Traitement du Signal
Vol. 39, No. 2, April, 2022, pp. 595-601
Journal homepage: http://iieta.org/journals/ts

Image Target Recognition Based on Multiregional Features under Hybrid Attention Mechanism

Hui Zhao

College of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723001, China

Corresponding Author Email: zhaohui@snut.edu.cn

https://doi.org/10.18280/ts.390221

Received: 5 December 2021
Accepted: 7 February 2022

Keywords:
hybrid attention, multiregional features, image target recognition

ABSTRACT

The growing volume of image data calls for better real-time performance of image feature extraction algorithms. To enhance the recognition accuracy of image targets, it is significant to build a more scientific deep learning network. Multimodal cross convolution or densely connected blocks have been introduced to classic deep learning networks, aiming to promote the recognition of image targets. However, these attempts fail to satisfactorily extract detailed features from the original image. To solve the problem, this paper explores the image target recognition based on multiregional features under hybrid attention mechanism. Specifically, a convolutional neural network (CNN) was established for extracting multiregional features based on the loss function of local feature aggregation. The model consists of three independent CNN modules, which are responsible for extracting the global multiregional features and the local features of different regions. Next, the channel domain attention mechanism and spatial domain attention mechanism were embedded in the proposed CNN, such that the model can recognize targets more accurately, without increasing the computing load. Finally, the proposed network was proved effective through the training and testing on a self-developed sample set of surveillance video images.

1. INTRODUCTION

With the development of deep learning, data mining, and artificial intelligence, computer vision has made remarkable progress in the fields of medical care, engineering construction, education, agriculture, and light industry [1-8], and displayed its strong ability and excellent effects in target recognition based on multiregional feature matching. The relevant innovative technologies provide strong support for accurate target recognition for all feature points in the captured images [9-14].

Traditionally, image features are extracted by principal component analysis (PCA), linear discriminant analysis (LDA), local linear embedding, etc. [15-20]. However, the growing volume of image data calls for better real-time performance of image feature extraction algorithms [21-24]. Although deep learning can learn the features of image targets, the network is too complex, and the computing load is too high. To enhance the recognition accuracy of image targets, it is significant to build a more scientific deep learning network.

Salient region detection plays an important role in image preprocessing. How to highlight salient regions evenly remains a difficulty in computer vision. Inspired by absorbing Markov chain, Zhang et al. [25] proposed a salient region detection method driven by multi-feature data. The method relies on super pixels to extract salient regions. Firstly, a function was constructed to calculate the absorption probability of each node on the absorbing Markov chain. Based on the image contrast and spatial relationship, the prior saliency mapping was modeled for the salient nodes in the foreground. Then, the saliency of each node was computed according to the absorption probability. To accurately detect the occluded regions in the video, Zhang et al. [26] developed a video occluded region detection approach based on multi-feature fusion. Drawing on light flow and brightness, three new occlusion-related features were designed, namely, brightness block matching, maximum flow difference, and flow residual, and the relevant calculation methods were defined.

It is always challenging to find information rich regions in the scene. Li et al. [27] presented a scene recognition method based on multi-scale salient regional features. Firstly, the multi-scale salient regions were detected in the scene. Then, the features of each region were extracted through transfer learning, using a convolutional neural network (CNN). Cheng et al. [28] extended the U-Net into a contour-aware semantic segmentation network for medical image segmentation. The network consists of a semantic branch and a detail branch, and introduces a spatial attention module to adaptively suppress redundant features. Compared with the latest strategies, their network performed remarkably on challenging public medical image segmentation tasks. Zhao et al. [29] devised a recurrent slice-wise attention network (RSANet) based on regional self-attention mechanism. The network focuses on the information flow through the entire image, rather than the local convolutional operations. It mines the relationship between adjacent pixels, contributing to a more logical understanding of image contents.

Multimodal cross convolution or densely connected blocks have been introduced to classic deep learning networks, aiming to promote the recognition of image targets. Some researchers suggested refining the structure with dilated convolution, or using the sequential extrusion module. However, these attempts fail to satisfactorily extract detailed features from the original image. To solve the problem, this paper explores the image target recognition based on multiregional features under hybrid attention mechanism. The main contents are as follows: (1) The authors set up a CNN for extracting multiregional features based on the loss function of local feature aggregation. The model consists of three independent CNN modules, which are responsible for extracting the global multiregional features and the local features of different regions. (2) The authors embedded the channel domain attention mechanism and spatial domain attention mechanism in the proposed CNN, such that the model can recognize targets more accurately, without increasing the computing load. (3) The authors verified the effectiveness of the proposed CNN through the training and testing on a self-developed sample set of surveillance video images.

