Research on Moving Target Behavior Recognition Method Based on Deep Convolutional Neural Network

Jian-fang Liu[1], Hao Zheng[1], and He Peng[2(✉)]

[1] Department of Computer and Software, Pingdingshan University, Pingdingshan 467000, China
babygongzuo@126.com
[2] Center of Engineering Practice Training, Tianjin Polytechnic University, Tianjin 300387, China
wwzzmmhhhh@126.com

Abstract. In order to solve the problem that the average recognition degree of moving target line is low by the traditional method of moving target behavior recognition. Therefore, a motion recognition method based on deep convolutional neural network is proposed. Construct a deep convolutional neural network target model, and use the model to design the basic unit of the network. The returned unit is calculated to the standard density map by the set unit, and the moving target position is determined by the local maximum method to realize the moving target behavior recognition. The experimental results show that The experimental results of the multi-parameter SICNN256 model are slightly better than other model structures. And the average recognition rate and the recognition rate of the moving target behavior recognition method based on deep convolutional neural network are higher than the traditional method, which proves its effectiveness. Since a single target is more frequent than multiple recognitions and there is no target similar recognition, similar target error detection cannot be excluded.

Keywords: Convolutional neural network · Moving target · Recognition · Depth

1 Introduction

The moving target recognition means that the computer simulates an eye to retrieve the target object of interest in the image. The recognition of the moving target is to complete the judgment of the target category and the calibration of the location of the target, which is a basic visual processing task for one, but it is very difficult for the computer [1]. An image is converted into a group after it is entered into the computer. The binary value of 0–255, the computer should abstract the high-level semantic information of the target category from this set of data, and determine the location of the target. The target will show different degrees of deformation due to the influence of angle of view, illumination, occlusion between objects and self-occlusion, noise, etc., which increases the difficulty of recognition of moving targets. Although there are

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 348, pp. 301–312, 2021.
https://doi.org/10.1007/978-3-030-67874-6_28

many difficulties in moving target recognition, it is the first step for the computer to "see the world" to handle advanced visual tasks [2]. Therefore, moving target recognition is of great significance in the field of computer vision and practical applications. Moving target recognition, also known as target extraction, combines the segmentation and recognition of the target to achieve the purpose of finding the target and identifying the target in the image. The speed and efficiency of moving target recognition is a very important evaluation criterion for the recognition system. Especially in complex scenes, when multiple targets are identified and processed, the target recognition ability becomes more important [3]. The research focuses on the research and development of moving target recognition methods based on deep convolutional neural networks. Through the analysis and comparison of these research work, the status quo of the development of moving target recognition is summarized, and some forward-looking research directions in moving target recognition are proposed.

2 Design of Moving Target Behavior Recognition Method Based on Deep Convolutional Neural Network

2.1 Deep Convolutional Neural Network Target Model

The moving target recognition method based on the deep convolutional neural network extracts the target features through the convolutional neural network. If the convolutional neural network is too shallow, its recognition ability is often inferior to that of ordinary SVM and boosting; If the convolutional neural network is deep, a large amount of data is needed for training, otherwise there will be over-fitting in the learning [4].

The moving target behavior recognition method of deep convolutional neural network has powerful characterization and modeling capabilities [5]. Through supervised, semi-supervised or unsupervised training methods, it is possible to learn the feature representation of the target layer by layer and automatically, and realize the abstraction and description of the object hierarchy. The moving target recognition process based on the deep convolutional neural network is shown in Fig. 1. The input image is subjected to pre-measurement, standardization, etc. Then the image is input into the deep convolutional neural network model. The convolutional neural network learns the target feature and the location of the target from a large amount of input data, and finally determines the category by softmax and other methods [6].

