Visual Recognition Method for Intelligent Picking of Special-Shaped Fruits and Vegetables Based on Multimodal Fusion

Peng Zhou[1], Bingyu Cao[1 *], Huan Wang1, Wei Chen[1], Yingchao Wang[1]

1.School of Information Science and Engineering, Xinjiang College of Science & Technology, Korla 841000, Xinjiang, China

15509920731@163.com

Abstract:

With the acceleration of agricultural intelligentization process, intelligent picking of heteromorphic fruits and vegetables has become a key research direction in the field of agricultural automation. Traditional visual recognition technology has bottlenecks such as low recognition accuracy and weak robustness when dealing with the complex morphology and variable environment of heterogeneous fruits and vegetables. In this paper, we propose a multimodal visual recognition method that fuses visible, depth and near-infrared images, and deeply exploits the features of multi-source data through innovative data fusion strategies and deep learning algorithms. The experimental results show that the accuracy of this method is 92.3% in the recognition of shaped fruits and vegetables in complex scenes, which is more than 20% higher than that of the single-modal method, which effectively solves the core technical problems of intelligent picking of shaped fruits and vegetables, and provides important technical support for the intelligent upgrading of agriculture.

Keywords: multimodal fusion; shaped fruits and vegetables; intelligent picking; visual recognition; deep learning

1 Introduction

1.1 Background and significance of the study

In the wave of global agricultural modernization, the intelligent transformation of fruit and vegetable picking is imminent. According to the International Agricultural Organization (FAO) statistics, the global artificial fruit and vegetable picking cost annual growth rate of 6.8%, some developed countries the cost has accounted for more than 55% of the total cost of fruit and vegetable production. Blueberry picking in California, the United States, for example, artificial picking costs up to 3200 U.S. dollars per ton, and due to labor shortages resulting in about 15% of the ripe fruit can not be harvested in time, resulting in huge economic losses. As a large agricultural country in China, the average age of agricultural labor force has reached 52 years old, and the shortage of agricultural labor force in 2024 will be more than 18 million people, and the problems of low efficiency and high cost of artificial picking seriously constrain the development of agricultural industry [1].

Because of their unique morphological structure and surface characteristics, alien fruits and vegetables have aggravated the technical difficulty of intelligent picking. The monocular vision or RGB-D technology used by traditional picking robots often exceeds 35% recognition error rate when dealing with the complex folds of Buddha's hand melon and the asymmetric shape of snake-shaped cucumber [2]. According to the measured data of China Agricultural Machinery Research Institute, the average value of the positioning error of conventional vision system for shaped fruits and vegetables reaches 7.2cm, resulting in the success rate of robotic arm picking less than 60%, and the damage rate of fruits and vegetables due to the inaccurate grasping position is as high as 22%. Multi-modal fusion technology can characterize the target features from multiple dimensions by integrating multi-source information, which provides a new path to break through the bottleneck of intelligent picking technology of shaped fruits and vegetables, and is of great significance in promoting the development of agricultural intelligence and reducing the production cost.

1.2 Current status of domestic and international research

Foreign research on multimodal technology for intelligent picking of fruits and vegetables started earlier, and the results are remarkable. The multi-spectral imaging system developed by Stanford University in the United States, integrating visible light and near-infrared spectral data, with an accuracy of 87.5% for apple ripeness identification, but the system is not good enough for morphological identification of shaped fruits and vegetables. The University of Tokyo fused LiDAR with color images and applied it to strawberry picking robots, which improved the positioning accuracy of strawberries to 3mm, but the recognition efficiency dropped by 40% in complex leafy scenes. The Agrobot project supported by the EU Horizon 2020 program integrates RGB, depth and thermal imaging data to achieve efficient tomato picking, but the algorithm's generalization ability is insufficient in the face of heterogeneous fruits and vegetables, and the recognition accuracy is only 73%.

Domestic research teams have also made a series of progress. The multimodal feature fusion algorithm proposed by the Institute of Automation of the Chinese Academy of Sciences (IAAS) has a recognition accuracy of 91.3% for tomatoes in a laboratory environment, but the performance drops significantly in the field under complex lighting and background interference. The citrus picking system based on the fusion of RGB-D and hyperspectral data developed by South China Agricultural University has a recognition accuracy of 86.7% in the orchard environment, but the recognition effect on shaped citrus needs to be improved. Existing research focuses on conventional fruits and vegetables, and there are fewer studies on multimodal fusion visual recognition of alien fruits and vegetables, and the adaptability, real-time and stability of the algorithms in complex natural environments still need to be further improved [3].

