Visual Image Feature Recognition Method for Mobile Robots Based on Machine Vision

Minghe Hu*, Jiancang He
School of Computing, Xinxiang Vocational and Technical College, Xinxiang, China

***Abstract*—With the continuous advancement of machine vision and computer technology, mobile robots with visual systems have received widespread attention in fields such as industry, agriculture, and services. However, the current methods for processing visual images of mobile robots are difficult to meet the requirements of practical applications. There are issues of low efficiency and low accuracy. Therefore, firstly, spatial information is integrated into the K-means algorithm and image spatial structure constraints are introduced for visual image segmentation. Then the dense connected network is added to the Convolutional neural network structure. This structure is combined with a bidirectional long-term and short-term memory network to achieve visual image feature recognition. The results show that the improved K-means algorithm has a maximum recall rate of 97.35% in the Berkeley image segmentation dataset, with a maximum Randall index of 86.18%. After combining with the proposed improved Convolutional neural network, the highest feature recognition rate for five scenes of mining, risk elimination, agriculture, factory and building is 96.1%, and the lowest error rate is 1.2%. It possesses a high degree of recognition accuracy and is capable of effectively being applied to visual feature recognition on mobile robots, providing a novel reference point for visual image processing on mobile robots.**

***Keywords*—Machine vision; mobile robots; image recognition; convolutional neural network; K-means algorithm**

I. Introduction

Mobile robots play an important role in modern technology, and their environmental perception and decision-making abilities are crucial for achieving autonomous navigation and task execution. Key technologies in a mobile robot's perception system are visual image processing and feature recognition, which provide robots with rich environmental information and accurate target recognition [1-3]. In recent years, the rapid development of deep learning technology has provided new solutions for visual image processing of mobile robots. Traditional Convolutional Neural Networks (CNN), as one of the core algorithms of deep learning, have strong feature extraction and recognition capabilities and have achieved outstanding results in fields such as image classification, object detection, and semantic segmentation. However, applying traditional methods to recognize visual images with mobile robots presents a challenge when using deep neural networks for image processing on the robot due to limited computational resources and power consumption [4-5]. Therefore, the integration of CNN technology with mobile robots for efficient image feature recognition has become a current research focus. Additionally, the K-means clustering algorithm serves as an unsupervised learning method that is widely used for clustering image features. By grouping image feature vectors, K-means aids in extracting key features of the image, ultimately resulting in image classification and target recognition. However, in the field of mobile robot vision, the application of machine vision is currently limited. Problems persist with low efficiency and accuracy in image feature recognition. In light of this, a study proposes a mobile robot visual image feature recognition method based on CNN and K-means technology. The method incorporates Recurrent Neural Network (RNN) to further improve the image recognition effect of mobile robots.

The paper is divided into four parts. The first part is an overview of the current development status of visual image recognition for mobile robots both domestically and internationally; the second part is to improve the image segmentation technology of K-means clustering and image feature recognition based on CNN and RNN, and constructs a robot image recognition system. The third part is the performance testing and application effectiveness of the system built by the study; the fourth part is a summary statement of the entire study.

II. Related Works

The CNN and K-means technologies' remarkable ability to identify image features has garnered significant interest from experts. In the context of mobile robot visual image feature identification, the aforementioned technologies are employed to boost accuracy. Jiang et al. proposed a casing infrared fault diagnosis method based on image segmentation and deep learning to effectively distinguish the fault area and background of the casing. During the process, a target detection system was constructed using the CNN framework, and K-means technology was introduced to classify and explore the positions and areas of the obtained images. The data indicates that the algorithm achieves an image classification accuracy rate of up to 98%, exhibiting superior performance [6]. Chen et al. developed a CNN and K-means-based method for scene perception to tackle CNN's low efficiency in different target recognition tasks. This method enabled coarse-grained classification of the input data and lowered complexity in structural design, boosting computational flexibility. Compared with traditional image recognition methods, this method improves accuracy by 36.65% [7]. To solve the time-consuming manual segmentation of brain tumors in magnetic resonance images, Ragupati and Karunakaran combined CNN with K-means to achieve a robust image feature segmentation method. CNN was used to classify images into normal and abnormal ones. Then K-means was used to segment brain tumor images from abnormal brain images. According to the findings, this method

has high accuracy and recognition efficiency [8]. The rapid development of highways and the increase in the number of vehicles require a safe and efficient transportation system for the automotive sector. Therefore, Chen and Zong proposed a license plate recognition model based on CNN-K-means. CNN was used for license plate detection and segmentation, followed by K-means for license plate number detection and segmentation, and finally for recognition. According to the findings, it is more effective and efficient than other models [9]. Rustam et al. proposed an image recognition method based on CNN-K-means to solve the low accuracy in lung cancer image diagnosis. All input data was checked through CNN, and image features were obtained through K-means and transmitted back to CNN for further recognition. The experimental results indicate that the highest performance measurement accuracy is 98.85%. The sensitivity is 98.32%, and the accuracy is 99.40%. This method has good results in lung detection [10].

