Image Recognition Method of Educational Scene Based on Machine Learning

Yingjian Kang(✉) and Lei Ma

Beijing Polytechnic, Beijing 100016, China
kangyingjian343@163.com

Abstract. The traditional image recognition method mainly relies on the similarity expansion calculation of the prominent features of the image to realize the image recognition. This method not only reduces the recognition accuracy of the image, but also makes the recognition efficiency of the image low due to the complex calculation process. In response to the above problems, this research designed an image recognition method for educational scenes based on machine learning. After performing normalization, denoising, and enhancement preprocessing on the educational scene image, the HOG, SIFT, and Haar features in the image are extracted. Then use the convolutional neural network model in machine learning technology to complete the recognition of educational scene images. Experimental results show that the effective recognition rate of this method is higher than 92%, and compared with traditional methods, the recognition efficiency of this method is significantly improved.

Keywords: Machine learning · Convolutional neural network · Educational scene image · Image recognition · Image enhancement · Feature extraction

1 Introduction

With the explosive growth of the number of digital images, the image processing method of using manual methods to manage and classify digital images has been difficult to meet people's needs because of its low efficiency. People urgently need a method that can automatically process and manage digital images, and computer vision task was born under this background [1].

Computer vision includes many disciplines, such as machine learning, image processing and pattern recognition. Through the powerful processing ability of the computer, it can analyze and process the obtained video images, and detect, recognize and analyze the targets in the video scene. Through the analysis and data mining of a large number of video data, the content information in video is analyzed, so as to replace some human work.

In the field of computer vision, how to use computers to realize the automatic recognition and management of digital images is an urgent problem to be solved. At present, the main development direction is to use machine learning methods to train computers

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 209–220, 2022.
https://doi.org/10.1007/978-3-031-18123-8_16

to have their own "thinking" so that they can automatically "recognize" and "understand" images. Indoor scene is more complex, which makes the research of indoor scene recognition more difficult.

The general process of scene image recognition includes feature selection and extraction and classifier learning. Among them, feature selection is the key step in the whole recognition task. The more abstract the feature, the richer its structural information and semantic information, and the stronger its expression ability. Therefore, in the scene recognition task, the feature of the target image is extracted first, and then input into the classifier for recognition. In previous studies, scene recognition methods can be divided into the following two types according to the nature of the extracted features: one is the scene recognition method based on the underlying features, and the other is the scene recognition method based on the middle and high-level semantics [2]. The above methods belong to scene recognition methods based on semantic features. In the process of feature extraction, the selection of features is artificially set by researchers, which is not only time-consuming and laborious, but also expensive. What features can accurately express scene information and achieve ideal recognition effect is still a difficult problem in the research.

In the early stage of scene recognition research, scholars often choose features such as color and texture to describe the scene. With the development of technology, related scholars have proposed the use of SIFT, HOG, GIST, CENTRIST and other features to recognize scenes. When faced with indoor scenes with complex structures, it is often impossible to accurately describe the most important features of the scene. Wang et al. used image segmentation algorithms to identify high-speed railway scenes. Sun et al. used the different importance of the information in the scene image, introduced the privilege information and attention mechanism, and realized the recognition of the special scene. However, when the above technology was applied to indoor scenes, the expected recognition result could not be obtained. Li et al. fused RGB images from multiple perspectives to realize indoor scene understanding and complete indoor scene recognition. This method had a large amount of calculation in practical application, and the efficiency of scene image recognition was low.

With the continuous development of machine learning, experts and scholars have proposed a large number of machine learning models. These models are used in different fields of machine learning according to different characteristics to solve different problems [6]. Applying machine learning to image recognition can improve the effect and accuracy of image recognition with the help of good learning.

Education scene image refers to the video image information of education and teaching collected by AR technology, which can reflect the teaching status and other information. By analyzing the education scene image, the improvement in the teaching environment can be analyzed. However, the existing image recognition methods are seldom used in educational scene images. Therefore, according to the above research background, aiming at the problem of low image recognition efficiency existing in traditional methods, this paper designs a new image recognition method for educational scenes based on the convolutional neural network model in machine learning technology.

