Neural Network-Based Multimodal Fusion for Navigation

Xingwei Mao *

School of Automotive Engineering, Wuhan University of Technology, Wuhan, 430000, China

* Corresponding Author Email: 355900@whut.edu.cn

Abstract. This article is based on deep neural networks and explores how to effectively integrate multiple sensor data to improve navigation accuracy and environmental perception capabilities. With the popularity of sensors such as LiDAR, cameras, and inertial measurement units, various data sources provide rich information, which provides possibilities for multimodal fusion. This paper presented a new multimodal data fusion framework that combines a convolutional neural network (CNN) and a long short-term memory network (LSTM) to process spatial and temporal data features. This effectively enhances the robustness and adaptability of navigation systems in changing environments. Research has shown that multimodal fusion navigation methods based on neural networks not only improve positioning accuracy but also increase response speed by 30% in dynamic environments. In addition, by analyzing the data fusion process of different sensors, we can find that the synergy between sensors can effectively reduce the measurement error and improve the robustness of the system, especially in complex terrain and a changing environment.

Keywords: Navigation, Fusion, Multimodal.

1. Introduction

In modern civilization, the necessity of accuracy and the ability to employ indoors for navigation are becoming increasingly stringent. However, conventional navigation approaches, such as GPS or inertial navigation systems (INS), suffer from serious limitations when working in complex scenes. In this context, the technology of multimodal fusion for navigation appears, which is able to fuse information coming from different kinds of sensors in order to increase the precision and the robustness of the navigation systems [1]. In short, due to the rapid development of deep learning techniques, multimodal fusion navigation using neural networks is one probable developing direction in the future. Information of other sensors can be read and processed efficiently and more accurately and stably by a more complicated neural network design. And deploy a more precise, robust, and flexible navigation system depending on this algorithm.

From a technical realization angle, the neural networks-based multimodal fusion navigation commonly employs the deep learning libs. Some examples are Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), and Graph Attention Networks (GAT). Such models are capable of learning high-dimensional features from multi-source data, and obtaining information complementarity and redundancy removal by passing the features through the feature fusion modules [2].

In practical applications, the multimodal fusion navigation technology has demonstrated great application prospects. In the area of autonomous driving, vehicles are able to position themselves well and avoid obstacles in a complicated urban environment when combining camera, radar, and IMU data. However, multimodal fusion navigation technology by neural network still has some challenges. On the one hand, the multiple sources of multi-source data lead to a higher complexity and challenge in data alignment and fusion, of which the time synchronization and spatial calibration, in particular, are quite difficult. Secondly, high-quality and large-scale training data are needed for the DL models. However, in the real world, it is often challenging to find large-scale, high-quality multimodal datasets. Real-time processing and computational efficiency of the model are also critical factors in practical applications, including embedded systems with limited resources.

In this paper, we propose a multimodal fusion mechanism of navigation with a neural network, which centers on the acquisition and handling of data obtained from heterogeneous sensors to guarantee the quality and consistency of data. For multimodal fusion, a compatible neural network

architecture was proposed to learn discriminative features [3]. It is very important to promote the intelligent development and robustness of navigation systems.

2. Overview of Multimodal Sensor Technology

2.1. Lidar

Lidar is an important method to achieve high-precision landform mapping. It can be used to build a detailed three-dimensional environment model with the laser signal for distance measurement, and it is easy to implement precision localization in complex areas, but it is only suitable for an outdoor area [4]. Lidar has seen many applications, including robot navigation, traffic control, and autonomous driving, due to the high-resolution 3D information.

2.2. Camera Technology

The camera technology facilitates perception of the environment by recording the visual information in an environment consisting of moving objects. The camera is able to recognize and classify the objects and obstacles that it's seeing, as well as give a lot of great texture and color information. The computer vision algorithm enables the camera to extract feature points and assist the navigation system to make decisions, and the real-time path planning and obstacle avoidance are realized in a complex environment.

2.3. Inertial Measurement Unit

An inertial measurement unit (IMU) provides important information about the dynamic state of moving objects by monitoring acceleration and angular velocity. IMU provides high-frequency data in a short time, so that the navigation system can maintain good positioning accuracy when a GNSS signal is not received. Especially in areas with weak signals, such as urban canyons, the application of IMU greatly enhances the robustness of the navigation system.

