

A Visible Light Positioning System Based on Monocular Vision

Jing Yang¹, Xinyao Fu², Yizhe Zhu³, Yu Zhang⁴

^{1,2,3,4}Xi'an Technological University, Xi'an, China

¹Corresponding author

E-mail: ¹3076602318@qq.com, ²3555299732@qq.com, ³2768938977@qq.com, ⁴3347791850@qq.com

Abstract

As a crucial component of location services, indoor positioning technology holds broad application prospects in areas such as mall navigation, warehousing logistics, and smart security systems. However, traditional indoor positioning methods suffer from limitations including insufficient accuracy, high costs, and susceptibility to environmental interference. This paper proposes a portable indoor positioning system based on multimodal data fusion, which integrates multi-dimensional features—including color, shape, and position—from optical imaging to construct a backpropagation (BP) neural network model enabling centimeter-level positioning. The system employs the K210 chip in conjunction with the CanMV framework, incorporating multimodal feature extraction techniques such as interference-resistant color recognition in the LAB color space, Hof transform-based circular detection, and rectangular contour fitting, achieving 95% color recognition accuracy under complex lighting conditions. Experimental results demonstrate a root mean square positioning error of 0.79 cm and an average percentage error of 3%, representing a significant improvement over traditional decimeter-level positioning accuracy. This study provides an effective solution for developing low-cost, highly reliable indoor positioning systems.

Keywords: Multimodal fusion, Indoor positioning, Visible light positioning, BP neural network, K210 chip.

1. Introduction

With the rapid advancement of mobile internet and IoT technologies, location services have become an indispensable infrastructure in modern society. While the Global Positioning System (GPS) performs exceptionally well in outdoor environments, its effectiveness is limited indoors due to building obstructions. Statistics indicate that people spend approximately 70% to 90% of their time indoors, driving growing demand for indoor location services and making indoor positioning technology a focal point of both academic and industrial research.

Currently, indoor positioning technologies are exhibiting diverse developmental trends. WiFi-based positioning estimates locations by measuring signal strength, offering lower costs but limited accuracy; Bluetooth positioning boasts advantages such as compact device size and low power consumption, yet suffers from shorter transmission ranges and poorer stability; ultra-wideband technology achieves centimeter-level precision but requires specialized hardware deployment, incurring higher costs. A common challenge shared by these technologies is the need for auxiliary nodes (e.g., range-measuring or angle-measuring devices) to enhance accuracy, thereby increasing system complexity and deployment costs. In contrast, image sensor-based positioning methods offer advantages like intuitive operation and rich information output. Visible light positioning technology utilizes widely distributed indoor LED light

sources as beacons, collecting light signals through image sensors for positioning without requiring additional specialized equipment. However, existing research faces challenges such as complex acquisition devices, substantial data processing volumes, and poor environmental adaptability, hindering the widespread adoption of these methods.

Multimodal data fusion technology offers a novel approach to addressing the aforementioned challenges. By integrating data features from multiple dimensions—including visual information such as color, shape, and position—it enables complementary enhancement of positioning accuracy. Compared to traditional single-modal methods, multimodal fusion fully leverages the strengths of diverse data sources, improving the systems robustness and precision in complex environments. This paper develops a multimodal data acquisition and fusion system based on the K210 chip and the CanMV framework, incorporating a backpropagation (BP) neural network to achieve end-to-end indoor positioning, thereby providing an innovative solution for cost-effective and highly reliable indoor positioning.

2. Overall System Architecture and Multimodal Acquisition Principles

2.1 Overall System Operating Principle

The visible light positioning system based on monocular vision achieves image recognition and localization through the collaborative operation of the K210 chip and the CanMV framework. The system workflow is as follows: the OV2640 monocular camera captures images; the K210 chip utilizes its built-in KPU hardware accelerator for image preprocessing and feature extraction; the CanMV framework performs target color recognition, contour detection, shape classification, and center coordinate calculation; the multidimensional features are then fed into the BP neural network model for training and prediction, ultimately enabling the system to predict actual spatial positions using features such as color, shape, and center coordinates.