2 Methodological Research

2.1 Image Preprocessing of Educational Scenes

The acquisition of digital images is the process of imaging objects, which refers to the process of obtaining image data in the scene through a specific imaging device and reconstructing the obtained data [7]. In this paper, the Baslel ACE series acA2500-149c industrial camera with a high-definition webcam is used to capture video in educational scenes to provide better depth images. Among them, the specific parameter information of the Basler camera is shown in Table 1.

Table 1. Specific parameters of Basler cameras

Parameter item	Specific value
Horizontal/vertical resolution	2590 × 1942px
Horizontal/vertical pixel size	2.2um × 2.2um
Frame rate	14fps
Black and white/color	Color
Interface	GigE
Pixel depth	12bits
Video output format	Mono 8, Bayer GB 1 2, YUV 4: 2: 2

Image normalization plays a very important role in processing digital images. It can eliminate the requirements of illumination and pose size for image processing. Generally, the pictures obtained by the camera contain a large part of background information, which will adversely interfere with the later image processing. In order to improve the algorithm performance of face detection and subsequent clothing segmentation in the classroom scene, the image needs to be processed as follows:

1) Image scale normalization. According to the image size obtained by the camera, in order to improve the speed of image calculation and processing and without affecting the loss of original image information, the image size is scaled to 640 × 480 pixels. In the face classifier training process, in order to segment the face region from the original image, we use the calibration software to calibrate and cut the face region of the original image, with a size of 20 × 20 pixels.
2) Noise reduction. For two-dimensional images, there are some noise points in both gray image obtained directly and gray image converted from color image. In this paper, considering the performance indicators of denoising effect and computational efficiency, median filtering is used to reduce the noise of the collected image.

The mathematical expression of median filtering for two-dimensional images is [8]:

$$y_{ij} = Mid_A\left\{f_{ij}\right\} \tag{1}$$

Among them, A represents the template formed by the neighborhood pixels selected during filtering; f_{ij} represents the pixel value at (i,j) in the two-dimensional image. Neighborhood median filtering can effectively filter out the isolated noise in the image, it is very effective for salt and pepper noise, and at the same time it can preserve the details of the edge and other information in the image to the greatest extent.

3) Image enhancement. High-pass filter can enhance image edge information and repair blurred details in the image. In this study, histogram equalization is used for image enhancement. For an image with a pixel area of A_0, it is assumed that the gray value of the image pixel is divided into N effective gray levels. If the number of pixels in each gray level of the image is A_0/N, then this image is called a histogram equalized image. Assuming that the gray level of the original image is g, the gray level histogram of the image before transformation is denoted as H_1, the gray level after transformation is denoted as G, the gray level of the transformed image is denoted as H_2, according to the principle of histogram equalization, it can be known that the transformed image has A_0/N pixels at each effective gray level. So there are the following expressions:

$$\begin{aligned} &G = F(g) \\ &\sum_{i=0}^{g} H_1(i) = \sum_{i=0}^{G} H_2(i) = G \times A_0/N \end{aligned} \tag{2}$$

On this basis, we can get:

$$G = F(g) = \frac{N}{A_0}\sum_{i=0}^{g} H_1(i) = \frac{N}{A_0}A(g) \tag{3}$$

$A(g)$ represents the cumulative function of pixel distribution, and the calculation formula is:

$$A(g) = A(g-1) + hist\left[g\right], g = 0, 1, \ldots, 255 \tag{4}$$

After preprocessing the education scene image according to the above process, the feature extraction of the education scene image is performed.

2.2 Image Feature Extraction of Educational Scenes

There are many features to be recognized in educational scene images. Based on HOG feature, SIFT feature and Haar feature, this study identifies scenes and people in the image.

