracking process, a time threshold ΔT has been designed. It is assumed that the elderly generally moves slowly. That is to say, there would plenty of common key points in sequential images if the elderly is in a normal situation. If the elderly is accidental falling, there would be hardly key points that can be tracked. So that the key point tracking between two frames in the threshold is used to evaluate whether the elderly is in danger of falling.

3 Binocular Positioning Algorithms

The binocular camera model is shown in Fig. 3. The focal lengths and the imaging planes of the two cameras are the same, the two optical axes are parallel, and the pixels in each row of the image are precisely aligned.

The camera coordinate of a certain point in space is P^c (X_c, Y_c, Z_c), where Z_c is the depth. The projections of the points on the left and right cameras are the pixel coordinates $p_l(u_l, v)$ and $p_r(u_r, v)$, and their coordinates in the camera coordinate system are $p'_l(x_l, y)$, $p'_r(x_r, y)$, respectively. The depth of the point can be solved by the principle of similar triangles:

$$\frac{Z_c}{Z_c - f} = \frac{B}{B - x_l + x_r} \tag{1}$$

After simplification:

$$\frac{Z_c}{B} = \frac{f}{dx(u_l - u_r)} \tag{2}$$

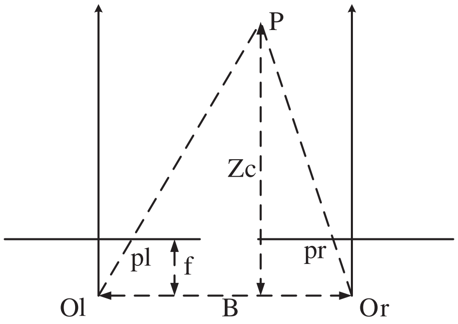


Figure 3: Ideal camera imaging model

Based on this model, the stereo alignment algorithm, feature extraction algorithm, feature point matching algorithm and motion estimation algorithm are studied gradually to achieve the indoor positioning of the elderly. In addition, the feature point matching algorithm has been employed to monitor the falling of the elderly.

3.1 Stereo Alignment Algorithm

In actual situations, it is often difficult for two cameras to achieve ideal coplanar and line alignment conditions. Generally, the rotation matrix $\boldsymbol{R}$ and the displacement vector $\boldsymbol{t}$ of the right camera relative to the left camera are used to represent the relative pose between the two cameras. For a spatial point P^w (X_w,Y_w,Z_w), the coordinates in the left and right camera coordinate systems are $P^{cl}(X_{cl},Y_{cl},Z_{cl})$ and $P^{cr}(X_{cr},Y_{cr},Z_{cr})$, respectively. Then the relationship between P^{cl} and P^{cr} can be expressed as

$$P^{cr} = \boldsymbol{R}P^{cl} + \boldsymbol{t} \tag{3}$$

The expressions of P^{cl} and P^{cr} relative to the world coordinate system coordinate P^w are:

$$\begin{aligned} P^{cl} &= \boldsymbol{R}_{wcl}P^w + \boldsymbol{t}_{wcl} \\ P^{cr} &= \boldsymbol{R}_{wcr}P^w + \boldsymbol{t}_{wcr} \end{aligned} \tag{4}$$

In Eq. (4), $\boldsymbol{R}_{wcl}$, $\boldsymbol{t}_{wcl}$ and $\boldsymbol{R}_{wcr}$, are external parameters of the left and right camera positions. The combination of three formulas can be solved:

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{R}_{wcr}\boldsymbol{R}_{wcl}{}^{T} \\ \boldsymbol{t} &= \boldsymbol{t}_{wcr} - \boldsymbol{R}\boldsymbol{t}_{wcl} \end{aligned} \tag{5}$$

Because of the projection errors, the rotation matrix $\boldsymbol{R}$ and displacement vector $\boldsymbol{t}$ calculated for each pair of points are different. To alleviate the problem, the camera is calibrated, such that the median of $\boldsymbol{R}$ and $\boldsymbol{t}$ calculated for each group of images is used as the initial value for the maximum likelihood estimation. By minimizing the reprojection error, accurate calibration results can be obtained.

$$\boldsymbol{E} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\|p_{ij} - \hat{p}_{ij}(\boldsymbol{K}, k_1, k_2, k_3, p_1, p_2, \boldsymbol{R}_{wcli}, \boldsymbol{t}_{wcli}, \boldsymbol{R}_{wcri}, \boldsymbol{t}_{wcri}, \boldsymbol{R}, \boldsymbol{t}, P_j)\right\|^2 \tag{6}$$

where $\boldsymbol{E}$ is the re-projection error, $\boldsymbol{K}$ is the internal parameter of the camera, k_1, k_2, k_3 are the radial distortion parameters, and p_1, p_2 are the tangential distortion parameters.

With the estimated rotation matrix $\boldsymbol{R}$ and the displacement vector $\boldsymbol{t}$ of the right camera relative to the left camera, the corrections on the binocular camera are performed. And $\boldsymbol{R}$ can be decomposed into two matrices $\boldsymbol{r}_l$ and $\boldsymbol{r}_r$ according to Eq. (7). $\boldsymbol{r}_l$ represents the half angle rotation matrix of the left camera, $\boldsymbol{r}_r$ represents the half angle rotation matrix of the right camera, the rotating direction of them is opposite.

$$\begin{aligned}&\boldsymbol{R} = \boldsymbol{r}_r^{-1}\boldsymbol{r}_l\\&\boldsymbol{r}_r\boldsymbol{r}_l = \boldsymbol{I}\end{aligned} \tag{7}$$

Afterwards, the left and right image planes are rotated about $\boldsymbol{r}_l$ and $\boldsymbol{r}_r$, respectively, to achieve co-planarity, and the effect is shown in Fig. 4.

