Pedestrian Motion Path Detection Method Based on Deep Learning and Foreground Detection

Meiman Li and Wenfu Xie

School of Artificial Intelligence, The Open University of Guangdong, Guangzhou 510091, Guangdong, China

Correspondence should be addressed to Meiman Li; mmli@gdrtvu.edu.cn

Received 19 January 2021; Accepted 23 May 2021; Published 30 May 2021

Academic Editor: Wei Wang

Copyright © 2021 Meiman Li and Wenfu Xie. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For the surveillance video images captured by monocular camera, this paper proposes a method combining foreground detection and deep learning to detect moving pedestrians, making full use of the invariable background of video image. Firstly, the motion region is extracted by the method of interframe difference and background difference. Then, the normalized motion region extracts the feature vectors based on the improved YOLOv3 tiny network. Finally, the trained linear support vector machine is used for pedestrian detection, and the performance of the fusion detection algorithm on caviar dataset is given, which proves the effectiveness of the proposed fusion detection algorithm. Experimental results show that the proposed method not only improves the practical application of pedestrian re-recognition but also reduces the detection range, computational complexity, and false detection rate compared with sliding window method.

1. Introduction

Pedestrian detection is an important research problem in computer vision. In recent years, with the development of machine learning, pedestrian detection has made great progress. The main content of pedestrian detection is to detect and locate pedestrians quickly and accurately in the image. This technology has a wide range of applications in the field of car driving assistance, human-computer interaction, and machine vision [1]. In the field of video surveillance, most of the pedestrian monitoring methods are still based on manual detection. This way needs heavy workload and is low efficient. With the advent of the era of big data, massive video data need to be processed. Artificial way has been unable to meet the current needs. In addition, intelligent and fast pedestrian detection in video image can not only estimate pedestrian flow density but also further analyze human behavior and warn dangerous scenes [2, 3]. Compared with manual detection, automatic pedestrian detection can not only improve the efficiency but also improve the accuracy. But for different people, their appearance, posture, and clothing have a great impact on the detection. Because people are nonrigid objects, their posture will change constantly when they walk [4]. Combined with the complex environment and shooting angle, these all increase the difficulty of detection. There are many pedestrian-like objects in real life, which may cause false detection. The size of pedestrian in the image is determined by the distance between the pedestrian and the camera. Multiscale pedestrian detection in the whole image will bring complex computation, and once affected by bad weather and light changes, the detection rate will decline. In addition, the moving pedestrian will appear occlusion; these are the difficulties of detection [5, 6].

At present, the commonly used pedestrian detection methods can be divided into two categories: pedestrian detection in static image and pedestrian detection in video image. Pedestrian detection methods in static images often use sliding window for multiscale window selection and then combine the corresponding features and classifiers for detection [7]. The method based on the detection template matching model is to first establish a human target template and then use the template to match the similarity of the possible target areas in the image, so as to determine whether
it is a pedestrian target [8]. Generally, this method is fast, but it needs to obtain a large number of templates manually. Some pedestrian poses cannot be accurately matched, so it is difficult to detect the pedestrian target under complex conditions. Statistical learning is the most widely used method, which is mainly composed of feature extraction and classifier training. Feature extraction mainly refers to the extraction of pedestrian features in the image and the selection of appropriate feature combination through machine learning knowledge [9, 10]. Classifier training refers to the learning of feature data to get a model that can distinguish pedestrians. This method is simple, efficient, and robust and has become the mainstream method of pedestrian detection.

Although there are many feature extraction methods, but too many candidate windows seriously affect the detection speed. Moreover, for video images, it has some unique motion information, which can be used to improve the detection effect. In order to cope with the massive growth of video surveillance data, the speed of pedestrian detection should be improved. Through the development process of pedestrian detection, although deep learning method has completely surpassed the traditional methods and become the main means to solve the problem of pedestrian detection, we cannot ignore the important basic role of traditional methods in pedestrian detection technology. It still has a great inspiration for pedestrian detection. On the other hand, although deep learning has made a breakthrough in the field of pedestrian detection, it still has great potential to be tapped, such as network structure optimization, network parameter optimization, and lack of training samples. Therefore, the method based on deep learning still has great room for improvement in solving pedestrian detection problems. In this paper, a pedestrian detection system based on foreground detection and deep learning is introduced. This method has the advantages of both real-time and accuracy.