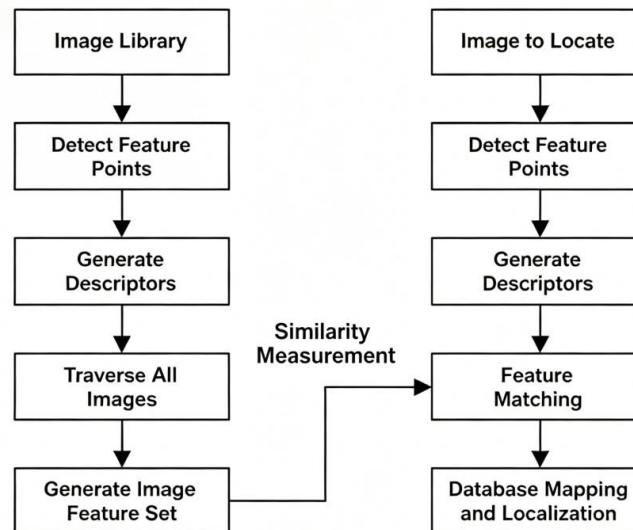


Fig. 1. Feature Matching-Based Visible Light Localization Algorithm^[1]

The systems core advantage lies in its collaborative working mechanism between the K210 and CanMV. The K210 employs a dual-core RISC-V CPU architecture equipped with a KPU neural network accelerator delivering 0.8 TFLOPS of computational power, providing robust support for

image processing tasks. CanMV, built on MicroPython, features a hierarchical architecture: the lower layer drives hardware, the middle layer provides algorithm APIs, and the top layer facilitates application development, significantly reducing the complexity of embedded vision application development. In the data processing workflow, camera-acquired images first undergo preprocessing by the KPU, followed by color recognition and shape detection using CanMV algorithms, with final coordinate calculations performed by the CPU. By sharing memory, both components minimize data transfer overhead while balancing low power consumption and real-time performance—the system operates at 20 FPS in QVGA resolution with a single-frame processing latency under 50 ms and standby power consumption below 200 mW.

2.2 Principles of Multimodal Data Acquisition

The core of multimodal data acquisition lies in extracting multi-dimensional features from optical imaging to establish a comprehensive target description system. The system employs an OV2640 camera to capture RGB565 images with a resolution of 160×120 , with automatic gain and white balance functions disabled to mitigate the impact of lighting variations on the imaging data.

For color recognition, the system converts images from RGB to the LAB color space. This conversion utilizes the XYZ color space as an intermediate bridge. The primary advantage of the LAB color space lies in the independence of luminance (L) and chrominance components (a, b), effectively isolating the impact of lighting variations on color information. Separate thresholds are established for red, green, blue, and yellow; a color is identified as the target when its similarity falls below the predefined threshold. Experimental results demonstrate that this method achieves a 95% color recognition accuracy in mixed indoor fluorescent and natural lighting environments, offering 40% improved interference resistance compared to traditional RGB threshold methods.

Shape feature extraction employs a dual-modal detection strategy. Circle detection is based on the Hough transform principle, mapping circular parameters from the image space to a three-dimensional parameter space (a, b, r) and determining parameter values through cumulative voting. When the voting function $V(a, b, r)$ exceeds 3000 and the radius r falls within the range of 2 to 100 pixels, a circular target is detected with an accuracy of up to 92%. Rectangle detection achieves a unified geometric description by fitting the circumscribed circle using contour corner points, and determines the circle's center and radius using the least squares method, with positioning error controlled within 5 pixels.

3. Construction of the BP Neural Network Localization Model

3.1 The Core Logic of Model-Driven Localization

Traditional localization methods rely on geometric algorithms or sensor fusion, requiring complex rule design. This system employs a data-driven backpropagation (BP) neural network model that takes the input feature vector $X = [\text{number of targets } n, \text{ four sets of colors, shape, center coordinates } X, \text{ center coordinates } Y, \text{ and circumscribed circle radius } R]$ and outputs $Y = [\text{actual coordinates } x, \text{ actual coordinates } y]$. The neural network automatically learns the nonlinear mapping relationship between features and actual coordinates, enabling end-to-end prediction of "feature input $\rightarrow$ coordinate output".