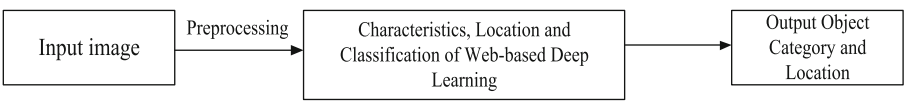

Fig. 1. Moving target detection framework based on deep convolution neural network

The advantage of the moving object recognition method based on deep convolutional neural network is that it can learn features from a large amount of data, and the

learned features are robust and have a strong generalization ability, which is very important for moving target recognition.

2.2 Network Basic Unit Design

The basic unit setting of the network is an important part of the method. By increasing the width and depth of the network, the network structure adjusts the parameters required for training and mitigates over-fitting problems [7]. In surveillance video or pictures, because of the distance, the height of the lens and the difference in the target, it usually causes a large difference in the size of the moving target. Therefore, it is necessary to extract features using three convolution layers of different convolution kernel sizes. The basic unit structure is shown in Fig. 2.

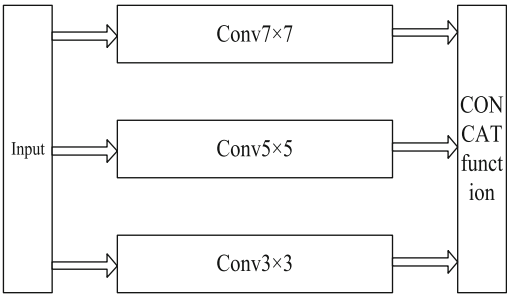

Fig. 2. Network basic unit

The input feature maps are extracted by three convolution layers of different convolution kernel sizes, and the obtained feature maps are stitched by Concat function to obtain a new feature map [8]. In general, the near-point target diameter is about 10, and the far point is about 5. Considering that the pooling layer is used for down-sampling operation, the convolution kernel size is selected by 7×7, 5×5, and 3×3 for feature extraction. Among them, the 7×7, 5×5 and 3×3 convolution kernels respectively extract the features of the moving targets of the near, middle and far levels, and the extracted feature maps are stitched by Concat function to obtain the feature map of the layer [9].

To use the convolutional neural network to identify the moving target behavior, first create a model and set the parameters of each layer of the deep convolutional neural network target model. The model includes a data layer, a feature extraction layer, an activation layer, a loss layer, and the like. In order to extract multiple features faster and better, the Inception structure is used as the basic structural unit of the network to extract different scale features. And after the convolutional layer, the largest pooling layer is added to downsample, and the multi-convolution kernel features at more scales are obtained [10]. Therefore, two models of convolutional neural network model based on serial Inception structure (Serial-Inceptions Based Convolutional Neural Network, SICNN) and convolutional neural network model based on composite

Inception structure (Multi-Inceptions Based Convolutional Neural Network, MICNN) are proposed. The SICNN model and parameters are shown in Fig. 3.