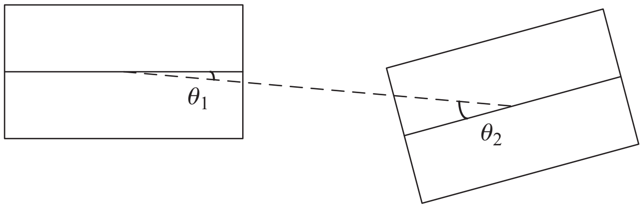


Figure 4: Two imaging planes in rotation

However, the baselines of the two image planes are not parallel, so that the transformation matrix $\boldsymbol{R}_{rect}$= $[e_1{}^T, e_2{}^T, e_3{}^T]^T$, has been constructed with the displacement vector to align the baselines. Where, e_1=$\boldsymbol{t}/||\boldsymbol{t}||$, e_2= $[-t_y, t_x, 0]^T/\sqrt{\mathrm{t}_\mathrm{x}^2 + \mathrm{t}_\mathrm{y}^2}$, e_3= $e_1 \times e_2$.

After all, the transformation matrix of the stereo correction can be expressed as

$$\begin{aligned}&\boldsymbol{R}_{left} = \boldsymbol{R}_{rect}\boldsymbol{r}_l\\&\boldsymbol{R}_{right} = \boldsymbol{R}_{rect}\boldsymbol{r}_r\end{aligned} \tag{8}$$

3.2 Feature Extraction Algorithm

In this paper, the feature matching of adjacent images has been completed by extracting ORB points, which are composed of key points and descriptors. Likewise, the feature detection algorithm can also be divided into two parts, namely, the key point extraction and descriptor construction. The key point extraction algorithm also known as the oFAST algorithm [16], which is evolved from the FAST algorithm [17], introduces the image pyramid and the gray-scale centroid method to guarantee the invariance of the FAST feature scale and rotation.

The main steps of the oFAST algorithm are presented as follows:

Step 1: A circle at point $\boldsymbol{P}$ with a radius of three pixels is determined as the center. The boundary of the circle passes through 16 pixel grids, marked as $\boldsymbol{P}_1 \sim \boldsymbol{P}_{16}$, and the brightness value of point $\boldsymbol{P}$ is $\boldsymbol{I}_p$, as shown in Fig. 5.
Step 2: Set the threshold $\triangle \boldsymbol{t}$ for brightness variation. From $\boldsymbol{P}_1$ to $\boldsymbol{P}_{16}$, the brightness of the 1st, 5th, 9th, and 13th pixels are first detected on the circle. Only if over three quarters' brightness of the pixels are greater than $\boldsymbol{I}_{p+t}$ or less than $\boldsymbol{I}_{p-t}$, the current pixel is considered as a corner point. With successive 12 pixels' brightness greater than $\boldsymbol{I}_{p+t}$ or less than $\boldsymbol{I}_{p\text{-}t}$, point $\boldsymbol{P}$ is recorded as a candidate feature point.
Step 3: Repeat the above two steps and perform the same operation on all pixels.
Step 4: Remove locally dense candidate feature points, and calculate the FAST credit of each candidate point through Eq. (9).

$$\boldsymbol{V} = \max\left(\sum_{x\in s_b} |\boldsymbol{I}_n - \boldsymbol{I}_p| - \Delta \boldsymbol{t}, \sum_{x\in s_1} |\boldsymbol{I}_n - \boldsymbol{I}_p| - \Delta \boldsymbol{t}\right) \tag{9}$$

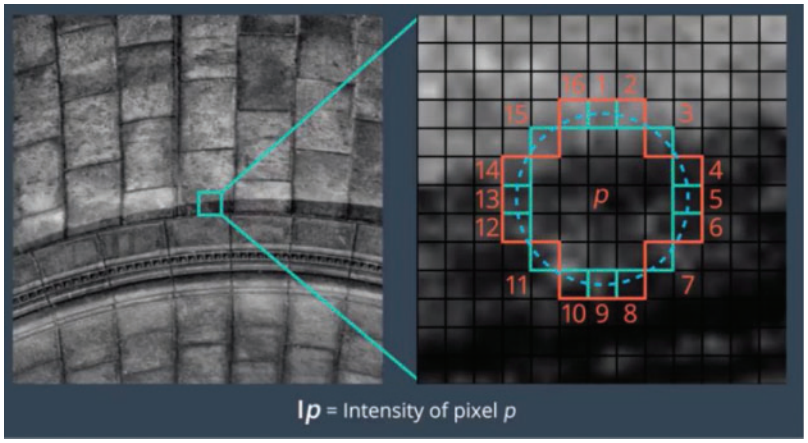


Figure 5: Pixel diagram

$\boldsymbol{I}_n$ is the value of the pixels on the circle, S_b is the set of pixels whose brightness value is greater than $\boldsymbol{I}_{p+t}$, and S_l is the set of pixels whose gray value is less than $\boldsymbol{I}_{p-t}$. Then the gray values of adjacent candidate feature points are compared, the candidate points with a larger $\boldsymbol{V}$ gray value are kept as key points, and the candidate points with a smaller gray value are removed.