The histogram can retain the edge and contour information of the image to the greatest extent, and is insensitive to light intensity and small offsets. Since the edge contour of the object is represented by the gradient distribution and the edge direction at the same time, the image can be divided into the same unit cell when extracting the HOG feature, and the gradient of all pixels in the image can be calculated. The gradient of the pixel in the image is [9]:

$$\begin{aligned} G_x(x, y) &= H(x+1, y) - H(x-1, y) \\ G_y(x, y) &= H(x, y+1) - H(x, y-1) \end{aligned} \tag{5}$$

Among them, $G_x(x, y)$ represents the horizontal gradient of the image pixel (x, y); $G_y(x, y)$ represents the vertical gradient of the image pixel (x, y); $H(x, y)$ represents the pixel value of the image pixel (x, y). Then the gradient magnitude and direction at (x, y) are:

$$G(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)} \tag{6}$$

$$\alpha(x, y) = \tan^{-1}\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \tag{7}$$

Then count the gradient histograms of each cell one by one, then cascade all the cells, and finally generate the HOG descriptor. The calculation of the HOG feature is small, and it can effectively represent the shape and contour of the target in the image within a specific range, and it is not sensitive to changes in illumination.

SIFT features include two steps: key point detection and key point feature description [10]. Among them, key point detection includes scale space extreme value detection and

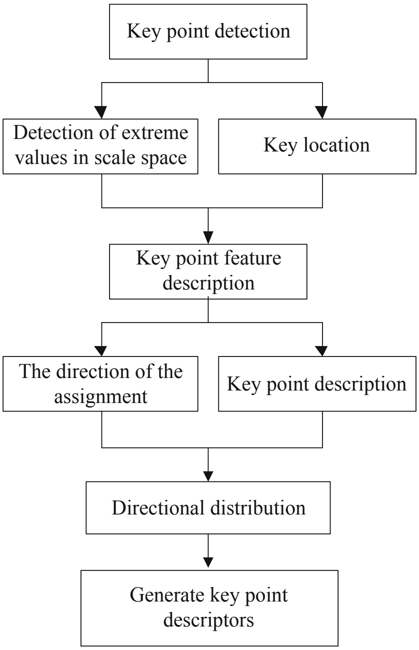

Key point detection
Detection of extreme values in scale space
Key location
Key point feature description
The direction of the assignment
Key point description
Directional distribution
Generate key point descriptors

Fig. 1. Schematic diagram of SIFT feature extraction process

key point positioning, and key point feature description includes direction assignment and key point description. The process of extracting SIFT features is shown in Fig. 1.

For the character feature extraction in the educational scene image, this paper uses Haar-like features to represent the character features in the educational scene. Figure 2 is a schematic diagram of the representation of Haar-like features.

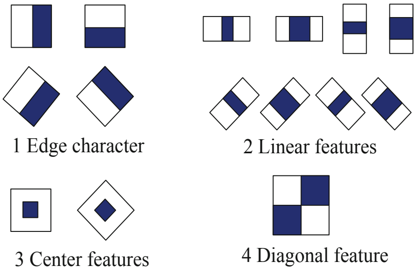


Fig. 2. Representation of Haar-like features

Haar feature is a performance feature that reflects the change of image gray level information. It is sensitive to some simple graphic structures with relatively large gray level changes, such as line segments and edge images. At the same time, it describes a structure with a specific trend [11]. By translating and telescoping the Haar feature template to change its position information and adjusting the size of the template, many rectangular area features can be obtained. Among them, the value of the rectangular feature is called the feature value. Denote the integral image of image I as I', where point (x, y) is represented as any pixel in image I, which is defined as follows:

$$I'(u, v) = \int_{x=0}^{u} \int_{y=0}^{u} I(x, y)dxdy \tag{8}$$

According to the definition of integral image, the integral value of any region can be obtained. When calculating the integral value of a rectangular area in the image, it can be calculated only by using the integral values of the current point and the three points in front of the current point. After extracting the features of education scene image, the convolution neural network in machine learning is used to realize education scene image recognition.

2.3 Using Convolutional Neural Networks to Recognize Educational Scene Images

Convolutional neural network is developed on the basis of artificial neural network. Convolutional neural network simulates the machine learning algorithm of biological visual nervous system, which consists of input layer, convolutional layer, pooling layer, full connection layer and output layer. Among them, the input layer can input 2d image data, the convolution layer and pooling layer are used to extract features, the full connection layer is used to integrate the features transferred from the convolution layer and pooling layer, and the output layer can judge the relevant images by category to achieve the goal of image recognition. The most important part of convolutional neural network is the convolutional layer and the pooling layer [12].