2. Moving Pedestrian Detection Algorithm Based on Deep Learning and Foreground Fusion

2.1. Fusion Detection Algorithm Flow. Through the above-mentioned analysis, it can be seen that the moving target detection algorithm based on background difference method can effectively detect the moving target and eliminate the false alarm target. The existing problem is that the detection area is not accurate, it is difficult to provide the judgment of target type, and it is greatly interfered by illumination, occlusion, overlap, and other factors [11–14]. The pedestrian detection algorithm based on improved YOLOv3 tiny network can adjust the accuracy and recall rate by setting different confidence thresholds. At the same time, it can judge the category of the target and get a more accurate single pedestrian detection frame. The existing problem is that it is very dependent on setting the confidence threshold artificially, so it is difficult to achieve the balance between accuracy and recall.

The fusion algorithm proposed in this paper combines the detection boundary frames obtained by the above two detection algorithms and uses deep learning method to describe the appearance of pedestrian object accurately and comprehensively, while mining the motion information of the pedestrian object [15–18]. The motion information is used to remove the false alarm target which may be produced by the deep learning method, and finally, the fusion detection result is obtained. The process of fusion detection algorithm is as follows: firstly, the parameters of pedestrian detection model based on the improved YOLOv3 tiny network are trained by using the dataset, and a frame only containing background in the video is taken as the background. Then, the original image is input into the background differential motion detection model and the improved YOLOv3 tiny network pedestrian detection model respectively to obtain two detection results. Finally, the detection results of the two algorithms are fused and output. The algorithm flow is shown in Figure 1.

2.2. Fusion Process Analysis. The two detection algorithms will give the coordinates of the boundary box of the detection results, and the fusion detection will process the position and size information of these boundary boxes to get the fusion results. Through a large number of experimental tests, three possible situations of two kinds of bounding boxes are summarized.

This article believes that the coincidence of the two detection boundary frames should be as high as possible. However, through a large number of experiments and theoretical analysis, it can be seen that because the background difference method needs to expand the binary image, the low degree of expansion processing may lead to incomplete pedestrian area, and the high degree of processing often leads to the outer rectangular frame to be larger than the actual pedestrian area [19–22]. Firstly, the coordinates, width, and height of the two kinds of bounding boxes are compared. If the upper left corner of the bounding box is close, it is determined that the two kinds of bounding boxes may belong to the same target. Then, it is determined whether the width and height of the two kinds of bounding boxes are close. If they are close, it is determined that the bounding box of the improved YOLOv3 tiny network contains real moving pedestrian targets.

When pedestrians are occluded, overlapped, or shadowed, it is difficult for moving pedestrians based on background difference method to distinguish multiple pedestrians. However, the pedestrian detection frame of the improved YOLOv3 tiny network may fall into the detection frame of the background difference method because of its high and low scattered and partial overlap. Whether the center point of the improved YOLOv3 tiny network detection boundary box falls into the center range of the background difference detection box is calculated. If it falls into the center range, it can be judged that the improved YOLOv3 tiny network detection box contains the correct moving pedestrian target.
When the connected domain is disconnected due to the shadow and other reasons, the detection frame of the background difference method will be incomplete. There is only one similar abscissa and ordinate of the upper left corner of the border box in the figure, and the center point of the improved YOLOv3 tiny network detection box does not fall into the center range of the background difference detection box [23–27]. At this time, the proportion of the intersection of the two kinds of bounding boxes in their respective areas is calculated. If the proportion is greater than a certain threshold, it can be determined that the detection box of the improved YOLOv3 tiny network contains real moving pedestrian targets.