2. DEEP LEARNING MODEL

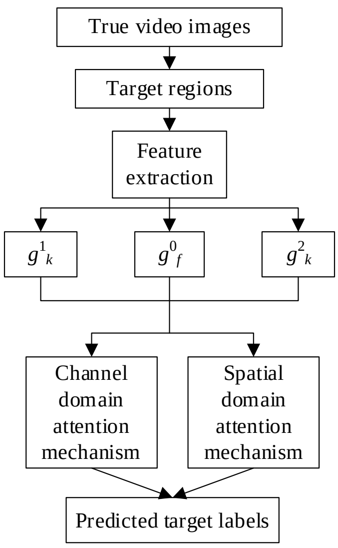


Figure 1. Process of image target recognition based on multiregional features

This paper focuses on real-world image sets of surveillance videos. Compared with the lab image sets with sample constraints, the real-world image sets of surveillance videos are highly flexible and stochastic, involving complex scenes, diverse angles, conclusion, and targets with rich process, expressions, and action. Figure 1 shows the process of image target recognition based on multiregional features for the real-world image sets of surveillance videos.

The traditional deep learning models, which are based on softmax loss function, perform poorly in feature extraction of such image sets. To solve the problem, this paper sets up a CNN for extracting multiregional features based on the loss function of local feature aggregation. The model consists of three independent CNN modules, which are responsible for extracting the global multiregional features and the local features of different regions. In addition, a supervision layer with the loss function of local feature aggregation was introduced to the three independent CNN modules. In this way, the proposed CNN could extract rich and diverse image features with strong complementarity, robustness, and identifiability, compared with single regional features. Figure 2 displays the framework of CNN for multiregional feature extraction.

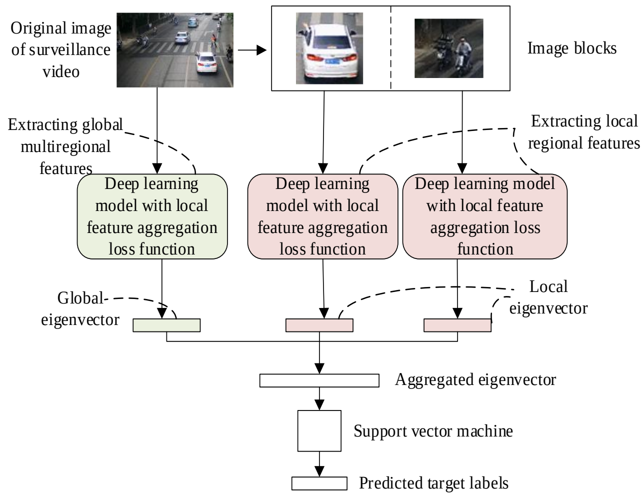


Figure 2. Framework of CNN for multiregional feature extraction

Let a_i be the i-th image eigenvector in the fully-connected layer before the supervision layer; $M_l\{a_i\}$ be the set of image features in the l nearest regions, which belong to the same class of a_i; $1/l\sum_{a\in Ml\{ai\}}a$ be the center of the set; m be the batch size. Then, the local loss function of the proposed CNN can be expressed as:

$$K_{kt}=\frac{1}{2}\sum_{i=1}^{m}\left\|a_i-\frac{1}{l}\sum_{a\in M_l\{a_i\}}a\right\|_2^2 \tag{1}$$

By minimizing K_{kt}, each image feature can gradually approximate the center of the set of image features in the l nearest regions, which belong to the same class as the feature. The minimization makes the said set more compact, reducing the intra-cluster feature difference. Let K_r be the softmax loss; K_{kt} be the local loss; μ be the weight coefficient to balance the two losses. Then, the final loss function can be expressed as $K=K_r+\mu K_{kt}$.