With the development of society, the visual image feature recognition of mobile robots requires higher precision recognition capabilities. More scholars are researching how to enhance image recognition abilities. Liu et al. believed that estimating the three-dimensional position and direction of objects in the environment using a single RGB camera is very challenging. Therefore, a new neural network module was introduced for detecting three-dimensional objects to achieve the normal movement and operation of mobile robots. The outcomes indicated that mobile robots achieve the most advanced performance [11]. Sungheetha and Sharma found that in the process of 3D image convergence, the projection of different planes was incorrectly recognized in image feature recognition. A machine learning algorithm was proposed as a preprocessing step to enhance the speed and accuracy of data handling. The outcomes indicated that the accuracy is 34.9% higher than that of ordinary digital visual target recognition [12]. Wang et al. used tensor based visual feature recognition methods to identify visual information generated in industrial processes. The research results indicate that this method has good recognition accuracy [13]. Jacob and Darney used DL methods to study user privacy and secure image recognition for IoT management. The research results indicate good performance in improving the appropriateness and robustness of the Internet of Things. Simultaneously, DL greatly improves the accuracy of image feature recognition [14]. Niu et al. used the generating adversarial networks (GANs) to identify defect images in actual production lines. Classical methods face challenges in obtaining adequate defect datasets due to the lack of diversity and quantity. This method repairs defect images through a large number of defect free images on industrial sites. The research results indicate that this method has high accuracy in image recognition. The entire image recognition system has high robustness [15].

In summary, the method for recognizing visual image features in mobile robots using improved CNN and K-means technology has promising applications. The method enhances the environmental perception and target recognition abilities of mobile robots and offers significant support for achieving intelligent navigation and task execution. At present, numerous scholars have researched this problem, but only a few academic achievements have utilized K-means clustering CNN in recognizing image features, and there exist application limitations. As a result, this paper proposes a method for recognizing mobile robot images that combines CNN and K-means clustering. It aims to offer technical guidance for robot image recognition and to highlight its potential in future research and application.

III. Robot Vision Image Feature Recognition Based on Machine Vision

This chapter consists of two parts. The first section improves the K-means algorithm to complete visual image segmentation. In the second section, CNN and RNN are combined to carry out image feature recognition.

A. Image Segmentation Based on K-Means Clustering

K-means is one of unsupervised learning, which does not need to provide label information. It has a very concise objective function. The operation process of the K-means algorithm mainly contains four steps. The first is to randomly initialize the cluster center. This step randomly obtains K points corresponding to the center of each class of clusters. The second step is to calculate the distance from the sample point to the corresponding center. By comparison, the center with the smallest distance is selected. This means that the sample points and their corresponding centers are all of the same class of clusters [16-17]. Based on the clusters created in the second step, the third step recalculates the centers of all the clusters. The process of the second and third steps is then repeated until the termination conditions are met, concluding the algorithm. Fig. 1 illustrates the operation process of the K-means.

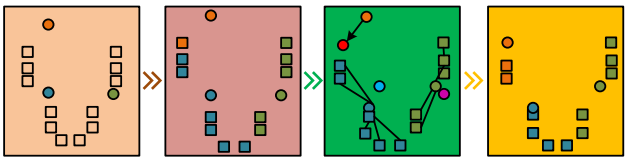


Fig. 1. Illustrate the operation process of the K-means.

If a dataset is $X=\{x_1,x_2,...,x_n\}$, the sample in that dataset is x_n. Usually, Euclidean distance is used to partition class clusters. The closest ones belong to the same cluster, while the points of different clusters are farther apart. The objective function of this algorithm is obtained as shown in Equation (1).

$$\min_{\gamma nk,\mu k}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma nk\left\|X_n-\mu_k\right\|^2 \quad (1)$$

In Equation (1), z is the samples in the dataset. E is the clusters. μ_k is the center of the k-th cluster. $\gamma_{nk}\in\{0,1\}$ represents whether the n-th sample point is in the k-th cluster. Among them, 0 indicates that the sample is not in the cluster, and 1 indicates that it is in the cluster. Meanwhile, each sample is only in a unique cluster. The K-means algorithm's objective function is to obtain the sum of squared errors. The goal is to minimize this error for each class cluster. This can achieve maximum compactness of cluster samples and ensure maximum distance between cluster samples. The Expectation–maximization is used to address the K-means algorithm. This algorithm is an efficient heuristic algorithm. It can ensure that

the algorithm converges to a local optimal solution in an extremely short time [18]. The Expectation–maximization algorithm is used to solve the objective function of K-means clustering. The first step is to take the derivative. The obtained content is shown in Equation (2).

$$-2\sum_{n1}^{N}\gamma_{nk}(x_n-\mu_k)=0 \tag{2}$$

The simplified expression of μ_k is shown in Equation (3).

$$\mu_k=\frac{\sum_{n=1}^{N}\gamma_{nk}x_n}{\sum_{n=1}^{N}\gamma_{nk}} \tag{3}$$

The specific values of all centers can be obtained through Equation (3). The cluster to which the sample point belongs can be reassigned, as shown in Equation (4).

$$\gamma_{nk}=\begin{cases}1, k=\arg\min_k\left\|x_n-\mu_k\right\|^2\\0\end{cases} \tag{4}$$

According to the solving process of the K-means in Equations (2) to (4), although it is difficult to ensure a global optimal solution, this algorithm can usually achieve the expected experimental objectives. When segmenting images with K-means, the clustering condition requires the presence of pixels with similar values. If such pixels are absent, the image belongs to a different cluster. Therefore, structural constraint information is relatively lacking. To address this issue, a K-means method for image spatial structure constraints was designed. This method adds image spatial structure constraints on the basis of the original algorithm. The color features and pixel position constraints in the image segmentation process are considered together to improve the segmentation effect [19]. The proposed K-means clustering model for image spatial structure constraints is shown in Fig. 2.