Step 5: Construct a Gaussian pyramid and add the scale invariance of key points. Set the number of pyramid levels $n = 8$, and the scale factor $s = 1.2$. The original image can be scaled and the pixel value $\boldsymbol{I}'$ of the image in each layer can be obtained with Eq. (10):

$$\boldsymbol{I}' = \boldsymbol{I}_p/1.2^k (k = 1, 2, \ldots, 8) \tag{10}$$

Step 6: The direction vector is constructed by the gray-centroid method, to strengthen the rotation invariance of key points. In the neighborhood image block of the key point, the moment of the image block $\boldsymbol{m}_{pq}$ is defined as:

$$\boldsymbol{m}_{pq} = \sum_{x,y\in B} x^p y^q \boldsymbol{I}(x,y) \quad p,q = \{0,1\} \tag{11}$$

The centroid of the image block can be determined by the moment $\boldsymbol{C}$:

$$\boldsymbol{C} = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{12}$$

Connecting the geometric center $\boldsymbol{O}$ and the centroid $\boldsymbol{C}$ of the image block to obtain a direction vector $\overrightarrow{\mathbf{OC}}$, the direction of the feature point can be defined as:

$$\theta = \arctan(m_{01}/m_{10}) \tag{13}$$

After the extraction of the oFAST key points, the descriptor of each point needs to be calculated, where the improved BRIEF is used. BRIEF is a binary descriptor. Its description vector consists of 0 and 1, which encode the size relationship between two pixels near the key point (denoted as p and q): if p is greater than q, it takes 1; otherwise, it takes 0. If we take 128 pairs of (p, q) on a key-point-center circle with a radius of a certain number of pixels, a 128-dimensional vector consisting of 0 and 1 can be obtained. BRIEF uses randomly selected points for comparison, which is very fast and convenient to store, and it is superior in real-time image matching. In ORB, the rotation-aware BRIEF descriptor is improved by adding a twiddle factor on the basis of the BRIEF descriptor.

3.3 Feature Point Matching Algorithm

For the feature point matching of the two images, the computation load is heavy through directly comparing the Hamming distance of each feature point, which hardly satisfies the real-time requirements. The approximate nearest-neighbor algorithm integrated in the Fast Library for Approximate Nearest Neighbors Open-source library is faster and adaptive for real-time occasions, but mismatches may occur more or less. The RANSAC algorithm can eliminate the mismatch by effectively calculating the homograph matrix. The homograph matrix is a conversion matrix that describes the mapping relationship between the corresponding points of two pictures on the plane, it is defined as follows [25,26]:

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \boldsymbol{H} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \tag{14}$$

where $(u, v, 1)^T$ and $(u', v', 1)^T$ are the pixel coordinates, $\boldsymbol{H}$ is the homograph matrix.

A pair of points can determine two equations, that is to say, at least 4 pairs of matching points are needed to determine the $\boldsymbol{H}$ matrix. The actual logarithm of the initial matching point is much greater than 4, so that the RANSAC is used to obtain the resolution. The main steps are as follows:

Step 1: Randomly select 4 pairs of matching points to fit the model (that is, estimate the homograph matrix $\boldsymbol{H}$);

Step 2: Due to the matching errors, the data points have certain fluctuations. Assuming that the error envelope is δ, taking the matrix $\boldsymbol{H}$ in step 1 as a benchmark, to calculate the residual matching error, find the points within the error envelope, and record the point number n;
Step 3: Randomly select 4 pairs of points again, and repeat the operations of Step 1 and Step 2 until the iteration stops;
Step 4: Find the homograph matrix $\boldsymbol{H}$ that satisfies the largest n.

When tracking the feature points in successive frames, a time threshold was set to count the feature point quantity in the interval to evaluate the falling risk of the elderly. It is considered that the elderly generally moves slowly, so that if there are plenty of common feature points in in successive frames, the elderly is thought be safe; otherwise, if there are few common feature points in in successive frames, the elderly has falling risk.

For the corresponding feature points matching of the left and right images, this paper uses the area-based algorithm based on the BRIEF descriptor. The traditional block matching algorithm is shown in Fig. 6. After stereo correction, the left image is used as a reference, and the pixel coordinate of a key point is set to (u_l, v), its gray scale is $I(u_l, v)$. This key point is taken as the center, and the $\boldsymbol{M}\times\boldsymbol{N}$ area (denoted as window $\boldsymbol{W}$) around center is viewed as the matching unit. According to the binocular imaging model, the corresponding point (u_r, v) in the right image must be on the left side of the key point. Taking the right image point which has the same coordinates in the left image as the start, sliding the window $\boldsymbol{W}$ from right to left along the row, to compare every pixel in turn, so that the image similarity can be calculated. The point with the largest similarity value is regarded as the matching point.

