Image Object and Scene Recognition Based on Improved Convolutional Neural Network

Guoyan Li
School of Information Engineering
Henan University of Animal Husbandry and Economy
China
81248@hnuahe.edu.cn

Fei Wang
School of Information Engineering
Henan University of Animal Husbandry and Economy
China
Fei_Wang2023@outlook.com

Abstract: *In recent years, due to the continuous optimization of network structure and the emergence of large-scale data, Convolutional neural network has made breakthroughs in a series of applications of computer vision. Based on this, the Convolutional neural network is improved and optimized. The improved convolutional neural network is introduced into image Object detection and scene recognition, and image object detection is carried out by combining sliding window fusion and Convolutional neural network. The image scene recognition model is constructed by using potential object area recognition and Convolutional neural network Transfer learning. Using different data sets to verify the algorithm, the research results show that in Group1 and Group2, the error rate of the multi column convolutional neural network fused by sliding window is reduced by about 25% compared with the single column convolutional neural network. As the group with the smallest decrease in error rate, Group3 also achieved a 9% decrease in error rate. The fitness rate of object detection algorithm is gradually stable after 7 runs, reaching about 9.8%, and its operation effect is obviously better than other algorithms. The multi column convolutional neural network fused by sliding window is more adaptive to the training data set, and gets better recognition effect in the algorithm operation. However, the image scene recognition model based on potential object area recognition algorithm and Convolutional neural network has good convergence. The average recognition time for image scenes is 1.5356s. The recognition speed is fast and stable, which can effectively solve the problem of multi-scale image scene recognition.*

Keywords: *Object recognition, deep learning, scene recognition, convolutional neural network, sliding window fusion.*

Received May 8, 2024; accepted September 2, 2024
https://doi.org/10.34028/iajit/21/5/13

1. Introduction

Nowadays, computer vision technology is becoming a key technology, which is also widely used in many industries. With the development of computer vision technology, feature learning methods based on Convolutional Neural Networks (CNN) have been widely applied in multiple fields of computer vision, such as face recognition, object detection, image classification, etc. For example, Ge *et al*. [12] studied the classification effect of CNN on hyperspectral images. By extracting image features and fusing them, several fully connected layers and a SoftMax layer were used to classify them and obtain the final result. This method has gradually replaced traditional features based on manual design and become a new research hotspot. Unlike artificially designed features, convolutional neural networks can automatically learn image features from massive image data after random initialization [6, 8]. With the development and iteration of technology, Multi column Convolutional Neural Networks (MCNN) have become the mainstream of computer vision technology today. Meanwhile, heterogeneous MCNN has also been widely applied. It can better capture the features of data, improve the accuracy and performance of the model. Based on this, this study aims to analyze the application of heterogeneous multi column CNNs in the research of key computer vision technologies [4, 27]. And from the perspective of network structure and loss function, this research proposes an algorithm combining Sliding Window Fusion (SWF) and CNN. Then, in response to some current problems, the two main structures of heterogeneous multi column convolutional neural networks were introduced and compared, and some relevant improvement suggestions were proposed. This is to improve the accuracy and performance of the model, and propose effective solutions for its advantages and disadvantages. In addition, on the basis of image Object detection, the research uses the potential object area recognition algorithm and Convolutional neural network to build a scene recognition model, hoping to further obtain the high-level semantic information of the image and improve the accuracy of image scene recognition.

2. Related Work

Many scholars have started to conduct in-depth research on computer vision. These studies involve the development of machine learning models to model applications, and provide corresponding research results. Maschler *et al*. [19] proposed an image recognition

method for deep industrial transfer learning at runtime. This method utilizes transfer learning technology and similarity detection algorithms for processing high-dimensional data to apply the algorithm to image recognition. This is to achieve higher accuracy. Bai *et al*. [4] introduced a dual discrimination autoencoder network for image zero-point recognition. The proposed method employs a series of deep neural network architectures, which integrate feature fusion, feature coding, and feature classification techniques, resulting in a significant improvement in the accuracy of image zero-point recognition. Daradkeh *et al*. [10] proposed a useful method for structural image recognition. This method combines the principles of data granularity and fuzzy logic to identify images by extracting structural and texture features. The characteristic of this method is that it can achieve image recognition under low computational load. And it has a high accuracy. Andriyanov *et al*. [3] explored the impact of visual attacks on neural networks in image recognition. The effect of visual interference on neural network in the process of image classification is studied, thereby affecting the accuracy of image recognition. They believe that visual attacks can alter the output by changing parameters and weights in the network, thereby undermining the accuracy and accuracy of the network model. Therefore, when designing neural network systems, special attention should be paid to the possibility of visual attacks. Effective defensive measures should be taken to prevent the occurrence of visual attacks. Hua *et al*. [14] proposed an artificial intelligence based gastroscopic image recognition model for the diagnosis of chronic atrophic gastritis. This model can effectively extract the structural features of gastroscopic images and effectively identify chronic atrophic gastritis in gastroscopic images. The experiment indicates that the model possesses a high accuracy and sensitivity in identifying chronic atrophic gastritis in gastroscopic images. It has a good clinical application prospect.

Long *et al*. [18] are based on PSO algorithm, combined with deep machine learning and PSO algorithm, to study a new intelligent education model. This mode can achieve intelligent teaching and management of image recognition and related tasks through image processing, preprocessing, and deep learning. Prajwalasimha *et al*. [22] proposed an iris image recognition method related to combined Hamming distance and cosine distance. This method can effectively extract features based on the structural features of iris images to achieve iris recognition.

Gautam *et al*. [11] proposed a novel coronavirus detection method based on machine learning and image recognition. This method can effectively detect novel coronavirus, especially in automatic detection. This method uses machine learning to identify coronavirus symptoms in lung images, and can effectively detect novel coronavirus related symptoms, such as pneumonia, nodules, etc.., Lee *et al*. [16] proposed a digital image recognition algorithm for maintenance data based on CNN and fully connected neural networks. This algorithm can effectively identify maintenance data, such as part numbers, customer information, and service times. This algorithm can effectively identify maintenance data and achieve good accuracy and performance. Kwon *et al*. [15] proposed a bumblebee image recognition method based on image fusion preprocessing and deep learning. This method can effectively extract the feature information of bumblebee images, and use this feature information for image classification and recognition. This method has high accuracy and short training time. It can be used for the recognition and classification of bumblebee images.

Domestic and foreign scholars have shown that computer vision technology will play a more important role in medical diagnosis, agricultural testing, industrial production control, and other fields. And it will also provide more convenient and efficient services for human society. Therefore, the research mainly proposes an algorithm combining SWF and CNN. This algorithm can improve the recognition rate and other issues existing in key technologies of computer visual recognition, thereby improving the efficiency of visual recognition.

3. Image Object Detection and Scene Recognition Technology Based on Improved Convolutional Neural Network

3.1. Construction of Recognition Model Based on Convolutional Neural Network Model

The conventional architecture of a CNN typically comprises several key components, including an input layer, one or more convolution layers, a pooling layer for down-sampling, a fully connected layer, and an output layer [9, 11, 17]. These components work together to enable feature extraction, parameter learning, and classification in a CNN. Such a design has become a standard framework for developing deep learning models in various computer vision tasks.