The output layer of the convolutional neural network is designed as a support vector machine classifier, the features are extracted through the convolutional neural network, and then the support vector machine classifier is used for classification to achieve a better image recognition effect [13]. The convolutional layer is a related operation used to extract feature surfaces in a convolutional neural network. Many abstract feature surfaces can be obtained through convolution operations. At the beginning, the input is an $M \times N$ pixel input image, and then the convolution process is performed according to the specific convolution and $P \times T$ convolution. If the sliding step size of the sliding window is O, then the size of the feature image obtained by the convolution operation is O Calculated as follows:

$$S = [(M - P)/O + 1] \times [(N - T)/O + 1] \tag{9}$$

The receptive field of the convolutional layer is to only connect the local area of the upper layer to the neurons of the next layer through the convolution kernel. It can be regarded as the first process of the image. The connection of the local area of the receptive field can significantly reduce the parameters. It can be assumed that the size of the output feature map is $R \times S$, then the number of fully connected parameters is:

$$N = M \times N \times R \times S \tag{10}$$

The number of parameters connected locally is:

$$N = P \times T \times R \times S \tag{11}$$

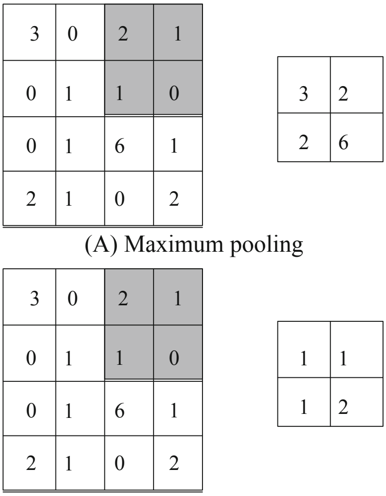
It can be seen from the comparison of the formulas that the parameter reduction is reduced from the multiple of the original image size to the multiple of the convolution kernel size. Subsequently, through weight sharing, all output neurons share the same weight, then the number of parameters is directly reduced to the number of parameters of convolution kernel, and the number of parameters shared through weight is:

$$N = P \times T \tag{12}$$

Pooling layer is the relevant operation of convolutional neural network to sample the feature surface. Through the sampling operation of pooling layer, the feature surface with smaller scale and fewer parameters can be obtained. Figure 3 is the schematic diagram of the two pooling methods selected in this paper.

The pooling layer immediately follows the convolutional layer and can play the role of secondary feature extraction. Full connection using the full connection method will cause more parameters, then the calculation of the entire network will become larger and it is prone to overfitting, which can be handled by the method of Dropout abstention. The waiver formula is:

$$r = m\left(a\left(\sum_t \omega_t v_t\right)\right) \tag{13}$$



3	0	2	1
0	1	1	0
0	1	6	1
2	1	0	2

3	2
2	6

(A) Maximum pooling

3	0	2	1
0	1	1	0
0	1	6	1
2	1	0	2

1	1
1	2

(B) Average pooling

Fig. 3. Schematic diagram of convolutional neural network pooling operation

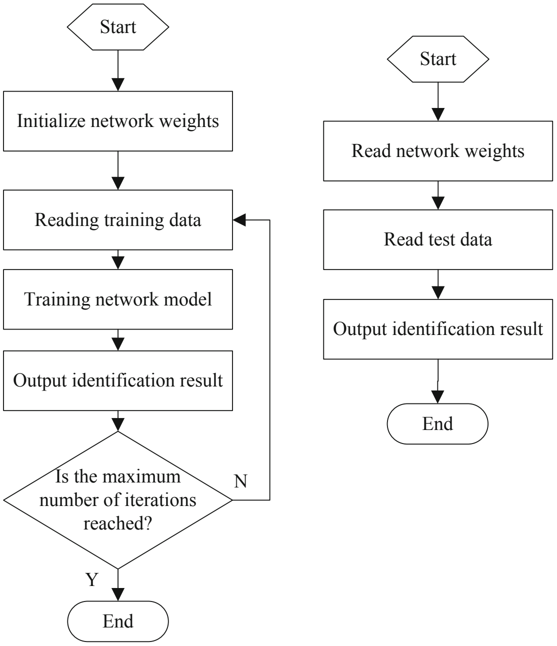
Regarding the waiver formula, consider a as the activation function, and then m as the waiver function. The value calculated by the activation function a represents the value of the neuron, and then the abstention function m represents whether the neuron is acting as an input. ω is the connection weight between the pooling layer and the fully connected layer in the convolutional neural network; v is the output weight of the convolutional layer.