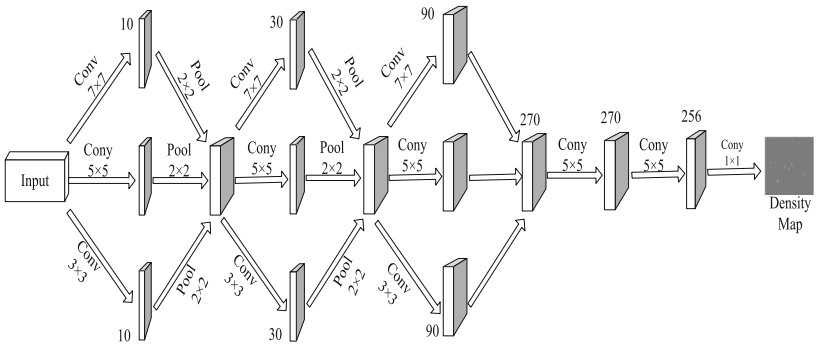

Fig. 3. SICNN_256 model

In the SICNN_56 model, the input data is input into the first Inception structure, and the convolution kernels are convolutional layers of 7 × 7, 5 × 5, and 3 × 3, respectively. And 10 feature maps are exported respectively. The 30 feature maps are integrated into 30 new feature maps by Concat function after 2X2 and the maximum pooling process with step size 2 [11]. The 30 feature maps and the first Inception structure process enter the second and third Inception structures, respectively, to obtain 90 and 270 feature maps, respectively. In this process, for example, the input image of pixel 480 × 640 is extracted by three convolution kernels of 7 × 7, 5 × 5, and 3 × 3 at three scales of 480 × 640, 240 × 320, and 120 × 160, respectively. It extracts information from different dimensions at different scales and obtains richer features. Then, the feature is further extracted by two convolution layers of 5 × 5 convolution kernel size; Finally, a 120 × 160 size density map is output through a 1 × 1 convolution kernel size convolutional layer. The MICNN_56 model and parameters are shown in Fig. 4.

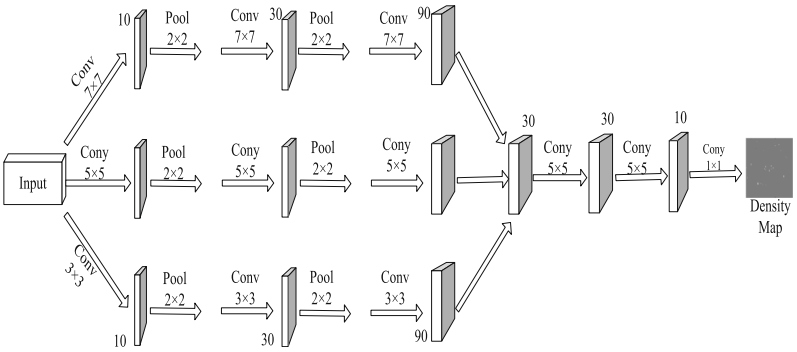

Fig. 4. MICNN_56 model

In the MICNN_56 model, there is only one Inception structure, and the depth is increased while keeping the width of the Inception structure unchanged [12]. At the same time, in order to eliminate the influence of parameters on the network structure, the number of convolution kernels in the MICNN_256 model is the same as that of the SICNN_56 model. The input data is input into the Inception structure, and is divided into three branches for operation. The first branch passes through the convolution layer with a convolution kernel size of 7 × 7, and outputs 10 feature maps. The feature map is subjected to a maximum pooling process of 2 × 2 and a step size of 2, and then a convolution layer having a convolution kernel size of 7 × 7 is output, and 30 feature maps are output [13]. The feature map is subjected to a maximum pooling process of 2 × 2 and a step size of 2, and then a convolutional layer with a convolution kernel size of 7 × 7 is output, and 90 feature maps are output. The second and third branches are similar to the first branch, but the convolution kernels are 5 × 5 and 3 × 3, respectively, and the three branches output 90 characteristic maps. In addition, in order to further explore the influence of convolution kernel parameters on the network, the structure and parameters of another set of experimental models SICNN_10 and MICNN_10 are designed when the network structure is unchanged and the number of convolution kernels in the network is changed. They are shown in Fig. 5 and Fig. 6, respectively.

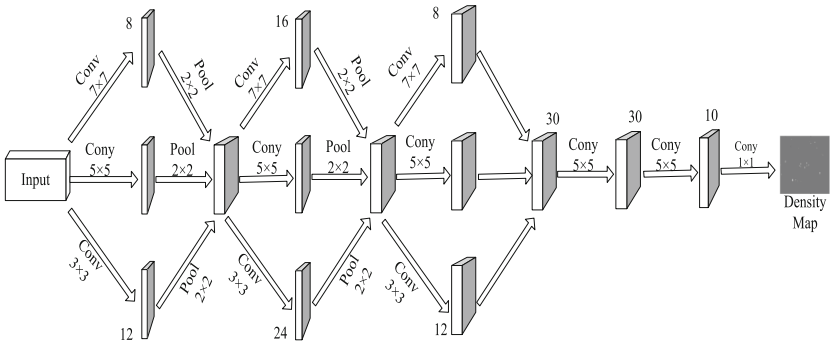

Fig. 5. SICNN_10 model

The activation layer activates the input data, that is, a function transformation, which is performed on an element-by-element basis. Commonly used activation functions are Sigm oid function, Tanh function, ReLU function, etc. The most commonly used function is ReLU function.