This paper needs to reduce the difference in the set of same class image features in the nearest region, while enlarging the difference between image features in different classes. Thus, it is not sufficient to consider the local loss alone. Hence, the between class feature difference K_x can be defined as:

$$K_x=\frac{1}{2}\sum_{i=1}^{m}\left\|a_i-\xi a_j-\left(1-\xi\right)a_d\right\|_2^2 \tag{2}$$

During the training of the CNN, it is necessary to search for the image feature a_j, which is the closest to a_i in the neighborhood l_1 but belongs to another class. The reduce the effect of noise on network training, the center a_d of set of image features in the neighborhood l_1, which belong to the same class of a_i, was introduced:

$$a_d=\frac{1}{l_2}\sum_{a\in M_{l_2}\{a_j\}}a \tag{3}$$

Let $M_{l2}\{a_j\}$ be the set of image features in the neighborhood l_1, which belong to the same class of a_i; l_2 be the number of image features in the set. Then, a parameter ξ was introduced to balance a_i with a_d. Then, K_x can be defined as:

$$K_x = \frac{1}{2}\sum_{i=1}^{m}\left\| a_i - \xi a_j - (1-\xi)\frac{1}{l_2}\sum_{a \in M_{l_2}\{a_j\}} a \right\|_2^2 \tag{4}$$

During the CNN training, the between-class feature difference K_x should increase continuously. Let μ_2 be the balancing weight coefficient. Then, the loss function of local feature aggregation for the proposed CNN can be defined as:

$$\begin{aligned} K_{LFA} &= K_{kt} + \mu_2 \frac{1}{K_x + \alpha} \\ &= K_{kt} + \\ &\mu_2 \sum_{i=1}^{m} \frac{1}{\frac{1}{2}\left\| a_i - \left(\xi a_j + (1-\xi)\frac{1}{l_2}\sum_{a \in M_{l_2}\{a_j\}} a \right) \right\|_2^2 + \alpha} \end{aligned} \tag{5}$$

where, K_{kt} is responsible for reducing the difference in the set of same class image features in the neighborhood; the other parts of the loss function are responsible for increasing the between class difference of image features. The parameter α only avoids exploding gradients and infinitely large loss, and does not affect the final effect of feature extraction. The final loss function of the deep learning network can be expressed as:

$$K = K_r + \mu_1 K_{LFA} \tag{6}$$

The optimal feature extraction model can be obtained through the forward and backward learning training of the proposed model. This paper imports the real-world video images into the global multiregional feature extraction module, extracts the global multiregional eigenvector of each image from the fully connected layer, and further obtains the multiregional feature points of image targets. Based on the multiregional feature points of image targets, the image was divided into two parts, in order to acquire local image targets with rich features.

After that, the local image targets were imported into the other two CNN modules for extracting features from a single region, aiming to obtain the local details of image targets. Except for the detailed configurations of CNNs, the structure and execution steps of the proposed network are the same as those of the global multiregional feature extraction module. Due to the size difference between input images, there are disparities between the extracted features. In this way, the global image target is complementary with local features. Hence, our model extracts more features from image targets, making the multiregional features of image targets more representative.

Let $g^0{}_f$ be the global multiregional feature; $g^1{}_k$ and $g^2{}_k$ be the single reginal local features. The two types of features are serial connected to obtain the aggregated feature g_x of image target. Then, we have:

$$g_x = \left(g_f^0; g_k^1; g_k^2 \right) \tag{7}$$

Importing g_x into the SVM for classification, the final predicted label of each image target can be obtained:

$$label_{out} = argmax\left(g_x \right) \tag{8}$$

3. HYBRID DOMAIN ATTENTION MECHANISMS

To further enhance the ability of our network to depict the details of image targets, this paper embeds the channel domain attention mechanism and spatial domain attention mechanism in the proposed CNN, that the model can recognize targets more accurately, without greatly increasing the computing load.

In the channel domain attention mechanism, the size of feature map A is denoted as $F'\times Q'\times D'$, and the size of feature map V is denoted as $F\times Q\times D$. Let h_i be the convolutional kernel of the i-th channel; A be the input; * be the convolutional operation; a^j be the j-th input feature map; v_i be the feature of the feature map V in the i-th channel, which is obtained after the convolution of feature map A. Then, we have:

$$v_i = h_i * A = \sum_{j=1}^{D'} h_i^j * a^j \tag{9}$$

Let G_{ap} be global average pooling; G_{cov} be a series of convolutional operations. Then, the value r_i of the i-th channel after global average pooling can be expressed as:

$$r_i = G_{\mathrm{cov}}\left(v_i \right) = \frac{1}{F \times Q}\sum_{t=1}^{F}\sum_{w=1}^{Q} v_i\left(t, w \right) \tag{10}$$