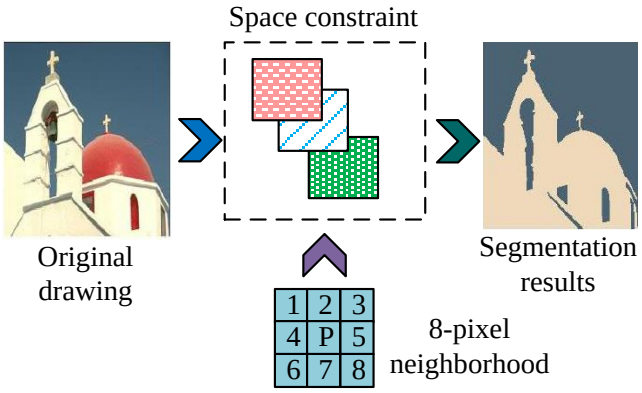


Fig. 2. Schematic diagram of the proposed K-means clustering model for image spatial structure constraints.

In Fig. 2, image segmentation is performed using the K-means algorithm, with sample points being the corresponding pixels in the color image. The clustering is completed under the specified number of clusters. Subsequently, the clustering center is used to replace the corresponding pixel points to achieve image reconstruction. Under the constraint of image spatial structure, K-means combines itself with spatial structure constraint information. The penalty constraint term corresponding to adjacent pixel points is added to the objective function. The new objective function is calculated, as shown in Equation (5).

$$\min_{\gamma_{nk},\mu k}\sum_{n=1}^{N}\sum_{k=1}^{K}(\gamma_{nk}\left\|x_n-\mu_k\right\|^2+\alpha\sum_{p=1}^{P}\left|\gamma_{nk}-\gamma_{pk}\right|) \tag{5}$$

In Equation (5), α greater than 0 is a hyperparameter. The main function is to balance the important relationship between the structural constraint term and the reconstruction error term. $KL(\bullet)$ is the neighborhood of the sample points, as shown in Fig. 3. It can be observed that for the internal, boundary, and corner pixels, the corresponding neighborhood pixels of the sample are 3, 5, and 8, respectively. The K-means algorithm for image spatial structure constraints adds the proposed spatial constraint term. As a result, similar color features can affect clustering performance. The position constraint relationship corresponding to adjacent pixel points is taken into account, greatly increasing the persuasiveness of image segmentation results.

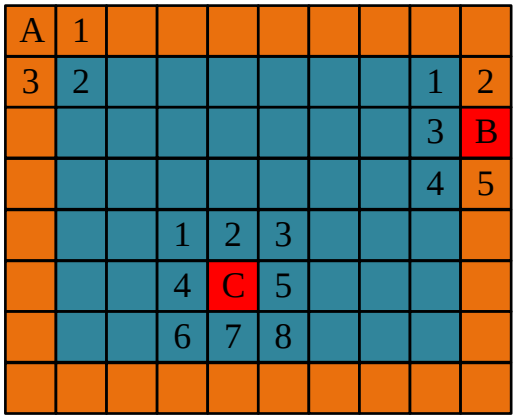


Fig. 3. Number of neighborhoods of sample points.

Then, the Expectation–maximization is used to address the proposed K-means objective function. The simplified expression of γ_{nk} is shown in Equation (6).

$$\gamma_{nk}=\begin{cases}1, k=\arg\min_k\left\|x_n-\mu_k\right\|^2+\alpha\sum_{p=1}^{P}\left|\gamma_{nk}-\gamma_{pk}\right|\\0\end{cases} \tag{6}$$

B. Image Feature Recognition Based on CNN and RNN

After completing image segmentation using the proposed K-means, CNN is further applied for image feature recognition. CNN is an adaptive abstract feature extraction model. The structure consists of an input layer, pooling layer, convolutional layer (CL), fully connected layer, and output layer. Among them, the pooling layer and CL connect adjacent nodes through sparse connections, which have the advantage of adaptive feature data extraction [13]. Fig. 4 illustrates the specific structure of CNN.

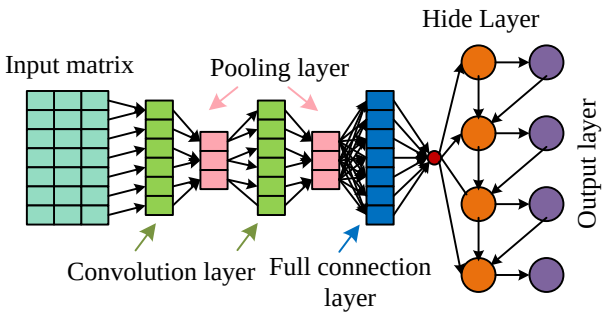


Fig. 4. CNN structure.