Figure 1 demonstrates that the input layer represents a node for receiving external input; The convolution layer represents a node that uses convolution checking to process input data to extract features. The pooling layer represents the use of pooling technology to sample features to reduce the amount of parameters, reduce the amount of computation, and enhance the robustness of features. The fully connected layer represents fully connecting features after the pooling layer to construct a multi-layer neural network. Down sampling represents the use of down sampling techniques to decrease the size of input data and decrease computational time before convolution. The output layer represents the nodes used to output the prediction results of the model. For input image X, a new method of image feature

extraction and dimensionality reduction based on CNN for convolutional and down sampling layers is proposed. In this study, $H_0=X$ is used to express the ^{i}-level characteristic graph of the network. If the ^{i} level is a convolutional layer, the H_i structure of the level is shown in Equation (1) [7].

$$H_i = f(W_{i\otimes}H_{i-1} + b_i) \tag{1}$$

Equation (1) indicates that W_i denotes the weight of the ^{i}-level convolution kernel number in the CNN; Research is conducted to convolution W_i onto the characteristic graph H_{i-1}("⊗") of layer i-1 of the network, and then add b_i to the offset vector of the current layer. Finally, the characteristic H_i of the current layer is obtained by excitation with a nonlinear function $f(x)$, as shown in Equation (2).

$$H_i = subsampling(H_{i-1}) \text{ SHAPE* MERGEFORMAT} \tag{2}$$

Equation (2) indicates that on multiple levels, the input image X is studied by performing convolution layers and down sampling layers on different levels, and then "dimensionality reduction" is performed on it. Finally, the study uses a fully connected network to classify the converted images and establish a corresponding relationship between the images and the typed probability distribution Y . The specific expression is in Equation (3).

$$Y(m) = P(L = l_m | \mathrm{H}_0; (W, b)) \tag{3}$$

Equation (3) indicates that m represents the index of the label category. Given the output of a model, the most commonly used loss function is cross entropy. By calculating the mean and standard deviation of the loss function, research can calculate the performance of the model. Therefore, the loss function can be estimated based on the maximum and minimum values. Among them, Negative Logarithmic Likelihood (NLL) and Mean Square Deviation (MSE) are two commonly used methods in the calculation of CNN sequences, which are expressed as Equations (4) and (5) respectively [5].

$$NLL(W, b) = -\sum_{m-1}^{|Y|} \log Y(m) \tag{4}$$

$$MSE(W, b) = \frac{1}{|Y|}\sum_{m=1}^{|Y|}(Y(m) - \hat{Y}(m))^2 \tag{5}$$

Super fitting is a common problem. The root cause is that this method has too many characteristics for the learned data sets, and the generalization performance is not ideal. All these have a certain impact on the feature extraction and determination of the sample. One way for CNN to alleviate over matching is to reduce network over connection or improve the random performance of the network [21]. On this basis, a (P_0, P_{n-1}) based adaptive learning method is proposed. This method improves existing learning methods and makes them have better learning abilities. Through different pre-processing, it provides a richer representation of the input image, which is conducive to the final determination of the image category by the network. Finally, the study uses the average values predicted by multiple neural networks to determine the target type. The specific situation is shown in Equation (6).

$$y^i_{MCDNN} = \frac{1}{N}\sum_{j=1}^{N} y^i_{DNN_j} \tag{6}$$

In Equation (6), y^i_{MCDNN} represents the possibility of the output of Class ^{i} in a multiline CNN; N denotes the number of convolutional networks; j denotes the subscript of the j th network in a multi column. Due to the significant differences in the identification capabilities between different networks, mixing these neural networks with the same weight will lead to a reduction in their identification and promotion capabilities. Therefore, studying a more flexible fusion method is essential [1, 23]. Therefore, the study intends to carry out research on data fusion methods based on sliding window multi row CNN, aiming at effectively processing multi row data. This is to achieve the fusion of multiple rows of data, thereby achieving effective data processing.

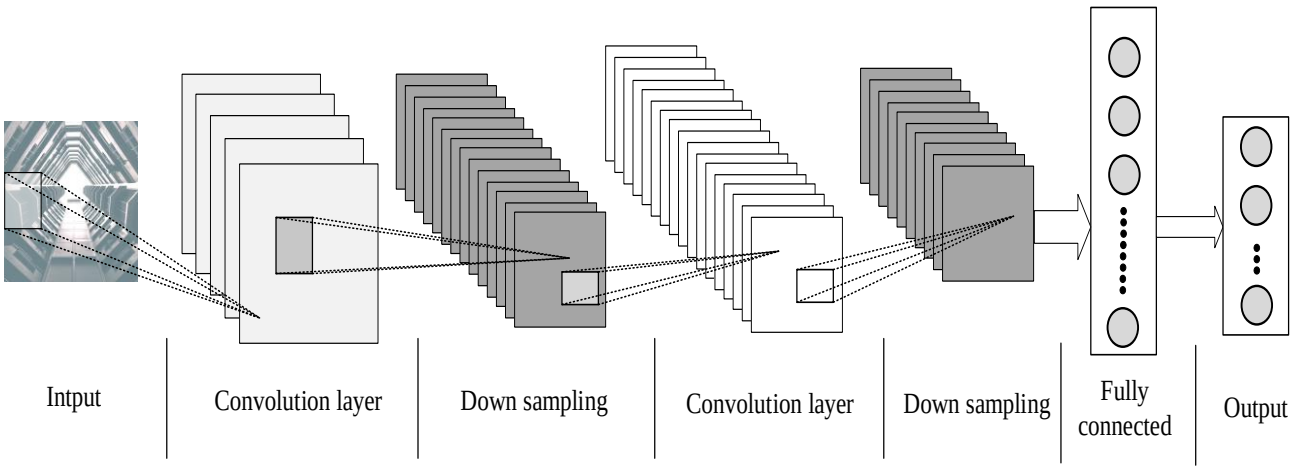


Figure 1. Typical structure of a convolutional neural network.

3.2. Construction of Object Detection Model Combining Sliding Window Fusion and Heterogeneous Multi Column Convolutional Neural Network

Heterogeneous MCNN consisting of multiple convolutional networks with different structures. This model's diagram is illustrated in Figure 2.

Figure 2 depicts the impact of folding cores on two aspects of pattern during the collapse process. The convolution core's size has a direct influence on both network size and performance. Additionally, convolution kernels' quantity affects the Eigen images' quantity. However, there is currently no established method for selecting the optimal convolution core size based on the number of feature graphs for a specific type of data. For solving this issue, a MCNN with various topological features is presented by varying the convolution kernel domain's size and the number of characteristic graphs. To enhance prediction accuracy, the outputs of multiple artificial neural networks are analyzed. An 8-cylindrical grid model is presented and an analysis of a model example is provided in this paper.

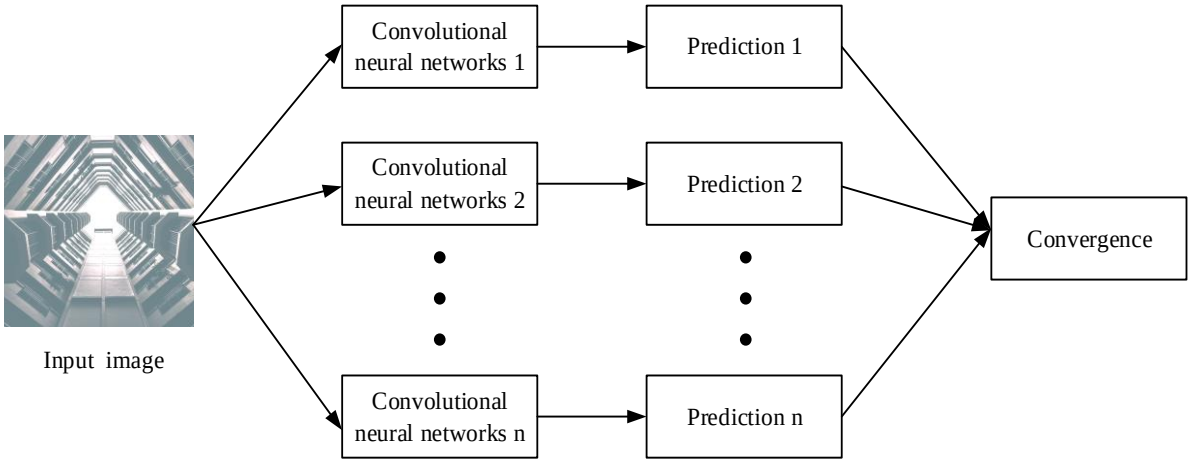


Figure 2. Heterogeneous multi-column convolutional neural networks.

