Research on Pedestrian Intrusion Detection Method in Coal Mine Based on Deep Learning

Haidi Yuan[1(✉)] and Wenjing Liu[2]

[1] Anhui Sanlian University, Hefei 230601, China
yyyhh22@yeah.net
[2] Student Affairs Office, Fuyang Normal University, Fuyang 236000, China

Abstract. Due to the complex background environment in coal mines, the timeliness and accuracy of pedestrian intrusion detection are low. In order to improve the detection accuracy and efficiency of pedestrian detection in complex coal mines, a deep learning-based pedestrian intrusion detection method in coal mines was studied. Build a pedestrian intrusion detection model in coal mine, the grayscale, denoising and illumination equalization processing is carried out for the surveillance video images of pedestrians in the coal mine. The image is preprocessed by nonlinear transformation method, gradient descriptor is obtained by gradient calculation method, HOG feature is obtained, and texture feature is obtained by LBP operator, and the features are used as input to construct a detection model using the restricted Boltzmann machine in deep learning to realize pedestrian intrusion detection in coal mines. The experimental results show that under the application of the research method, the average accuracy rate is higher, reaching more than 90%, and the FPS value is greater, reaching more than 40fps, indicating that the research method has higher detection accuracy and faster detection speed.

Keywords: Deep Learning · Restricted Boltzmann Machine · Underground Coal Mine · Pedestrian Intrusion · Intrusion Detection

1 Introduction

Coal mine is a high-risk industry, and its production safety has always been highly valued by the society. With the development of digital technology and the continuous advancement of smart mine policies, deep learning technology has great potential for development in coal mine safety protection. Combining computer vision with the coal mining industry has important research value and social significance in improving work efficiency, improving production environment, and ensuring production safety. Pedestrian detection, a subtask of object detection, aims to identify the precise location of people in video images using computer vision techniques [1]. Pedestrian intrusion detection is a technology that can judge whether there are illegal or illegal pedestrians by inputting pictures or video frames, and express the location information of pedestrians. This technology belongs to an important branch of target detection. At present, in

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 534, pp. 169–183, 2024.
https://doi.org/10.1007/978-3-031-50577-5_13

the pedestrian intrusion monitoring system in coal mines, accurate target detection and identification are the key to monitoring technology and the basis for subsequent target tracking and behavior analysis.

In terms of intrusion detection, reference [2] proposed a video image moving target detection based on improved background subtraction. The background model is reconstructed using the image block mean method based on GMM. In the target detection stage, the method of combining mathematical morphology and wavelet semi-soft threshold function is used to denoise the detected moving targets. In the background update stage, an adaptive background update method is used to update the background. Reference [3] proposed a UAV moving target detection method combining single Gaussian and optical flow method. An improved single Gaussian model is used to model the background of the image captured by the action camera, and multiple Gaussian models of the previous frame image are fused to perform motion compensation. The obtained foreground image is used as a mask to extract feature points and perform optical flow tracking, and perform hierarchical clustering of the motion trajectories of sparse feature points. Reference [4] proposed a deep learning target recognition simulation study incorporating the inter-frame difference method. Under the framework of deep learning theory, the inter-frame difference method is integrated into the recognition process to supplement and enhance the candidate frame segmentation image, and the candidate frame is screened by the NMS algorithm. Reference [5] proposed a dynamic pedestrian intrusion detection method based on PIDNet. A special PID task for feature sharing, a module for feature clipping and a branch network for feature compression are designed, and a benchmark data set is established to evaluate the proposed method. Reference [6] proposed a pedestrian detection method based on roadside light detection and distance measurement. In order to improve the real-time performance of detection, the octree selected by ROI is introduced and improved to filter the background in each frame, thus improving the clustering speed. The detection is completed by combining the adaptive distance Euclidean clustering search radius method. The above method realizes the detection and recognition of moving objects to a certain extent. However, in the mine environment, the detection and recognition rate of moving objects in video images is low, and the detection effect is poor, which in turn affects the subsequent behavior analysis.

In recent years, with the improvement of hardware equipment, deep learning technology has developed rapidly. Deep learning has better robustness, and has achieved great results in the field of computer vision such as image classification and target detection, and a large number of detection algorithms based on deep learning have appeared. In this context, a deep learning-based pedestrian intrusion detection method is proposed.