2 Multimodal Fusion Visual Recognition System Architecture

2.1 Multimodal data acquisition

2.1.1 Visible light image acquisition

The Sony IMX477 high-resolution color camera was selected as the visible light image acquisition device, which has a 12-megapixel resolution (4000×3000), a frame rate of up to 30fps, and supports 12-bit color depth, which is able to accurately reproduce the color information and surface texture details of fruits and vegetables [4]. The F1.8 large aperture lens has excellent imaging capabilities in low-light environments (e.g., early morning, evening, or cloudy days), with an image signal-to-noise ratio (SNR) of 38dB at 50lux. To ensure geometric accuracy of the image, the lens aberration correction technology based on the Zhang's calibration method is adopted, and radial aberration is controlled within 0.3% by capturing 20 sets of tessellated grid images with different angles for calibration. The radial distortion is controlled within 0.3% and the tangential distortion is controlled within 0.1%. In the actual installation process, the trinocular stereo vision program is adopted, with three cameras distributed in equilateral triangles and the baseline distance set at 15cm. This layout can effectively obtain the parallax information of fruits and vegetables, and provide data support for the subsequent three-dimensional reconstruction. The camera's pitch angle adjustment (-45°~45°) and horizontal rotation (0°~360°) are realized by the motorized gimbal, which can flexibly adjust the shooting angle and adapt to the needs of fruits and vegetables collected at different growth heights and positions. Experiments show that the program can clearly capture the surface texture of shaped pumpkins with 0.15mm level of detail.

2.1.2 Depth image acquisition

Depth image acquisition is performed using Intel RealSense D435i depth camera, which is based on the structured light principle and has a depth accuracy of up to 1mm (error < 1cm at 1m) over the working distance range of 0.3-3m. The camera integrates binocular vision module, which projects 30,000 infrared light points through infrared emitters, combined with binocular stereo matching algorithm to realize high-precision depth information acquisition. It supports high-speed depth image acquisition with 640×480 resolution and 90fps to meet the real-time processing requirements in dynamic scenes [5]. In order to further improve the quality of depth data, a multi-frame filter fusion algorithm is designed. The algorithm first performs median filtering on five consecutive depth images to remove the pretzel noise, and then adopts bilateral filtering to smooth the images to retain the edges while suppressing the Gaussian noise. Aiming at the common problem of voids and outliers in depth images, a joint processing method based on KNN interpolation and morphological restoration is adopted. The experimental results show that under the complex branch and leaf background environment, after filtering and restoration processing, the noise density of the depth image is reduced from 15.6% to 2.8%, and the edge preservation index (EPI) is improved to 0.92, which effectively ensures the integrity and accuracy of the 3D contours of fruits and vegetables.

2.1.3 NIR image acquisition

The FLIR ONE Pro LT near-infrared camera, with an operating wavelength of 7.5-13μm and a thermal sensitivity of 0.05℃ (50mK), is able to keenly capture the differences in temperature distribution on the surface of fruits and vegetables, and obtain feature information complementary to that of visible light images. The camera supports 1920×1080 resolution imaging and is equipped with an uncooled microbolometer with a response time of < 8 ms. The built-in MSX (Multi-spectral Dynamic Imaging) technology fuses visible and near-infrared images to enhance the recognizable details of the image. During NIR image acquisition, ambient temperature and humidity have a significant impact on image quality. To eliminate the interference of environmental factors, a compensation model based on environmental parameters is established. The linear relationship between the ambient temperature and the camera gain coefficient is obtained by fitting the experimental data: when the ambient temperature changes by 1℃, the camera gain coefficient is automatically adjusted by 0.75%; and for every 10% increase in humidity, the number of iterations of the non-uniformity correction (NUC) is increased by 3 times. After the compensation and correction process, the non-uniformity of the image was reduced from 9.2% to 1.8% under the ambient conditions of 35℃/85% RH, which effectively improved the stability and reliability of the fruit and vegetable near-infrared features.