In the structure of CNN, the input layer receives preprocessed network data through appropriate dimensions. The convolution layer mainly performs convolution operations on the input values of the Receptive field. The pooling layer is responsible for dimensionality reduction of feature data and filtering out excess information. The fully connected layer combines the local abstract features extracted from the convolutional and pooling layers and reflects them into the label space. The output layer is responsible for outputting the prediction results of the network. Among them, the convolution operation is shown in Equation (7).

$$X_j^{(l)} = f\left(B_j^{(l)} + \sum_{i \in N_l} W_{ij}^{(l)} \times X_i^{(l-1)}\right) \quad (7)$$

In Equation (7), $X_j^{(l)}$ represents the j-th feature output of the l-th CL. N_l represents the set of inputs from layer l. $X_i^{(l-1)}$ refers to the data extracted by the convolutional kernel. $W_{ij}^{(l)}$ stands for the weight of the convolutional kernel. $B_j^{(l)}$ is the bias term. The increase of network layer in DL models is beneficial for feature extraction. However, excessive network layers can lead to overly complex model parameters, making it difficult for errors to be transmitted through gradient backpropagation. Therefore, based on the CNN structure, the DenseNet structural model is introduced. This model can extract abstract features at different levels and merge them. Feature information can be utilized to the greatest extent possible. At the same time, each CL has a fast channel connecting the input and output layers, making it easier for errors to update network parameters through gradient backpropagation. The DenseNet structural model is shown in Fig. 5.

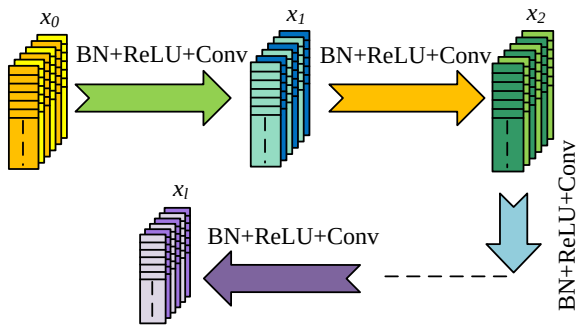


Fig. 5. DenseNet model structure.

In the DenseNet structure, when the CL is L, the connection channels between layers are L (L+1)/2. These channels are mainly used for the transmission of feature information. All previous layer output features are combined to complete each subsequent layer input. The calculation is shown in Equation (8).

$$x_l = H_l\left(\left[x_0, x_1, ..., x_{l-1}\right]\right) \quad (8)$$

In Equation (8), x_0 represents the input data of the first layer. $\left[x_0, x_1, ..., x_{l-1}\right]$ is the combination of output data from layer l to layer $l-1$. H_l refers to the feature extraction and transformation of the layer, including ReLU Activation function, standardization and convolution operations.

RNN is a special model with "memory" function. The nodes in its hidden layer can receive both the current input signal and the previous output signal. Therefore, the current state information is determined by the hidden layer node input and the previous node output. The calculation is shown in Equation (9).

$$p^{(n)} = q + Ux^{(n)} + Wh^{(n-1)} \quad (9)$$

In Equation (9), $p^{(n)}$ represents the intermediate variable. $h^{(n)}$ refers to the hidden state of node n. W and U represents the weights of the previous hidden state and the current input, respectively. $x^{(n)}$ is the current input information. q is the offset term for the hidden layer. The calculation of $h^{(n)}$ is shown in Equation (10).

$$h^{(n)} = f(p^{(n)}) \quad (10)$$

In Equation (10), f stands for the Activation function. The state information value of the hidden layer is shown in Equation (11).

$$o^{(n)} = c + Vh^{(n)} \quad (11)$$

In Equation (11), $o^{(n)}$ is the node output. V represents the weight of the output. c represents the bias term. The target value corresponding to $o^{(n)}$ is shown in Equation (12).

$$y^{(n)} = Soft\max(o^{(n)}) \quad (12)$$

In Equation (12), $y^{(n)}$ refers to the target value of $o^{(n)}$ mapped to the probability space through the Softmax function. RNN networks have significant advantages in processing sequence information. The traditional LSTM, in contrast, only facilitates one-way memory, relying on previous information to predict output results. It cannot meet the encoding requirements in reverse order. Therefore, a Bidirectional Long Short-term Memory (BLSTM) is introduced, which can model from front to back and from back to front, and output results based on contextual information. If the length of the input sequence is T, the structural model of BLSTM is shown in Fig. 6.

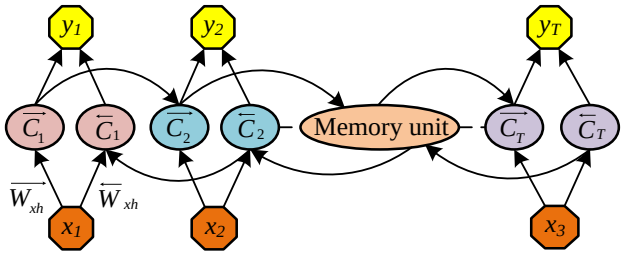


Fig. 6. BLSTM network structure.

In the BLSTM network, the memory units in the forward layer iterate from time step 1 to T, as shown in Equation (13).

$$\overrightarrow{C_n} = H(\overrightarrow{q_c} + \overrightarrow{W_C} g\overrightarrow{C_{n-1}} + \overrightarrow{W_x} g x_n) \quad (13)$$

In Equation (13), $\overrightarrow{C_{n-1}}$ and $\overrightarrow{C_n}$ respectively represent the previous and current state of the unit. x_n represents the current input data. The backward layer transmits information from the end to the front. The calculation is shown in Equation (14).

$$\overleftarrow{C_n} = H(\overleftarrow{q_c} + \overleftarrow{W_C} g\overleftarrow{C_{n+1}} + \overleftarrow{W_x} g x_n) \quad (14)$$

In Equation (14), $\overleftarrow{C_{n+1}}$ refers to the next state of the unit. The output is determined by the forward and backward cell propagation. The output is shown in Equation (15).

$$y_n = q_y + \overrightarrow{W_y} g\overrightarrow{C_n} + \overleftarrow{W_y} g\overleftarrow{C_n} \quad (15)$$

In Equation (15), y_n represents the output of time step n. $\overrightarrow{W_y}$ and $\overleftarrow{W_y}$ refer to the output weight. q_y represents the bias term. Compared to unidirectional RNN, BLSTM structure can extract more complete temporal features. DenseNet and BLSTM were further combined to construct the DenseNet BLSTM model. This model can adaptively extract multi-level features from signals. BLSTM is used to predict results from both the front and back directions to obtain more complete feature information. Simultaneously, it can maximize the utilization of feature information extracted from each layer of network.