Figure 3 illustrates the prediction probability of the CNN with Pr1-Pr8 for each row under a certain category after forward transmission through each row based on the input target image. First, the study ranked the probabilities of each factor in descending order, and then added a sliding window to the arranged prediction distribution. Finally, the predicted results are obtained. The work to be done for sliding windows includes the following two parts. First, it is essential to analyze the proposed predictions to determine which networks will be used for the final merge processing. This function depends on the start position of the window (head) and the selected area (area). Secondly, the selected predicted values are processed, and the corresponding data set is established, and through the processing of each data set, the corresponding data set is obtained. Sliding windows have optional areas and optional mixing methods. Therefore, in practical applications, sliding windows can better solve the mixing problem [20, 25, 26]. Therefore, fusing sliding Windows and constructing heterogeneous multi-column convolutional neural networks can increase the diversity of models and provide more refined fusion strategies, which is expected to improve the overall performance and robustness. There are two main reasons for this. First, by constructing heterogeneous multi-column convolutional neural networks, each column network has different structures and parameters, and can capture different features of input data, thus improving the expression and generalization ability of the model. This diversity helps to reduce overmatching, as different networks may have different tendencies for misclassification, and by fusing their outputs, these errors can be offset by each other. Second, the sliding window fusion method provides a flexible mechanism to integrate the prediction results of different networks. By adjusting the parameters of the sliding window, it is possible to dynamically select which network's prediction results are more valuable for the final decision. This method is more refined than simple average or Max/min rules because it takes into account the ordering and distribution of the predicted results, allowing for more efficient use of the strengths of each network. Therefore, a heterogeneous multi-column convolutional neural network fusion algorithm based on sliding window is proposed. The specific description of this algorithm is shown in Algorithm (1).

Algorithm 1: Sliding window fusion.

Description of the Sliding Window Fusion algorithm
Input:P: P(i,j) represents the prediction probability of the i column of the convolutional neural network for the j category;
h:Start position of the sliding window(head);
r:Range of sliding windows(range);
o:Operation of the sliding window(opratopn，can be set to 'sum' and 'product');
Output: T: T(i) represents the predicted probability of the j category after sliding window fusion
1. Number of columns of a convolutional neural network:m←size(P,1)
2. Number of predicted categories:n←size(P,2)

3. for j=1:n
4. Splice the predictions of a multi-column convolutional neural network into a vector: t←[P(1,j)...P(m,j)
5. Sort the elements in vector t:ts←sort(t)
6. switch o
7. case: o ==‘sum’，T(j)←accumulate r elements from ts(h) onwards
8. case: o ==‘product’，T(j)←cumulative multiplication of the next r elements starting from ts(h)
9. end
10. end

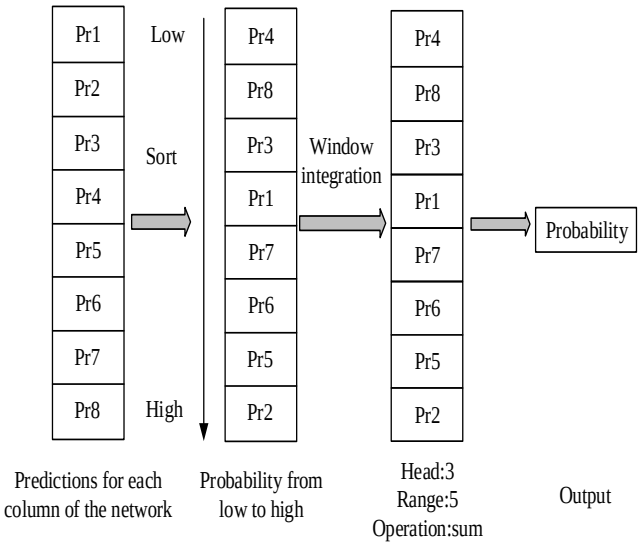


Figure 3. Sliding window fusion process.

Algorithm (1) shows that there are three parameters for sliding windows: “head”, “range”, and “operation”. Head and range determine the prediction probability of independent CNN. And it is used to calculate the overall forecast probability. The head represents the starting point of the possibility of being classified; “Range” refers to the calculation from this starting point to the last possibility. If the range extension exceeds the number of CNN, then the selection continues from the first prediction after the permutation. The parameter “operation” can be selected from Sum and Product. The prediction possibility selected by each sliding window is summed up in a straight line, and the final prediction value is obtained. The sorted prediction probability is calculated as shown in Equation (7).

$$y^{i}_{MCDNN} = \sum_{j=head}^{M} y^{i}_{sort-CNN_j} \tag{7}$$

Equation (7) indicates that y^{i}_{MCDNN} represents the prediction probability after ranking the prediction results. Therefore, a classifier mixing method based on SWF is proposed. For example, the parameter head and range of a sliding window are both 1, because equation $y^{i}_{sort-CNN_j}$ is the result of ranking the prediction probability from low to high. The specific situation is shown in Equation (8).

$$y^{i}_{MCDNN} = \min_{j\in[1,N]} y^{i}_{CNN_j} \tag{8}$$

SWF can evolve into traditional classifier fusion methods. Its core is to get the final prediction effect by comparing several basic learning algorithms. Details are listed in Table 1.

Table 1. Some special cases of sliding window fusion.

Head	Range	Operation	Integration methods
1	1	-	min rule
N	1	-	max rule
-	N	Sum	average rule
-	N	Prouduct	product rule

As shown in Table 1, two different training samples are used to train the training samples, and the best training samples are obtained. The maximum criterion refers to the optimal solution obtained by comparing the optimal solution of each basic learning algorithm with the final optimal solution. The average criterion is used to calculate the average score of each basic learning algorithm, and it is calculated as the final predicted score. The product method is the last operator obtained by multiplying each basic learning operator. In order to make the output of the model as close as possible to the actual label, the cross-entropy loss function $loss(x)$ is used for training, and the equation is shown in Equation (9).

$$loss(x) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln y_{ip} + (1-y_i)\ln(1-y_{ip})] \tag{9}$$

In Equation (9), y_i and y_{ip} denotes the label values and algorithm prediction values of image^i.

3.3. Scene Recognition Model Construction Based on Potential Object Region Recognition and Convolutional Neural Network Migration Learning

Image scene recognition requires acquiring scene feature information so as to extract high-level semantic information of the image to realize scene classification and judgment [18, 24]. Both scene recognition and object recognition use the mapping of image and category to obtain image information, but the features of recognized objects in scene recognition are more abstract and complex, including the feature information of multiple objects, so the object recognition technique cannot be directly used for image scene analysis [2, 13, 28]. Based on the object recognition model, the study proposes a multi-scale scene recognition model based on potential object region recognition and convolutional neural network migration learning for the scene recognition problem of images, and the scene recognition model is shown in Figure 4. The potential object region recognition algorithm is used to define and extract the significant regions of the image, obtain the main semantic information of the image, and use convolutional neural network to achieve scene feature representation and recognition.