2 Pedestrian Intrusion Detection Model in Coal Mine

As a high-risk industry in coal mines, a large number of surveillance cameras are installed at the entrance, exit, and underground tunnels. However, a large number of video resources have not been effectively utilized at present. The video images in the mine have problems such as complex environment, dim light, and large noise interference. In addition, the installation position of the underground camera in the mine is high, and the pedestrians detected in the surveillance video have problems such as small

size, low resolution, scale change, and pedestrian overlap. Due to the special environment of the underground, the underground image contains the common target distortion, multi-scale, occlusion, illumination and so on in the target detection and pedestrian detection problems. Therefore, downhole pedestrian detection has high research value and significance.

Pedestrian detection is a typical target detection problem in the field of computer vision. Specifically, the training sample set is learned by a classification algorithm to obtain a classification model, and then the model is used to detect and classify the test samples. It mainly includes two steps: target positioning and target classification. Target positioning is mainly to determine the location of pedestrians, and target classification is to determine whether it is a pedestrian target. The framework of pedestrian detection is shown in Fig. 1.

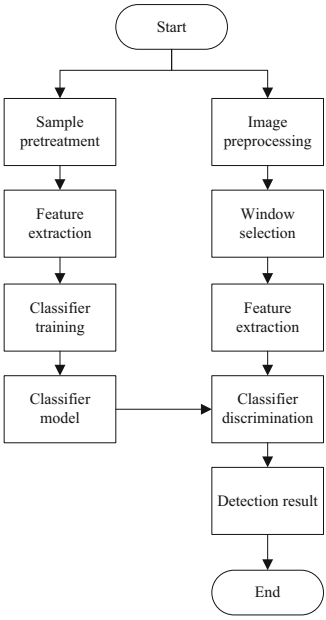

Fig. 1. Framework for pedestrian detection

In the training phase, a training sample set is first given, in which the positive sample set is composed of pedestrian targets, and the negative sample set is collected from background images. Then, a specific feature extraction algorithm is used on the sample set to convert the feature dimension from the image space to the feature space, and then the machine learning algorithm is trained to obtain a classification model to determine the sample category. In the detection process, the pre-calibrated detection window is mainly obtained through the sliding window strategy, and the feature extraction algorithm is also used to convert it into the feature space, and the trained classification detection model is used for screening. Therefore, feature selection and extraction is one of the key factors to determine the detection performance. By improving the feature extraction algorithm,

it is an important means to improve the detection performance to make the feature have stronger representation and discrimination ability.

2.1 Preprocessing of Pedestrian Images in Coal Mines

The underground environment of coal mines often contains noise and dust, which affects the quality of image shooting. The camera shake will also make the image unclear and blurred, which in turn affects the effect of detection and recognition. Therefore, image preprocessing is required before pedestrian target detection [7]. The preprocessing stage is to select images that are not clear or of poor picture quality in the video frame, perform a series of preliminary processing to make them meet the requirements of subsequent feature extraction, and then perform target detection.

(1) Grayscale image

Each color image is composed of three basic colors, red, green, and blue, which are represented by R, G, and B, respectively. Each component has 256 representation values, ranging from 0 to 255, so that 16.7 million colors can be combined. The grayscale of the image is to let R = G = B, then the grayscale value has 256 levels, 0 ~ 255, which can represent 256 colors. It can be seen that after the color image is grayed, the pixel value range will be greatly reduced, and a lot of workload will be reduced in the subsequent image processing process, which can effectively improve the processing speed. The component method selects the R, G, and B values as the gray value for grayscale, and then determines it according to the actual use.

$$\begin{cases} F_1(x, y) = R(x, y) \\ F_2(x, y) = G(x, y) \\ F_3(x, y) = B(x, y) \end{cases} \tag{1}$$

Among them, $F_i(x, y)$, $i = 1, 2, 3$ is the gray value at the pixel point (x, y).