Through the abovementioned analysis, it can be found that there is a certain relationship between the two types of boundary frames in terms of coordinates, side length, and area. In order to reduce the time consumption of fusion detection algorithm as much as possible, the principle of more addition and subtraction and less multiplication and division is followed to fuse the bounding box. The i-th target in the pedestrian detection results is i, and the boundary box of the target detected by the pedestrian detection algorithm based on the improved YOLOv3 tiny network is \((x_i^1, y_i^1, w_i^1, h_i^1)\). The moving object detection algorithm based on the background difference method detects the object’s bounding box as \((x_i^2, y_i^2, w_i^2, h_i^2)\), which are the coordinates of the upper left corner and the width and height of the bounding box, respectively. The specific process of integration is as follows:

1. As shown in the equation, when the coordinates of the upper left corner of the border and the difference between the width and height are within a certain threshold \(\alpha, \beta, \gamma, \xi\), it is directly determined that the target is a real target. When at least one or at most three of the four differences meet the threshold, it is still considered that there may be a real target and it goes to the next step:
   \[
   \begin{cases}
   |x_i^1 - x_i^2| < \alpha, \\
   |w_i^1 - w_i^2| < \beta, \\
   |y_i^1 - y_i^2| < \gamma, \\
   |h_i^1 - h_i^2| < \xi.
   \end{cases}
   \]
   (1)

2. As shown in equations (2) whether the center point of the pedestrian detection boundary box \((x, y, w, h)\) based on improved YOLOv3 tiny network is in a certain height range of the moving object detection boundary box based on background difference method is judged, which is determined by \(\mu\). If it is in this range, it will be judged as a real target; otherwise, it will enter the next step:
   \[
   \begin{cases}
   x_i^2 < x_i^1 + \frac{w_i^2}{2} < x_i^1 + w_i^1, \\
   y_i^2 + \mu h_i^2 < y_i^1 + \frac{h_i^2}{2} < y_i^1 + \mu h_i^1.
   \end{cases}
   \]
   (2)

3. As shown in formulas (3), pedestrian detection algorithm based on improved YOLOv3 tiny network and moving object detection algorithm based on background difference method get the area of boundary box as IC and ID. If IA is greater than a certain threshold, the bounding box is considered to be a real target. If it is less than the threshold, it will continue to judge whether IB is greater than the threshold \(\eta\). If it is greater than the threshold \(\eta\), it will be considered that the bounding box is a real target; otherwise, it will be considered as a false alarm target or an error target.
Through the abovementioned experiments, it can be found that the pedestrian detection algorithm based on the improved YOLOv3 tiny network can achieve a higher recall rate when the lower confidence threshold is set, which means that most of the real pedestrian targets can be marked. At the same time, some false alarm targets and duplicate detection boxes are also included in the detection results. The fusion detection algorithm is used to fuse the two kinds of results under a low confidence threshold.

The two most important indexes to evaluate the performance of target detection algorithm are accuracy and recall. At present, experts and technicians in the field of computer vision are eager to improve the accuracy of target detection without reducing the recall rate. Most of the research focuses on the optimization of deep convolution neural network and its associated technology. The fusion detection algorithm proposed in this paper provides a new idea to improve the two indicators at the same time. The moving target detection algorithm based on background difference method is used to select and modify the detection results of pedestrian detection algorithm based on improved YOLOv3 tiny network under low confidence threshold. The fusion detection algorithm does not rely on artificial confidence threshold and can effectively improve the detection accuracy and recall rate at the same time.