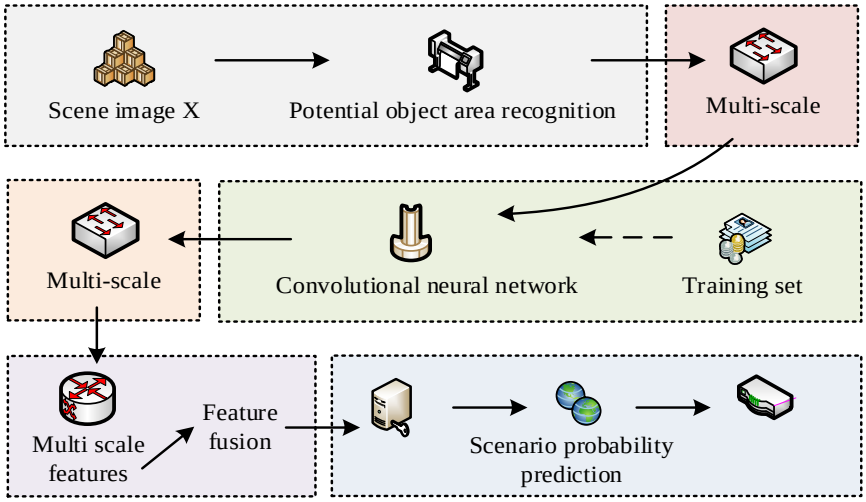


Figure 4. Multi scale scene recognition model based on potential object area recognition and CNN Transfer learning.

Initial image X Object density of potential areas in the scene M is calculated as shown in Equation (10).

$$\begin{cases} M(h,u)=\dfrac{\sum_{t=1}^{num(L)} g(X(h,u),L(t))}{num(L)} \\ g(X(h,u),L(t))=\begin{cases}1, X(h,u)\in L(t)\\ 0, X(h,u)\notin L(t)\end{cases}\end{cases} \quad (10)$$

In Equation (10), h and u denote the coordinate indexes of the image pixel points, L denotes the potential object regions, and t is the potential region index. Combined with the object density in the image scene, the sliding window is used to count the object distribution density in different areas of the image at a limited scale, and the area with the highest density is selected as the significant area of the image, and the formula is shown in Equation (11).

$$a_{\max}=\arg\max_{a\in[1,num(B)]}\psi(M,B(a)) \quad (11)$$

In Equation (11), a indicates the index of sliding window, $\psi(M,B,(a))$ indicates the calculation of the density sum of object distribution in the sliding window, $B(a)$ indicates the sliding window area, and $\psi(M,B,(a))$ is calculated as shown in Equation (12).

$$\psi(M,B(a))=\sum_{(h,u)\in B(a)} M(h,u) \quad (12)$$

The region with the highest object density in the sliding window region $B(a_{\max})$ is solved and taken as the salient region of the image, and the image scene feature information is expressed through the salient region feature extraction, which is helpful to improve the accuracy of image scene recognition. The salient region is the location with the richest semantic information in the image, and the study uses convolutional neural network to feature extraction of semantic information for multi-scale salient region, so as to realize multi-scale image scene recognition. Let the complete scene in the image be Q_1, Q_2, and Q_3 are the salient regions at two different scales, then the multi-scale salient region in the image X is (Q_1, Q_2, Q_3), and the trained convolutional neural network is used for region feature extraction to obtain the region feature expressions U_p, p=1, 2, 3, and the feature calculation formula is shown in Equation (13).

$$U(p)=g(Q(p);(W,\delta)) \quad (13)$$

In Equation (13), both W and δ are trainable neural network parameters. $g(Q(p);(W, \delta))$ indicates that the forward conduction is computed for the image region $Q(p)$, and the parameter (W, δ) is adjusted with the conduction during the computation. The multi-scale feature expression of the whole image U is obtained by integrating different region expressions as shown in Equation (14).

$$U=U(1)+U(2)+U(3) \quad (14)$$

The spatial mapping between the feature space and the probability distribution space is achieved using Multi-Layer Perceptron (MLP), and the loss function of the multilayer perceptron is shown in Equation (15).

$$E=-\sum_{p=1}^{num(Z)} \log Z(v)+\frac{\lambda}{2}W_{mlp}^{L}W_{mlp} \quad (15)$$

In Equation (15), $Z(v)$ is the predicted probability value of the MLP for the v class of scenes, W_{mlp} is the training parameter, and λ is the control parameter to control the overfitting strength of the MLP. The MLP is trained using stochastic gradient descent to reduce the model loss by back-passing the residuals and iteratively optimizing W_{mlp}. The final scene category is determined by combining the MLP prediction, and the category with the highest prediction probability is judged as the scene category of the image, and the formula is shown in Equation (16).

$$class=\arg\max_{n\in[1,num(z)]} Z(n) \quad (16)$$

4. Verification of Image Object Detection and Scene Recognition Model

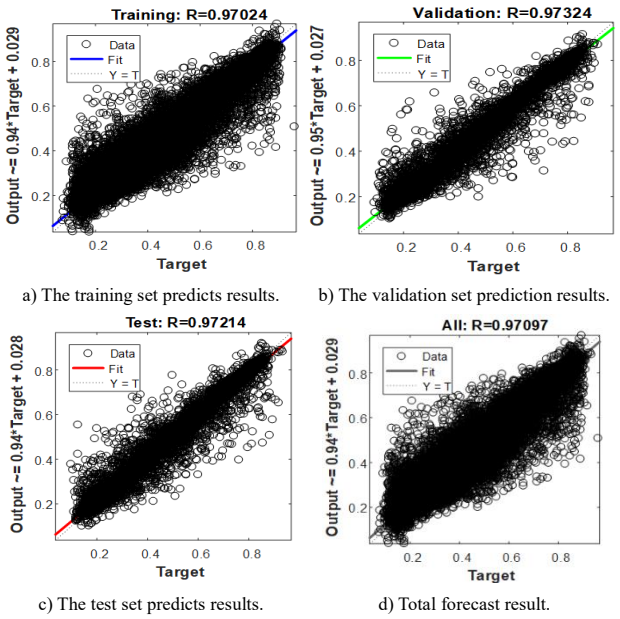
In the experimental analysis of object recognition algorithm based on heterogeneous multi-column convolutional neural network, three classical datasets of MNIST, CIFAR-10 and Caltech-256 were selected for verification, which ensured the universality and reproducibility of experimental results. Among them, the MNIST dataset is a benchmark dataset in the field of handwritten digit recognition, consisting of 60,000 training images and 10,000 test images, each of which is a 28x28 pixel grayscale map representing a number between 0 and 9. This dataset is the first choice for beginners and algorithm validation due to its size and simplicity. The diversity of MNIST is mainly reflected in the change of handwriting style, which requires the generalization ability of the algorithm. The CIFAR-10 dataset is a small image dataset of 60,000 32x32 color images, divided into 10 categories of 6,000 images each. The images cover a variety of natural and man-made objects such as airplanes, cars, birds, cats, and more, taken from a variety of angles, with a certain variety of poses, viewing angles, and lighting conditions. CIFAR-10 is widely used for training and evaluation of image classification tasks, especially for entry-level and medium-complexity computer vision models. The Caltech-256 dataset is an object recognition dataset of more than 30,000 images divided into 256 categories, each containing at least 80 images. These images come from the real world and have large size variations and complexity, which puts higher requirements on recognition algorithms. Because of its broad category coverage and image diversity, Caltech-256 is an important data set for testing algorithm generalization and robustness. These three datasets have their own emphasis on scale, diversity and specific features, which together provide a comprehensive and rich verification platform for the experimental analysis of heterogeneous multi-column convolutional neural network object recognition algorithms. The network training environment for the experiment is the Ubuntu 16.04 system, CPU: Inter Xeon E5-2620, and GPU: NVIDIA TITAN X. The testing environment is the Ubuntu 16.04 system, with CPU: Inter i7-7700 and GPU: NVIDIA GTX 1080Ti. From MNIST to CIFAR-10 to CALTECH-256, the difficulty of classifying these data increases successively. The correctness and generalization performance of the method are proved by experiments. In the construction of multi-row heterogeneous CNN framework, 10 arrays are constructed through experiments, with a variety of network structures. Its construction and accuracy are listed in Table 2.