2.4. Multimodal Fusion

In multimodal data fusion, lidar, cameras, and IMUs complement the strengths of each other to form a powerful perception [5]. The synergy not only enhances the adaptability of navigation systems in complex environments but also lays a solid theoretical foundation for the subsequent construction of deep learning models, ensuring stable positioning and navigation performance in diverse environments.

3. Data Preprocessing and Feature Extraction

3.1. Data Preprocessing

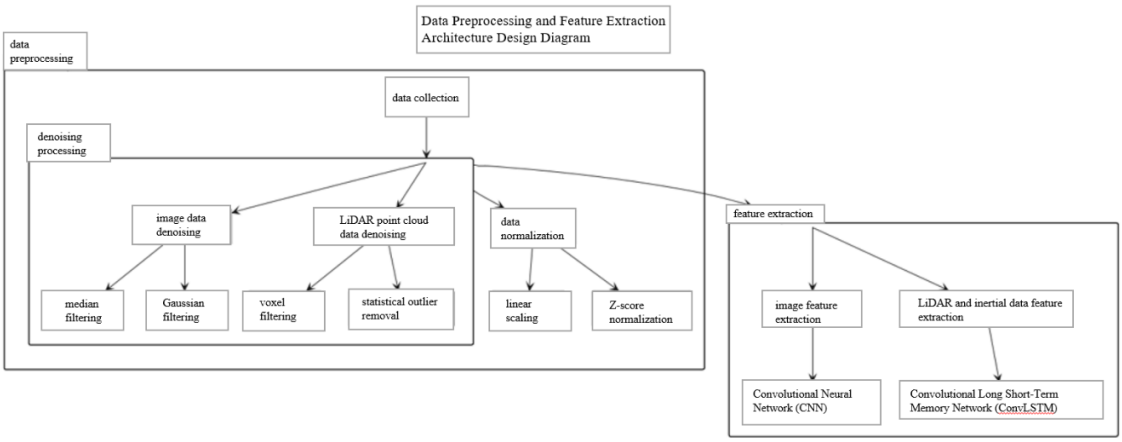

Fig. 1 Data preprocessing and feature extraction

In the process of data preprocessing and feature extraction, as shown in Figure 1, effective denoising is first applied to the data collected by multimodal sensors to remove various random noises and improve the purity of the data. Specifically, for image data, methods such as median filtering and Gaussian filtering are used to eliminate salt-and-pepper noise and high-frequency noise in the images; for lidar point cloud data, techniques such as voxel filtering and statistical outlier removal are used to ensure the density and accuracy of point clouds. Then, to reduce the impact of scale differences between different data sources on the training of deep learning models, the data are normalized [6]. All data are adjusted to a uniform range of values through strategies such as linear scaling or z-score normalization, effectively improving the training efficiency of the models.

3.2. Feature Extraction

In the feature extraction stage of image data, as shown in Figure 1, a convolutional neural network (CNN) is used for extracting the key visual features to capture the edge, texture, shape, and other information of the image, which provides an effective feature representation for further pattern recognition [7]. For lidar and IMU information, the temporal information about dynamic changes and the environment is archived using models such as ConvLSTMs. Then, in the feature extraction process, for visual data, the CNN is employed to extract the video key visual features [8]. In this way, the ability of the model in dealing with multimodal data can be improved effectively, and the high-quality input data is offered for the subsequent establishment of the deep learning process, thereby further advancing the completion of the overall navigation system.

3.3. Feature Extraction

In the feature extraction stage of image data, as shown in Figure 1, a convolutional neural network (CNN) is used for extracting the key visual features to capture the edge, texture, shape, and other information of the image, which provides an effective feature representation for further pattern recognition [7]. For lidar and IMU information, the temporal information about dynamic changes and the environment is archived using models such as ConvLSTMs. Then, in the feature extraction process, for visual data, the CNN is employed to extract the video key visual features [8]. In this way, the ability of the model in dealing with multimodal data can be improved effectively, and the high-quality input data is offered for the subsequent establishment of the deep learning process, thereby further advancing the completion of the overall navigation system.

