Research on Behavior Recognition Method Based on Machine Learning and Fisher Vector Coding

Xing-hua Lu(✉), Zi-yue Yuan, Xiao-hong Lin, and Zi-qi Qiu

Huali College Guangdong University of Technology, Guangzhou 511325, China
luxinghua5454565@163.com

Abstract. Aiming at the problem that the existing behavior recognition method can not extract the human body interaction area, resulting in low recognition rate, a behavior recognition method based on machine learning and Fisher vector coding is proposed. Constructing artificial neural network based on machine learning, designing the main steps of backward propagation neural network, making the cost function minimum; using the depth continuity of the image to extract the foreground part of the video motion, multiplying with the corresponding 2D video frame to detect the time domain motion Behavior; Solving the dual quadratic programming problem of Fisher support vector machine, obtaining its optimal solution and completing behavior learning; segmenting the current frame image, solving the normal vector to extract the moving target, and completing the behavior recognition method based on machine learning and Fisher vector coding the study. In order to verify the effectiveness of the design method, a comparative experiment was designed. The experimental results show that the average recognition accuracy of the design method is 7.6% higher than the traditional method.

Keywords: Machine learning · Fisher vector coding · Behavior recognition

1 Introduction

Behavior recognition is a technique that uses computer vision techniques to determine whether a particular behavior exists in an image or video sequence. Machine learning has a wide range of applications, especially in the fields of intelligent video surveillance, traffic, and criminal investigation. Therefore, the behavior recognition of machine learning appearance image matching has received more and more attention. Since the two-dimensional image itself is a projection of a three-dimensional object and cannot reflect the space of the real world, an error may occur when extracting the human interaction area. Machine learning and Fisher vector coding have the characteristics of clustering, and the depth values of the same object or objects in contact with each other are continuous. The method in [1] was used to construct a risk prediction model of acute kidney injury (AKI) in severe burn patients, and the prediction performance of machine learning and logistic regression models were compared. Methods The clinical data of 157 patients who met the inclusion criteria for severe burns during the aluminum dust explosion in the “82” Kunshan plant were collected. Univariate

Y.-D. Zhang et al. (Eds.): ICMTEL 2020, LNICST 326, pp. 136–147, 2020.
https://doi.org/10.1007/978-3-030-51100-5_12

analysis was used to screen for factors that may be related to the occurrence of AKI, including patient gender, age, admission time, basic injury, initial admission score, treatment status, and mortality at 30, 60, and 90 days after injury. Mann-Whitney U test, $\chi 2$ test, and Fisher's exact probability test were performed on the data. The P value of 0.1 in univariate analysis and variables that may have clinical significance were included in the construction of the prediction model, and the logistic regression analysis and XGBoost machine learning algorithm were used to construct the AKI prediction model. Calculate the area under the receiver's working characteristic curve (AUC), and the sensitivity and specificity under the optimal threshold. This method, though, identifies clinical symptoms in patients with statistical significance. However, the sample data analyzed by it has screening indicators, so it negates the significance of individual cases and has inaccurate problems. At present, there are many designs for the first and second floors in the smart home field, but in order to realize the true sense of intelligence, in addition to obtaining information and transmitting information, it is necessary to analyze the living data. The development of smart homes can not only rely on the improvement of technology level, but also from the perspective of simple and environmental protection. Therefore, designing smart homes with harmonious interaction has more development prospects.

The existing behavior recognition method can not extract human body interaction area, resulting in low recognition rate. Based on this, this paper proposes a behavior recognition method based on machine learning and Fisher vector coding [2]. Constructing artificial neural network based on machine learning, designing the main steps of backward propagation neural network, making the cost function minimum; using the depth continuity of the image to extract the foreground part of the video motion, multiplying with the corresponding 2D video frame to detect the time domain motion Behavior; Solving the dual quadratic programming problem of Fisher support vector machine, obtaining its optimal solution, completing behavior learning; segmenting the current frame image, solving the normal vector to extract the moving target, and completing the behavior recognition method based on machine learning and Fisher vector coding the study.