Table 2. A 10-column convolutional neural network trained on the MNIST dataset.

Network	**Cnn_1**	**Cnn_2**	**Cnn_3**	**Cnn_4**	**Cnn_5**	**Cnn_6**	**Cnn_7**	**Cnn_8**	**Cnn_9**	**Cnn_{10}**
L_1–C	6(5)	6(9)	6(13)	6(5)	6(9)	6(13)	12(5)	12(9)	12(13)	16(5)
L_2–P	2	2	2	2	2	2	2	2	2	2
L_3–C	12(5)	12(5)	12(5)	24(5)	24(5)	24(5)	24(5)	24(5)	24(5)	24(5)
L_4–P	2	2	2	2	2	2	2	2	2	2
error rates (%)	0.98	1.00	1.46	0.91	0.85	1.06	0.78	0.66	0.92	0.65

Table 2 shows that L_i represents the i layer of the network; C and P denote convolution layer and pooling layer respectively; As can be summarized from the table, each of the 10 CNN contains 2 convolutional layers and 2 down sampling layers. The difference is mainly reflected in the size of the convolution core on each line and the number of feature graphs on each line. Finally, each column network classifies the handwritten characters in the network through a complete connectivity layer, and analyzes them in this way. Based on the training data of MNIST, all 10 different kinds of networks have achieved 500 batch gradient descent training, and the training parameters are set to: learning rate=1, batch size=50. Each train of trained networks has passed the MNIST test set to test the error rate of the network. The results are indicated in the last column of Table 2. Through the test of a large number of MNIST samples, it indicates that there are some differences in the testing effect of the model based on the construction of the network. For CNN10 with smaller bit error and CNN3 with larger bit error, the difference in accuracy is about 1%. Consequently, the divergence in network accuracy is not solely based on the construction of the network, but also on the training and testing process. The performance of the CNNs depends on many factors, such as the choice of hyperparameters, the architecture of the network, the number of training epochs, and the size and quality of the training data. Moreover, the networks' function can also be affected by the presence of noise, distortion, and occlusion in the input images. Overall, Table 2 provides a detailed description of the architecture of the 10-column CNNs trained on the MNIST dataset, as well as their performance in terms of error rates. The results demonstrate that the choice of architecture and training parameters can greatly affect the function of the CNNs, and that the performance of the networks can be improved by optimizing these factors.

Figure 5 showcases that the predicted values of the training set are high, with deviations in a few cases. The R-value of each of its set's hovers around 0.97, which is a good fit, and the R-value of the training set is 0.97024, which can indicate that the model proposed in the study is a good fit.



a) The training set predicts results. b) The validation set prediction results.

c) The test set predicts results. d) Total forecast result.

Figure 5. Fitted curve after training.

Table 3 provides insight into the error detection rates of multi-row CNN in the MNIST dataset across four groups. The results indicate that the error rate of MCNN fused through sliding windows has decreased by about 25% compared to single-column CNN in Group1 and Group2. Group3, which had the smallest error rate decrease, still achieved a 9% error rate reduction. These findings suggest that the sliding window approach used in multi-row CNN can improve object classification accuracy. Building on these findings, this paper proposes a hybrid learning method that utilizes a sliding window approach with multi-row datasets. Our experiments indicate that this proposed algorithm can significantly improve object classification accuracy. Hence, characteristic graphs' quantity in the multi-row convolution network significantly impacts the performance of the multi-column convolution network. To account for this, the study divided the 10-column convolution network into three groups, including group 1 (CNN2, CNN5, CNN8, CNN11) and group 3 (CNN3, CNN6, CNN9, CNN10). Experiments are conducted on multi-row CNN using the MNIST database and observed significant improvements in performance.

Overall, the results of our experiments demonstrate the potential of the sliding window approach in multi-row CNN for improving object classification accuracy. The proposed hybrid learning method has the potential to be applied to other datasets and can contribute to the development of more effective and efficient neural networks.

Table 3. Research on error detection of multi-row convolutional neural networks in Mnist data set.

-	Group1	Group2	Group3	Group4
Single column convolutional networks	0.98%	0.85%	0.66%	0.65%
Multi-column convolutional networks	0.72%	0.64%	0.60%	0.54%
Reduced error rate	26.53%	24.71%	9.09%	16.92%

Figure 6 displays the results of testing two and three samples, indicating that the sliding window CNN exhibits higher identification accuracy compared to the single-column network. The error rate of the multi column convolutional neural network after sliding window fusion is reduced compared to a single network, which is consistent with the conclusion in Figure 7. However, it is important to note that the first set of data suggests that the performance of the MCNN does not enhance when compared to the single-column network. This may be attributed to the small size (5×5) of the convolution kernel used in the CNN of the first group, which leads to insufficient diversity among the networks. As a result, the convergence of multi-column networks does not result in any performance improvements. The first team concluded that, in all experiments conducted, there was only one instance where the performance of multi-row CNNs was not better than that of single-column networks.

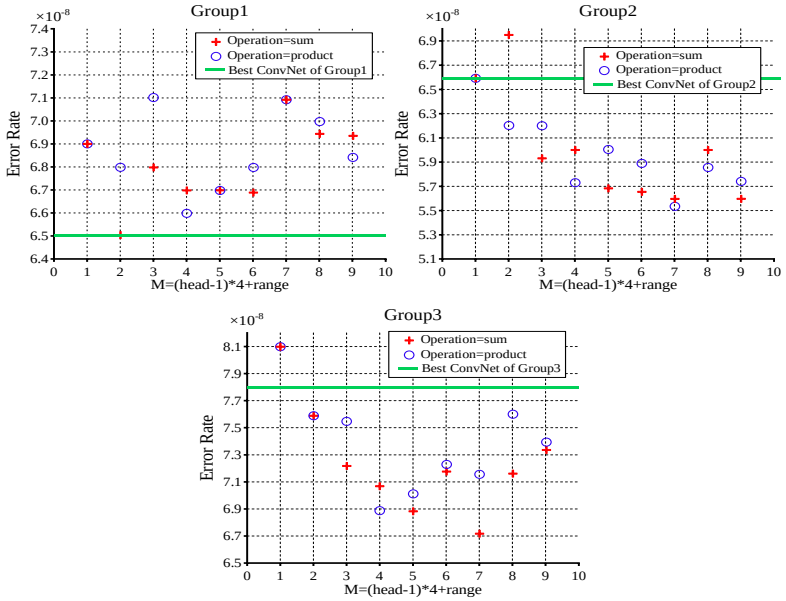


Figure 6. Multi-column convolutional neural network feature map quantity analysis.