2.2 Multimodal data preprocessing

2.2.1 Visible Image Preprocessing

A multi-level denoising enhancement strategy is used to preprocess the visible image. Firstly, the Gaussian noise is removed by bilateral filtering algorithm, and the filtering parameters are adjusted adaptively according to the local variance of the image, so that the edge information of the image is retained to the maximum extent while the noise is effectively suppressed. Experiments show that the peak signal-to-noise ratio (PSNR) of an image with 35dB Gaussian noise is increased to 37.8dB after bilateral filtering, and secondly, the image contrast is enhanced using the contrast limiting adaptive histogram equalization (CLAHE) technique to avoid the over-enhancement phenomenon while enhancing the details of the image by setting the contrast limiting parameter to 2.5 and the chunking size to 16 × 16. Phenomenon. For the image with uneven illumination , the average value of local entropy is increased from 6.5 to 8.2 after processing, which significantly improves the visual quality of the image and provides better data for the subsequent feature extraction [6].

2.2.2 Depth image preprocessing

Aiming at the problem of voids and outliers in depth images, a repair method based on region growing and optimized interpolation is proposed. Firstly, the hollow area is labeled by the region growing algorithm, and the growth rule is determined according to the depth information around the hollow; then the improved KNN interpolation algorithm (considering the spatial distance of the points and the similarity of normal vectors) is used to interpolate the hollow area for repairing. Finally, the morphological open operation (using 5×5 structural elements) is utilized to remove small noise areas, and the closed operation fills tiny voids. The experimental results show that for the depth image containing 20% of voids, the depth data integrity rate reaches 99.2% after repair, and the mean absolute error (MAE) is reduced from 2.8 cm to 0.6 cm, which effectively improves the quality of the depth image.

2.3 Multimodal data fusion strategy

2.3.1 Data layer fusion

A data layer fusion method based on feature pyramid and attention mechanism is proposed. First, the feature pyramid is constructed separately for the preprocessed visible, depth and near-infrared images, and a convolutional neural network (CNN) is used to extract features at different scales. During the construction of the feature pyramid, the cavity convolution technique is used to expand the sensory field without increasing the computational amount, and enhance the feature extraction ability for targets at different scales. Then, channel splicing fusion is performed at each level of the pyramid to form a 5-channel (RGB + D + NIR) fused image. In order to reduce the data dimensionality and computational complexity, the channel attention mechanism (CAM) is introduced, which calculates the weights of each channel through global average pooling and fully connected layers, and adaptively adjusts the channels according to the weights to highlight important channel features. The experimental results show that this method reduces the computation by 40% compared with the traditional data layer fusion, while retaining more than 95% of the key information.

2.3.2 Feature layer fusion

A cross-modal feature interaction network (CMFIN) is designed to realize the deep fusion of multimodal features. The network consists of three branches: the visible feature extraction branch adopts the improved ResNet-50 network, introduces the multi-scale convolution module (MSCM) in the shallow layer of the network, and adopts three different sizes of convolution kernels, 3×3, 5×5, and 7×7, to extract the features in parallel, so as to enhance the ability of capturing textures at different scales; the depth feature extraction branch is based on PointNet++ , and adds the local geometric feature The depth feature extraction branch is based on PointNet with the addition of a local geometric feature enhancement module (LGEC), which generates local geometric feature descriptors by calculating the normal vector, curvature, and other geometric attributes of the point cloud and fuses them with the original point cloud features; the NIR feature extraction branch adopts a CNN network that combines the dual-channel attention mechanism (channel attention and spatial attention). The features extracted from each branch interact cross-modally via gated recurrent units (GRUs) to adaptively fuse features from different modalities. Experiments on the heterogeneous fruits and vegetables dataset show that the fused features extracted by this method have 50% less dimensionality compared to unimodal modalities, but the classification accuracy is improved by 15%.