After the G_{ap} operation, the feature map is dimensionally reduced from $F\times Q\times D$ to $1\times 1\times D$. Let ρ be the weight of learning for each channel. Then, the input features are activated by G_{ef}:

$$\rho = G_{ef}\left(r, W \right) = \varepsilon\left(h\left(r, W \right) \right) = \varepsilon\left(W_2 \psi\left(W_1 r \right) \right) \tag{11}$$

After being processed by the fully connected layer W_1 and the activation function of rectified linear unit (ReLU), the feature r of image target is reduced to the dimension of $1\times 1\times d/s$. Then, the feature dimension is increased to $1\times 1\times D$, after being processed by the fully connected layer W_2 and sigmoid activation function $\varepsilon(\cdot)$. Let $\dot{a}$ be the product between the learning weight ρ_i of the i-th channel, and the i-th feature map of the original video image. Then, the channel weights of the feature map of the original video image targets can be redistributed by:

$$\dot{a} = G_{wd}\left(v_i, \rho_i \right) \tag{12}$$

The channel weights can be re-calibrated through the above operations.

In the spatial domain attention mechanism, the input feature map V can be expressed as $V=[v^{1,1}, v^{1,2}, \ldots, v^{i,j}, \ldots, v^{F,Q}]$ by spatial dimensions, where $v^{i,j}$ is all the features at position (i, j). Firstly, the features are compressed through a convolution with the kernel U_{rw}. The projection tensor t obtained by convolution is of the size $F\times Q$:

$$t = U_{rw} * V \tag{13}$$

The size of U_{rw} is $1\times 1\times 1$. Suppose the t_{ij} of each projection position represents the linear combination of all channel features at position (i, j). The projection tensor t is nonlinearly activated by Sigmoid function. The activation results are multiplied with each position of the original video image features. Let $\delta(t_{ij})$ be the importance of all the spatial information of the feature at position (i, j) relative to the image. Then, the output of the spatial domain attention mechanism can be expressed as:

$$\tilde{V}_{SAA} = G_{SAA}(V) = \left[\delta(t_{11})v^{1,1}, \cdots, \delta(t_{ij})v^{i,j}, \cdots, \delta(t_{FQ})v^{F,Q}\right] \quad (14)$$

After re-calibration of $\delta(t_{ij})$, the region positions not highly correlated with the target recognition task are eliminated, highlighting the region positions strongly corelated with the target.

The channel domain and spatial domain attention mechanisms are jointly introduced to the proposed network, i.e., the features solved by the two attention mechanisms are superimposed channel by channel and pixel by pixel:

$$\tilde{V}_{MAA} = \tilde{V}_{SAA} + \tilde{V}_{PAA} \quad (15)$$

If the input at position (i,j,l) of image target feature map V is important in both attention mechanisms, it would be assigned a high activation value by our model. Figure 3 displays the hybrid attention mechanism.

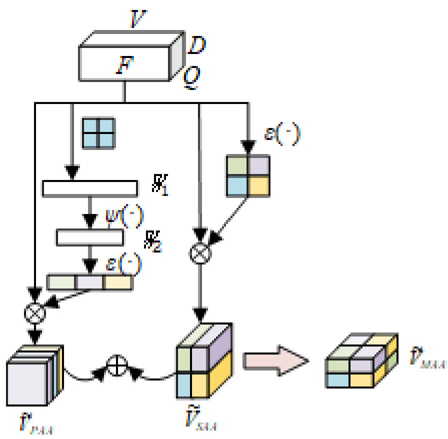


Figure 3. Hybrid attention mechanism

4. EXPERIMENTS AND RESULTS ANALYSIS

The configuration of relevant parameters affects the final target recognition accuracy of images. Hence, this paper carries out four experiments on the four parameters involved in model calculation, including the number of centers in the set of image features in the nearest regions, the weight coefficient ξ that balances a_i and a_d, the weight coefficient μ that balances K_r and K_{kt}, and the weight coefficient μ_2 that balances $1/K_x+\alpha$ and K_{kt}. The experimental results are shown in Figure 4.

The results of the first experiment are displayed in Figure 4 (1). During the first experiment, ξ, μ, and μ_2 were fixed, while the number of centers in the set of image features in the nearest regions was changed. The optimal target recognition accuracy was achieved at around 30 centers.