3. Design of Pedestrian Movement Path Detection System

3.1. Network Structure. Figure 2 shows the network structure of an end-to-end pedestrian detection and recognition system proposed in this paper. The network accepts the full scene image of the camera as the input, and two branches are used to extract the image features. One branch is to extract the depth feature through convolution neural network, and the other branch is to extract the LOMO feature of the pedestrian image processed by Retinex algorithm. Next, the two features are sent into the pedestrian detection network, respectively, and are fused through a full connection layer in the pedestrian detection network. On the basis of this fusion feature, the pedestrian detection and the prediction of the boundary box of the traveler are carried out by using the YOLO target algorithm through regression [28–30]. This pedestrian detection method is the third chapter of the pedestrian detection method based on feature fusion. After getting the pedestrian boundary box, that is, the coordinates of pedestrians in the image, they are sent into the ROI pooling layer. The operation of ROI pooling can map the precise pedestrian coordinates obtained by pedestrian detection to the depth feature map obtained after a series of feature extraction operations and the depth feature of the landmark is obtained. Then, we continue to use this feature for pedestrian recognition, so as to achieve the end-to-end effect. The network consists of two branches. One is feature extraction based on convolution neural network, the other is feature extraction based on traditional manual design LOMO. Similar to the strategy of feature fusion in pedestrian detection, we use fusion layer to fuse the two features to construct a robust pedestrian feature. In the training phase, in order to train the pedestrian reidentification network with supervision, we use random sampling Softmax loss function to train the pedestrian reidentification network. In the test phase, we take each pedestrian’s ID as a category and then multiclassify the extracted and fused pedestrian features to authenticate each detected pedestrian, that is, assign an ID.

Pedestrian rerecognition technology is often used to deal with the problem of video surveillance. In the face of massive video processing, it is very important to ensure the real-time performance of the system. Therefore, we design pedestrian detection and recognition system based on ResNet-50 network structure. The first layer of ResNet-50 network is convolution layer, which has a convolution kernel of $7 \times 7$ and is called conv1. The remaining network structure is divided into four levels. The first layer contains three residual blocks, the second layer contains four residual blocks, the third layer contains six residual blocks, and the last layer contains three residual blocks. We use conv1, conv2, conv3, conv4_1, conv4_2, and conv4_3, and these layers make up the depth feature extraction network. After the input pedestrian image passes through the depth feature extraction network, 1024 channel convolution feature maps are generated, and the resolution of these feature maps is $1/16$ of the original image. At the same time, we also extract the LOMO features of traditional manual design through LOMO method. In order to get accurate pedestrian boundary box, we use the YOLO algorithm to establish a pedestrian detection network. In the pedestrian detection network, the full connection layer is used to map the 1024 channel depth feature map to a 4096-dimension feature vector. A robust pedestrian feature is constructed by fusing the same 4096-dimension LOMO feature vector. The full connection layer is used as the feature fusion strategy, and the back propagation mechanism of the network can be used for feature fusion adaptively. According to the YOLO target algorithm, we generate 9 anchors on each feature graph and obtain the fusion features in each anchor and use the obtained features to train the classifier and linear regression for the network. The classifier is used to predict whether each anchor is a pedestrian, and the linear regression is to adjust its boundary box. In order to reduce the redundancy of the target frame, we first optimize the generated anchor with NMS and then classify and regress the rest of the anchor. Finally, accurate pedestrian detection results are obtained.

In order to solve the problem of pedestrian reidentification based on pedestrian detection, we use ROI pooling operation to combine the designed pedestrian reidentification network and pedestrian detection network into an end-to-end pedestrian detection and reidentification system. The image coordinate position of the target person is obtained in the pedestrian detection network. We build a pedestrian identification recognition network, further extract pedestrian features which combine CNN features and traditional hand-made LOMO features, and use these
features to train pedestrian identification classification network, and finally achieve the effect of pedestrian identification classification. First, we use the ROI pooling layer to extract $14 \times 14 \times 1024$ feature maps from each pedestrian frame. Then, the extracted feature map is sent to the pedestrian identification network to generate 2048d feature vector. Finally, the feature vector of 2048d is reduced to 256d by L2 normalization, and the pedestrian feature is used for training and testing. In the training phase, we use random sampling Softmax (RSS) loss function to train pedestrian reidentification network. This loss function can effectively classify a large number of pedestrians with similar targets.

3.2. ROI Pooling. ROI pooling is a variant of pooling operation, which is to pool the region of interest and get the fixed size region of interest feature map on different scale feature maps. This idea is well applied in fast RCN, which uses RPN to generate region proposals and then continues to design target detection network to complete the target detection task, so it is difficult to train the network end-to-end. In order to solve this problem, this paper uses ROI pooling operation to realize the end-to-end pedestrian detection system. This operation can not only realize the end-to-end training of the network but also accelerate the network training and improve the detection accuracy. Inspired by this idea, this paper introduces ROI pooling into the end-to-end pedestrian detection and recognition task. After the pedestrian image coordinates, that is, the region of interest (ROI), are obtained by the pedestrian detection network, the depth feature map of the pedestrian image is obtained by the pooling operation on the depth feature map as the input of the pedestrian identification re-recognition network.