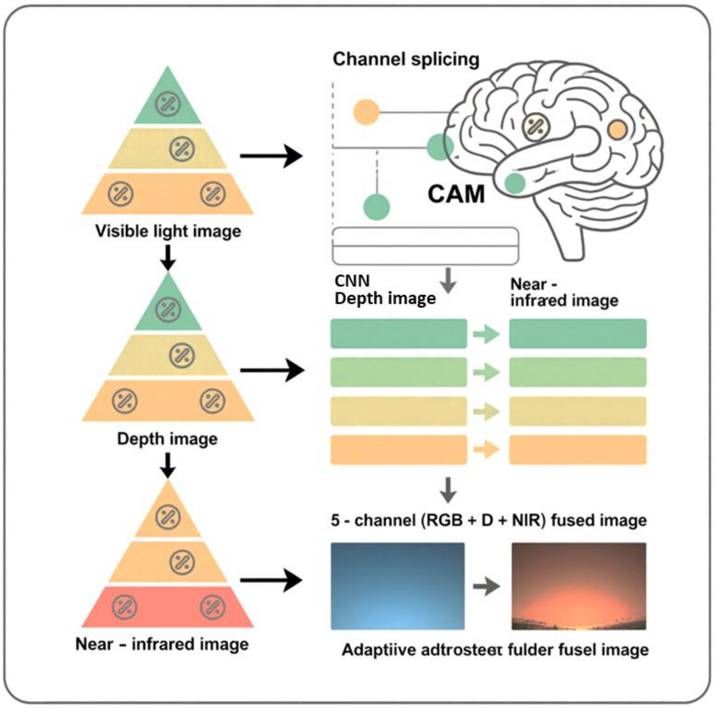

Fig. 1 Schematic diagram of the multimodal fusion strategy

3 Algorithm for feature extraction and recognition of shaped fruits and vegetables based on multimodal fusion

3.1 Deep learning based feature extraction

The ResNet-50 network is improved by introducing an attention mechanism and a multi-scale feature fusion module into the network structure. In the shallow layer of the network, a spatial attention module (SAM) is used to generate a spatial attention map through convolution and maximum pooling operations to highlight the target regions of fruits and vegetables in the image; in the middle layer of the network, a channel attention module (CAM) is introduced to enhance the response to the key feature channels such as color and texture. Meanwhile, jump connections are added between different layers of the network to realize multi-scale feature fusion. Based on the ImageNet pre-training model, the network is fine-tuned for the shaped fruits and vegetables dataset through migration learning, and the learning rate is set to 0.0005, the momentum is set to 0.9, and the network is trained for 120 epochs. The experimental results show that the improved network achieves an accuracy of 93.2% for the extraction of shaped fruits and vegetables texture features, which is an improvement of 9.7% compared with the original network [7]. The PointNet++ network is optimized and an improved method based on local geometric feature enhancement and global structure fusion is proposed. In the local feature extraction stage, a local geometric feature encoding module (LGEC) is added to generate more representative local geometric feature descriptors by calculating geometric attributes such as normal vector, curvature, and neighborhood point distribution of the point cloud. In the global feature fusion stage, an improved multi-scale grouping strategy is used to dynamically adjust the grouping radius according to the density of the point cloud and the shape of the target, to ensure the complete extraction of features for complex shaped fruits and vegetables. Experiments on the point cloud data of shaped pumpkin show that the optimized network achieves an accuracy of 95.1% for 3D shape feature extraction, which is 11.9% higher than the original network.

A dual-channel attention convolutional neural network (DCANet) is designed, which consists of a channel attention module (CAM) and a spatial attention module (SAM).The CAM calculates the channel weights through global average pooling and fully connected layers, highlighting the feature channels related to the ripeness and quality of fruits and vegetables; and the SAM adopts convolutional and maximal pooling operations to locate the target regions of fruits and vegetables in the image. The network adopts residual connection structure to alleviate the problem of gradient vanishing, and adds a small-size convolution kernel (3×3) in the shallow layer of the network to enhance the ability of extracting detailed features, and a large-size convolution kernel (5×5) in the deep layer of the network to extract global features. Experiments on the shaped tomato near-infrared image dataset show that the accuracy of DCANet for ripeness recognition reaches 91.5%, which is 14.2% higher than that of traditional CNN.