The results of the second experiment are displayed in Figure 4 (2). During the second experiment, the number of centers, μ, and μ_2 were fixed, while ξ was changed. The target recognition accuracy peaked at ξ=0.9.

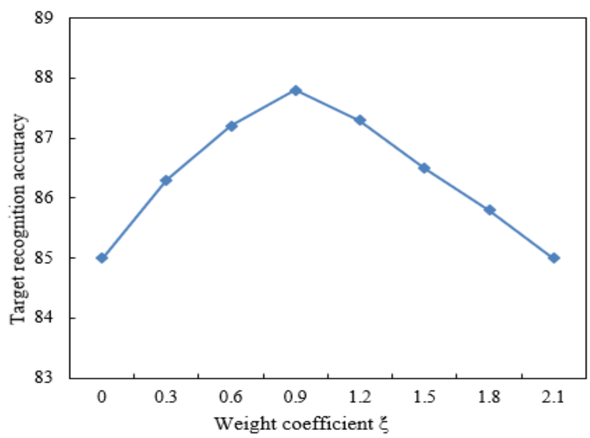
The results of the third experiment are displayed in Figure 4 (3). During the third experiment, the number of centers, ξ, and μ_2 were fixed, while μ was changed. The highest target recognition accuracy was observed at $\mu=1\times10^{-6}$.

The results of the fourth experiment are displayed in Figure 4 (4). During the fourth experiment, the number of centers, ξ, and μ were fixed, while μ_2 was changed. The best target recognition accuracy was observed at μ_2=0.3.

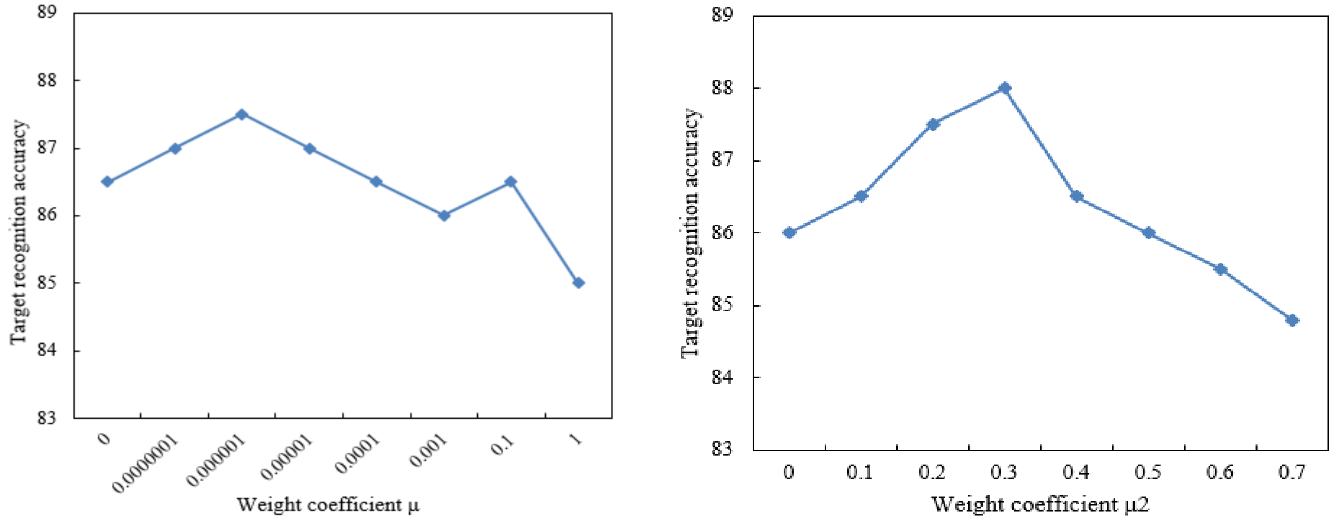
To verify the effectiveness of the hybrid domain attention mechanism in our model, this paper carries out training and verification on a self-developed sample set of surveillance video images. Figures 5 and 6 show how the loss function of local feature aggregation and Dice coefficient of our model varies with the number of iterations in training. It can be seen that the loss and Dice coefficient both tended to be stable, when the number of iterations reached 800. This may be attributed to the insufficiency of samples.