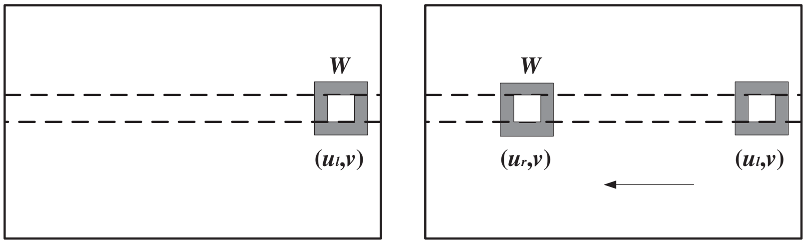


Figure 6: Block matching principle

Because of the existence of sheltering, the key points in the left image may not be able to find matching points in the right image. In the matching process, the sum of absolute differences S_1 is used as the similarity function to measure the matching degree of two points and the surrounding window, it is expressed in Eq. (15).

$$S_1 = \sum_{i,j\in W} |\boldsymbol{I}_l(u_l + i, v + j) - \boldsymbol{I}_r(u_r + i, v + j)| \tag{15}$$

In which, $\boldsymbol{I}_l$ and $\boldsymbol{I}_r$ are the brightness values of the left and right pixels, respectively. However, owing to the illumination impact, the brightness of the pixel is susceptible to external interference, which may introduce errors to affect the matching accuracy. In the process of extracting ORB feature points, the BRIEF descriptor has been obtained, and the BRIEF vector is used as the feature information instead of the brightness value to construct the cost function, which can effectively help to improve the matching accuracy. The similarity measurement function S_2 is shown in Eq. (16):

$$S_2 = \sum_{i,j \in W} |\boldsymbol{L}(u_l + i, v + j) - \boldsymbol{R}(u_r + i, v + j)| \tag{16}$$

where $\boldsymbol{L}$ and $\boldsymbol{R}$ mainly refer to the BRIEF features of the left and right pixels. When the value of the similarity measurement function is at the peak, the matching is ended, and the matching point is obtained.

3.4 Motion Estimation Algorithm

With the tracks of the key points, the matching relations of them can be obtained to further estimate the camera motion. That is, after the stereo matching, the three-dimensional camera coordinate $\boldsymbol{P}_c$ of the key points can be obtained. For the k-1*th* frame image with known external parameters, the world coordinates of the key points can be recovered accordingly. The key points are then tracked to obtain the pixel coordinates of the key points in the subsequent image frames. According to the correspondence between the pixel coordinates and the world coordinate system, the camera pose at the k-*th* frame can be restored. Therefore, the camera motion recovery is depicted as a 2D-3D multi-point perspective problem, or the PnP problem [27].

The reprojection error, comparing the pixel coordinates (observed projection after matching) with 3D point projection of real-time pose estimation, is produced. For which, a nonlinear optimization has been adopted to find a possible solution. As shown in Fig. 7, through feature point matching, p_1 and p_2 are the projections of the same spatial point $\boldsymbol{P}$ on the two images before and after, while the pose of the camera is unclear. After substituting the initial value, there is a certain distance between the projection p_2^ of $\boldsymbol{P}$ on the next frame of image and the actual p_2. Therefore, the pose of the camera needs be adjusted to reduce this difference. And there are many points to deal with, the error of each point is hardly zero.

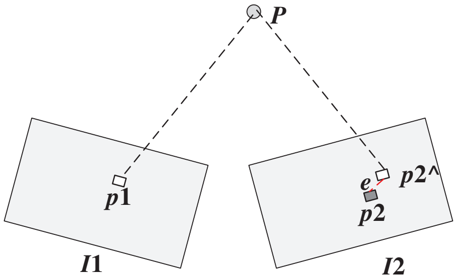


Figure 7: Schematic diagram of reprojection error

In Fig. 7, the homogeneous coordinates of the spatial point $\boldsymbol{P}$ are $P = [X, Y, Z, 1]^T$, and the pixel coordinates of its projection in the image $\boldsymbol{I}_1$ are $p_1 = [u_1, v_1]^T$, the pixel coordinates of the reprojection in image $\boldsymbol{I}_2$ are $p_2\hat{} = [u_2', v_2']^T$, and the observation value of the spatial point $\boldsymbol{P}$ in the image $\boldsymbol{I}_2$ is $p_2 = [u_2, v_2]^T$, while $\boldsymbol{e} = p_2\text{-}p_2\hat{}$ represents the reprojection error. The ideal re-projection process is expressed by Eq. (17):

$$s_2 u_2 = \boldsymbol{K} \exp(\boldsymbol{\xi}^\wedge) \boldsymbol{P} \tag{17}$$

where s_2 represents the depth of the spatial point $\boldsymbol{P}$ in the camera coordinate system where the image $\boldsymbol{I}_2$ is located, and $\boldsymbol{K}$ represents the camera internal parameters, which represents the posture transformation matrix of the camera from image $\boldsymbol{I}_1$ to image $\boldsymbol{I}_2$, which can also be represented by $\boldsymbol{T}$, and $\boldsymbol{\xi}$ represents the Lie algebra corresponding to $\boldsymbol{T}$. There is usually a certain error with the true value during reprojection. The definition of this error is shown in Eq. (18):

$$e_2 = u_2 - \frac{1}{s_2}\boldsymbol{K}\exp(\boldsymbol{\xi}^\wedge)p \tag{18}$$