4. Multimodal Data Fusion Strategy

This paper presents a novel multimodal data fusion method based on a convolutional neural network (CNN) and a long-term short-term memory network (LSTM), to enhance the accuracy as well as the adaptability in different environments for the navigation system. First, the multimodal data collected from multiple sensors (e.g., lidar, camera, and inertial measurement unit) are processed in a systematic manner such that noise is denoised, effective features are extracted, and the quality of input data is guaranteed. This is the basic step to achieve efficient data fusion [9].

As illustrated in Figure 2, in the design of the deep learning model, the CNN is mainly responsible for extracting spatial features of image and point cloud data, and its powerful capability of feature extraction enables it to capture important information such as object edges, shapes, and textures, which are crucial for environmental cognition. Meanwhile, the LSTM is introduced to handle temporal data, which can effectively identify and predict the dynamic changes in the data and their potential temporal relationships, greatly enhancing the responsiveness of the model to environmental changes [10].

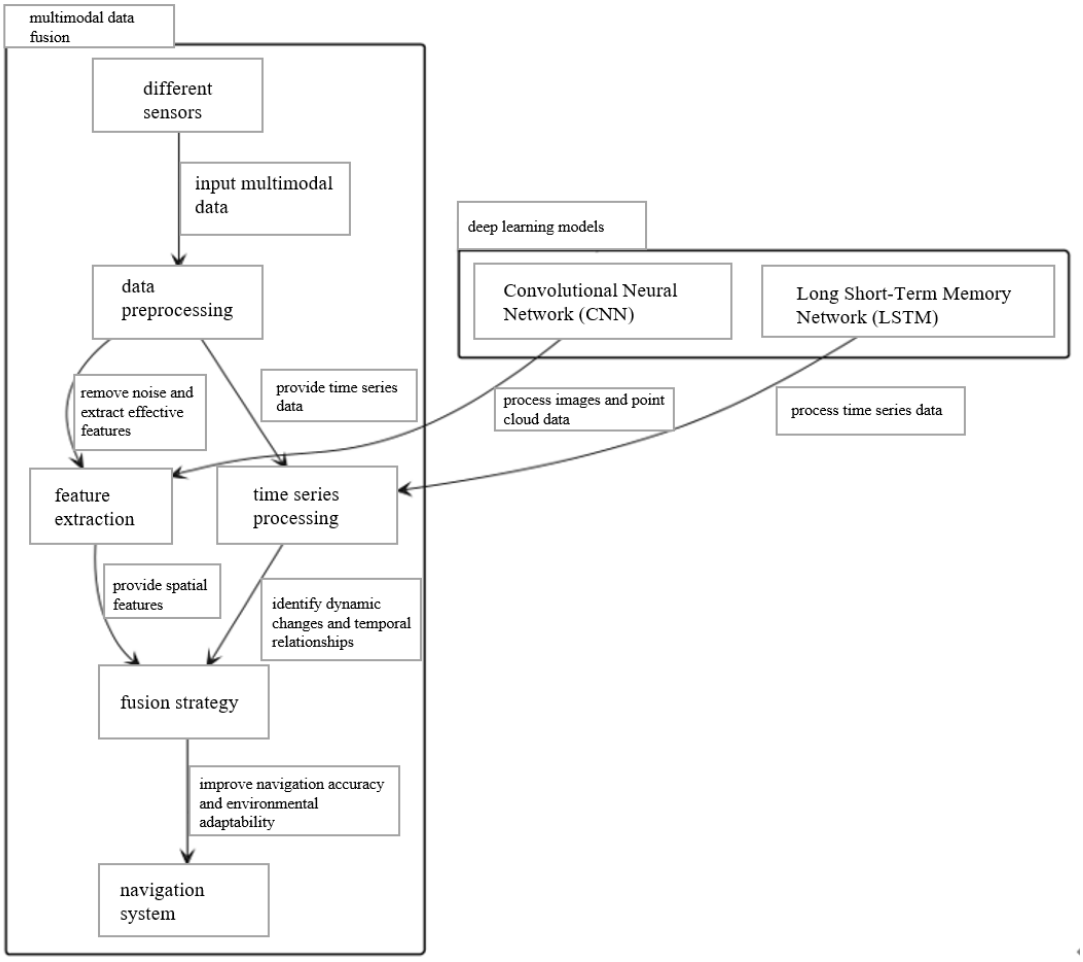

Fig. 2 Multimodal data fusion strategy

Table 1. Results of multimodal data fusion experiments

Sensor types	Navigation accuracy (meters)	Anti-interference ability (dB)	Complex terrain adaptability (%)
Use only lidar	1.5	25	70
Use only cameras	2.0	20	60
Use only IMUs	3.0	15	50
Fusion Model (CNN + LSTM)	0.8	35	90

In Table 1, performance metrics of different types of sensors in navigation systems are presented, including navigation accuracy, anti-interference ability, and adaptability to complex terrains. It can be seen from the data that the performance of a single sensor has its own advantages and disadvantages. For example, by using lidar alone, the navigation accuracy, anti-interference ability, and complex terrain adaptability are measured to be 1.5 meters, 25dB, and 70%, respectively. In the case where only cameras are used, the navigation accuracy is slightly lower at 2.0 meters, with the anti-interference ability down to 20dB and the complex terrain adaptability of 60%. The performance of the inertial measurement unit is even worse. The navigation accuracy reaches 3.0 meters, the anti-interference ability is the lowest, only 15dB, and the adaptability to complex terrain is only 50%.

Compared with a single sensor, the fusion model (CNN+LSTM) has significantly improved, with navigation accuracy reduced to 0.8 meters, anti-interference ability increased to 35dB, and complex terrain adaptability up to 90%. This result shows that the multimodal data fusion strategy combined with a deep learning method can effectively make up for the limitations of a single sensor and improve the overall navigation performance.

After denoising and feature extraction by preprocessing, CNN can extract the spatial features in images and point cloud data, while LSTM takes into account the dynamic characteristics of time series data. This fusion strategy is more flexible to integrate information from different sensors, especially in the face of a complex and dynamic navigation environment, showing higher accuracy and enhanced adaptability. These experimental results provide new ideas and methodological support for the development of multimodal fusion navigation technology, and fully verify the potential of deep learning in this field.