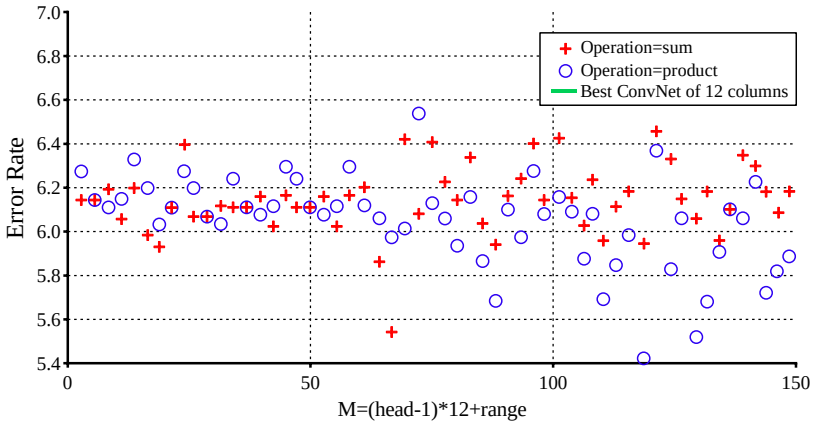


Figure 7. Sliding window fusion for 10-column convolutional neural networks.

In the research, 10 types of heterogeneous multi-sequence CNNs with multiple topological characteristics are used to construct a new non-uniform multi-sequence CNN. In the MNIST dataset, a SWF algorithm is used and a result is gotten similar to Figure 6. Here the title=CEIL (Mhand 10), the range=mod ((Mmur1), 10)+1, and M [1150]. For single MNIST data, the experimental effect of 10-column CNN is the best, with an error of 0.65%, which is CNN10. The use of a 10-row CNN has been implemented for the purpose of classifying multiple data, resulting in a classification accuracy of 0.54%. By comparing it with a row convolution network with the minimum error probability, it has been observed that the error probability of a multi-row convolution network using the sliding window method is reduced by 16.92%. This suggests that the performance of the non-uniform multi-row CNN with sliding window outperforms that of traditional single-row neural networks. To verify whether the performance of multi-column convolutional neural network combined with sliding window fusion algorithm is significantly improved in object recognition tasks, the paired sample t test is used to compare the error rates of single convolutional neural network and multi-column convolutional neural network combined with sliding window fusion algorithm on MNIST dataset. The paired sample T-test is used to evaluate whether there is a significant difference between two sets of data (single network and multi-column network), and the 95% confidence interval of the error rate of the multi-column convolutional neural network combined with the sliding window fusion algorithm is also calculated during the T-test to evaluate the confidence degree of performance improvement. The specific results are shown in Table 4.

Table 4. Error rates and statistical test results of different algorithms on MNIST data sets.

Group	Single network error rate	Multi-row network error rate	P	Lower bound of 95% confidence interval	Upper 95% confidence interval
Group1	0.80%	0.60%	0.001	0.15%	0.25%
Group2	0.75%	0.56%	0.002	0.14%	0.24%
Group3	0.70%	0.63%	0.05	0.02%	0.12%
All 12 columns	0.65%	0.54%	<0.001	0.09%	0.13%

As can be seen from Table 4, the multi-column convolutional neural network combined with sliding window fusion algorithm significantly reduces the error rate compared with a single convolutional neural network. The error rate for Group1 and Group2 dropped by about 25 percent, and for Group3 by 9 percent. This shows the effectiveness of multi-column structure and sliding window fusion algorithm. When the convolutional network with 12 columns of different structures is constructed into a heterogeneous multi-column convolutional neural network, the error rate is reduced from 0.65% to 0.54% by sliding window fusion, which is 16.92%. This significant performance improvement further verifies the effectiveness of the multi-column structure and sliding window fusion algorithm. The efficacy of this algorithm has been further evaluated through a fitness comparison curve, which compares the algorithm's performance with that of CNN algorithm and other improved algorithms. Figure 8 illustrates the results of this comparison.

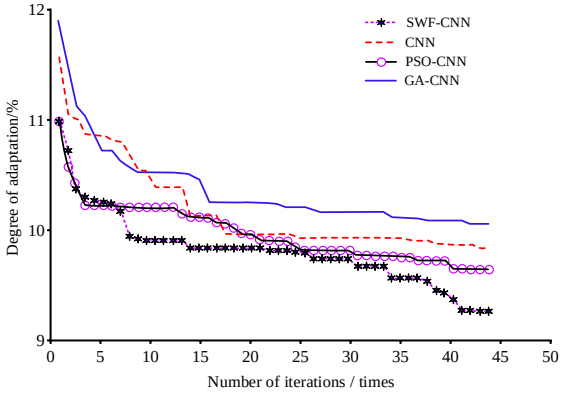


Figure 8. Comparison of convergence curves of different algorithms.

Figure 8 provides a graphical representation of the adaptation rate for various algorithms. Specifically, the CNN algorithm was observed to gradually stabilize at around 10% after 19 runs, while the PSO-CNN algorithm reached a similar stabilization point of around 10.1% after 25 runs. The GA-CNN algorithm, on the other hand, achieved a stabilization point of approximately 10.5% after 16 runs. The SWF-CNN algorithm demonstrated superior performance, with a convergence speed that was faster and a lower fitness when compared to the CNN, PSO-CNN, and GA-CNN algorithms. Additionally, multi-column model fusion was conducted using three fusion methods: complete SWF, trained SWF, and conventional single fusion. A quantitative comparison was then carried out to determine the efficacy of each approach. Table 5 illustrates the results based on the MNIST data group. For each group, one way to integrate multiple neural networks is to exhaust SWF, train SWF, maximum rules, minimum rules, and rules, and accumulation rules.

In Table 5, the results of the two most successful fusion algorithms are highlighted in boldface for each multi-column convolution network set. Upon examination, it is evident that the exhaustive SWF method outperforms other fusion methods in every group of MCNN. This is due to the fact that other fusion strategies are a special form of SWF, and as a result, the results obtained from the exhaustive SWF method are not inferior to those of the traditional single fusion method. However, significantly, while the exhaustive SWF method yields better results, it may not be as practical for real-world applications as the trained SWF strategy. This is because the complexity of the testing process is lower for the latter approach, making it more suitable for practical application requirements. Set up a comparative experiment to verify the effectiveness of the image scene recognition model. The experiment compared the Faster R-CNN and SSD network models, and trained the three network models. The training results are shown in Figure 9.

Table 5. Error rate test results of various multi-column convolutional neural network fusion methods on mnist dataset.

Rules	Three-column network convergence				Four column network convergence			Twelve columns of network convergence
	Group1	Group2	Group3	Group4	Group1	Group2	Group3	-
Exhaust	0.865%	0.395%	0.621%	0.566%	0.692%	0.495%	0.697%	0.753%
Train	0.443%	0.688%	0.593%	0.591%	0.692%	0.568%	0.688%	0.638%
Max	0.910%	0.639%	0.572%	0.593%	0.655%	0.533%	0.693%	0.722%
Min	0.364%	0.745%	0.709%	0.597%	0.711%	0.697%	0.795%	0.741%
Sum	0.779%	0.615%	0.603%	0.668%	0.698%	0.495%	0.708%	0.603%
Proudct	0.782%	0.693%	0.691%	0.665%	0.709%	0.721%	0.698%	0.705%

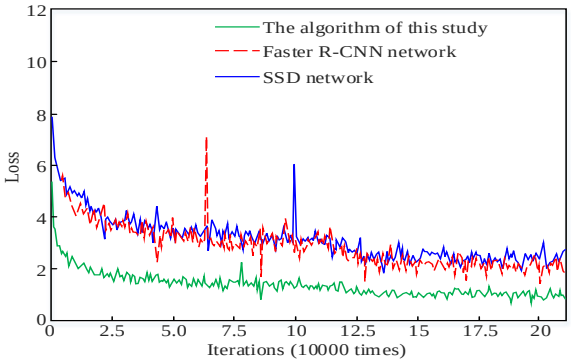


Figure 9. Network model training process and model comparison analysis.