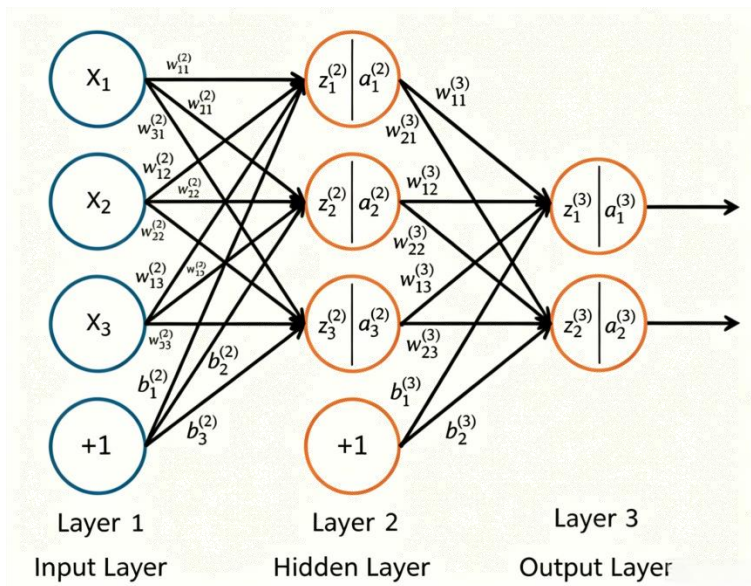


Fig. 2 Principle of the BP Neural Network

The construction process of the BP neural network constitutes the core of model-driven localization. The number of nodes in the input layer is determined to be 41 based on the input feature dimension, while the output layer has 2 nodes. By iteratively exploring different numbers of hidden layer nodes (from 4 to 13), the optimal network architecture is automatically identified using the mean squared error of the training set as the evaluation metric. Experimental results demonstrate that a hidden layer size of 16 achieves the lowest mean squared error of 0.010086, thus selecting this configuration as the final network model. Training was performed using the Levenberg-Marquardt algorithm (trainlm) with a maximum training iteration limit of 1000, a learning rate of 0.01, and an error tolerance of $1e-6$. During training, the model continuously adjusts weights and thresholds to minimize the discrepancy between predicted and actual values, progressively learning the mapping relationship between features and coordinates.

Input layer nodes: 41, Output layer nodes: 2

Hidden layer node range: 7 to 16

Determining the optimal hidden layer nodes:...

When the number of hidden layer nodes is 7, The training set mean squared error is: 0.031041

When the number of hidden layer nodes is 8, The training set mean squared error is: 0.012791

When the number of hidden layer nodes is 9, The training set mean squared error is: 0.018127

When the number of hidden layer nodes is 10, The training set mean squared error is: 0.015116

When the number of hidden layer nodes is 11, The training set mean squared error is: 0.017642

When the number of hidden layer nodes is 11, The training set mean squared error is: 0.017642

When the number of hidden layer nodes is 12, The training set mean squared error is: 0.013877

When the number of hidden layer nodes is 13, The training set mean squared error is: 0.027808

When the number of hidden layer nodes is 14, The training set mean squared error is: 0.031534

When the number of hidden layer nodes is 15, The training set mean squared error is: 0.022787

When the number of hidden layer nodes is 16, The training set mean squared error is: 0.010086

The optimal number of hidden layer nodes is: 16, mean squared error: 0.010086

Dimension of test Simu: 24 2

Dimension of outputtest: 24 2

Fig. 3. Process for determining the optimal hidden layer nodes

3.2 Model Evaluation and Test Results

To comprehensively evaluate the models performance, the system calculated multiple error metrics. The test set results showed a root mean square error of 0.79 cm and an average percentage error of 3%, achieving centimeter-level positioning accuracy. Error analysis revealed an error sum of 165.34, an average absolute error of 1.21 cm, and a mean square error of 0.62 cm². The correlation coefficient R values were 0.972 for the training set, 0.949 for the validation set, and 0.933 for the test set, indicating that the model demonstrates good data fitting performance with high correlation between predicted and actual values.