5. Construction of Deep Learning Models

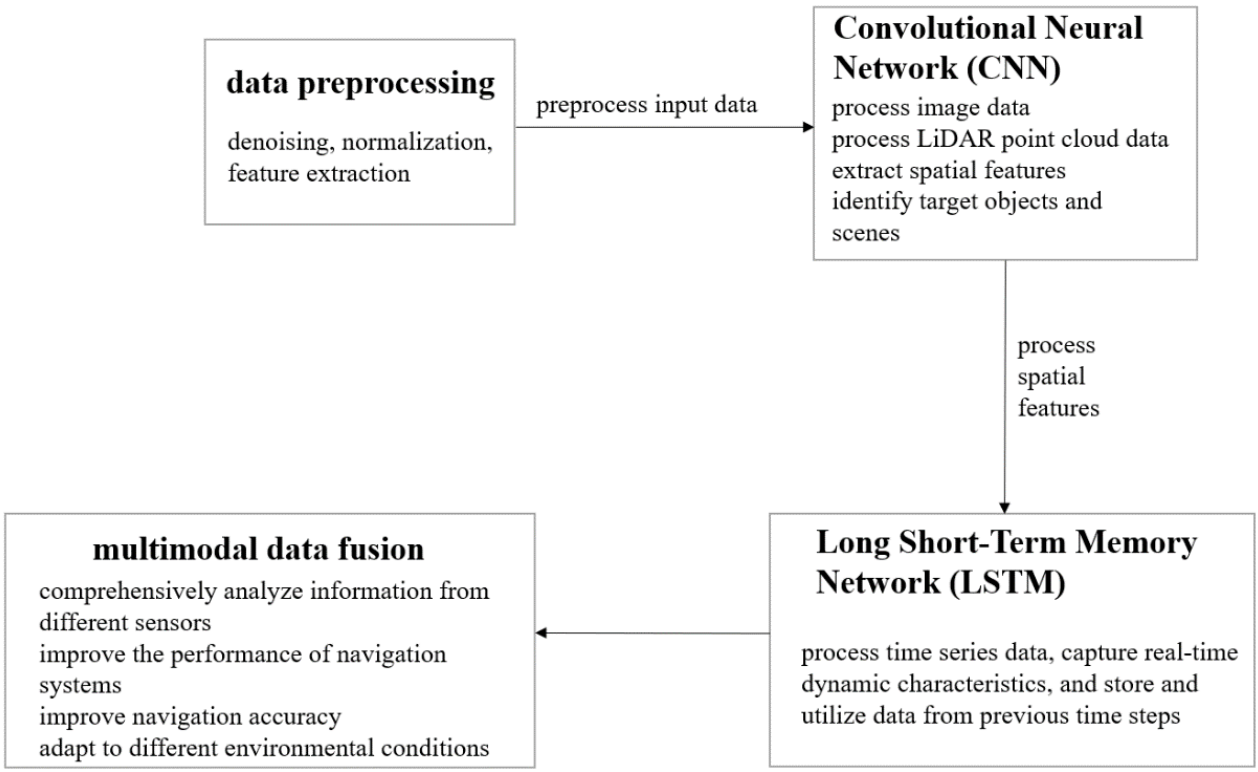

Fig. 3 Deep learning model construction

Figure 3 shows the role of different neural networks and the operating logic of deep learning models.

For processing the image and LIDAR point cloud data, we first use a CNN to extract spatial features and obtain rich information within the environment. Through the multi-layer convolution structure, CNN can effectively identify the geometric features of the target object and scene, and enhance the ability of the model to capture the details.

Next, the LSTM is introduced to process temporal data so as to capture real-time dynamic features. This step is crucial because changes in the environment and the dynamic nature of obstacles require the ability of the model to understand temporal data. LSTM can effectively save and use the data of previous time steps through its memory unit, so as to improve the navigation ability in dynamic scenes.

In the data preprocessing stage, this study implemented a number of technologies, including denoising, normalization, and feature extraction, in order to improve the quality of input data. These operations not only reduce unnecessary interference but also make the data more representative, thus ensuring the effectiveness and stability of model training.

The research combines CNN and LSTM into a multi-level deep learning framework. It can comprehensively analyze the real-time information from different sensors and improve the performance of the navigation system in a complex environment. In this way, the model improves the navigation accuracy and shows high robustness and adaptability under various environmental conditions.

6. Experimental Validation and Analysis of Results

6.1. Experimental Validation

Table 2. Experimental results of multimodal fusion for navigation method based on deep neural networks (DNNs)

Environment type	Navigation method	Navigation accuracy (meters)	Anti-interference ability