A linear rectification function, also known as a modified linear unit, usually refers to a nonlinear function represented by the ramp function $f(x) = \max(0, x)$ and its variants.

The loss layer calculates the loss value by calculating the difference between the training sample output and the real sample value. At present, there are three main Loss layer loss functions: Sigmoid, Softmax and Euclidean. Where Sigmoid is mainly used

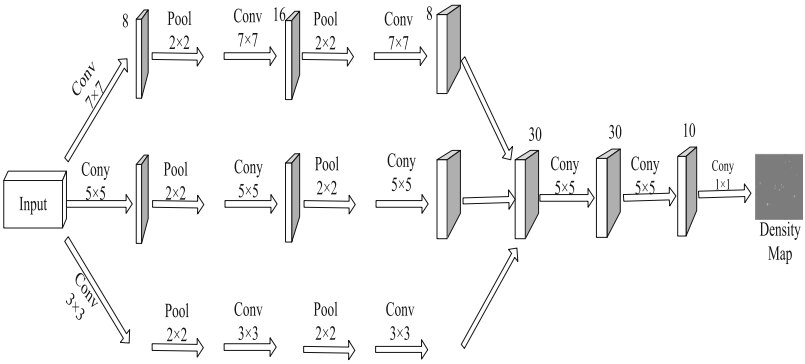

Fig. 6. MICNN_10 model

for the two-classification problem, Softmax can be used for the multi-classification problem, and Euclidean is the commonly used loss function for linear regression. The designed network structure model uses the Euclidean loss function as the Loss layer to return the density map of the network output to the standard density map. The loss function is:

$$L = \frac{1}{2N} \sum_{i=1}^{N} \left\| F(X_i) - D(X_i) \right\|_2^2 \tag{1}$$

In the formula (1), N is the number of training pictures, X_i represents an input image, D represents a label density map corresponding to standard data, and F represents a density map generated by a network structure; L is the calculated loss value, and the network judges according to the magnitude of the loss value and feedbacks the relevant parameters to obtain better experimental results.

2.3 Implementation of Moving Target Behavior Recognition

After obtaining the estimated density map output by the convolutional neural network model, the local maximum value is extracted from the estimated density map, and the extracted local maximum is the position where the target of the image is likely to be located. In the process of target behavior recognition, the density map needs to be denoised, and the target should be located within a certain range according to the target size. Because the size of the target is different due to the influence of the height of the lens and the distance from the target to the lens, the positioning should be selected according to certain classification measures; Due to the complexity of the background and the degree of light and darkness, the numerical values of the target points in the estimated density map obtained by learning are not the same. It is necessary to use a histogram to set a threshold to remove the background and remove the wrong target point to perform the target positioning. The specific operation process is shown in Fig. 7.

Fig. 7. The flow chart of motion target behavior recognition algorithm

Degree histogram, calculate the pixel value size P with the highest proportion, and set a certain threshold value T, the pixel point whose pixel value $F_k(\mathrm{x},\ \mathrm{y})$ is smaller than $P+T$ in the estimated density map is defaulted to the image background, and the interference term is removed, and is set to 0, and the estimated density map $D_k(\mathrm{x},\ \mathrm{y})$ of the background is obtained, as shown in Eq. 2:

$$D_{\mathrm{k}}(\mathrm{x},\mathrm{y}) = \begin{cases} 0, & F_{\mathrm{k}}(\mathrm{x},\mathrm{y}) < P+T \\ F_{\mathrm{k}}(\mathrm{x},\mathrm{y}), & F_{\mathrm{k}}(\mathrm{x},\mathrm{y}) \geq P+T \end{cases} \tag{2}$$