In convolutional neural networks, both the S function and the R function are widely used, and the two functions have their own excellent performance. For the S function, the function is closer to the expression of the neuron and has good smoothness, but it has the disadvantages of gradient disappearance and slow convergence speed; for the R function, the function can have sparseness and faster convergence speed, which just compensates for S The lack of function, so combining the advantages of the two functions, a new activation function SR function is designed. The function formula is as follows:

$$f(x) = \max\left(\frac{1}{1+e^{-x}}\left(x+\frac{1}{2}\right)\right) \tag{14}$$

The training samples of convolutional neural network are composed of education scene images with typical characteristics, and the convolutional neural network established above is used for image recognition of the training sample set. According to the recognition results of the sample set, the parameters of the convolutional neural network are continuously adjusted to minimize the recognition error of the convolutional neural network. The specific training and testing process of convolutional neural network is shown in Fig. 4.

The parameters of neural network with minimum identification error are taken as the final identification parameters. According to the above contents, the recognition results of education scene images are output by convolution neural network, and the content


Start
Initialize network weights
Reading training data
Training network model
Output identification result
Is the maximum number of iterations reached?
N
Y
End
Start
Read network weights
Read test data
Output identification result
End

A)The training process

B)Testing process

Fig. 4. Flow chart of training process and testing process of convolutional neural network

research of education scene image recognition method based on machine learning is completed.

3 Experimental Research

In order to verify the effectiveness of the above machine-learning-based educational scene image recognition method, the following experiments are designed. In order to avoid the uniformity of experimental results, the traditional recognition method based on image segmentation algorithm is compared with the image recognition method based on feature fusion.

3.1 Experiment Content

In the experiment, the recognition method designed in this paper was taken as the experimental group, the recognition method based on image segmentation algorithm was taken as the comparison group 1, and the image recognition method based on feature fusion was taken as the comparison group 2.

From the monitoring of different types of educational places in a university, 5000 educational scene images are taken as experimental data set. All images to be recognized have the same resolution and size to avoid interference to the experimental results. Of these, 2,700 were used for machine learning training and 2,300 for comparative testing.

Due to the large difference in the number of sample data, the training and testing of the model will be affected, and the scenario category with a large number of data is prone to overfitting, while the scenario category with a small number of data has little influence on the training process. Therefore, data enhancement operations such as magnification rotation, horizontal flip and vertical flip were selected for the scene image in the experiment.

The data set is divided into training set, validation set and test set according to the ratio of 0.6, 0.2 and 0.2. The training set is used to train the network model, the validation set is used to adjust the parameters of the network model, and the test set is used to test the recognition performance of the network model. Perform unified preprocessing on the experimental images, and annotate the images manually.

On three experimental platforms with exactly the same configuration, three image recognition methods of the experimental group and the comparison group were used to recognize the images. The recognition results of the 3 groups of recognition methods are compared with the real image annotations, and the correct recognition numbers of the recognition methods under different recognition numbers are counted, and the effective recognition rate of the recognition methods is calculated. It takes time to record the recognition of 3 groups of recognition methods at the same time, so as to compare the recognition efficiency of the methods. Comprehensive analysis of the two index data, evaluation of the performance of the identification method.

3.2 Experimental Results

Figure 5 shows the comparison of the effective recognition rates of the three groups of methods.