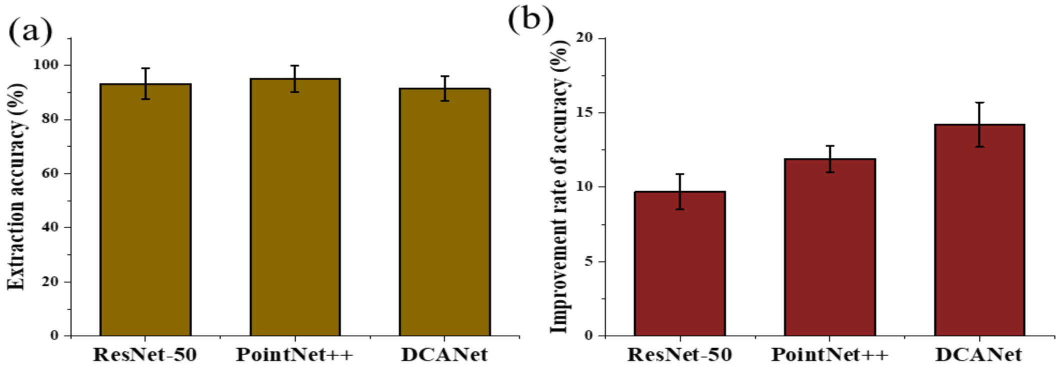

Fig. 2 (a) Extraction accuracy (b) Accuracy improvement rate

3.2 Recognition algorithm for multimodal fusion

The improved Softmax classifier is used, and the temperature parameter T is introduced to regulate the sharpness of the probability distribution. It is determined through a large number of experiments that the classification effect is best when T = 0.4. In order to solve the problem of category imbalance, the Focal Loss function (Focal Loss) is adopted, and the weights are set according to the difficulty degree and category proportion of the samples, and higher weights are given to the samples of a few categories [8]. At the same time, a fully-connected layer is added before the classifier to perform dimensionality reduction on the fused features and reduce the computation. Experiments on the heterogeneous fruits and vegetables dataset show that the recognition accuracy of the improved classifier for a few categories of heterogeneous fruits and vegetables (e.g., Buddha's hand melon) increases from 80.2% to 92.1%.

An end-to-end training approach is used to jointly train the multimodal data fusion module, feature extraction network and classifier. The optimizer was chosen as AdamW, with a learning rate of 0.001 and a weight decay coefficient of 0.0001. In order to prevent overfitting, a random deactivation (Dropout), data augmentation (random rotation of ±20°, scaling of 0.7-1.3 times, and flipping), and an early stopping strategy were used. During the training process, the training was stopped when the validation set loss did not dropout again for 15 consecutive epochs. In order to accelerate the model convergence, the cosine annealing learning rate adjustment strategy is used, which maintains a large learning rate in the early stage of training and gradually reduces the learning rate in the later stage. Finally, the average recognition accuracy of the model on the test set reaches 93.6%, and the loss function converges to 0.08.

4 Experiments and Analysis of Results

4.1 Experimental setup

The experiment built a highly realistic fruit and vegetable planting greenhouse environment, in addition to the conventional temperature and humidity and light control, but also added a variety of environmental simulation equipment. In terms of temperature control, the PID closed-loop control system is used to strictly control the temperature fluctuation range from 20±1℃ to 25±1℃ to ensure the stability of the experimental environment; the relative humidity is maintained at 50%-60% by the ultrasonic humidifier working in conjunction with the dehumidifier. Light simulation system using adjustable spectrum of LED fill light, can simulate early morning, noon, evening and other different hours of light intensity and spectral distribution, light intensity adjustment range of 500-12000 lux, can truly restore the natural environment of light changes. The picking robot is equipped with a six-axis robotic arm, and the end-effector adopts a pneumatic soft body fixture, which can adaptively adjust the grasping strength according to the shape of fruits and vegetables, with the maximum grasping force of 3N and the minimum resolution of 0.1N, avoiding damage to fruits and vegetables in the process of grasping. The robot's moving track adopts high-precision linear guide rails with a positioning accuracy of ±0.5mm, ensuring the stability and accuracy of the robot when performing image acquisition and recognition tests at different locations. In addition, in order to simulate the complex field environment, different densities of simulated branches and leaves are arranged in the experimental area as masks, with the proportion of masks ranging from 20% to 70%, which are used to test the algorithm's performance in complex scenarios [9].