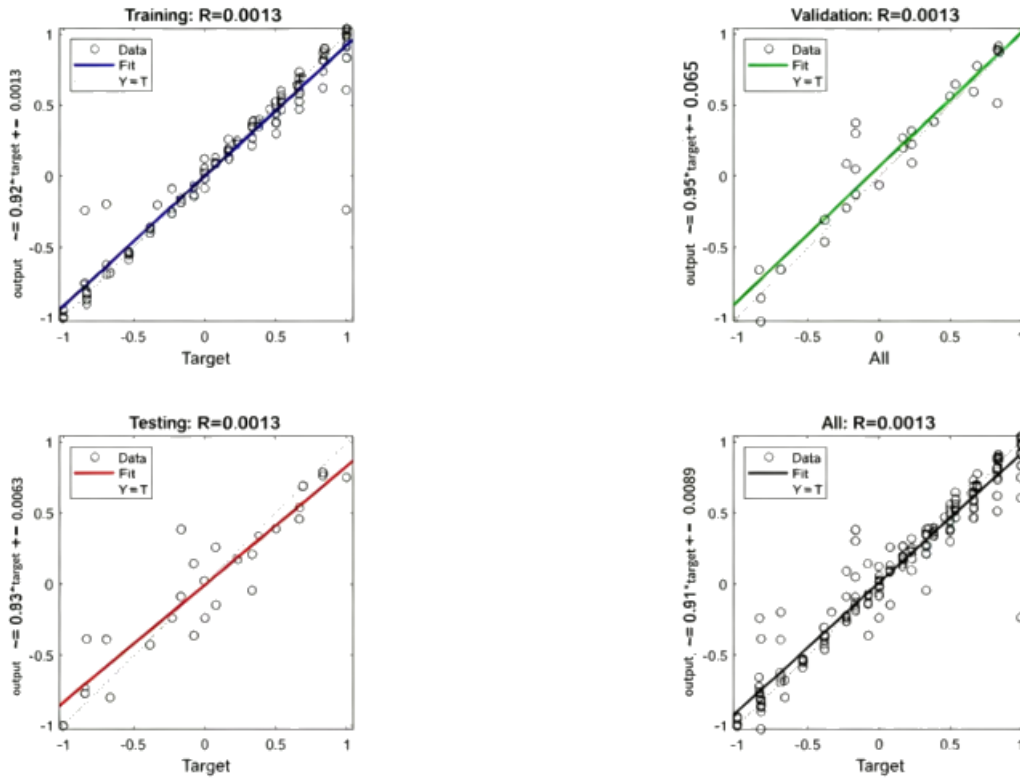


Fig. 4. Relationship between target values and model output values for the training set, validation set, test set, and the entire dataset

Using the actual test points (55 cm, 30 cm) as an example, the model predicted values were (59.13 cm, 28.80 cm), with an X-direction error of 4.13 cm and a Y-direction error of 1.20 cm, both within acceptable limits. Most test samples exhibited errors within ± 5 cm; larger errors in some samples may be attributed to insufficient coverage of training data or environmental noise interference. Key system performance metrics include: hardware standby power consumption below 200 mW, single-frame processing latency under 50 ms, color recognition accuracy of 95%, rectangular positioning error within 5 pixels, and circular detection accuracy of 92%.

4. Experimental Environment and Results

4.1 Experimental Environment Setup

This experiment was conducted in an indoor setting, focusing on establishing a comprehensive experimental environment that includes a mobile vehicle platform, designated target areas, and data

acquisition and processing terminals to ensure the authenticity, stability, and reproducibility of the experimental data.