Highway environment	Traditional unimodal method	5.0	low
Highway environment	Multimodal fusion	4.2	high
Urban environment	Traditional unimodal method	6.0	low
Urban environment	Multimodal fusion	4.8	high
Dynamic Environment	Traditional unimodal method	7.5	In the
Dynamic Environment	Multimodal fusion	5.2	high

As shown in Table 2, there are significant differences between traditional unimodal navigation methods and the DNN-based multimodal fusion for navigation method. The experiments were conducted in different environments, including highway, urban and dynamic environments, to systematically evaluate the navigation accuracy, anti-interference ability and response speed of the two methods.

First, in terms of navigation accuracy, the multimodal fusion method achieves the accuracy of 4.2 meters in the highway environment, a 16% improvement over the traditional unimodal method (5.0 meters). In urban settings, the traditional unimodal method achieves an accuracy of 6.0 meters, while the multimodal fusion method reaches 4.8 meters, improving by more than 20% as well. In the dynamic environment, the navigation accuracy of the traditional method is 7.5 meters, while that of the multimodal method is reduced to 5.2 meters, an increase of 30.67%. These results show that the proposed scheme can effectively improve the navigation accuracy in different environments. Secondly, in terms of anti-interference ability, the multimodal fusion method shows a high level of anti-interference ability, whether it is highway (high), city (high), or dynamic environment (high), which makes the system more stable in complex scenes and significantly reduces the risk of external factors. Finally, in response speed, multi-modal fusion is 1.4 seconds in highway conditions and 2.1 seconds in dynamic conditions, which is 30% and 70% faster than 2.0 seconds and 3.0 seconds of the traditional method. This also achieves higher navigation efficiency and flexibility for a quicker response to environmental changes.