2 Research on Behavior Recognition Method Based on Machine Learning and Fisher Vector Coding

The machine learning-based behavior recognition mainly includes the construction of artificial neural network, detecting the time domain motion behavior, using the coding vector to complete the learning, and finally implementing the behavior recognition based on machine learning and Fisher vector coding. The specific process of behavior recognition is shown in Fig. 1:

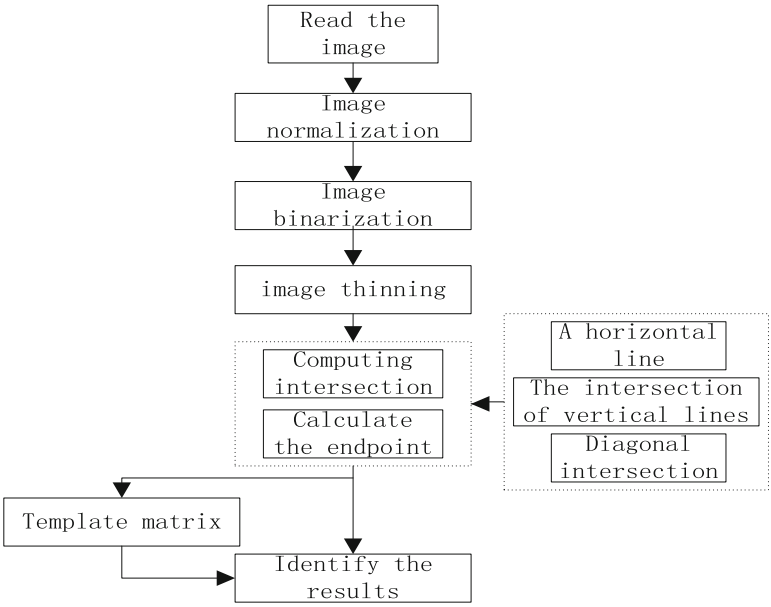

Fig. 1. Behavior recognition process

According to the above process of behavior recognition, the behavior recognition method based on machine learning and Fisher vector coding is studied.

2.1 Building Artificial Neural Network Based on Machine Learning

Machine learning is a scientific rule that uses computers to learn data and to obtain new knowledge and technology by turning data into useful information. This paper uses backward propagation, and the neural network structure is shown in Fig. 2:

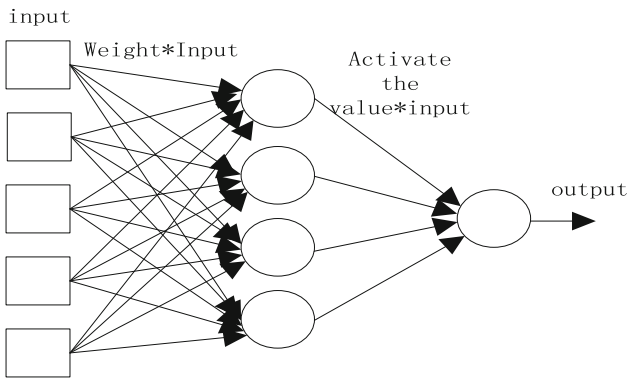

Fig. 2. Back propagation neural network structure

The artificial neural network based on machine learning is a statistical learning algorithm inspired by the biological field. A backward propagation neural network consists of the forward and backward propagation of the input layer, hidden layer and output layer [3]. In detail, backward propagation means that the input signal comes from the input layer, passes through the hidden layer, and is passed to the output layer. If the output layer achieves the expected result, the algorithm ends. Otherwise, the wrong result will propagate back to the output layer through the original connection path and loop continuously to get the error that minimizes the cost function. Before these steps, in order to calculate the activation value, an activation function is set between the input layer and the hidden layer, the hidden layer, and the output layer, as shown in Eq. (1):

$$f(x) = \frac{1}{1 + e^{-\theta x}} \tag{1}$$

In the above formula, $e^{-\theta x}$ represents the training precision during the activation process. The activation function is monotonically increasing and continuous smooth. By setting the parameters, the function can balance the linear and nonlinear relationship, and the function is biased. The calculation of the wrong backpropagation is very important. The number of hidden layers is determined by the number of input and output layers. The gradient descent algorithm is used to mediate the sum of the minimum error squares, which is a common method for training backward propagation neural networks. The training steps for the entire machine learning are shown in Fig. 3:

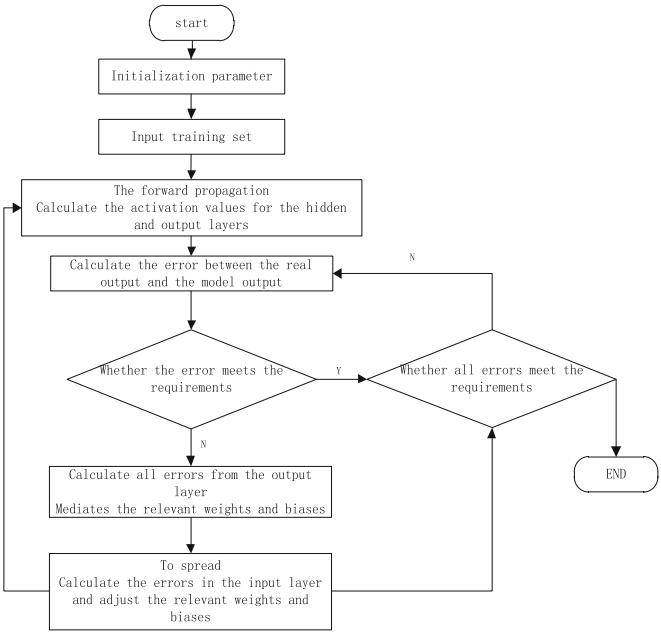

Fig. 3. The main steps of the backward propagation neural network

It is critical to choose a reasonable number of hidden layers. Generally speaking, empirical formulas that are generally accepted by the academic community can be used to obtain parameters. Weight values can be connected to two adjacent layers. The bias term is the mediation input in each layer. Useful items. So far, the construction of the structure and algorithm steps of the artificial neural network based on machine learning is completed.

2.2 Detecting Time Domain Motion Behavior

In order to effectively extract the motion template of the characters in the video stream, this paper uses the weighted cumulative frame difference method to extract the motion time domain. Since the RGB image itself is a two-dimensional projection of a three-dimensional object, in which the depth information of the object is lost, it is inevitable that the segmentation of the RGB video frame will result in over-segmentation or under-segmentation [4]. According to the depth continuity of the depth image, the foreground portion of the video motion can be roughly extracted, and then multiplied by the corresponding two-dimensional video frame to remove the background. On this basis, the weighted cumulative frame difference calculation is performed, and the interference of background noise is effectively removed by this method, and the calculation efficiency is improved. The motion area is then detected in the 2D video frame from which the background is removed. Calculate the time domain difference using the weighted cumulative frame difference method:

$$\begin{aligned} A(t) = {} & \omega_{t-n}|f_{t-n}(x,y) - |f_t(x,y)| + \\ & \omega_{t-n+1}|f_{t-n+1}(x,y) - |f_t(x,y)| + \cdots + \\ & \omega_{t-1}|f_{t-1}(x,y) - |f_t(x,y)| + \\ & \omega_{t+1}|f_{t+1}(x,y) - |f_t(x,y)| + \cdots + \\ & \omega_{t+n-1}|f_{t+n-1}(x,y) - |f_t(x,y)| + \\ & \omega_{t+n}|f_{t+n}(x,y) - |f_t(x,y)| \end{aligned} \tag{2}$$

In formula (2), $f(x, y, t)$ represents the video sequence [5] in the GRB image, and the t time frame is $f_t(x, y)$, ω_i is used to indicate the influence of the n frames before and after the tth time [9]. The value of ω_i is expressed as:

$$\omega_i = \frac{|i|^{-1}}{2\sum_{i=1}^{n} i^{-1}} \tag{3}$$

The moving target time_motion region of the time domain portion of the video can be obtained by the above formula. Although multiplication of RGB images with depth images can eliminate some of the interference of complex backgrounds, the depth maps are also affected by noise and other factors in the acquisition. At the same time, there are also parts of the background and the moving human body that have continuous depth, so this part is extracted. The boundary of the human motion area is not clear, and there is also a hollow phenomenon in some motion areas, so it is necessary to extract the human boundary image of the video static frame.

2.3 Using Fisher Vector Coding to Complete Behavior Learning

The Fisher vector uses the GMM model to model the feature points, and calculates the likelihood function gradient to represent the vector of each image according to the parameters of the model. GMM refers to the estimation of the probability density distribution of the image samples [6], usually using two Gaussian distributions are fitted simultaneously. The GMM data is composed of K Gaussian model, and the image is divided into T descriptors. It is assumed that each feature x_i of the image is independently and identically distributed. Both sides take the logarithm at the same time, and the linear combination of Gaussian distribution is used to approximate the description. p_i refers to the Gaussian distribution of feature x_i:

$$\lambda = \{\omega_i, \mu_i, \sigma_i, i = 1, \ldots, K\} \tag{4}$$

In the above formula, ω_i is the weighting coefficient of the i model, μ_i is its first-order mean, and σ_i is the second-order standard deviation. In the process of behavior recognition using Fisher vector coding, the field of pattern recognition technology is involved. The main step of adopting this method is to introduce Fisher's regularity into the support vector machine method and use it for behavior recognition to obtain high-dimensional data for small samples. A better recognition rate, and thus a better recognition of behavior. First, the behavior training video is processed to obtain the sample set:

$$\begin{gathered}\{(x_i, y_i)|x_i \in R_n, y_i \in \{-1, +1\}\} \\ i = 1, \ldots, l\end{gathered} \tag{5}$$