Prepare two boards with coordinate paper printed on the reverse side, covering the expected movement range of the K210. Draw a 5 cm $\times$ 5 cm coordinate grid with the X-axis and Y-axis labeled accordingly. Manually move the K210 slowly and steadily to various positions on the coordinate paper. After each movement, wait for the K210 system to stabilize, then record the X and Y coordinates of its current position in an Excel sheet.

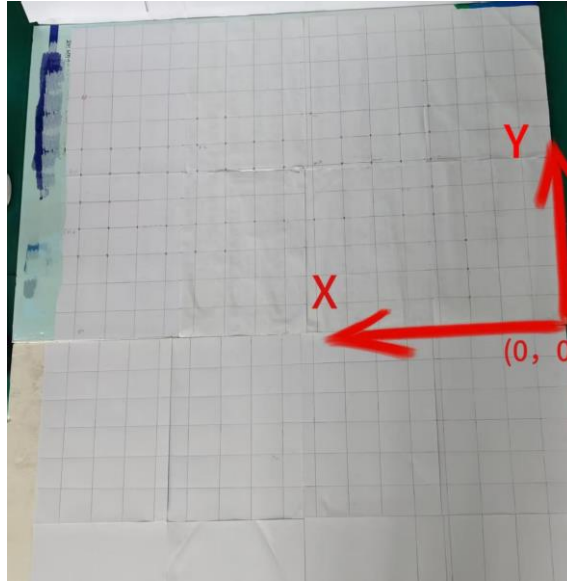


Fig. 5. Plotting coordinate paper

Use the OV2640 camera equipped with the K210 to capture images containing this image. To secure the K210, place it on a cart and ensure the camera remains stationary during acquisition. Use a single light source to provide uniform illumination for consistent image data. Position the left front wheel of the cart as the movement reference point, read the current data in CanMV, and record both the data and the actual position in an Excel sheet to complete the dataset collection.

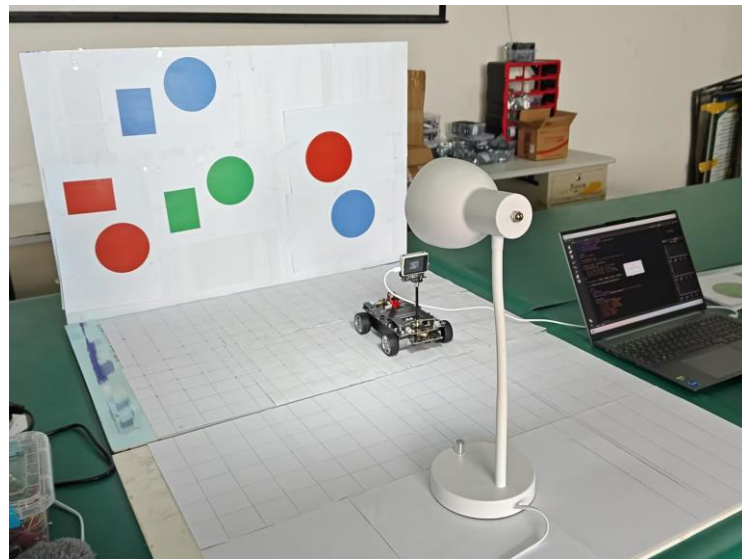


Fig. 6. Final site setup

The K210 terminal runs programs developed using CanMV IDE to perform image acquisition, target detection, and data transmission. By setting color thresholds and shape detection parameters, it accurately identifies circular and rectangular targets, collecting information such as target count, color, shape, and center coordinates into an Excel spreadsheet. The system also displays edge contours and the radius of the circumscribed circle.