Selecting a moderately sized sliding window $M_k(\mathrm{x},\mathrm{y})$ selects the target point by local maximum in the estimated density map $D_k(\mathrm{x},\ \mathrm{y})$ of the removed background. The maximum point is set to 0, and the estimated position map $R_k(\mathrm{x},\ \mathrm{y})$ of the target is obtained, as shown in Eq. 3.

$$R_{\mathrm{k}}\ \mathrm{x}\ \mathrm{y} = \begin{cases} 0,\ D_{\mathrm{k}}\ \mathrm{x}\ \mathrm{y} < \max\ M_{\mathrm{k}}\ \mathrm{x}\ \mathrm{y} \\ 255\ D_{\mathrm{k}}\ \mathrm{x}\ \mathrm{y} = \max\ M_{\mathrm{k}}\ \mathrm{x}\ \mathrm{y} \end{cases} \tag{3}$$

The target point obtained by using the local maximum value sometimes has the same value of two similar points while retaining the information of the two position points, thereby causing adhesion, resulting in a re-inspection when the target is marked. In order to avoid such things, the obtained target center point position coordinate map will eliminate some wrong target points according to the law of two points, in order to obtain better recognition results. First, input the labeled coordinates and estimated coordinates of the test set; secondly, calculate the distance between each point in the estimated coordinate data and each point in the labeled coordinate data; Again, find the nearest point in the estimated coordinates by the labeled coordinates, and set the nearest distance threshold as the search range; finally, retrieve the target.

3 Simulation Experiment

3.1 Experiment Data

The algorithm is validated using the Mall data set. The Mall dataset contains different target densities and lighting conditions and is widely used in target counting work. The dataset uses a surveillance camera to collect data. It is a continuous video sequence. The video sequence in the dataset consists of 2000 frames of 640 × 480 color images, which are tagged with more than 60,000 moving targets. A total of four experiments were performed, including SICNN_256 model, SICNN_10 model, MICNN_256

Fig. 8. Recognition results

model, and MICNN_10 model. Each model targeted the test images with 27 real targets, and the results are shown in Fig. 8.

Figure 8(a) shows the target recognition result of the SICNN_256 model, in which 31 target targets are estimated, 23 targets are correctly identified, 8 targets are misidentified, and 4 targets are missing. Figure 8(b) shows the target recognition result of the SICNN_10 model, in which 32 targets are estimated, 22 targets are correctly identified, 10 targets are misidentified, and 5 targets are missing. Figure 8(c) shows the target recognition result of the MICNN_256 model, in which 36 targets are estimated, 23 targets are correctly identified, 13 targets are misidentified, and 5 targets are missing. Figure 8(d) shows the target recognition result of the MICNN_10 model, in which 31 targets are estimated, 21 targets are correctly identified, 10 targets are misidentified, and 6 targets are missing.

3.2 Average Literacy Rate and Comparison Rate

In this experiment, the training time of the less parameter network is 0.39 s/5 times, and the training time of the network with more parameters is 1.68 s/5 times.

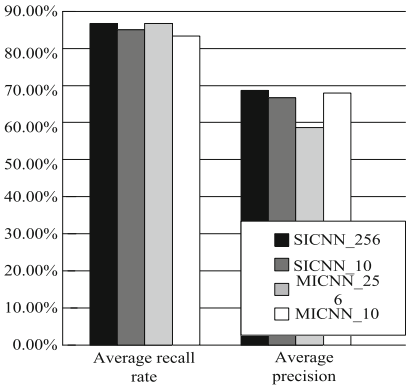

Fig. 9. Comparison of average recall rate and precision rate

SICNN_256 model, SICNN_10 model, MICNN_256 model, MICNN_10 model 4 groups of experiments on the average recognition accuracy and time-consuming situation of 400 test set pictures under the same number of iterations, as shown in Fig. 9.