(2) Image denoising

At present, the processing of video images by computer is all digital. In the process of image acquisition or conversion, some interference factors, such as random noise, will be introduced, which will undoubtedly bring difficulties to the detection and affect the subsequent processing. Image denoising can also be called smoothing, and there are two processing methods in the frequency domain and the spatial domain. Frequency domain processing requires Fourier transform of the image, analysis of its spectral components, and then inverse transformation to obtain the result. This processing method requires a large amount of calculation and is difficult to meet the needs of timeliness [8]. Here, the neighborhood averaging method is used for denoising. Neighborhood averaging method, also known as mean filtering, is to process the local spatial domain. The expression is:

$$D(x, y) = \frac{\sum\limits_{(x,y)\in S} d(x, y)}{N} \tag{2}$$

In the formula, $d(x, y)$ is the original image, S is the predetermined neighborhood, N is the number of pixels, and $D(x, y)$ is the processed pixel value.

(3) Lighting equalization processing

The underground environment of coal mines is complex, and many images will be affected. For example, the background color is similar to the target color, and the light source is mostly localized lighting, etc., so that the captured image will have extremely bright and extremely dark divisions of light and dark areas. When extracting feature points, due to the very low signal-to-noise ratio in extremely dark areas, almost no effective feature points can be detected, which brings trouble to the image registration step. According to the performance of the image with uneven illumination, it can be divided into: the overall gray value of the image is low, and the image analysis and recognition are more difficult. The light in the mine is relatively dark. In order to facilitate underground work, lighting equipment will be installed. The intensity of these lights is large and concentrated, which affects the detection effect of moving targets and increases the false detection rate. To this end, it is necessary to perform illumination equalization processing on the image.

The key problem in the preprocessing of pedestrian images under uneven illumination is how to remove the influence of illumination factors, and then enhance the details of the dark areas of the original image to a certain extent. Histogram equalization has a good enhancement effect on the overall dark image, but it is not ideal for images with a large overall grayscale range and high light areas. According to the homomorphic filtering and algorithm theory, the original image can be decomposed into two parts: illumination component and detail component, but both have their own shortcomings in the decomposition process. Two-dimensional decomposition is a completely data-driven process that processes images in the spatial domain, overcoming the shortcomings of homomorphic filtering and algorithms. The two-dimensional decomposition process can decompose the image into a series of residual components corresponding to different frequencies and corresponding to the overall change trend of the image. Since the illumination component corresponds to low-frequency information, the detail component corresponds to high-frequency information. Therefore, the obtained residual component can be used as an approximate estimation of the illumination component. After the illumination component is removed from the image, the remaining IFM component is overall dark due to the removal of the illumination factor. At this time, the histogram equalization can be used to process such images, and the histogram equalization is applied to them, and then appropriate illumination compensation is given to obtain a railway image with uniform illumination distribution and enhanced details.

Aiming at the image quality problem caused by uneven illumination of pedestrian images downhole, the algorithm in this paper is as follows:

Step 1: First determine whether the input original pedestrian image is an image with uneven illumination;

Step 2: Use the two-dimensional empirical mode decomposition method to extract the illumination components and detail components of the original downhole image with uneven illumination of people. The specific process is as follows:

Step 3: Enhance the detail components of the pedestrian image through the histogram equalization algorithm to obtain a detail-enhanced image. The process is as follows:

1) List the gray level $A_p, p = 0, 1, ..., 255$ of the original image. A_p is the p-level gray value.

2) Count the number of pixels M_p of each gray level, and M_p is the number of pixels of the gray value A_p.
3) Calculate the histogram. Calculated as follows:

$$B(A_p) = \frac{M_p}{n} \tag{3}$$

where n is the total number of image pixels.
4) Calculate the cumulative histogram. Calculated as follows:

$$C_p = \sum_{p=1}^{255} B(A_p) \tag{4}$$

5) Determine the mapping relationship between A_p and C_p, and then count the pixel M_p.
6) Calculate a new histogram $B(C_p)$.

Calculate the equalized mapping value corresponding to each gray-level pixel in the area to be equalized according to $B(C_p)$. And the mapping value replaces the pixel value before equalization of the gray-level pixel point, so as to obtain a brand-new gray-level equalized image, and then the image is obviously enhanced.

Step 4: Determine whether illumination compensation needs to be performed on the detail-enhanced image. If necessary, the method of adjusting the grayscale range of the light intensity is used to obtain a suitable light compensation component;

Step 5: Superimpose the detail enhancement component that needs illumination compensation and the modified illumination component to obtain an enhanced pedestrian image. If illumination compensation is not required, the detail-enhanced image is the final processing result.