IV. The Application Effect Analysis of Image Generation and Recognition

This chapter analyzes the application effect of the proposed method in visual image feature recognition of mobile robots. This includes improving the visual image segmentation performance of the K-means and the recognition performance of the DenseNet BLSTM model.

A. Visual Image Segmentation Effect

Firstly, the proposed K-means algorithm incorporating image spatial structure constraints is validated for the visual image segmentation effect of mobile robots. The dataset used in the experiment is the Berkeley image segmentation dataset, which includes benchmark codes, real human annotations, and 500 natural images. The proposed method is compared with classic Otsu algorithm, Laplacian of Gaussian (LOG), Normalized Cuts (Ncuts), NormTree, and traditional K-means algorithm. Five different evaluation indicators are used for quantitative evaluation of image segmentation effectiveness, i.e., F1 measure, accuracy, precision, recall, and Rand index (RI). All images are repeated 10 times in the Berkeley image dataset. Table I illustrates the results.

From Table I, in the comparison of F1 values, the NormTree algorithm is (56.28±13.22) %. The proposed K-means algorithm is (63.45±10.12) %, which is significantly superior to the other five methods. In the comparison of Recall, RI, and ACC, the proposed methods were (85.82±11.53) %, (74.85±11.33) %, and (61.47±12.58) %, respectively. All are the best of the five methods, indicating that they can effectively balance accuracy and recall. The comparison results obtained by all methods on different datasets are significantly different. This could stem from notable disparities amid the data samples within the two datasets, as well as the differing collection methods for the data in each dataset. This method has better performance. To further verify the performance, experiments are conducted again by changing the hyperparameter α. The F1-measures of the proposed method and K-means under different α conditions are shown in Fig. 7.

From Fig. 7, as the hyperparameter α increases, the F1 value of the traditional K-means does not change and remains stable at 0.55. For the proposed K-means, when α is 0, the F1 value is 0.55. As the value of α increases, the F1 value of this method also increases. When the value of α is 10, F1 reaches a maximum of 0.83. Then F1 descends. When the value of α is 20, the F1 value is 0.72, which is still significantly better than the K-means algorithm. The changes in ACC and RI indicators under different hyperparameters α are illustrated in Fig. 8.

TABLE I. Test Results of Six Methods in Berkeley Image Segmentation Dataset/%

Methods	F1-measure	Precision	Recall	RI	ACC
NormTree	56.28±13.22	45.17±15.36	72.0±17.06	62.62±8.44	58.75±11.74
LOG	47.37±14.27	35.28±16.26	81.28±4.43	42.54±11.36	46.58±15.32
Ncuts	51.76±12.57	72.12±10.04	48.47±21.07	72.73±14.07	54.57±12.59
Otsu	54.39±13.22	46.14±16.72	81.51±14.44	61.59±7.85	55.34±11.32
K-means	47.14±8.54	74.43±13.15	41.39±7.56	72.11±11.43	45.88±8.67
Ours	63.45±10.12	61.57±12.86	85.82±11.53	74.85±11.33	61.47±12.58

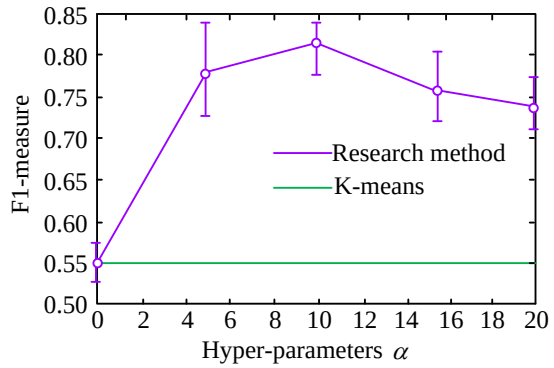


Fig. 7. The impact of different α value on F1 measure and its comparison with K-means clustering method.

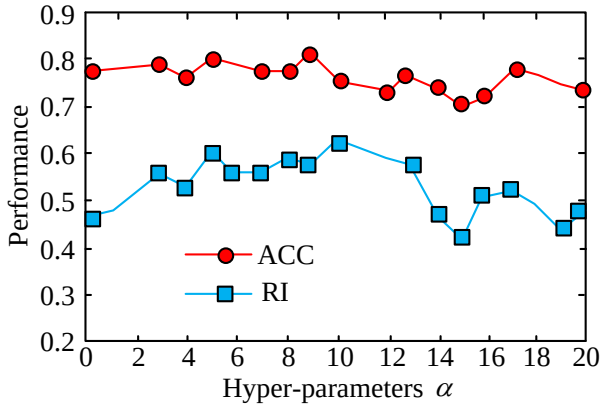


Fig. 8. The impact of different α values on accuracy (ACC) and Randall Index (RI).