There are often more than one feature point observed in a camera pose. Assuming that there are N feature points, it constitutes the least squares problem of finding the camera pose $\boldsymbol{\xi}$:

$$\xi^* = \arg\min_{\xi} \frac{1}{2}\sum_{1}^{N} ||e_i||_2^2 = \arg\min_{\xi} \frac{1}{2}\sum_{i=1}^{N} ||(u_i - \frac{1}{s_i}K\exp(\xi^\wedge)p_i)||_2^2 \tag{19}$$

4 Experimental Results and Analysis

The experimental platform of this paper uses a laptop computer (Lenovo Xiaoxin Air15), and the running environment is the Ubuntu 16.04 operating system under the VirtualBox virtual machine. Using Intel's RealSense camera D435i, the camera is a global shutter, the frame rate is 30 fps, the image resolution is 1280 × 720, and the camera baseline is 5 cm. The experimental scene is indoor, and the sensor is handheld to move in the scenario to estimate the pose. The camera installation and its connection with computer is shown in Fig. 8.

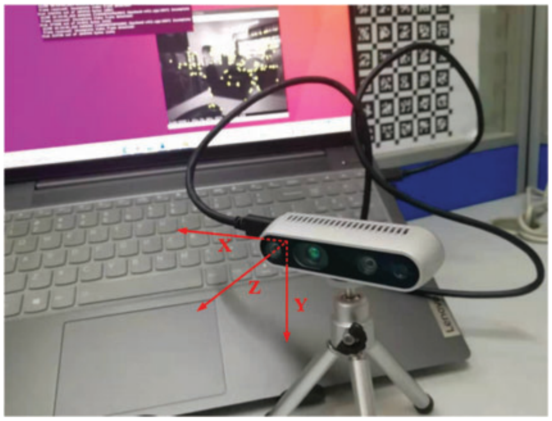
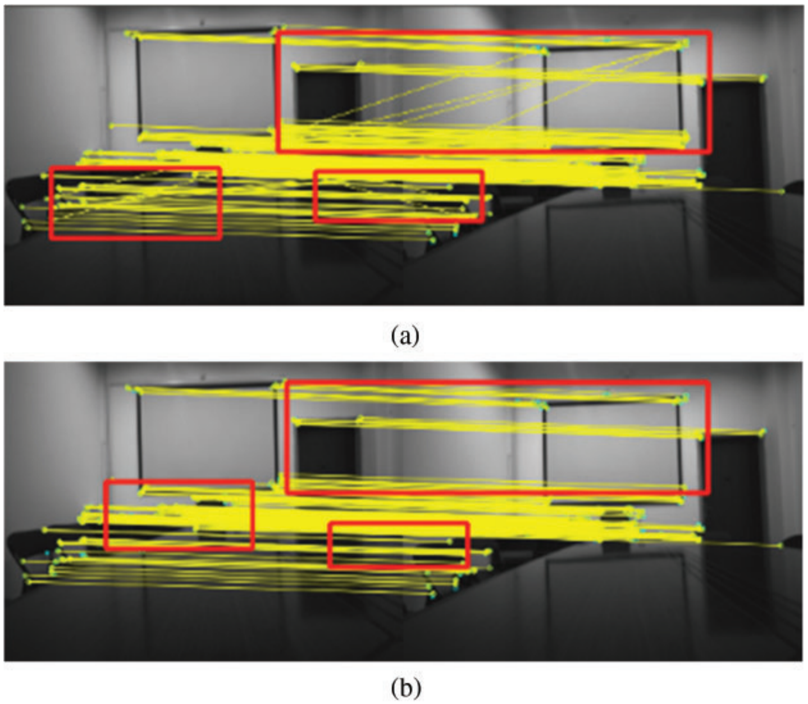


Figure 8: The camera installation

4.1 ORB Feature Matching Experiments

During the experiment, the ORB feature points were extracted and matched on two adjacent frames of images. Fig. 9(a) shows the result with the approximate nearest neighbor algorithm where the mismatch has not been eliminated, and Fig. 9(b) shows the result with the combination of the nearest neighbor and RANSAC algorithm, where the mismatch has been eliminated. It can be seen that there are many matching errors as circled by the red boxes in Fig. 9(a), but fewer mismatches in the same area in Fig. 9(b). Therefore, it illustrates that the combination algorithm has better matching ability, which is beneficial to the improvement of the accuracy of the binocular positioning.


(a)

(b)

Figure 9: (a) Feature extraction with approximate nearest neighbour, (b) Feature extraction with nearest neighbor

Then, the hand-held camera performed a slow linear motion, the sampling interval was 1 s. From the observation of the feature tracking between two frames, as shown in Fig. 10a, it is clearly that plenty of key points had to be tracked.