2.2 Image Feature Extraction

In reality, the environment in which pedestrians live is complex and changeable, and is easily disturbed by factors such as posture, clothing, lighting, and occlusion. Therefore, it has the characteristics of nonlinearity and high noise, which brings challenges and difficulties to the detection task. How to select more robust features in different scenarios, so that they can be better used to describe pedestrians, and show the greatest degree of discrimination between pedestrians and backgrounds, is the key issue to improve the detection performance. Commonly used feature descriptors include simple low-level features and complex high-level features. The simple underlying features refer to the edge contour information, color change information and structural texture information of the image [9]. The edge contour information mainly refers to the area where the brightness changes sharply in the image, which can effectively describe the contour features in the image, and can provide a lot of useful information for pedestrian targets. Structural texture information refers to patterns that change regularly in a certain area, which can describe the essential characteristics of the image and better characterize the pedestrian target. The color change information is the basic element of the image, and the appearance of the target can be described by the change of the color. The advantage of simple low-level features is that the feature is single and the calculation speed is fast, but

the disadvantage is that the image information contained is too single, the discrimination ability is poor, and it is difficult to achieve high detection accuracy. Based on this, two types of features are extracted here, namely texture features and HOG features.

(1) HOG features

HOG mainly cascades multiple gradient direction histograms that describe local edge information, so that the overall edge structure of the target can be characterized, and the shape and appearance information of the target can be effectively described. The HOG feature is the statistics of the local information of the target, so when the target has a small deformation, it can also have strong anti-interference and good robustness.

HOG features are very robust and effective for the representation of pedestrian overall structure. The specific extraction process is as follows: first, the image is divided into small continuous regions by pixel points, which are called cells, and multiple cells are combined into blocks. By counting the gradient directions of all pixels in each cell, the cells in the same block are compared and normalized to reduce the influence of illumination. The final HOG feature descriptor can be expressed as a concatenation of gradient histogram vectors in all overlapping intersecting blocks. The specific calculation process is as follows:

1) Standardize the color space. In order to make the extracted features not affected by lighting factors, it is necessary to use the standardization method of nonlinear transformation to preprocess the image to enhance the contrast of the image, thereby reducing the impact of lighting factors. This step is not particularly important for the calculation process of the entire HOG feature, because the normalization of the extraction block can also reduce the influence of lighting well.
2) Gradient calculation. When counting the gradient directions corresponding to all pixels in the image, it is necessary to first calculate the gradient size corresponding to each pixel. The gradient direction reflects the local edge and structural information of the image, which is very effective for detection. The calculation of the gradient size is mainly determined by the gradient of the abscissa and ordinate of the pixel, and the final gradient descriptor is obtained through the use of the first-order symmetric template.

Let $L_x(x, y)$ and $L_y(x, y)$ represent the horizontal and vertical gradients at the image pixel (x, y), respectively, which can be expressed as:

$$\begin{cases} L_x(x, y) = J(x + 1, y) - J(x - 1, y) \\ L_y(x, y) = J(x, y + 1) - J(x, y - 1) \end{cases} \tag{5}$$

Then the gradient size and gradient direction corresponding to the pixel point (x, y) can be expressed by the following formula:

Gradient size:

$$L(x, y) = \sqrt{\left[L_x(x, y)\right]^2 + \left[L_y(x, y)\right]^2} \tag{6}$$

Gradient direction:

$$V = \arctan \frac{L_y(x, y)}{L_x(x, y)} \tag{7}$$

It can be seen from formula (5) that the gradient of each component is mainly calculated by the first-order symmetric template [-1,0,1] matrix operator. Then the final gradient magnitude and direction are calculated by formula (6) and formula (7).

3) Gradient direction histogram construction

The image is divided to form multiple non-intersecting small regions (cells), and then adjacent 2×2 cells form a block. The gradient information in each cell is counted, and 9 projection channels are usually generated according to the different gradient directions. The gradient size is weighted and projected into these 9 different channels (bins), and the 9 bins divide the gradient direction into 9 direction blocks on an average of 360 degrees. That is, 0–180 degrees are divided into 9 blocks, and 180–360 degrees are mapped to 9 direction blocks in a diagonal manner. If the gradient direction of a pixel statistics is 60–80 degrees, it is mapped to the fourth bin. The weight of the weighted projection is determined by the gradient size of the pixel point. By counting all the pixel points in a cell, a 9-dimensional feature vector value is obtained in a cell. Concatenate the statistical vectors of all cells in the block, and finally obtain a $2 \times 2 \times 9 = 36$-dimensional feature vector.