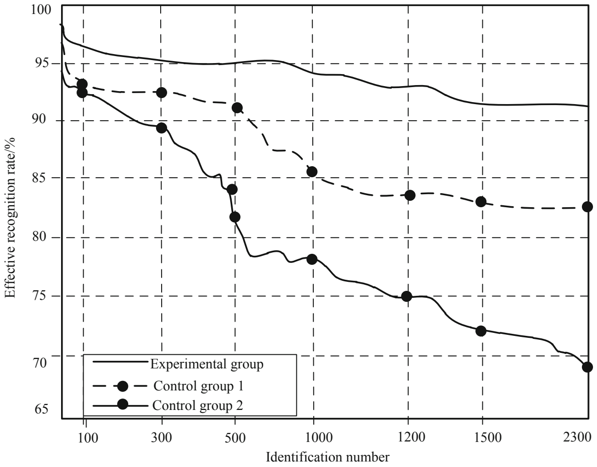


Fig. 5. Comparison of effective recognition rate of methods

Table 2 shows the identification time of the three groups of methods. The longer the time, the lower the identification efficiency.

Table 2. Recognition time-consuming comparison/s

Identification number	Test group	Comparison group 1	Comparison group 2
100	1.62	1.78	1.85
500	1.71	1.99	2.13
1000	1.84	2.07	2.54
1500	1.95	2.36	2.91
2300	2.16	2.63	3.52

According to the analysis of Fig. 5, with the increase of the number of samples to be identified, the effective recognition rate of comparison group 2 first decreased, followed by the effective recognition rate of comparison group 1. However, the effective recognition rate of control group 1 will not be reduced after it is reduced to 83%. The effective recognition rate of the experimental group method has been maintained at more than 92%.

According to the data in Table 2, under the condition of high effective recognition rate, the identification time of the experimental group method is the shortest and the identification efficiency is the highest.

To sum up, the education scene image recognition method based on machine learning designed in this paper has the advantages of high recognition accuracy and high recognition efficiency.

4 Conclusion

Image recognition is the research focus of machine learning and the foundation of machine vision. This paper studies the image recognition method of educational scenes based on machine learning. Correlative experiments prove that this method effectively improves the recognition effect of pictures. However, the amount of data in this experiment is relatively small. Since the larger the amount of data in the image library, the more essential the characteristics of things can be learned. Therefore, follow-up studies will be applied to larger-scale image recognition tasks.

References

1. Zhu, X., Fu, Y.: Development of Image–based classroom attendance system. Hubei Agric. Sci. **58**(03), 111–115 (2019)
2. Fang, G., Hu, Q., Fang, S., et al.: Face image quality evaluation in video stream and its application in classroom attendance system. Comput. Appl. Softw. **35**(10), 140–146+251 (2018)
3. Wang, Y., Zhu, L., Yu, Z., et al.: Segmentation and recognition algorithm for high-speed railway scene. Acta Optica Sinica **39**(06), 119–126 (2019)
4. Sun, N., Wang, L., Liu, J., et al.: Scene recognition based on privilege information and attention mechanism. J. Zhengzhou Univ. (Eng. Sci.) **42**(01), 42–49 (2021)

5. Li, X., Zhang, B., Sun, F., et al.: Indoor scene understanding by fusing multi-view RGB-D image frames. J. Comput. Res. Dev. **57**(06), 1218–1226 (2020)
6. Cai, X.: Automatic recognition simulation of non-orthogonal building images in complex scenes. Comput. Simul. **36**(10), 339–343 (2019)
7. Zhang, K., Hou, J.: A new image recognition method in improved convolution neural network. Sci. Technol. Eng. **20**(01), 252–257 (2020)
8. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)
9. Ma, Y., Ma, H., Wang, Y.: An image recognition method based on CD-WGAN. J. Nat. Sci. Heilongjiang Univ. **38**(03), 348–354 (2021)
10. Xia, C., Meng, Q.: Fuzzy edge recognition of ship monitoring video image based on machine learning. Ship Sci. Technol. **42**(18), 85–87 (2020)
11. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-Learning and e-Education. Mob. Netw. Appl. **26**(5), 2123–2126 (2021). https://doi.org/10.1007/s11036-021-01777-7
12. He, X., Yang, F., Chen, Z., et al.: The recognition of student classroom behavior based on human skeleton and deep learning. Mod. Educ. Technol. **30**(11), 105–112 (2020)
13. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)