The specific operation of ROI pooling is as follows:

1. According to the input image, the ROI is mapped to the corresponding position of feature map;
2. The mapped region is divided into sections of the same size, and the number of sections is the same as the output dimension;
3. Max pooling operation is performed for each section. The specific operation of ROI pooling is shown in Figure 3.

3.3. Random Sampling Softmax. One of the key problems in training pedestrian recognition network is to adopt appropriate classification loss function. Our training set has 5532 identities, so the classification targets in this task are very intensive. Secondly, due to the high cost of computing on large images, each small batch of the training set consists of only two scene images, which usually contains no more than ten different training identities. Therefore, the label distribution of small batch labeled images and datasets does not match significantly, and the training datasets lack diversity. These two problems make the traditional Softmax difficult to solve such a large-scale pedestrian identification classification problem. In practice, we find that if we use the ResNet-50 model of ImageNet pretraining to transfer the network directly, it cannot only speed up the network convergence but also reduce the training loss. We use random sampling Softmax (RSS) loss function to replace the traditional Softmax. Because of the loss of traditional Softmax, it may only support a few classes that appear in a small batch and severely inhibit other classes. RSS loss function solves this problem by randomly selecting a subset of Softmax neurons of each input sample to calculate the loss and gradient. We assume that the target is classified as $C + 1$, in which there are $C$ categories of pedestrian identity. $C + 1$ is the pedestrian background, and $(x, s)$ is used to represent each data sample, so the traditional Softmax loss function can be expressed as

![Figure 2: End-to-end pedestrian detection network.](image)
\[ P_l = - \sum_{i=1}^{C+1} s_i \log (x_i + \delta), \quad (4) \]

\[ x_m = \frac{e^{\gamma_m}}{\sum_{n=1}^{C+1} e^{\gamma_n}} + \ln e^{\delta + \gamma}. \quad (5) \]

The RSS loss function selects \( K (k < C) \) categories to calculate the loss and gradient of the function. If the selected classification is expressed as \( x_i \), then the data samples are expressed as \( (x_1, x_2, x_3, \ldots, x_m) \); then, the RSS loss function can be expressed as

\[ \bar{P}_l = - \sum_{i=1}^{C+1} \bar{s}_i \log (\bar{x}_i + \delta), \quad (6) \]

\[ \bar{x}_m = \frac{e^{\bar{\gamma}_m}}{\sum_{n=1}^{C+1} e^{\bar{\gamma}_n}} + \ln e^{\delta + \bar{\gamma}}. \quad (7) \]

In order to optimize this problem, we need to set a good starting point for the random sampling Softmax classifier. Specifically, we cut the real boundary box of the ground for each trainer and randomly extract the same number of background boxes. Then, we adjust the pedestrian frames to 224 \( \times \) 224 and then classify the pedestrian frames with batch size of 256 by ResNet-50 model. Due to the diversity of tags in each small batch, this network processing process is very flexible, and the obtained model is used as the starting point of training the whole framework.

### 4. Experiment and Analysis

#### 4.1. Performance Evaluation of Fusion Detection Algorithm

In order to test the effectiveness of the proposed method in removing false alarm information, the scene with low resolution, strong or weak illumination, overlapping occlusion, and false alarm target is selected for testing. The experimental data come from the caviar project. The test set provides 26 groups of monitoring videos and corresponding labeling information, such as pedestrian shopping, meeting, entering and leaving the store, and carrying luggage in a shopping mall aisle. The video resolution is 384 \( \times \) 288, the video size is between 6 MB and 12 MB, and the frame rate is 25 frames per second, a total of 36222 frames.