Our CNN extracts multiregional features based on the loss function of local feature aggregation. The weighted mean and direct mean of the target recognition accuracy of our model were 90.21% and 82.6%, respectively. To demonstrate its superiority, our model was compared with several other typical models, in terms of recognition accuracy. The results are shown in Table 1. The contrastive models are U-Net based on multi-scale and attention mechanism (MA-UNet), shape attentive U-Net (SAU-Net), attention-based nested U-Net (ANU-Net), multi-region ensemble CNN (MRE-CNN), Attention-UNet, three-dimensional U-Net (3D U-Net), three-dimensional volumetric fully convolutional neural network (3D V-net), multi-channel fusion CNNs (MCF-CNNs), no-new-Net (nnU-Net), and three-dimensional universal U-Net (3D U2-Net).

(1) Results of the first experiment



(2) Results of the second experiment



(3) Results of the third experiment

(4) Results of the fourth experiment

Figure 4. Target recognition accuracies at different parameter settings

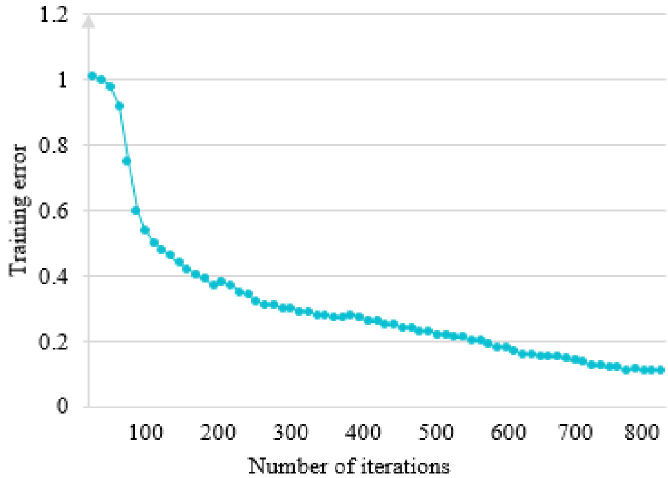


Figure 5. Loss curve of our model

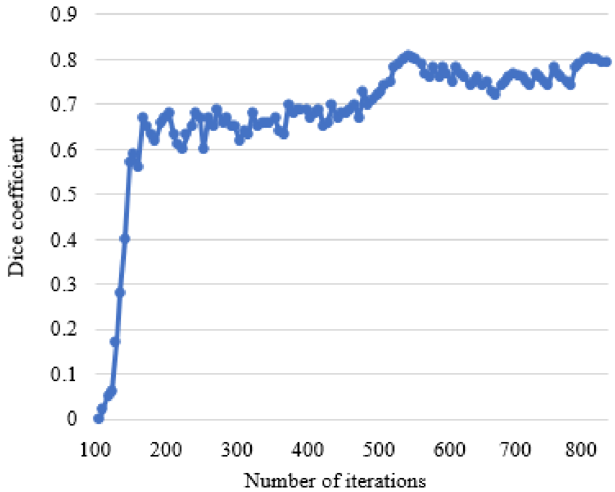


Figure 6. Dice coefficient curve of our model

Table 1. Recognition accuracies of different models

Model	Weighted average (%)	Direct average (%)
MA-UNet	82.61	45.02
SAU-Net	81.35	51.92
ANU-Net	80.49	73.35
MRE-CNN	83.62	71.81
Attention-UNet	84.14	79.62
3D U-Net	85.92	73.95
3D V-net	83.06	70.62
MCF-CNNs	87.19	73.68
nnU-Net	86.73	75.13
3D U2-Net	85.64	75.6
Our model	90.21	82.6

Table 2. Confusion matrix of 7 classes of expressions

	Cars	Bikes	Adults	Kids	Trees	Buildings	Others
Cars	87.48	1.62	1.95	2.51	0.94	1.69	4.18
Bikes	22.63	59.17	3.2	5.37	6.85	4.72	4.94
Adults	2.2	3.6	63.29	9.58	12.63	6.31	7.41
Kids	0.48	0.19	0.72	48.52	0.37	0.48	2.68
Trees	0.39	0.45	2.81	4.15	65.64	0.63	7.05
Buildings	2.84	1.96	5.38	7.31	1.57	46.91	5.12
Others	1.82	1.03	2.95	3.08	4.39	0.16	4.25

As shown in Table 1, our model achieved higher weighted mean and direct mean of target recognition accuracy than the other 10 models. Finally, the confusion matrix of different classes of image targets is listed in Table 2.