Fig. 7. Display screen during image acquisition

5. Innovations and Future Prospects

The innovation of this system is primarily manifested in three key aspects. First, the multimodal feature fusion mechanism: The system integrates multi-dimensional data—including color, shape, and position—from optical imaging, enabling complementary enhancement of positioning information through collaborative processing. The combined application of the LAB color space and the Hough transform effectively addresses stability issues in feature extraction under complex lighting conditions. Second, the model-driven positioning method: Utilizing a BP neural network instead of traditional geometric algorithms, the system automatically learns the mapping relationship between features and coordinates from the data, eliminating the need for complex rule designs while optimizing network architecture through dynamic parameter tuning to enhance generalization capabilities. Third, low-power embedded implementation: Leveraging the dual-core RISC-V architecture of the K210 chip and the KPU hardware accelerator, the system achieves low-power operation while maintaining real-time performance, laying the foundation for deployment in portable devices.

Although the system has achieved satisfactory positioning performance, it still faces several challenges. Firstly, the differences in characteristics and scales among multimodal data increase the complexity of feature extraction and data integration. Secondly, the models adaptability to extreme lighting conditions or occlusion scenarios needs improvement, as evidenced by the significant errors in some test samples, indicating room for enhancing its generalization capability. Additionally, the coverage and diversity of the training dataset require further expansion.

Future research can be advanced in the following directions: First, strengthen interdisciplinary collaboration by integrating the latest advancements in computer vision and embedded systems to optimize multimodal fusion algorithms. Second, incorporate additional modal data such as inertial navigation and Bluetooth positioning to develop multi-source fusion positioning models that enhance system robustness. Third, explore lightweight neural network models to reduce computational overhead while maintaining accuracy, making the systems suitable for resource-constrained scenarios. Fourth, establish standardized datasets and testing protocols to promote the standardized development of indoor positioning technologies. With the deep integration of IoT and AI technologies, indoor positioning systems based on multimodal data fusion will play a more significant role in smart cities, intelligent security, healthcare, and wellness applications.

Acknowledgments

This paper is completed relying on the project Research on Key Technologies of Portable Indoor Positioning System Based on Image Situation Awareness (Project No. 202510702041). Authors sincerely thank my supervisor for the careful guidance during project research, system design and thesis writing, which helped me successfully complete the overall construction, algorithm optimization and experimental verification of the monocular vision visible light positioning system.

The future is long and promising. In the days ahead, the authors will further delve into professional fields, keep studying diligently, apply what we have learned into practice, and steadily improve my comprehensive capabilities.

References

- [1] Ke, Chenghu, Yuting Shu, Jingyuan Liang. "Research Progress of Indoor Visible Light Positioning". China Illuminating Engineering Journal 34.03 (2023): 80-89.
- [2] Daxiong, Bei. "Mutual Conversion Between RGB and Lab Color Spaces". CSDN Blog, 17 Aug. 2017, <https://blog.csdn.net/lz0499/article/details/77345166>.
- [3] AI, Research Community. "A Detailed Explanation of Neural Network BP Algorithm Principle and Python Implementation". AI Research Community, 28 Mar. 2018, <https://cloud.tencent.com/developer/article/1076834?shareByChannel=link>.
- [4] Neural Network and Deep Learning Matlab. "BP Neural Network Prediction and Regression MATLAB Code". CSDN Blog, 9 Sep. 2024, <https://blog.csdn.net/baoliang12345/article/details/128065676>.
- [5] Ji, Shengjie. Research on Visible Light Indoor Positioning Algorithm Based on Image Semantic Matching. Xi'an University of Technology, 2024.
- [6] Meng, Xiangyan, Xin Zhang, Feng Zhang. "Bio-inspired Network Integrating Transfer Feature Learning for Visible Light Imaging Indoor Position Sensing". Chinese Journal of Lasers 50.10 (2023): 176-183..
- [7] Quan J, Bai B, Jin S, et al. Indoor positioning modeling by visible light communication and imaging[J]. Chinese Optics Letters, 2014, 12(5): 052201.
- [8] Guan W, Huang L, Hussain B, et al. Robust robotic localization using visible light positioning and inertial fusion[J]. IEEE Systems Journal, 2022, 22(6): 4882-4892.
- [9] Chen C, et al. An Up-View Visual-Based Indoor Positioning Method via Deep Learning[J]. Remote Sensing, 2024, 16(6): 1024.