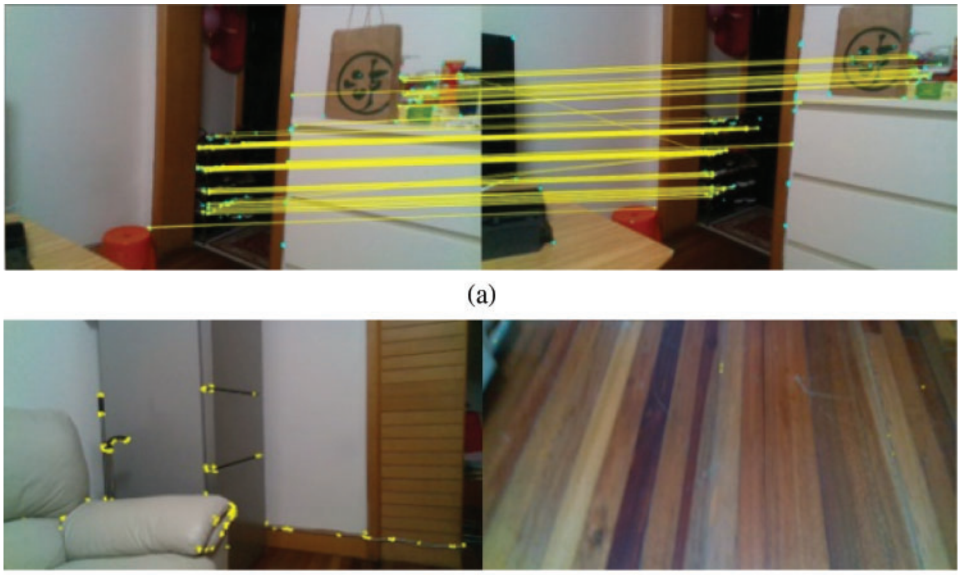
Afterwards, the hand-held camera had been swung quickly to simulate the falling situation of the elderly, and two image frames were sampled at 1 s. The result in Fig. 10b shows that there are almost no key points tracked in the latter image, indicating that the two images with an interval of 1 s have almost no common viewpoint, which is an abnormal movement of the elderly. At this time, it can be judged that the elderly might be in falling danger.

4.2 Indoor Positioning Experiment

The positioning experiment was also carried out indoor, including the linear reciprocating motion and arbitrary motion, to evaluate the positioning capability of the binocular scheme designed in this paper.

4.2.1 Linear Motion Scenario

In this scenario, the handheld camera was kept at a certain height while moving straightly between two points at three different distances: 1 m, 3 m, and 5 m, respectively. The camera coordinate of the first image after the system initialization was defined as the world coordinate. The camera coordinate system was defined as: the facing direction of the camera lens was the positive direction of the *z*-axis, the *x*-axis pointed right the camera, the *y*-axis and *x*, *z* constitute the right-hand coordinate system. And the *y*-axis and *x*-axis, and the *z*-axis constitute the right-hand coordinate system, as shown in



(a)

(b)

Figure 10: (a) Feature point tracking of a slow linear motion, (b) The tracking of feature points of a quickly swung

Fig. 8. The starting point coordinates were (0, 0) m, the end point coordinates were (0, 1) m, (0, 3) m, (0, −5) m. The images collected during the experiments are shown in Fig. 11.

The trajectories of the linear motion at the three distances are shown in Fig. 12. The positioning results are shown in Tab. 1.

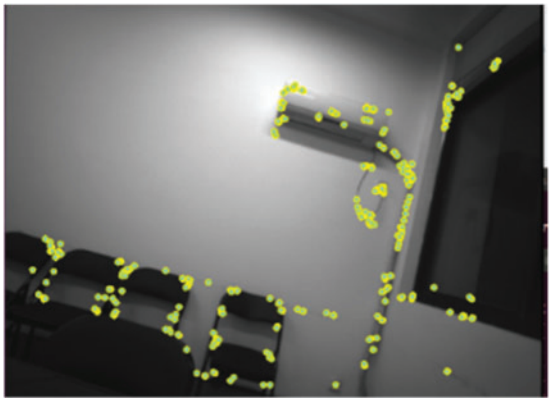


Figure 11: Effect picture collected during the experiment

It can be seen from Tab. 1 that when the carrier performs linear reciprocating motions of different lengths, the positioning error is within 65 cm. As the running length increases, the positioning error does not diverge significantly.

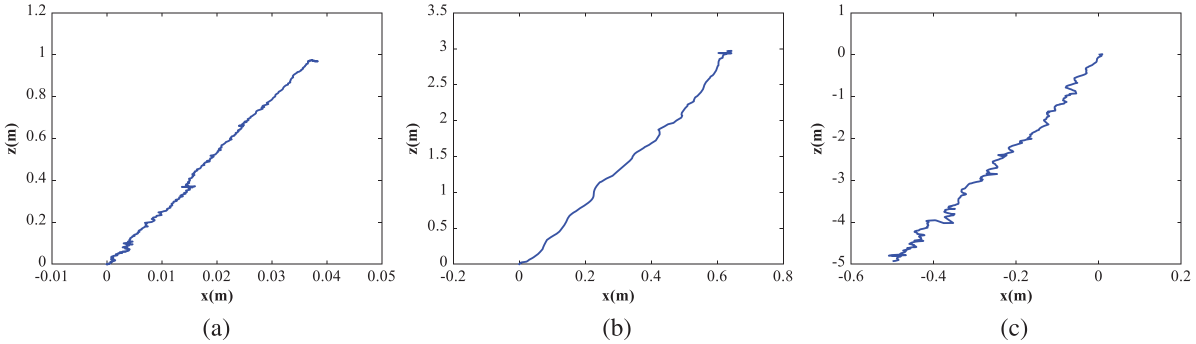


Figure 12: Trajectory diagram, (a) 1 m movement (b) 3 m movement (c) 5 m movement

Table 1: Linear motion measurement positioning error

Movement process	End point positioning coordinates	Positioning error (m)
First linear motion	(0.03832, 0.9687)	0.0495
Second linear motion	(0.6433, 2.9735)	0.6438
The third linear motion	(−0.4996, −4.928)	0.5048

4.2.2 Arbitrary Movement Scenario

After holding the camera indoor for arbitrary movement, it returns to the start point, and the result of ORB feature point extraction of the collected images is shown in Fig. 13. The pose estimation of the odometer is shown in Fig. 14. Assuming the position of the start is (0, 0, 0) m, the calculated position of the end is (−0.03765, −0.1251, −0.01099) m, the positioning error is 0.1311 m.