As shown in Figure 9, as the iterations continue to increase, the losses of all three network models continue to decrease. When the number of iterations reaches 63000, the loss of Faster R-CNN has a short mutation. When the number of iterations reaches 10000, the loss of SSD model also has a short mutation. The scene recognition model based on Convolutional neural network has no mutation. And after the number of iterations reached 150000, the losses of the three network models gradually converged, and it is not difficult to see that the overall loss of the scene recognition model based on Convolutional neural network decreased significantly faster than the other two network models, and the convergence process was smoother. In order to further test the performance of the scene recognition technology in this study, the traditional algorithm is also compared with the scene recognition model based on Convolutional neural network, and the time spent in the actual scene recognition process is analyzed. The results are shown in Figure 10.

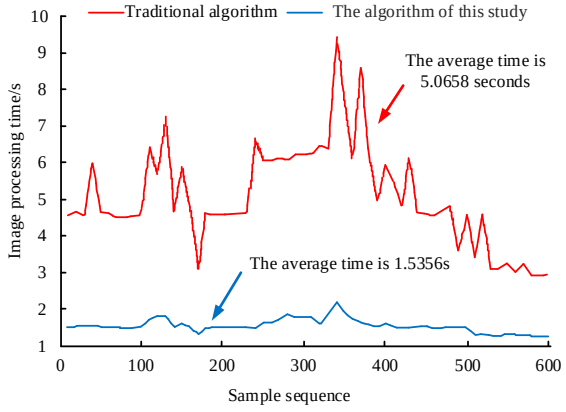


Figure 10. Comparison of image recognition detection time consumption between traditional algorithms and our research algorithm.

Figure 10 shows the comparison of recognition time consumption between the two algorithms. The horizontal axis represents the scene image sequence, and the vertical axis is the sample sequence. As shown in Figure 9, there is a significant difference in time consumption between the two recognition and detection

methods. Traditional recognition and detection methods take longer, with an average recognition and detection time of 5.0658 seconds. However, the scene image recognition and detection method based on the potential object area recognition algorithm and CNN is faster and more efficient, and its time consumption is far lower than the former, the average value is only 1.5356s, and the fluctuation is small, more stable. To comprehensively evaluate the performance of the proposed heterogeneous multi-column convolutional neural network object recognition algorithm, the nu-GSNG algorithm, BING-Frac algorithm and G2L-Net+Center Net algorithm are compared. Comparison indicators include error rate, fitness, convergence speed, and computational complexity. The specific comparison results are shown in Table 6.

Table 6. Performance comparison of the four algorithms.

Algorithm	Error rate	Fitness	Number of iterations	Computational complexity
Nu-GSNG	0.70%	10.0%	19	High
BING-Frac	0.68%	9.5%	25	Medium
G2L-Net+Center Net	0.67%	10.5%	16	Medium
SWF-CNN	0.54%	9.8%	7	Low

As can be seen from Table 6, SWF-CNN has the lowest error rate of 0.54%, which is 0.16%, 0.14% and 0.13% less than Nu-GSNG, BING-Frac and G2L-Net+Center Net, respectively. The results show that the error rate of SWF-CNN is significantly reduced compared with other algorithms, indicating its superiority in object recognition tasks. In addition, SWF-CNN's fitness of 9.8% was lower than G2L-Net+Center Net's 10.5% and Nu-GSNG's 10.0%, but slightly higher than BING-Frac's 9.5%. The results show that SWF-CNN has a good fitness, but there is still room for improvement. Finally, it can be found from Table 6 that SWF-CNN has the fastest convergence speed and the lowest computational complexity. It only takes 7 iterations to achieve stable fitness, which is significantly better than the comparison algorithm. The results show that SWF-CNN is more suitable for the resource-limited application scenarios. In summary, the proposed SWF-CNN algorithm shows significant advantages in multiple performance indexes.

5. Conclusions

The combination of computer vision technology and pattern recognition technology has been widely applied in fields such as intelligent transportation, food safety, medical diagnosis, and industrial testing. Based on this, this paper proposes an image Object detection technology combining sliding window fusion and heterogeneous multi column Convolutional neural network for the field of image computer vision processing. The research also studies the super parameters in the sliding window fusion algorithm, and proposes a sliding window fusion based on the exhaustion and training of super parameters. And the research uses the potential object domain recognition algorithm and Convolutional neural network to identify the image scene, and solves the image multi-scale scene recognition problem through the potential area object density analysis. Experiments based on the public dataset MNIST, CIFAR-10 and Caltech-256 show that the heterogeneous multi column Convolutional neural network based on sliding window fusion can improve the accuracy of Object detection by 10% to 25% compared with a single Convolutional neural network structure. SWF-CNN algorithm has gradually stabilized its adaptation rate after seven runs, reaching about 9.8%, and its operation effect is significantly better than other algorithms. The sliding window fusion is carried out through the prediction results of 12 column Convolutional neural network, and the error rate is reduced to the lowest 0.54%. Compared with the single column Convolutional neural network with the lowest error rate, the error rate of the multi column Convolutional neural network after sliding window fusion decreases by 16.92%. Although the effectiveness of the trained sliding window fusion method in practical testing is slightly lower than that of the exhaustive sliding window fusion method, compared to other traditional fusion strategies, the trained sliding window fusion still exhibits stronger network fusion capabilities. Comparing the scene recognition model with Faster R-CNN and SSD network models, the overall loss reduction speed of the proposed scene recognition model is significantly faster than the other two network models, and the convergence process is smoother. And the image scene recognition efficiency of the model is higher, with an average recognition time of 1.5356 seconds. In order to further improve the recognition effect of image objects and scenes, future research will consider further optimizing the structure of Convolutional neural network, incorporating a larger scene dataset for model training, and improving the network's feature learning and discriminant generalization performance.

Statements and Declarations

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Fundings

The research is supported by The Education Department of Henan Province in 2021: Application Research on Real-Time Cattle Status Recognition and Disease Early Warning based on Convolution Representation Flow (No.212102210138).

References

[1] Afif M., Ayachi R., Said Y., and Atri M., "Deep Learning-Based Application for Indoor Scene Recognition," *Neural Processing Letters*, vol. 51, no. 3, pp. 2827-2837, 2020. https://doi.org/10.1007/s11063-020-10231-w

[2] Anami B. and Sagarnal C., "Influence of Different Activation Functions on Deep Learning Models in Indoor Scene Images Classification," *Pattern Recognition and Image Analysis*, vol. 32, no. 1, pp. 78-88, 2020. https://doi.org/10.1134/S1054661821040039

[3] Andriyanov N., Dementiev V., and Kargashin Y., "Analysis of the Impact of Visual Attacks on the Characteristics of Neural Networks in Image Recognition," *Procedia Computer Science*, vol. 186, no. 12, pp. 495-502, 2021. https://doi.org/10.1016/j.procs.2021.04.170

[4] Bai H., Zhang H., and Wang Q., "Dual Discriminative Auto-Encoder Network for Zero Shot Image Recognition," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 12, pp. 1-12, 2021. https://doi.org/10.3233/JIFS-201920

[5] Carpenter C., "Machine-Learning Image Recognition Enhances Rock Classification," *Journal of Petroleum Technology*, vol. 72, no. 10, pp. 63-64, 2020. https://doi.org/10.2118/1020-0063-JPT

[6] Chen A., Hong S., Wang Y., Li C., Yang C., and Chen H., "Rapid Assessment of Gasoline Quality by Near-Infrared (NIR) Deep Learning Model Combined with Fractional Derivative Pretreatment," *Analytical Letters*, vol. 55, no. 11, pp. 1745-1756, 2022. https://doi.org/10.1080/00032719.2021.2024219