4) Block normalization processing

In order to further reduce the influence of light and shadow, the feature vector in the block needs to be locally normalized, so as to obtain a better detection effect and make the HOG feature extraction more robust. Commonly used normalization methods are: L2-norm, L1-norm, L1-sqrt, etc. Among them, the L2-norm method is mainly used in pedestrian detection.

5) HOG feature generation

By concatenating the gradient direction histograms obtained in each block, the HOG feature descriptor of the entire image is obtained. There are overlapping cell parts between blocks, so the features in the same cell will be counted multiple times. Although there are redundant calculations, they cannot be simplified. Because overlapping cell features can better reflect the context information of pedestrians, it can better represent the pedestrian target and achieve better detection results.

(2) Texture features

Texture is the regular arrangement of pixels on the surface of all objects in nature, and countless texture primitives are regularly distributed on the surface of things according to their own internal structure or natural laws. Generally speaking, from the tactile point of view, the textures are mostly large particles and have a certain roughness. But from a scientific point of view, even smooth marbles and clouds flowing in the air have certain texture characteristics on their surfaces, but they cannot be seen due to the limitations of the visual range of the human eye. At present, researchers have conducted a lot of research on texture related work, and found that it has strong stability, strong robustness to external lighting and noise, and does not change with the movement and rotation of the original image. The disadvantage is that the texture feature belongs to a surface pixel arrangement feature. As the feature information used to uniquely identify the image, it is not convincing enough to quickly and efficiently identify the category of the image, which affects the accuracy of image recognition. Therefore, based on this deficiency, researchers often combine a variety of image features to jointly play their

respective advantages, so that the extracted features are more representative, which is also conducive to the gradual development of later work.

The LBP operator is another magic weapon for extracting texture information and is widely used in many related fields of machine vision. The core idea is to first count the neighborhood information of each local pixel point, and then normalize the gray level information of all pixel points to obtain the texture information of the entire image. The advantage of LBP operator to extract texture features lies in its strong stability. In most cases, it is basically not affected by external light and noise, and has strong rotation invariance. Moreover, the calculation is simple, and it has achieved good results in many application fields of machine vision. This paper will improve the encoding calculation of the original LBP mode, and propose to add regular features in the image texture space, consider adding them to the calculation process of LBP encoding values, improve the representation ability of texture features, and further improve the accuracy of image recognition.

The acquisition of the texture features of the whole image is the sum of the LBP values of multiple local regions and the statistics. Therefore, the slight difference in the features of each local small region will have a great impact on the subsequent analysis and processing of the computer. The original LBP operator is defined based on a 3*3 square kernel, and the central pixel value of the kernel is defined as the threshold of the local area, multiplied by 8 pixel values are the neighborhood values under this basis. The grayscale comparison between the neighborhood value and the threshold value is used to calculate the LBP value of the local area. The calculation process is as follows:

Step 1: Divide the converted grayscale image into multiple square local areas to facilitate the calculation of the LBP value. Each local area can be regarded as a 3*3 square kernel, and the pixel values of 9 positions are obtained;

Step 2: Take the central pixel as the value of 1 ~ 7, and compare the grayscale with the 8 neighboring pixel values respectively. Finally get an 8-bit binary sequence (01111100);

Step 3: According to the conversion relationship between different binary digits, and combined with the weight corresponding to each binary bit, convert binary to decimal, that is, the feature value of the central pixel of the local area.

The texture feature of the entire pedestrian image is the sum of the LBP feature values of each local area, that is, the number of occurrences of all different LBP values is counted, and data processing and analysis are performed to obtain the texture information of the image.