Through comparison between different experiments, it can be seen from Fig. 9 that the experimental results of the multi-parameter SICNN256 model are slightly better than those of other model structures. However, the model structure has a large amount of computation and takes a long time, and performance and timeliness cannot be simultaneously provided.

At the same time, in the test set, select some pictures to compare the recognition results, and compare the results of the full rate and the identification rate parameters, as shown in Fig. 10, respectively:

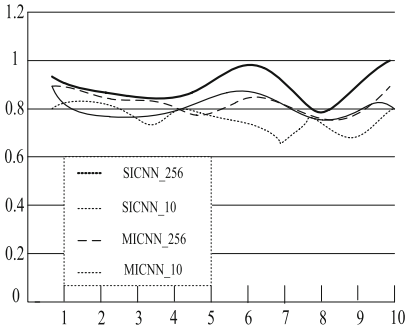

Fig. 10. Test set picture recognition rate

Based on the comprehensive experimental data, it can be concluded that the omission recognition caused by the occlusion situation in the test picture and the decrease in the recognition rate caused by the excessive number of samples are the main reasons for the poor experimental results. It can be seen from Fig. 10 that the missed detection of the sample mostly occurs in the area with more obstructions or edges, and the misdetection of the sample mostly occurs on the same target body or the like. Since the individual is replaced by a single head as the recognition target, although the recognition number is accelerated, since there is no further recognition judgment. Therefore, the single target is recognized more frequently than many times, and since there is no target similar recognition, the similar target false detection cannot be excluded.

3.3 Comparative Experiment

In order to verify the effectiveness of the proposed motion target behavior recognition method based on deep convolution neural network, a comparative experiment was conducted. The method based on spatiotemporal semantic information and the method based on intelligent video analysis were selected as the experimental comparison methods. The recognition accuracy and recognition time were selected as the experimental indicators. The results are as follows.

(1) Recognition Accuracy

Three methods are selected to test the recognition accuracy, and the results are shown in Fig. 11.

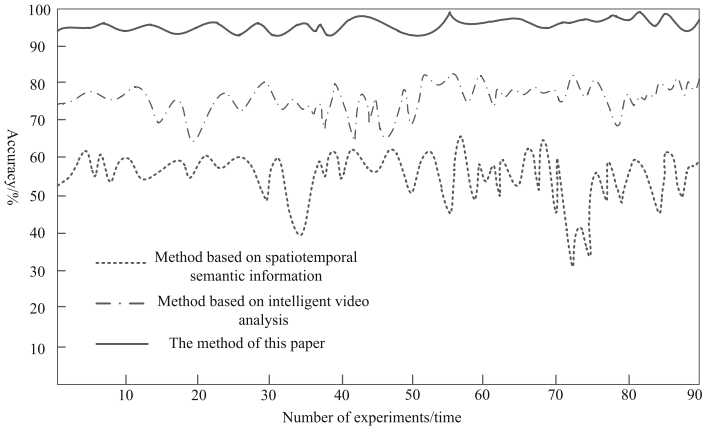

Fig. 11. Identification accuracy comparison

Analysis of Fig. 11 shows that, compared with the experimental comparison method, the recognition accuracy of this method is more than 94%, which shows that the method can accurately identify the behavior of moving targets.