From Fig. 8, with the continuous increase of hyperparameter α, the ACC and RI indicators of the proposed method have fluctuated to varying degrees. Among them, the ACC indicator curve has relatively small changes. Most of them are stable between 0.7 and 0.8. When the hyperparameter α values are 9 and 15, the maximum and minimum values of the ACC index appear, which are 0.82 and 0.69, respectively. In terms of RI indicators, when the hyperparameter α is less than 10, the overall trend shows an upward trend, reaching a maximum of 0.65. When the α exceeds 10, the overall RI index shows a downward trend. However, the minimum is maintained above 0.4, and the performance is still relatively good.

B. Analysis of Visual Image Feature Recognition Results

After verifying the proposed image segmentation method, the image feature recognition performance of the constructed DenseNet BLSTM model is further analyzed. The ImageNet dataset is selected for experiments. This dataset is a large visualization database used for visual object recognition software research, which contains more than 20000 categories. It has a large number of pictures. 200 images are randomly selected from four batches for image feature recognition, denoted as groups A, B, C, and D. The proposed DenseNet-BLSTM model is compared with four methods, namely, CNN, RNN, CRNN, and DenseNet. The obtained results are shown in Fig. 9.

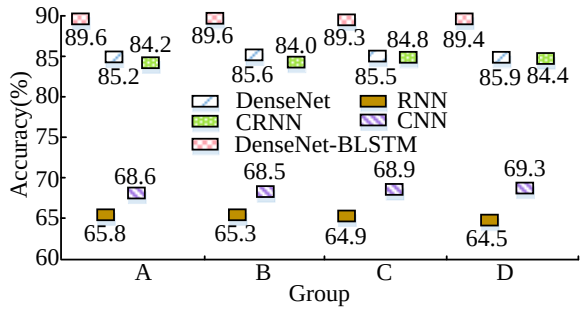
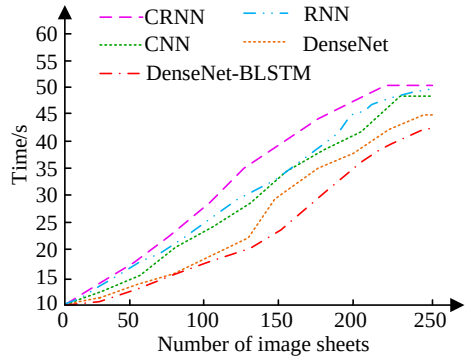
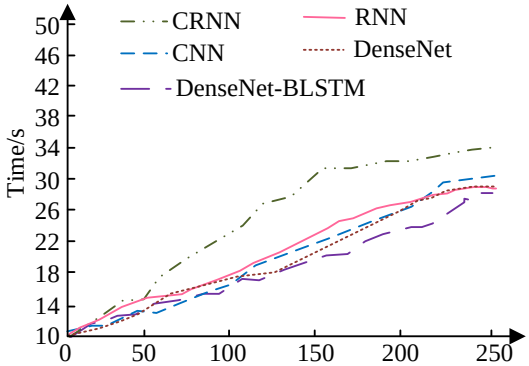


Fig. 9. Image feature recognition results of five models in ImageNet dataset.

From Fig. 9, in the four selected image feature recognition tests, the accuracy of the RNN model is 65.5%, 65.3%, 64.9%, and 64.5%, respectively, concentrated around 65%. The feature recognition accuracy of CNN is 68.6%, 68.5%, 68.9%, and 69.3%, all around 68%. The accuracy of CRNN and DenseNet models is relatively similar. The two fluctuate around 84% and 85% respectively. The four test results of the proposed DenseNet-BLSTM model are 89.6%, 89.6%, 89.3%, and 89.4%, all approaching 90%. Compared with the other four methods, this model has high accuracy and significant performance advantages. Further experiments are conducted on the efficiency of image feature recognition. The Pascalvoc2012 and Cityscapes datasets are used for testing. Among them, the proportion of training and testing sets is 70% and 30%. The recognition time changes of the five methods in the two selected datasets are shown in Fig. 10.



(a) Pascalvoc 2012



(b) Cityscapes

Fig. 10. Five methods for identifying time changes in the two selected datasets.

From Fig. 10(a), in the dataset Pascalvoc2012, as the images increases, the recognition time of all five models increases. Among them, CRNN takes the longest time. When the number of images reaches 250, the time consumption increases to around 50s. The time difference between CNN, RNN, and DenseNet models is relatively small, but they are better than the CRNN model. Among them, the DenseNet model takes the highest time of around 45s. The proposed DenseNet BLSTM model has a maximum duration of 40s, which is the shortest among the five models. From Fig. 10(b), in the dataset Cityscapes, CRNN takes up to 34s, which is the longest among the five methods. The proposed method still takes the shortest time, with a maximum of only 28s. The feature recognition efficiency is high and the performance advantage is significant. Finally, the improved K-means clustering combined with the DenseNet-BLSTM model is applied to the visual image feature recognition of mobile robots. To enhance the persuasiveness of the results, five main scenarios are selected for experiments, namely, mining, risk management, agriculture, factories, and construction. The image feature recognition results before and after the combinations are compared, as shown in Fig. 11.