2.3 Intrusion Detection Based on RBM

Deep learning is the intersection of neural network, artificial intelligence, graphical modeling, optimization, pattern recognition, signal processing and other research fields. Deep learning is a multi-level machine learning algorithm that simulates complex relationships between data and is based on representation learning. An observation can be represented in a number of ways, such as pixels represented by a matrix of intensity values, some representations can make learning tasks easier for algorithms. The goal of representation learning is to seek better representations and build better models to learn these representations. There are many structures of deep learning, and most of them are

branches of some original structures. Since they are not implemented on the same dataset and have different applicability and conditions, it is not always possible to compare the performance of multiple architectures [10]. Deep learning is a rapidly developing field, with new systems, structures, and algorithms emerging with each passing day. A deep network is a network with at least one hidden layer. Similar to shallow networks, deep networks can also model complex nonlinear systems. However, the multi-level structure provides a higher level of abstraction for the model, thereby improving the expressiveness of the model. Restricted Boltzmann Machine (RBM) is a two-layer directed acyclic graph, a special structure of Boltzmann Machine (BM), which is usually used as the basic unit for building deep structures. It consists of a series of binary state hidden layer units g and binary or true value visible layer units q. There is only connection between the visible layer and the hidden layer, but there is no connection between the visible layer units and between the hidden layer units. In an RBM, let the pixels correspond to the visible layer unit q, there are n nodes, the extracted features correspond to the hidden layer g, there are m nodes, and the system (q, g) composed of the visible layer and the hidden layer has energy:

$$R(q,g) = -\sum_{i=1}^{n} \alpha_i q_i - \sum_{j=1}^{m} \beta_j q_j - \sum_{i=1}^{n} \sum_{j=1}^{m} q_i w_{ij} g_j \tag{8}$$

Among them, α_i and β_j are the corresponding biases of the visible layer and the hidden layer, respectively, and w_{ij} is the weight between the visible layer and the hidden layer.

It can be clearly seen from the energy formula that, given the visible layer, the hidden layer units are independent of each other, whereas, given the hidden layer, the visible layer units are independent of each other. In particular, the unit of the binary layer is an independent Bernoulli random variable. If the visible layer is a true value, the visible unit is a Gaussian variable of diagonal covariance. Usually, the expected value of the unit is used as the activation value.

The joint probability distribution where both the visible layer unit vector and the hidden layer unit vector state are 1 is obtained by exponentiating and normalizing the energy function:

$$G(q,g) = \frac{e^{-R(q,g)}}{H} \tag{9}$$

Among them, H is a normalization constant, that is, all pairs of visible and hidden layers are added together.

The probability that the network assigns to the visible layer vector is to sum up all the visible layer vectors to get $G(q)$.

$$G(q) = \sum_g e^{-R(q,g)} \Big/ H \tag{10}$$

Since the visible layer units and the hidden layer units are independent of each other, there are:

$$\begin{cases} G(g|q) = \prod_{j}^{m} G(g_j|q) \\ G(q|g) = \prod_{i}^{n} G(q_i|g) \end{cases} \tag{11}$$

When the state of the visible layer q is given, the probability that the binary state of the hidden layer unit g is 1 is $G(g_j = 1|q)$;

$$G(g_j = 1|q) = f\left(\beta_j + \sum_{i=1}^{n} q_i w_{ij}\right) \tag{12}$$

When the state of the hidden layer g is given, the probability that the binary state of the visible layer unit q, is 1 is $G(q_i = 1|g)$.

$$G(q_i = 1|g) = f\left(\alpha_i + \sum_{j=1}^{m} g_j w_{ij}\right) \tag{13}$$

where $f(\cdot)$ is the activation function.

It can be seen from formulas (12) and (13) that the learned weights and biases directly determine the conditional distributions $G(g|q)$ and $G(q|g)$, and indirectly determine the joint distribution $G(q, g)$ and the marginal distributions $G(q)$, $G(g)$. Sampling from a joint distribution is difficult, but it can be achieved by "alternating Gibbs sampling". Start with a random image, then update all features in parallel with formula (12) and all pixels with formula (13), alternating the two processes. After Gibbs has sampled long enough, the network reaches "thermal equilibrium". At this point, the states of the pixel and feature detectors are still changing, but the probability of finding a system under any particular binary structure does not change.

For an RBM network, there are only input layers and output layers. In the process of training RBM, the number of visible layer units is generally the original input data, and the number of hidden layer units needs to be given. The training process of RBM and the optimization of weights are as follows:

Step 1: First, randomly initialize the weight matrix w and the bias vector β of the visible layer and the bias vector α of the hidden layer to the network.