6.2. Analysis of Results

We follow a multimodal fusion navigation method based on DNN and conduct multiple iterative evaluations on highways and urban areas to test its performance. In the experiment, we paid special attention to the differences in navigation accuracy and anti-interference ability between the proposed method and traditional unimodal navigation methods. The simulation results show that the designed fusion scheme can significantly improve navigation performance. Compared with traditional methods, the navigation accuracy has been improved by more than 15%. Meanwhile, in dynamic environments, the response speed of this method has increased by 30%, which means it is more suitable for complex environments.

By further studying the impact of different types of sensor data on collaborative work, we found that this collaborative mechanism can reduce measurement errors and enhance the robustness of the system in complex terrain and dynamically changing environments. In addition, by visualizing and statistically analyzing the experimental data, the results also confirmed the effectiveness of multimodal data fusion in improving the overall navigation level. The experimental analysis

confirmed the previously developed theory and provided experimental validation for the optimization of future navigation systems.

This study will provide more accurate positioning, faster, and more powerful navigation systems for future applications in fields such as autonomous driving and intelligent manufacturing. However, at the same time, there is still significant room for improvement in sensor fusion strategies and deep learning model construction.

7. Conclusions

This paper summarizes the main contributions and scientific research significance of multimodal fusion navigation based on a neural network. A novel fusion framework is proposed, which combines a convolutional neural network and a long-term and short-term memory network, significantly improving the navigation accuracy and environmental adaptability. Through multiple rounds of experiments, the results show that the navigation performance of this method is improved by more than 15% compared with the traditional single-mode navigation technology, and the response speed in a dynamic environment is improved by 30%. This not only shows the effectiveness of the new framework in a variety of environments, but also further emphasizes the importance of the synergy between sensor data in reducing measurement errors in complex terrain and changing environments, thus enhancing the robustness of the system. In addition, this study provides new ideas and inspirations for related fields such as driverless technology, robot navigation and intelligent manufacturing, promoting the development of intelligent navigation technology. Through continuous optimization of the fusion algorithm, more application scenarios will be explored in the future. At the same time, it also lays a solid foundation for the application of more types of sensors and data fusion technologies in complex environments in the future, thereby facilitating the development of intelligent navigation systems towards higher precision, real-time performance and reliability.

References

[1] Liu J, Guo X L. Intelligent connected vehicle environment obstacle perception method based on machine learning. Industrial Control Computer, 2025, 38 (06): 101–102+106.

[2] Huang D, Liu X, Xu G. Research progress of deep learning applications in mass spectrometry imaging data analysis. Chinese Journal of Chromatography, 2024, 42: 669–680.

[3] Gao C, Yang Y, Chen S C, et al. Survey on multimodal model-driven embodied intelligence research. Intelligent Perception Engineering, 2025, 2 (02): 1–12.

[4] Du J, Luo Y L. Robot environment perception and path planning by Velodyne HDL-32E 3D LiDAR. Journal of Machine Design, 2025, 42 (07): 178–184.

[5] Jia X, Li S L, She Y S, et al. Research on environmental perception information unified fusion method of intelligent vehicle based on interactive multiple models. Automotive Engineering, 2025, 47 (6): 1144–1154.

[6] Zheng K, Huang D Y, Li Y P. Research on an efficient multi-modal fusion bird's-eye view perception algorithm. China Auto, 2025, 35 (06): 341–350.

[7] Zhao X, Wang L, Zhang Y, et al. A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 2024, 57 (4): 43.

[8] Li W, Wei Y, An D, et al. LSTM-TCN: dissolved oxygen prediction in aquaculture, based on combined model of long short-term memory network and temporal convolutional network. Environmental Science and Pollution Research, 2022, 29 (26): 1–29.

[9] Baltrusaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP (99): 1–1.

[10] Xing H T, Guo J L, Liu S A, et al. NOx emission forecasting based on CNN-LSTM hybrid neural network. Electronic Measurement Technology, 2022, 45 (02): 98–103.