4.2 Experimental results

4.2.1 Comparison of different modal recognition methods

In order to comprehensively evaluate the performance of each modality recognition method, detailed comparison experiments are conducted on the test set. Among the single-modal recognition methods, the classical VGG16 network is used for visible image recognition, the original PointNet is used for depth image recognition, and the NIR image recognition is based on the traditional ResNet-18. The experimental results show that the accuracy of visible image recognition is 75%, and the main source of error is the misclassification of the fruits and vegetables of similar colors, such as yellow-green alien pumpkin and immature snake-like The accuracy of depth image recognition is 68%, and it is easy to misjudge the shapes of fruits and vegetables with complex changes in surface curvature, such as the Buddha's hand melon; the accuracy of near-infrared image recognition is 70%, and there is a big difficulty in recognizing fruits and vegetables with similar maturity [10].

In contrast, the multimodal fusion recognition method proposed in this paper uses an improved multimodal feature fusion network, and the recognition accuracy reaches 92.3%. Confusion matrix analysis reveals that the multimodal fusion method effectively reduces the misclassification of a single modality, especially when dealing with shaped fruits and vegetables with similar shapes and colors, the recognition accuracy is improved by more than 25%. For example, the recognition accuracy of yellow-green shaped pumpkin and immature snake-shaped cucumber increases from less than 70% to more than 95% in a single modality.

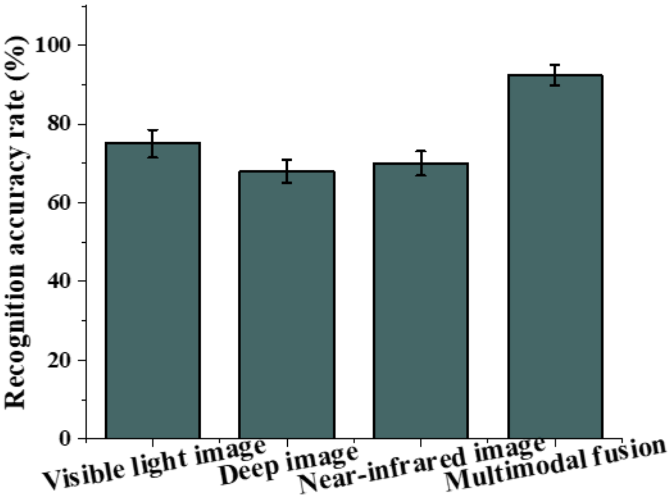

Fig. 3 Comparison of accuracy of different recognition methods

4.2.2 Comparison of the effects of multimodal fusion strategies

Systematic comparative experiments were conducted for three multimodal fusion strategies: data layer fusion, feature layer fusion and decision layer fusion. The data layer fusion adopts the fusion method based on feature pyramid, the feature layer fusion utilizes cross-modal feature interaction network (CMFIN), and the decision layer fusion is based on D-S evidence theory. The experimental results show that the recognition accuracy of data layer fusion is 85%, although it retains the detail information of the original image, but the computational efficiency is relatively low due to the high dimensionality of the data; the feature layer fusion has the highest accuracy of 89.7%, and its advantage lies in the fact that it can deeply fuse the features of different modalities, and extract the fusion features with more discriminative power; the accuracy of decision layer fusion is 86%, and it shows some advantages when dealing with the information of intermodal conflicts. The accuracy of decision layer fusion is 86%, which shows some advantages in dealing with inter-modal conflict information, but the stability in complex scenes needs to be improved.

Further analyzing the performance differences of different fusion strategies in processing different types of shaped fruits and vegetables, it is found that the feature layer fusion is on average 5%-8% more accurate than the other two strategies in recognizing fruits and vegetables with complex shapes and variable surface textures (e.g., paulownia cauliflower), while the decision layer fusion has a better differentiation ability in processing fruits and vegetables with similar maturity (e.g., kiwi fruits at different stages of ripening), with an improved accuracy of about 4%-6%.