Combining the experimental results of the carrier's linear reciprocating motion and arbitrary trajectory motion in the room, it can be seen that the binocular positioning algorithm designed in this paper has a good positioning capability for indoor environments. As for the fluctuating magnitude in y-axis is mainly caused by the walking up and down of the human body.

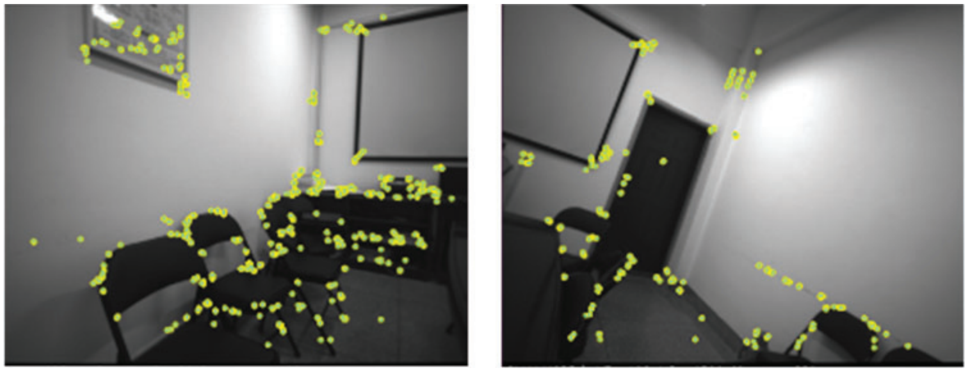


Figure 13: Effect picture collected during the experiment

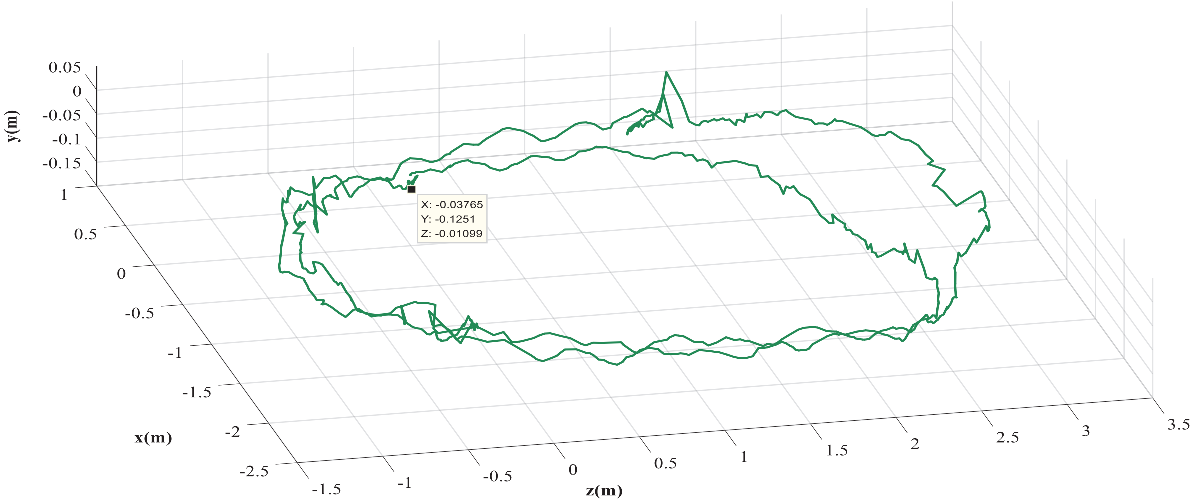


Figure 14: Three-dimensional view of the motion estimation

5 Concluding Remark

Aiming at the pressing concern of the indoor monitoring of the alone-living elderly, this paper proposed a positioning algorithm based on the binocular visual scheme through feature extraction, feature matching, and motion estimation, to finally obtain a high accuracy location of the indoor elderly. And feature matching is focused and modified. On one hand, the RANSAC algorithm has been adopted to eliminate the mismatch caused by the approximate nearest neighbor method; on the other hand, a cost function based on the BRIEF descriptor has been proposed as the feature information to improve the stereo matching accuracy. On this basis, the feature point comparison of two image frames within a certain time interval is used to determine whether the elderly is in falling danger. Three sets of experiments are carried out to verify the feasibility of the proposed method. Through the feature matching experiment, it can be intuitively seen that the RANSAC algorithm can effectively eliminate the mismatch; the contrast of the walking and falling situations in feature matching experiment also demonstrates the tracking efficacy of two images in the designed time threshold and verifies the feasibility of the falling danger evaluation; furthermore, the effectiveness and accuracy of the improved method with the BRIEF descriptor is verified by the indoor positioning experiments in different situations. It is worthy to mention that, there were position and attitude drifts due to accumulated errors in the measurement system. The further study will continue to work on this issue by adding auxiliary navigation, taking the inertial measurement unit for instance, to improve accuracy.