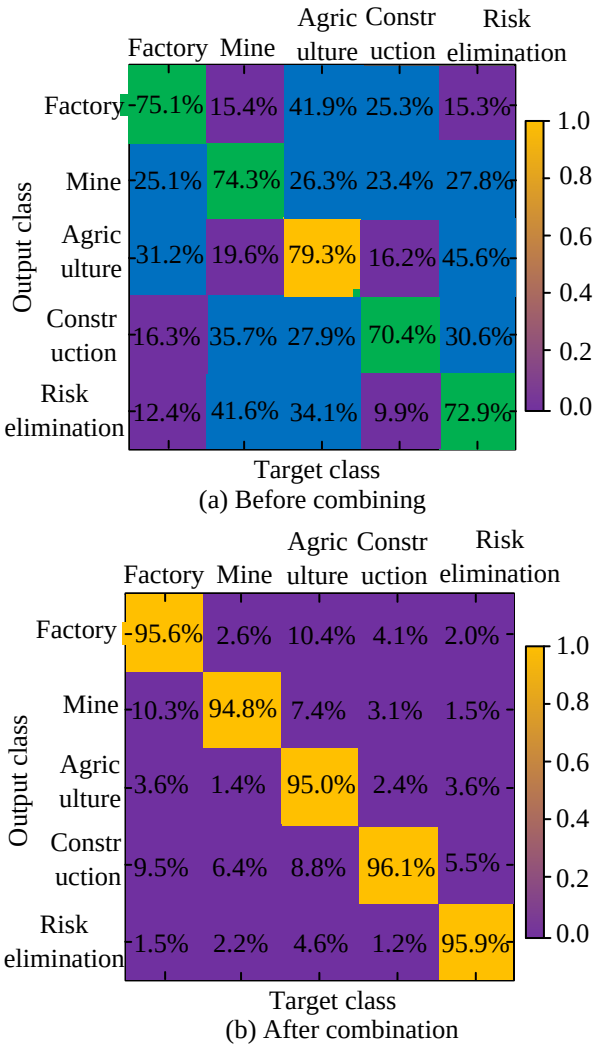


Fig. 11. The recognition results of image features before and after the combination of the proposed method.

From Fig. 11(a) in the image feature recognition of mining, risk management, agriculture, factories, and construction scenes, the combination of the previous method achieves the highest recognition accuracy of 79.3% and the lowest score of 70.4% occurs in agricultural and construction settings. Meanwhile, the lowest error is 9.9%, and the highest value is 45.6%, indicating poor classification performance. From Fig. 11(b), after the combination of the proposed method, the accuracy of image feature recognition for buildings is the highest, reaching 96.1%. The lowest value appears in mining scenarios, at 94.8%. Compared with the recognition results before the combination, the combined method has significant performance advantages in visual image feature recognition of mobile robots. The accuracy is between 94.8% and 96.1%. The application effect is better. The results above suggest that the study's method, consisting of CNN and clustering algorithms, is better suited for unsupervised learning tasks, large datasets, and image processing for data clustering.

V. Discussion

With the emergence of artificial intelligence and big data technology, an increasing amount of unlabeled data has become available. It is crucial to utilize the information within the data to explore its potential value. Unsupervised learning employs clustering algorithms to effectively address these challenges, with the K-means clustering algorithm being a classic example. Additionally, in real-world scenarios, data can possess unique structural constraints. Merely applying the K-means clustering algorithm to solve problems without accounting for the data's inherent features frequently results in suboptimal outcomes. Therefore, in response to this issue, we combine the K-means clustering algorithm with the structural characteristics of the data itself and introduce a combination of CNN and K-means clustering to develop a method for recognizing visual images in mobile robots. Throughout the entire experimental process, the study first utilized K-means to process spatial information and introduced image spatial structure constraints for visual image segmentation. Next, a densely connected network is added to the CNN, and combined with a bidirectional long short-term memory network to achieve recognition and segmentation of visual image features. The proposed models in the research have clear and objective functions, all solved using the maximum expectation algorithm. The effectiveness of the algorithm has been verified through a large number of experiments. The K-means clustering algorithm considers spatial constraints in images by integrating positional relationship information between adjacent pixels into the algorithm. This creates an overall framework with a unified objective function. The K-means clustering algorithm with hierarchical constraints achieves hierarchical clustering by establishing a hierarchical tree that can be globally iterated and updated, thereby better mining the hierarchical structure of the data itself. In addition, in complex and dynamically changing scenarios, deep learning methods introduce convolutional neural network models which demonstrate higher accuracy and robustness.

However, due to the continuous changes in environmental factors, regularly updating models has become the key to ensuring the long-term and efficient operation of mobile robots. Online learning and incremental learning technologies provide

effective methods for this. Overall, machine vision based image feature recognition for mobile robots is a versatile challenge that involves algorithm selection, computational efficiency, environmental factors, real-time requirements, and model training and updating. With the advancement of technology, this study aims to develop more efficient and robust methods in the future to meet the practical application needs of mobile robots in various environments.