As shown in Table 2, the highest recognition accuracy (87.48%) was achieved on cars, and the lowest (4.25%) was achieved on others. The difference in target recognition accuracy mainly comes from the size imbalance between different types of targets in the sample set of surveillance video images. The number of appearances for the targets of the others class peaked at 546, and that for the targets of the kids and buildings classes peaked at 1,487. Both are much smaller than the number of appearances for the targets in the other four classes. Meanwhile, almost 17.24% the “others” targets were misidentified as trees and buildings in the image background. This is because the local similarity between the “others” targets and the “trees” and “buildings” targets.

5. CONCLUSIONS

Based on hybrid attention mechanism, this paper develops a method for the image target recognition based on multiregional features. Firstly, a CNN was constructed for extracting multiregional features based on the loss function of local feature aggregation. There are three independent CNN modules in the network, which are responsible for extracting the global multiregional features and the local features of different regions. After that, the authors embedded the channel domain attention mechanism and spatial domain attention mechanism in the proposed CNN, such that the model can recognize targets more accurately, without greatly increasing the computing load. Next, this paper carries out four experiments on the four parameters involved in model calculation. The experimental results verify the influence of the parameter configuration on the final recognition accuracy of image targets, and show the correct values of these parameters. In addition, the authors tested how the loss function of local feature aggregation and Dice coefficient of our model varies with the number of iterations in training. It can be seen that the loss and Dice coefficient both tended to be stable, when the number of iterations reached 800. Finally, the superiority of our model was confirmed through comparison with several other typical models, in terms of recognition accuracy.

REFERENCES

[1] Martynenko, A. (2017). Computer vision for real-time control in drying. Food Engineering Reviews, 9(2): 91-111. https://doi.org/10.1007/s12393-017-9159-5

[2] Mangaonkar, S.M., Khandelwal, R., Shaikh, S., Chandaliya, S., Ganguli, S. (2022). Fruit harvesting robot using computer vision. 2022 International Conference for Advancement in Technology, ICONAT 2022.

[3] Bochkarev, K., Smirnov, E. (2019). Detecting advertising on building façades with computer vision. Procedia Computer Science, 156: 338-346. https://doi.org/10.1016/j.procs.2019.08.210

[4] Rakhshan, V., Okano, A.H., Huang, Z., Castelnuovo, G., Baptista, A.F. (2022). Biomedical applications of computer vision using artificial intelligence. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/9843574

[5] Jiang, Z., Wang, L., Wu, Q., Shao, Y., Shen, M., Jiang, W., Dai, C. (2022). Computer-aided diagnosis of retinopathy based on vision transformer. Journal of Innovative Optical Health Sciences, 15(2): 2250009.

[6] Moeslund, T.B., Thomas, G., Hilton, A., Carr, P., Essa, I. (2017). Computer vision in sports. Computer Vision and Image Understanding, 159: 1-2.

[7] Le, N., Rathour, V.S., Yamazaki, K., Luu, K., Savvides, M. (2021). Deep reinforcement learning in computer vision: a comprehensive survey. Artif Intell Rev, 55: 2733-2819. https://doi.org/10.1007/s10462-021-10061-9

[8] Duke, B., Salgian, A. (2019). Guitar tablature generation using computer vision. In International Symposium on Visual Computing, 11845: 247-257. https://doi.org/10.1007/978-3-030-33723-0_20

[9] Liu, D., Teng, W. (2022). Deep learning-based image target detection and recognition of fractal feature fusion for BIOmetric authentication and monitoring. Network Modeling Analysis in Health Informatics and Bioinformatics, 11(1): 1-14. https://doi.org/10.1007/s13721-022-00355-5

[10] Chen, X., Peng, X., Duan, R., Li, J. (2017). Deep kernel learning method for SAR image target recognition. Review of Scientific Instruments, 88(10): 104706. https://doi.org/10.1063/1.4993064

[11] Wang, S., Jiao, L., Yang, S., Liu, H. (2016). SAR image target recognition via complementary spatial pyramid coding. Neurocomputing, 196: 125-132. https://doi.org/10.1016/j.neucom.2016.02.059

[12] Huan, R., Wang, C., Pan, Y., Guo, F., Tao, Y. (2016). New structure for multi-aspect SAR image target recognition with multi-level joint consideration. Multimedia Tools and Applications, 75(13): 7519-7540. https://doi.org/10.1007/s11042-015-2674-6