Step 2: Assign the original input data to the visual layer unit, and forwardly propagate the visual layer input matrix q. According to formula (12) and $G(g|q)$ of formula (11), the activation probability $G(g|q)$ of the output matrix g of the hidden layer unit is calculated by the visible layer. The activation probability of the input matrix q and g corresponds to the matrix of the node product to obtain the probability of forward propagation.

Step 3: At this time, the $G(g|q)$ output in step 2 is the g probability value, and it is randomly binarized into a binary variable.

Step 4: Use the g probability value binarized in Step 3 to propagate in the reverse direction. According to formula (13) and formula (11) $G(q|g)$, the activation probability

of the matrix q' of the visible layer is calculated, that is, a reconstruction of the visible layer is obtained.

Step 5: Perform forward propagation on q' again, and calculate the activation probability of the matrix g' of the hidden layer according to formula (12) and $G(g|q)$ of formula (11). As in step 1, the activation probability of the input matrix q' and g' corresponds to the matrix of the node product to obtain the probability of back propagation.

Step 6: Subtract the activation probability of g' obtained in step 5 from the activation probability of the hidden layer g obtained in step 2, and the result is used as the increment of the bias β corresponding to the input layer g. The activation probability of q' is subtracted from the activation probability of the visible layer q, and the result is used as the offset α corresponding to the visible layer q. The probability vector of back propagation obtained in step 5 is subtracted from the probability vector of forward propagation obtained in step 2, and the result is used as the weight increment between the input layer and the output layer. In each iteration, the update of the weights and the update of the bias are performed simultaneously, so they should converge at the same time. Combined with its corresponding learning rate, the weights and biases are updated.

In addition, in order to alleviate the contradiction between the speed of the learning rate and the stability of the algorithm, the momentum l is introduced, and the initial value is multiplied by the momentum m before updating. For example, the initial value of momentum l is set to 0.5, when the reconstruction error gradually stabilizes from a large initial decline process, the momentum l is increased to 0.9.

Step 7: Repeat steps 2 to 7 until convergence or the maximum number of iterations is reached. In this way, the training of an RBM is completed.

3 Experimental Tests

In this paper, the above structure is evaluated on the downhole pedestrian detection dataset to verify the performance of our algorithm. And the RBM-based pedestrian intrusion detection algorithm in coal mines is evaluated on the VOC 07 public beta data set.

3.1 Experimental Dataset

The coal mine underground data set in this paper comes from a coal mine underground monitoring video. The entire data set contains a total of 23,210 pictures, all of which are 1280×720 in size. 11,605 pictures are selected as the training set and 11,605 pictures are used as the test set. The number of pedestrians in each picture ranges from 1 to 20, of which 80% are pedestrians without intrusion behavior, and 20% are pedestrians with intrusion behavior.

3.2 Detection Indicators

E was used as the evaluation index in the experiment. E is the average accuracy rate of all categories in the intrusion detection model. The larger the value of E is, the closer to 1 it is, indicating that the detection accuracy rate of the method is higher; The smaller

the value of E, the lower the accuracy of the method detection. The calculation formula is:

$$E = \frac{\sum_{k=1}^{u} \overline{h}_k}{u} \tag{14}$$

In the formula, u is the number of target categories in the target detection task, $\overline{h}_k$ is the average accuracy of various detection targets.

In practical applications, deep learning models need to meet the requirements of timeliness. Therefore, the detection speed needs to be evaluated, and the commonly used evaluation method is to evaluate the detection speed by the time required to process a single image. The number of detectable frames per second can also be used as the evaluation standard, and the latter is selected to evaluate the detection speed of the model. FPS is the number of frames per second, generally 24 fps for movies, 30 fps for TVs, and 60 fps for LCD monitors. In order to make the picture smooth, the FPS generally needs to be kept above 30 fps. In deep learning, FPS can represent the number of frames detected by the model per second, and can intuitively express the timeliness of the model. The larger the FPS value, the more frames can be detected per second, and the better the timeliness.

3.3 Experimental Results and Analysis

The research method is compared with the improved background subtraction method, the combination of single Gaussian and optical flow method, and the inter-frame difference method, respectively, and the comparison results of the average accuracy of different methods are shown in Fig. 2.