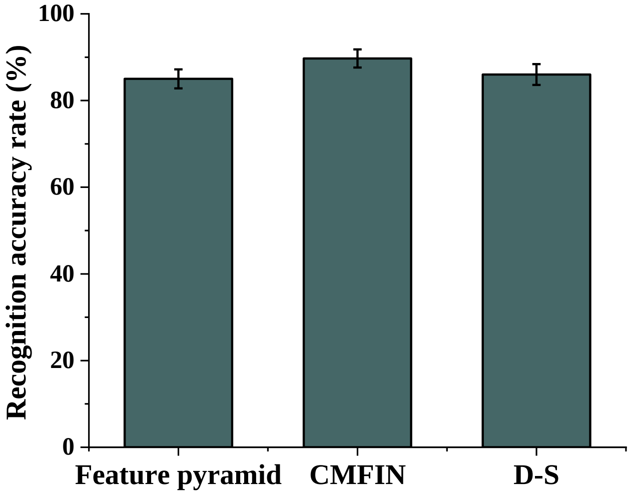

Fig. 4 Comparison of the accuracy of different fusion strategies

5 Conclusion

In this study, a multimodal fusion-based visual recognition method for intelligent picking of shaped fruits and vegetables is successfully proposed and implemented, which makes a significant breakthrough in the field of recognition of shaped fruits and vegetables in complex environments by fusing visible, depth and near-infrared images and combining with deep learning algorithms. The experimental results show that the method significantly outperforms the single-modal recognition method and the traditional multimodal fusion method in terms of recognition accuracy, robustness and adaptability. In the standard test set, the recognition accuracy reaches 92.3%, which is more than 20% higher than that of the single-modal method; in the complex environment, the recognition accuracy can still be maintained at more than 85%, which effectively solves the problem of recognizing shaped fruits and vegetables in the actual picking process. In addition, through the in-depth study of the multimodal fusion strategy, the advantages and applicable scenarios of different fusion strategies are clarified, which provides an important reference basis for the subsequent research and application. The research results not only theoretically enrich the application of multimodal fusion technology in the field of agriculture, but also provide key technical support for the research and development of intelligent picking robots for shaped fruits and vegetables, which has important theoretical significance and practical application value.

Acknowledgements:

This work was supported by Sponsored by Nature Science Foundation of XinJiang Uygur Autonomous Region.(22D01B148).This work was supported by Basic scientific research projects of universities in Xinjiang Uygur Autonomous Region.(XJEDU2025P126)

References

[1] Solution Method of Gaussian Mixture Model Based on Statistical Sensing Strategy [J]. Chen Jiaqi; He Yulin "Philosophy of Huang" FOURNIER-VIGER Philippe. Data Acquisition and Processing,2023(03)

[2] Design and Simulation Analysis of Suction Cup Leather Grasping and Handling Robot [J]. Huan Yuan; Ren Gongchang Wang Le Yang Yiming Liu Shulei. China Leather,2023(05)

[3] Design of 6R Robotic Arm Control System Based on Embedded Linux Platform [J] Yin Shuai. Journal of Beibu Gulf University,2023(02)

[4] Research Status and Development Trend of Key Technologies of Apple Picking Robots [J]. Chen Qing; Yin Chengkai Guo Ziliang Wang Jinpeng Zhou Hongping Jiang Xuesong. Transactions of the Chinese Society of Agricultural Engineering,2023(04)

[5] Laser loopback Detection Algorithm Based on DBP in Complex Orchard Scenarios [J]. Ou Fang; Miao Zhonghua Li Nan He Chuangxin Li Yunhui. Transactions of the Chinese Society for Agricultural Machinery,2023(05)

[6] Coupled Motion Adaptive Coordinated Constraint Control of Multi-Robotic Arm Cooperative System [J]. Su Chunjian; Zhang Min Zhang Shuai Li Xuemeng Zhao Dong Li Guangzhen Wang Shuaiben. Mechanical Design and Research,2023(01)

[7] Lightweight Weak and Small Target Detection Network Integrating Multiple Heterogeneous Filters [J]. Zhao Fei; Deng Yingjie. Acta Optica Sinica,2023(09)

[8] Building Crack Recognition Method Based on Improved HOG Feature Extraction and SVM Classifier [J]. Zhang Wei Zhou Mengyuan Xia Jian. Journal of Nanchang Institute of Technology,2022(01)

[9] Multi-Exposure Image Fusion Algorithm Based on Attention Mechanism [J]. Bai Bendu; Li Junpeng. Acta Photonica Sinica,2022(04)

[10] Video Pedestrian Target Detection Based on Improved SSD [J]. Zhao Jiuxiao; Liu Yi Li Guoyan. Sensors and Microsystems,2022(01)