In the above formula, x_i indicates the behavior characteristic in a video, that is, one training sample in the n dimensional real space, and y_i is its identification, indicating that the category is not +1 or −1, and l is the number of samples. The process of obtaining samples is to stack the behaviors directly into an n-dimensional feature [7], or use a feature extraction method to extract n features for training learning. Select the regular factors λ_1 and λ_2, select the positive definite kernel function and its parameters, and calculate the kernel matrix. The role of λ_1 is to control the complexity of the hypothesis space, and the role of λ_2 is to control Fisher's regular influence. The positive definite kernel function is a kernel function that satisfies the positive definite condition and is expressed as:

$$k(x, x^{'}) = (\phi(x))^T \phi(x^{'}) \tag{6}$$

In the above formula, $x, x^{'}$ represents two learning samples, ϕ represents an implicit nonlinear mapping function, and superscript T represents the transposition of a matrix or vector, which can determine a reproducing kernel space in which space, there is a representation theorem that the hypothetical function can be expressed as:

$$\begin{aligned} f(x) &= w^T \phi(x) + b \\ &= \sum_{i=1}^{l} \alpha_i k(x, x_i) + b \\ &= K\alpha + b \end{aligned} \tag{7}$$

In the above formula, α and b are the spheroidal coefficients in the hypothesis function, and w is the weight vector, which can be calculated by the following expression:

$$w = \sum_{i=1}^{l} \alpha_i \phi(x_i) = \Psi\alpha \tag{8}$$

The dual quadratic programming problem of Fisher support vector machine can be solved [8], and the optimal solution is obtained to complete the behavior learning.

2.4 Behavior Recognition

In order to realize behavior recognition, it is necessary to segment the current frame image by spatial domain target detection, thereby extracting a moving target. At the same time, the depth image is a new data source, which embodies the depth information of objects in the image, and the depth information belonging to the same object has continuity characteristics [9]. Therefore, the depth information can be combined to effectively segment the moving target. In the depth map, the normal vector of the surface of the object can effectively represent the geometric information of the object, and the normal vector of the same object surface has only a gradual change from far and near, and a strong mutation information is generated for different objects. Using the normal vector information can effectively extract the edge features of the object. The schematic diagram of the normal vector is shown in the following figure:

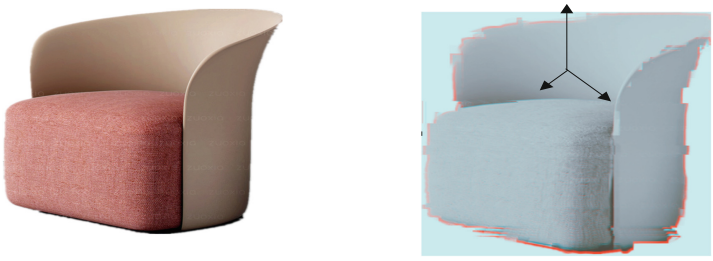

Fig. 4. Schematic diagram of surface normal vector

In Fig. 4, the right image is a depth map, and the position of any point in the figure is $P = (x, y)$, and the depth value of the coordinate value is $d(x, y)$, so each pixel in the depth map can be represented by $P = [x, y, d(x, y)]$, which is The method describes an object by combining spatial information in a three-dimensional coordinate system,

which is essentially different from the intensity value and spatial information of the RGB image. The normal vector of point $P = (x, y)$ in the depth map can be expressed in the form of a vector cross product:

$$N = S_x \times S_y \tag{9}$$

In the above formula, S_x is the horizontal vector cross of the point, S_y is the longitudinal vector cross of the point, and the two vector forks are calculated in detail to obtain the formula:

$$S_x = \frac{\upsilon}{\upsilon x}\begin{bmatrix} x \\ y \\ d(x,y) \end{bmatrix} \tag{10}$$

$$S_y = \frac{\upsilon}{\upsilon y}\begin{bmatrix} x \\ y \\ d(x,y) \end{bmatrix} \tag{11}$$