VI. Conclusion

The improvement of visual image processing technology is an important foundation for the wider application of mobile robots. Firstly, an improved K-means algorithm using image spatial structure constraints is proposed. This method is applied to visual image segmentation. The increase in the layers in the CNN network results in complex parameters. Therefore, the DenseNet structural model is introduced and combined with BLSTM to achieve visual image feature extraction. According to the findings, in the comparison of F1 values, the NormTree algorithm is (56.28±13.22) %. The proposed K-means algorithm is (63.45±10.12) %, which is significantly superior to the other five methods. As the hyperparameter α increases, the F1 value of the traditional K-means algorithm stabilizes at 0.55. The proposed K-means algorithm has an F1 value of 0.55 when α is 0. As the value of α increases, the value of F1 also increases. It reaches the maximum of 0.83 when the α is 10. In terms of RI indicators, when the α is less than 10, the overall trend shows an upward trend, reaching a maximum of 0.65. When the α value of the hyperparameter exceeds 10, the overall RI index shows a downward trend. However, the minimum is above 0.4. The performance is still relatively ideal. The image feature recognition performance of the DenseNet BLSTM model constructed is analyzed. In the ImageNet dataset, the four test results are 89.6%, 89.6%, 89.3%, and 89.4%, all approaching 90%. In the dataset Pascalvoc2012, the maximum time consumption of this model is only 40s. In the image feature recognition of five scenes including mining, risk elimination, agriculture, factory and building, the accuracy of the recognition model combined with the improved K-means is between 94.8% and 96.1%, with a high accuracy. However, the proposed method has poor real-time performance. In the future, dimensionality reduction technology is needed to reduce computational complexity. The recognition speed of the algorithm will be further improved.

References

[1] P. Rosenberger, A. Cosgun, R. Newbury, J. Kwan., V. Ortenzi, P. Corke, and M. Grafinger, "Object-independent human-to-robot handovers using real time robotic vision", IEEE Robot. Autom. Lett., vol. 6, pp. 17-23, January 2021.

[2] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms", J. Comput. Cogn. Eng., vol. 2, pp. 5-9, April 2023.

[3] L. Jiang, W. Nie, J. Zhu, X. Gao, and B. Lei, "Lightweight object detection network model suitable for indoor mobile robots", J. Mech. Sci. Technol., vol. 36, pp. 907-920, February 2022.

[4] H. Kim, H. Kim, S. Lee, and H. Lee, "Autonomous exploration in a cluttered environment for a mobile robot with 2D-Map segmentation and object detection", IEEE Robot. Autom. Lett., vol. 7, pp. 6343-6350, July 2022.

[5] J. Zan, "Research on robot path perception and optimization technology based on whale optimization algorithm", J. Comput. Cogn. Eng., vol. 1, no. 4, pp. 201-208, March 2022.

[6] J. Jiang, Y. Bie, J. Li, X. Yang, G. Ma, Y. Lu, and C. Zhang, "Fault diagnosis of the bushing infrared images based on mask R-CNN and improved PCNN joint algorithm", High Voltage, vol. 6, pp. 116-124, December 2021.

[7] K. C. Chen, Y. W. Huang, G. M. Liu, J. W. Liang, Y. C. Yang, and Y. H. Liao, "A hierarchical k-means-assisted scenario-aware reconfigurable convolutional neural network", IEEE Trans. Very Large Scale Integrat. (VLSI) Syst., vol. 29, pp. 176-188, January 2021.

[8] B. Ragupathy and M. Karunakaran, "A deep learning model integrating convolution neural network and multiple kernel K means clustering for segmenting brain tumor in magnetic resonance images", Int. J. Imaging Syst. Technol., vol. 31, pp. 118-127, September 2021.

[9] D. J. I. Z. Chen, "Automatic vehicle license plate detection using K-means clustering algorithm and CNN", J. Electrical Eng. Autom., vol. 3, pp. 15-23, March 2021.

[10] Z. Rustam, S. Hartini, R. Y. Pratama, R. E. Yunus, and R. Hidayat, "Analysis of architecture combining convolutional neural network (CNN) and kernel K-means clustering for lung cancer diagnosis", Int. J. Adv. Sci. Eng. Inf. Technol, vol. 10, pp. 1200-1206, June 2020.

[11] Y. Liu, Y. Yixuan, and M. Liu, "Ground-aware monocular 3D object detection for autonomous driving", IEEE Robot. Autom. Lett., vol. 6, pp. 919-926, April 2021.

[12] A. Sungheetha and R. Sharma, "3D image processing using machine learning based input processing for man-machine interaction", J. Innov. Image Process. (JIIP), vol. 3, pp. 1-6, February 2021.

[13] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and M. J. Deen, "A tensor-based multiattributes visual feature recognition method for industrial intelligence", IEEE Trans. Ind. Inform., vol. 17, pp. 2231-2241, March 2021.

[14] I. J. Jacob and P. E.Darney, "Design of deep learning algorithm for IoT application by image based recognition", J. ISMAC, vol. 3, pp. 276-290, September 2021.

[15] S. Niu, B. Li, X. Wang, and H. Lin, "Defect image sample generation with GAN for improving defect recognition", IEEE Trans. Autom. Sci. Eng., vol. 17, pp. 1611-1622, July 2020.

[16] S. Oslund, C. Washington, A. So, T. Chen, and H. Ji, "Multiview robust adversarial stickers for arbitrary objects in the physical world", J. Comput. Cogn. Eng., vol. 1, pp. 152-158, September 2022.

[17] L. Wührl, C. Pylatiuk, M. Giersch, F. Lapp, T. von Rintelen, M. Balke, and R. Meier, "DiversityScanner: robotic handling of small invertebrates with machine learning methods", Mol. Ecol. Resour., vol. 22, pp. 1626-1638, May 2022.

[18] Y. Wang, Y. Liu, W. Feng, and S. Zeng, "Waste haven transfer and poverty-environment trap: evidence from EU", Green Low-Carbon Econ., vol. 1, pp. 41-49, February 2023.

[19] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition-A review", Inform. Process. Agr., vol. 8, pp. 27-51, February 2021.