[7] Chen Z., Su Y., Liu Y., Huang J., and Cao W., "Key Technologies of Intelligent Transportation Based on Image Recognition," *International Journal of Advanced Robotic Systems*, vol. 17, no. 3, pp. 110-120, 2020. https://doi.org/10.1177/1729881420917277

[8] Cooper M., Krishnan R., and Bhat M., "Deep Learning and the Future of the Model for End-Stage Liver Disease-Sodium Score," *Liver Transplantation*, vol. 28, no. 7, pp. 1128-1130, 2022. DOI:10.1002/lt.26485

[9] Corti E., Khanna A., Niang K., Robertson J., Moselund K., Gotsmann B., Datta S., and Karg S., "Time-Delay Encoded Image Recognition in a Network of Resistively Coupled VO_2 on Si Oscillators," *IEEE Electron Device Letters*, vol. 41, no. 4, pp. 629-632, 2020. DOI:10.1109/LED.2020.2972006

[10] Daradkeh Y., Tvoroshenko I., Gorokhovatskyi V., Latiff L., and Ahmad N., "Development of Effective Methods for Structural Image Recognition Using the Principles of Data Granulation and Apparatus of Fuzzy Logic," *IEEE Access*, vol. 9, no. 99, pp. 13417-13428, 2021. DOI:10.1109/ACCESS.2021.3051625

[11] Gautam S. and Dharv G., "Detection of Novel Corona Virus Using Machine Learning and Image Recognition," *International Journal for Modern Trends in Science and Technology*, vol. 6, no. 12, pp. 394-397, 2020. https://doi.org/10.46501/IJMTST061274

[12] Ge Z., Cao G., Li X, and Fu P., "Hyperspectral Image Classification Method Based on 2D-3D CNN and Multibranch Feature Fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5776-5788, 2020. DOI:10.1109/JSTARS.2020.3024841

[13] Han X., "An Improved Classification Model for English Syntax Error Correction Design of DL Algorithm," *The International Arab Journal of Information Technology*, vol. 21, no. 4, pp. 560-570, 2024. DOI:10.34028/iajit/21/4/2

[14] Hua W., Guan X., and Jiang X., "Clinical Study on Gastroscopy Image Recognition Model Based on Artificial Intelligence in Diagnosis of Chronic Atrophic Gastritis," *Chinese Journal of Gastroenterology*, vol. 12, pp. 588-593, 2020. DOI:10.3760/cma.j.cn112148-20200420-00123

[15] Kwon K. and Lee H., "Vespa Mandarinia Image Recognition Using Image Fused Preprocessing and Deep Learning," *Journal of Digital Contents Society*, vol. 21, no. 10, pp. 1855-1862, 2020. http://journal.dcs.or.kr/xml/26359/26359.pdf

[16] Lee K., Na J., Sohn J., Sohn S., and Lee S., "Image Recognition Algorithm for Maintenance Data Digitization: CNN and FCN," *Transactions of the Korean Society for Noise and Vibration Engineering*, vol. 30, no. 2, pp. 136-142, 2020. DOI:10.5050/KSNVE.2020.30.2.136

[17] Li Z., Zhou A., and Shen Y., "An End-To-End Trainable Multi-Column CNN for Scene Recognition in Extremely Changing Environment," *Sensors*, vol. 20, no. 6, pp. 1556-1571, 2020. DOI:10.3390/s20061556

[18] Long S. and Zhao X., "Smart Teaching Mode Based on Particle Swarm Image Recognition and Human-Computer Interaction Deep Learning," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 4, pp. 5699-5711, 2020. https://doi.org/10.3233/JIFS-179762

[19] Maschler B., Kamm S., and Weyrich M., "Deep Industrial Transfer Learning at Runtime for Image Recognition," *Automatisierungstechnik*, vol. 69, no. 3, pp. 211-220, 2021. https://doi.org/10.1515/auto-2020-0119

[20] Matsuzaki S., Miura J., and Masuzawa H., "Multi-Source Pseudo-Label Learning of Semantic Segmentation for the Scene Recognition of Agricultural Mobile Robots," *Advanced Robotics*,

vol. 36, no. 19, pp. 1011-1029, 2022. https://doi.org/10.48550/arXiv.2102.06386

[21] Moradipour K., Fallah M., Abdali E., and Asadi S., "Efficiency Evaluation in Hybrid Two-Stage Network DEA with Non-Discretionary Inputs and Shared Discretionary Inputs," *International Journal of Computer Mathematics: Computer Systems Theory*, vol. 7, no. 1, pp. 33-41, 2022. https://doi.org/10.1080/23799927.2021.1983876

[22] Prajwalasimha S., Sahana G., and Vaani K., "Iris Image Recognition Based on Combined Hamming and Cosine Distances Approach," *International Journal of Advanced Science and Technology*, vol. 29, no. 4, pp. 6708-6719, 2020.

[23] Rafique A., Gochoo M., Jalal A., and Kim K., "Maximum Entropy Scaled Super Pixels Segmentation for Multi-Object Detection and Scene Recognition Via Deep Belief Network," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 63-64, 2023. DOI:10.1007/s11042-022-13717-y

[24] Rehman A., Saleem S., Khan U., Jabeen S., and Shafiq M., "Scene Recognition by Joint Learning of DNN from Bag of Visual Words and Convolutional DCT Features," *Applied Artificial Intelligence*, vol. 35, no. 9, pp. 623-641, 2020. https://doi.org/10.1080/08839514.2021.1881296

[25] Seong H., Hyun J., and Kim E., "Fosnet: An End-To-End Trainable Deep Neural Network for Scene Recognition," *arxiv Preprint*, vol. arXiv:1907.075702020, pp. 1-11, 2019. DOI:10.48550/arXiv.1907.07570

[26] Tan K., Xu Y., Zhang S., Yu M., and Yu D., "Audio-Visual Speech Separation and Dereverberation with a Two-Stage Multimodal Network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542-553, 2020. DOI:10.1109/jstsp.2020.2987209

[27] Wang C., Peng G., and Lin W., "Self-Weighted Discriminative Metric Learning Based on Deep Features for Scene Recognition," *Multimedia Tools and Applications*, vol. 79, no. 3-4, pp. 2769-2788, 2020. https://doi.org/10.1007/s11042-019-08486-0

[28] Yu D., Xu Q., Guo H., Zhao C., Lin Y., And Li D., "An Efficient and Lightweight Convolutional Neural Network for Remote Sensing Image Scene Classification," *Sensors*, vol. 20, no. 7, pp. 63-64, 2020. DOI:10.3390/S20071999

Guoyan Li was born in Zhengzhou, Henan Province, CHN in 1979. He received the M.S. degree in computer science and technology from Henan University, Henan, China, in 2008. From 2003 to 2008, he was a Teaching Assistant, in the School of Information Engineering, Henan University of Animal Husbandry and Economy, Zhengzhou, China. Since 2008, he has been a Lecturer in the School of Information Engineering, Henan University of Animal Husbandry and Economy, Zhengzhou, China. He is the author of two books, more than 10 articles. His research interests include data mining and deep learning

Fei Wang was born in Zhengzhou, Henan Province, CHN in 1982, Han nationality, male, Ph.D., 2006, Computer Science and Technology, Engineering degree, Zhengzhou University of Aeronautics; 2010, Machine Learning, Engineering degree, Henan University of Technology; 2019, Machine Learning, Engineering degree, Donghua university. Research interests: machine vision and deep learning. 2020-2024 he works School of Information Engineering, Henan University of Animal Husbandry and Economy, Deputy Director of teaching and research. He has published over 10 academic papers; Academic works 3; 2 Project leaders in Provincial level project; Patents 4.