Equation (9) can be transformed to get the normal vector of the point:

$$N = S_x \times S_y = \begin{bmatrix} \frac{-\upsilon d(x,y)}{\upsilon x} \\ \frac{-\upsilon d(x,y)}{\upsilon y} \\ 1 \end{bmatrix} \tag{12}$$

After obtaining the normal vector of depth, it is also necessary to fuse the extracted edge features. In this paper, the principle of blending edges is: when a point has only θ or φ detected edges, the feature is accepted, and when there are two corner edges, the simultaneous acceptance and weighted average is adopted [10]. By detecting the edge of the object in the depth map, the resulting human contour is space_motion region. Through the obtained outline, the recognition of the behavior is completed. So far, the research on behavior recognition method based on machine learning and Fisher vector coding has been completed.

3 Experimental Study

In order to verify the effectiveness of the identification method designed in this paper, it is necessary to carry out experiments together with the traditional identification methods, and analyze and compare the experimental results.

3.1 Experimental Preparation

In this paper, the CAD-120 data set was selected experimentally. Setting up the data set consists of 4 experimenters demonstrating 10 different actions, each performed by a different experimenter 3 times. Actions include taking medicine, picking up things, eating, cleaning things, etc. The experimenter performs the same activities with

different objects, and the same action of different experimenters also has a background change. The specific parameters of the experiments that need to be used are shown in Table 1:

Table 1. Experimental parameters

The serial number	Project	Parameter
1	Operating system	Windows 7.0
2	Processor	Intel core i5
3	System environment	8 G Lenovo V310
4	Software platform	Matlab2015(a)

Under the above experimental preparation, the experiment was carried out by the conventional method and the method designed in the present paper.

3.2 Experimental Method

In order to improve the credibility of the experiment, the experiment is divided into three parts: In test 1, for each action, one of each experimenter's demonstration is selected as the training set, and the rest is used as the test set. In Test 2, for each action, select two of each experimenter's presentation as the training set, and the rest as the test set. In Test 3, for each action, the actions demonstrated by two experimenters are selected as the training set, and the actions of the remaining two experimenters are used as the test set. For each of the above tests, each test was repeated 100 times, and the confusion level table was obtained, and the recognition rate was analyzed according to the confusion level table, and the accuracy of each test was averaged.

3.3 Experimental Results and Analysis

Under the above experimental environment and experimental method, the experimental results of the two methods in the three tests are obtained. In order to more accurately and intuitively reflect the accuracy of the identification of the two methods, it is necessary to pass the Matlab2015(a) software to the test. Identify the situation for analysis. Due to the similar situation in the test, it is necessary to analyze the degree of confusion between the two methods, and the results of the confusion of the results are shown in Figs. 5 and 6:

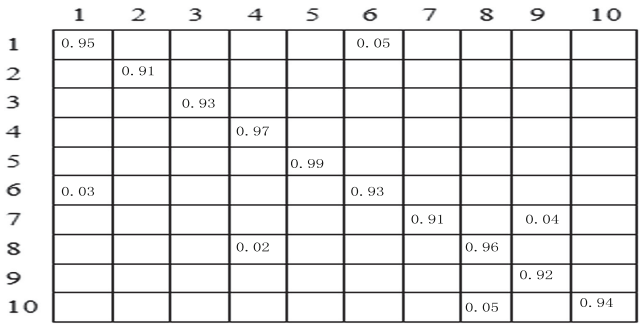

Fig. 5. Action confusion map of the method

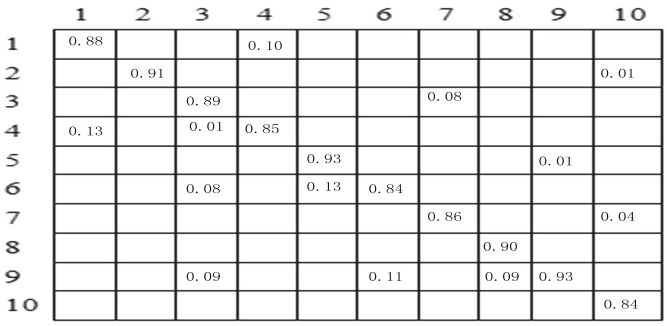

Fig. 6. Traditional method of action confusion map

In the above two figures, the 1–10 actions are: 1 for planting cereals, 2 for taking medicine, 3 for stacking items, 4 for split piles, 5 for microwave food, 6 for picking things, 7 for cleaning items, 8 For eating, 9 is for items, and 10 for eating. By analyzing the two confusion degree maps, the average recognition rate of the two methods is obtained, as shown in Table 2:

Table 2. Average recognition rate of the two methods

The serial number	Methods of this paper	The traditional method
Test 1	91.3%	90.3%
Test 2	95.1%	86.4%
Test 3	82.6%	69.7%
Average recognition rate	89.7%	82.1%

In this test, action 1 (planting grain) and action 6 (squatting things) are similar, action 2 (medication) and action 8 (eat) are similar. Table 1 shows that the method is identified in test 2. The highest accuracy rate is the lowest in the test three. In Test 2, there are more data sets for training, so the accuracy rate is higher than that of the test. In Test 3, the data sets for training and testing are respectively demonstrated by

different experimenters, so there are intra-class changes of the same action. Larger, resulting in lower accuracy. The average recognition accuracy of this paper is 89.7%, and the average recognition accuracy of traditional methods is 82.1%, which verifies the effectiveness of the proposed method.

4 Conclusion

The existing behavior recognition method can not extract the human body interaction area, resulting in low recognition rate. Therefore, the behavior recognition method based on machine learning and Fisher vector coding is proposed. Constructing artificial neural network based on machine learning, designing the main steps of backward propagation neural network, making the cost function minimum; using the depth continuity of the image to extract the foreground part of the video motion, multiplying with the corresponding 2D video frame to detect the time domain motion Behavior; Solving the dual quadratic programming problem of Fisher support vector machine, obtaining its optimal solution and completing behavior learning; segmenting the current frame image, solving the normal vector to extract the moving target, and completing the behavior recognition method based on machine learning and Fisher vector coding the study. In order to verify the effectiveness of the design method, a comparative experiment was designed. The experimental results show that the average recognition accuracy of the design method is 7.6% higher than the traditional method.

5 Fund Project

2019 “climbing plan” Guangdong University Student Science and technology innovation and cultivation special fund project, project name: Research on behavior recognition method based on dense trajectory and Fisher vector coding, project number: pdjh2019b0615.

References

1. Tang, C., Li, J., Xu, D., et al.: Comparison of machine learning method and logistic regression model in prediction of acute kidney injury in severely burned patients. Chin. J. Burns **34**(6), 343–348 (2018)
2. Li, L., Luo, H.-q.: Human action recognition with weighted feature fusion based on STIP and dense trajectory feature. Microelectron. Comput. **34**(04), 110–114 (2017)
3. Qian, M., Zhang, D., Jiang, H.: Recognizing construction worker activities based on accelerometers. J. Tsinghua Univ. (Sci. Technol.) **57**(12), 1338–1344 (2017)
4. Li, Z., Wan, Q., Liu, Y., et al.: Non rigid 3D model retrieval method based on fisher vector encoding and distance learning. J. Comput. Aided Des. Comput. Graph. **30**(7), 1297 (2018)
5. Nguyen, X.S., Nguyen, T.P., Charpillet, F., Vu, N.-S.: Local derivative pattern for action recognition in depth images. Multimedia Tools Appl. **77**(7), 8531–8549 (2017). https://doi.org/10.1007/s11042-017-4749-z

6. Perota, A., Lagutina, I., Quadalti, C., et al.: 203 Single-step gene editing of 3 xenoantigens in porcine fibroblasts using programmable nucleases. Reprod. Fertil. Dev. **29**(1), 210 (2017)
7. He, J., Xue, Y., Li, S., et al.: Head re-identification by local Fisher vector encoded and cross-view quadratic discriminant analysis. J. Tianjin Univ. Technol. **35**(01), 9–15
8. Chen, Z., Wu, S., Fan, B.: Fisher discriminant analysis based on stacked autoencoders. J. S. China Normal Univ. (Nat. Sci. Edn.) **49**(03), 117–122 (2017)
9. Wu, Q., Zhao, X.: Incremental learning algorithm of support vector machine based on vector projection of Fisher discriminant. J. Xi'an Univ. Posts Telecommun. **23**(01), 79–84 (2018)
10. Li, K., Zhang, X., Zhang, M., et al.: n/γ Pulse shape discrimination in Cs2 LiYCl6:Ce3 + crystal using fisher linear discriminant. At. Energy Sci. Technol. **51**(11), 2069–2074 (2017). Author, F.: Article title. Journal 2(5), 99–110 (2016)