[13] Ding, J., Liu, H.W., Chen, B., Wang, Y.H. (2016). SAR image target recognition in lack of pose images. Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University, 43(4): 5-9. https://doi.org/10.3969/j.issn.1001-2400.2016.04.002

[14] Guo, M., Li, B., Shao, Z., Guo, N., Wang, M. (2022). Objective image fusion evaluation method for target recognition based on target quality factor. Multimedia Systems, 28(2): 495-510. https://doi.org/10.1007/s00530-021-00850-1

[15] Li, P., Li, T., Yao, Z.A., Tang, C.M., Li, J. (2017). Privacy-preserving outsourcing of image feature extraction in cloud computing. Soft Computing, 21(15): 4349-4359. https://doi.org/10.1007/s00500-016-2066-5

[16] Liu, S., Ma, J., Yang, Y., Qiu, T., Li, H., Hu, S., Zhang, Y.D. (2022). A multi-focus color image fusion algorithm based on low vision image reconstruction and focused feature extraction. Signal Processing: Image Communication, 100: 116533. https://doi.org/10.1016/j.image.2021.116533

[17] Huang, G., Sun, J., Lu, W., Peng, H., Wang, J. (2021). ECT image reconstruction method based on multi-exponential feature extraction. IEEE Transactions on Instrumentation and Measurement, 71. https://doi.org/10.1109/TIM.2021.3132829

[18] Li, Z., Qian, Y., Wang, H., Zhou, X., Sheng, G., Jiang, X. (2022). A novel image-orientation feature extraction method for partial discharges. IET Generation, Transmission & Distribution, 16(6): 1139-1150. https://doi.org/10.1049/gtd2.12356

[19] Zhang, Z., Yang, D.S. (2017). Image point feature extraction algorithm of circumferential binary descriptor. Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics, 29(8): 1465-1476.

[20] Bai, Z., Ma, S., Li, G. (2022). A WeChat official account reading quantity prediction model based on text and image feature extraction. IEEE Access, 10: 28348-28360. https://doi.org/10.1109/ACCESS.2022.3157715

[21] Eisserer, C. (2015). Portable framework for real-time parallel image processing on high performance embedded platforms. In 2015 23rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, pp. 721-724. https://doi.org/10.1109/PDP.2015.31

[22] Deng, Y., Bi, F., Chen, L., Long, T. (2010). Research on high-performance remote sensing image real-time processing system. In 2010 International Conference on Computer Design and Applications, 1: V1365-V1369. https://doi.org/10.1109/ICCDA.2010.5540848

[23] Bielski, C., Lemoine, G., Syryczynski, J. (2009). Accessible high performance computing solutions for near real-time image processing for time critical

applications. In Image and Signal Processing for Remote Sensing XV, 7477: 107-115. https://doi.org/10.1117/12.830356
[24] Grossauer, H., Thoman, P. (2008). GPU-based multigrid: Real-time performance in high resolution nonlinear image processing. In International Conference on Computer Vision Systems, pp. 141-150. https://doi.org/10.1007/978-3-540-79547-6_14
[25] Zhang, W., Xiong, Q., Shi, W., Chen, S. (2016). Region saliency detection via multi-feature on absorbing Markov chain. The Visual Computer, 32(3): 275-287. https://doi.org/10.1007/s00371-015-1065-3
[26] Zhang, S., He, H., Kong, L. (2015). Fusing multi-feature for video occlusion region detection based on graph cut. Acta Optica Sinica, 35(4): 0415001. https://doi.org/10.3788/AOS201535.0415001
[27] Li, Y.D., Lei, H., Hao, Z.B., Tang, X.F. (2017). Scene recognition based on feature learning from multi-scale salient regions. Dianzi Keji Daxue Xuebao/Journal of the University of Electronic Science and Technology of China, 46(3): 600-605. https://doi.org/10.3969/j.issn.1001-0548.2017.03.020
[28] Cheng, Z., Qu, A., He, X. (2022). Contour-aware semantic segmentation network with spatial attention mechanism for medical image. The Visual Computer, 38(3): 749-762. https://doi.org/10.1007/s00371-021-02075-9
[29] Zhao, D., Wang, C., Gao, Y., Shi, Z., Xie, F. (2021). Semantic segmentation of remote sensing image based on regional self-attention mechanism. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. https://doi.org/10.1109/LGRS.2021.3071624