Funding Statement: This work was supported by the National Natural Science Foundation of China (No. 61803203).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Liu, “Study on indoor positioning technologies based on Wi-Fi,” *Information and Communication,* no. 2, pp. 259–260, 2018.

[2] H. Li, “Low-cost 3D bluetooth indoor positioning with least square,” *Wireless Personal Communications,* vol. 78, no. 2, pp. 1331–1344, 2014.

[3] S. Alghamdi, R. S. van and I. Khalil, “Accurate positioning using long range active RFID technology to assist visually impaired people,” *Journal of Network and Computer Applications,* vol. 42, pp. 135–147, 2014.

[4] H. Y. Yu, J. Cao and H. M. JI, “An indoor navigation system method based on UWB,” *Digital Technology and Application,* vol. 38, no. 3, pp. 138–142, 2020.

[5] A. Alhussain, H. Kurdi and L. Altoaimy, “A neural network-based trust management system for edge devices in peer-to-peer networks,” *Computers,” Materials & Continua,* vol. 59, no. 3, pp. 805–815, 2019.

[6] T. Sattler, B. Leibe and L. Kobbelt, “Fast image-based localization using direct 2D-to-3D matching,” in *Int. Conf. on Computer Vision*, Barcelona, Spain, pp. 667–674, 2011.

[7] A. J. Davison, I. D. Reid and N. D. Molton, “MonoSLAM: Real-time single camera SLAM,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1052–1067, 2007.

[8] R. F. Guo and J. Rong, “Research on multi-information fusion target tracking algorithm based on LK optical flow method,” *Modern Electronics Technique,* vol. 42, no. 18, pp. 55–59, 2019.

[9] Y. F. Guan, “Reaearch on positioning technology combining binocular vision and inertial measurement unit,” M.A. dissertation. Harbin Institute of Technology, China, 2020.

[10] D. Nistér, “An efficient solution to the five-point relative pose problem,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, no. 6, pp. 756–770, 2004.

[11] O. Naroditsky, X. S. Zhou and J. Gallier, “Two efficient solutions for visual odometry using directional correspondence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 4, pp. 818–824, 2011.

[12] I. Parra, M. A. Sotelo and D. F. Llorca, “Robust visual odometry for vehicle localization in urban environments,” *Robotica,* vol. 28, no. 3, pp. 441–452, 2010.

[13] G. Huang, “Visual-inertial navigation: a concise review,” in *2019 Int. Conf. on Robotics and Automation (ICRA)*, Montreal, Canada, pp. 9572–9582, 2019.

[14] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91–110, 2004.

[15] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, “Speeded-up robust features (SURF),” *Computer Vision and Image Understanding,* vol. 110, no. 3, pp. 346–359, 2008.

[16] E. Rublee, V. Rabaud and K. Konolige, “ORB: an efficient alternative to SIFT or SURF,” in *Int. Conf. on Computer Vision*, Barcelona, Spain, pp. 2564–2571, 2011.

[17] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *European Conference on Computer Vision,* Springer, Berlin, Heidelberg. pp. 430–443, 2006.

[18] A. Chhabra, G. Singh and K. S. Kahlon, “Qos-aware energy-efficient task scheduling on hpc cloud infrastructures using swarm-intelligence meta-heuristics,” *Computers, Materials & Continua,* vol. 64, no. 2, pp. 813–834, 2020.

[19] A. Janarthanan and D. Kumar, “Localization based evolutionary routing (lober) for efficient aggregation in wireless multimedia sensor networks,” *Computers, Materials & Continua,* vol. 60, no. 3, pp. 895–912, 2019.

[20] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *International Journal of Computer Vision,* vol. 47, no. 1, pp. 7–42, 2002.

[21] L. Hong and G. Hen, “Gment-based stereo matching using graph cuts,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, vol. 1, pp. I–I, 2004.

[22] L. D. Stefano, M. Marchionni and S. Mattoccia, “A fast area-based stereo matching algorithm,” *Image and Vision Computing,* vol. 22, no. 12, pp. 983–1005, 2004.

[23] X. Z. Huang, J. B. Wang, F. Q. Gao and Y. Y. Ran, “Location algorithms of machine binocular vision odometer,” *Radio Communication Technology,* vol. 45, no. 6, pp. 676–681, 2019.

[24] F. Xiao, W. Liu, Z. T. Li, L. Chen and R. Wang, “Noise-tolerant wireless sensor networks localization via multi-norms regularized matrix completion,” *IEEE Transactions on Vehicular Technology,* vol. 67, no. 3, pp. 2409–2419, 2018.

[25] W. K. Zhao and G. Li, “Visual odometry based on binocular camera,” *Computer Engineering and Design,* vol. 41, no. 4, pp. 1133–1138, 2020.

[26] Q. Y. Deng, Z. T. Li, J. B. Chen, F. Z. Zeng, H. M. Wang *et al.* “Dynamic spectrum sharing for hybrid access in OFDMA-based cognitive femtocell networks,” *IEEE Transactions on Vehicular Technology,* vol. 67, no. 1, pp. 10830–10840, 2018.

[27] C. Tang, X. Zhao and J. Chen, “Fast stereo visual odometry based on LK optical flow and ORB-SLAM2,” *Multimedia Systems,* vol. 26, no. 3, pp. 1–10, 2020.