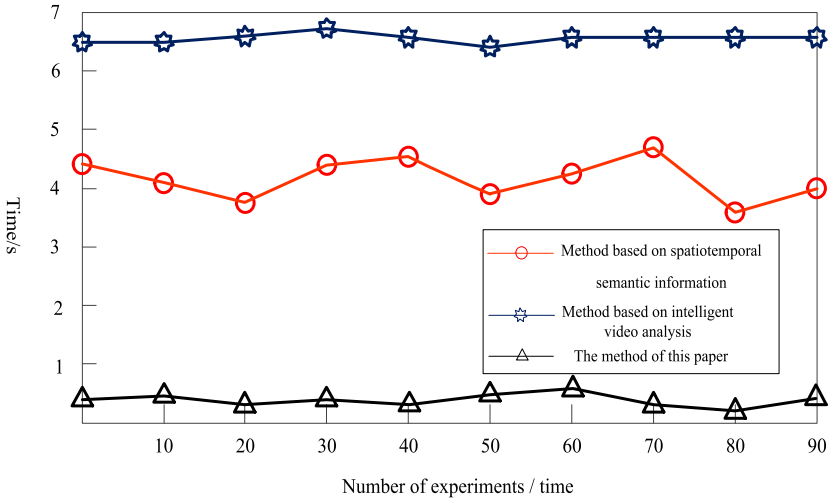

Fig. 12. Identification time comparison

(2) Identification Time

Three methods are selected to test the recognition time, and the results are shown in Fig. 12.

Analysis of Fig. 12 shows that the recognition time of this method is always below 0.9 s, which is far lower than the experimental comparison method, which shows that the method can realize fast recognition of moving target behavior.

4 Conclusion

In order to solve the problem that the average recognition degree of moving target line is low in the traditional method of moving target behavior recognition, a motion recognition method based on deep convolution neural network is proposed. The method mainly uses the deep convolution neural network target model to realize the behavior recognition of moving objects. The experimental results show that the method has excellent performance and can be further applied in practice.

Acknowledgments. This work is supported by Henan provincial department of science and technology planning project social development (NO.182102310040), and Pingdingshan University Youth Scientific Research Fund Project (PXY-QNJJ-2018005).

References

1. Liu, Z., Huang, J., Feng, X.: Constructing behavior recognition model of multiscale depth convolution neural network. Opt. Precis. Eng. **25**(3), 799–805 (2017)
2. Tang, Z., Zhang, K., Li, C., et al.: Motion imagination classification based on deep convolution neural network and its application in brain-controlled exoskeleton. J. Comput. Sci. **40**(6), 1367–1378 (2017)
3. Zhou, T., Xu, Y., Zheng, W.: Classification and recognition of tomato main organs based on deep convolution neural network. J. Agri. Eng. **33**(15), 219–226 (2017)
4. Li, C., Qin, P., Zhang, J.: Research on image denoising based on deep convolution neural network. Comput. Eng. **43**(3), 253–260 (2017)
5. Wang, Z., Min, H., Zhu, Q.: Optical flow detection of moving objects based on deep convolution neural network. Optoelectron. Eng. **45**(8), 1–9 (2018)
6. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (2020). https://doi.org/10.1007/s40747-020-00161-4
7. Yuan, G., Tang., Y., Han., W., et al.: Vehicle type recognition method based on deep convolution neural network. J. Zhejiang Univ. Eng. Edn. **17**(4), 12–25 (2018)
8. Liu, Z., Ho, D., Xu, X., et al.: Moving target indication using deep convolutional neural network. IEEE Access **6**, 1 (2018)

9. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genetics **10**, 120 (2019)
10. Zhang, Y., Li, J., Guo, Y., et al.: Vehicle driving behavior recognition based on multi-view convolutional neural network (MV-CNN) with joint data augmentation. IEEE Trans. Veh. Technol. **68**(5), 1 (2019)
11. Shuai, L., Gelan, Y.: Advanced Hybrid Information Processing, pp. 1–594. Springer, New York (2019). https://doi.org/10.1007/978-3-030-36402-1
12. Fei, G., Teng, H., Jinping, S., et al.: A new algorithm of SAR image target recognition based on improved deep convolutional neural network. Cogn. Comput. **17**(6), 1–16 (2018)
13. Liu, G., Liu, S., Khan., M., et al.: Object tracking in vary lighting conditions for fog based intelligent surveillance of public spaces. IEEE Access **6**, 29283–29296 (2018)