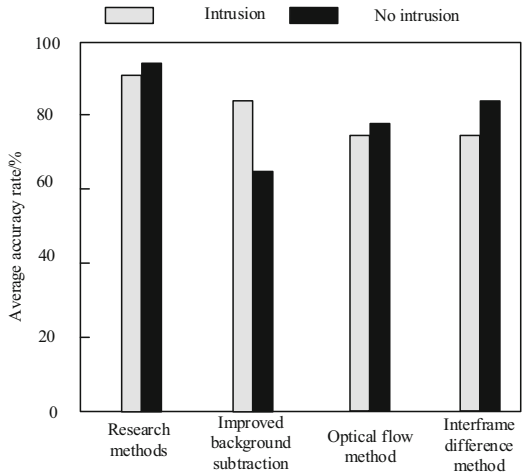

Fig. 2. Comparison of the average accuracy of different methods

The FPS comparison results of different methods are shown in Table 1.

Table 1. FPS comparison table of different methods

Method	FPS/fps
Research methods	42.36
Improved background subtraction	30.21
Combining single Gaussian with optical flow method	31.52
Interframe difference method	24.58

As can be seen from Fig. 2 and Table 1, compared with improved background subtraction, combined single Gaussian and optical flow method, and inter-frame difference method, the average accuracy rate is higher and the FPS value is larger. It shows that the detection accuracy of the research method is higher and the detection speed is faster.

4 Conclusion

The mining industry is transforming into a smart mine, accelerating the development of automation and intelligence, but its safety has not been well guaranteed. The underground environment is dark and the mine is complicated, and the staff can easily enter the dangerous area. At present, it is mainly through manual shifts to monitor each work point to determine whether someone has entered the dangerous area by mistake. However, there are many monitoring points, and the monitor screen is small, and the requirements for the staff are relatively high. They are prone to fatigue and inattention for a long time, and they cannot detect the occurrence of danger in time or misjudgment certain behaviors. To this end, a deep learning-based pedestrian intrusion detection method in coal mines is studied. After testing, the accuracy and detection speed of pedestrian intrusion detection in coal mines have been improved to a certain extent. However, the algorithm in this paper does not consider the problem of multiple pedestrians occluding each other. Therefore, in the next research, the main consideration is to solve the problem of pedestrian occlusion in the coal mine.

Acknowledgement. This work was supported by Anhui Provincial Education Department Foundationunder grant no.KJ2021A1176.

References

1. Astolfi, G., Rezende, F.P.C., Porto, J.V.D.A., et al.: Syntactic pattern recognition in computer vision: A systematic review. ACM Computing Surveys (CSUR) **54**(3), 1–35 (2021)
2. Junhui, Z., Zhenhong, J., Jie, Y., et al.: Moving object detection in video image based on improved background subtraction. Comp. Eng. Design **41**(05), 1367–1372 (2020)
3. Changjun, F., Lingyan, W., Quanyong, M., et al.: Detection of moving objects in UAV video based on single gaussian model and optical flow analysis. Comp. Sys. Applicat. **28**(02), 184–189 (2019)

4. Hui, W., Lijun, Y., Rong, S., et al.: Research on simulation of deep learning target recognition based on inter-frame difference method. Experim. Technol. Manage. **36**(12), 178–181 and 190 (2019)
5. Sun, J., Chen, J., Chen, T., et al.: PIDNet: An efficient network for dynamic pedestrian intrusion detection. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 718–726 (2020)
6. Gong, Z., Wang, Z., Zhou, B., et al.: Pedestrian detection method based on roadside light detection and ranging. In: SAE International Journal of Connected and Automated Vehicles **4**(12-04-04-0031), 413–422 (2021)
7. Yang, Y., Su, W., Qin, Y., et al.: Research on object detection method of high-speed railway catenary image based on semantic label. Comp. Simula. **37**(11), 146–149 and 188 (2020)
8. Wang, R., Chen, H., Guan, C., et al.: Research on the fault monitoring method of marine diesel engines based on the manifold learning and isolation forest. Appl. Ocean Res. **112**(2), 102681 (2021)
9. Calvo-Bascones, P., Sanz-Bobi, M.A., Welte, T.M.: Anomaly detection method based on the deep knowledge behind behavior patterns in industrial components. Application to a hydropower plant. Computers in Industry **125**(5), 103376 (2021)
10. Shatalin, R.A., Fidelman, V.R., Ovchinnikov, P.E.: Incremental learning of an abnormal behavior detection algorithm based on principal components. Comput. Opt. **44**(3), 476–481 (2020)