Using the basic idea of dataset segmentation, the training set and test set are divided according to the ratio of no less than 95%:5%. Due to the differences of pedestrian movement in each group of videos, it is necessary to ensure that the pedestrian detection algorithm based on the improved YOLOv3 tiny network can learn the features of the dataset to the maximum extent. Therefore, a certain number of pictures are randomly selected in each group of video frames, and a total of 30000 pictures are extracted as the training dataset. The remaining 6222 images were used as the test dataset. The performance of pedestrian detection algorithm and fusion detection algorithm based on improved YOLOv3 tiny network are analyzed, respectively. The following is a comparative analysis of the accuracy, recall, and F1 indicators of the modified output layer and the improved network based on the caviar dataset, as shown in Figure 4.

The improved output layer network is compared with the network in this paper. It can be seen from Figure 4(a) that under the lower confidence threshold, the accuracy advantage of this network is obvious, and the recall rate is slightly improved. It shows that the number of targets detected by the two methods is close, but the number of correct targets detected by the network in this paper is more. Under the high confidence threshold, the recall rate advantage of this network is obvious, and the accuracy rate is slightly improved. It shows that the proportion of the correct targets detected by the two is similar to that of all the detected targets, but the network in this paper can detect more
targets. It can be seen from Figure 4(b) that the overall performance of the network in this paper is improved obviously when the confidence threshold is low and still has a certain improvement when the confidence threshold is high. To sum up, compared with the YOLOv3 tiny network, which modified the output layer and is trained by caviar dataset, this network has a certain improvement in accuracy, recall, and F1 score.

The PR curve is drawn to illustrate the comprehensive detection performance, as shown in Figure 5.

It can be seen from Figure 5 that the area under the improved network PR curve is larger, and the two curves are closer when the recall rate is lower than 0.8. When the recall rate is greater than 0.8, the gap between the two increases. This shows that the proposed network can also improve the detection performance on the caviar dataset to a certain extent. The model performance under different confidence thresholds is shown in Table 1.

It can be seen from the abovementioned analysis that the pedestrian detection algorithm based on improved YOLOv3 tiny network can detect more real targets with higher recall rate, but it will bring some false alarm targets into the detection results. Fusion detection algorithm can use the motion information to filter out part of the false alarm information, so as to improve the detection accuracy. Table 2 shows the performance of fusion detection algorithm under several groups of low confidence threshold.

It can be seen from Figures 6 and 7 that the fusion detection algorithm improves the detection results of pedestrian detection algorithm based on improved YOLOv3 tiny network. Under the same confidence threshold, the fusion detection algorithm can effectively improve the detection accuracy at the cost of losing a small part of the recall. With the improvement of the confidence threshold, the improvement effect of accuracy slightly decreased, but F1 index still maintained a high score. This shows that the fusion detection algorithm can improve the accuracy and recall rate at the same time to a certain extent.

4.2. Analysis of Experimental Results. In this paper, 3000 images of the target region and background region will be cropped from the test datasets of INRIA and Caltech datasets, respectively, including 1500 positive samples with pixel size of 96×160 and 1500 negative samples to form the test sample group. Figures 8 and 9 show the experimental results of several algorithms. This paper compares the performance of Dagnet Kelm algorithm and Dagnet-SVM, traditional CNN algorithm, and classical algorithm HOG-SVM on INRIA dataset and Caltech dataset. The positive detection rate of this paper on INRIA dataset is 97.9%, which is better than traditional CNN algorithm (95.8%) and classical HOG + SVM algorithm (92.5%). In Caltech dataset of 3000 images test, the experimental results are shown in the figure, the positive detection rate of this algorithm is 94.7%, higher than the other three algorithms.
In order to improve the detection speed, in the detection stage, this paper first uses the trained DAG network to obtain the feature map of the first stage fusion of the test image. The linear fusion method is used to fuse the features of each channel, and then the obtained image is scaled to the size of the original image and input into GBVS saliency detection algorithm to get the corresponding saliency map. The original image and the fused feature image are input into GBVS saliency detection algorithm, and the saliency image is compared, as shown in Figure 9. It can be seen that the detection effect of the fused feature image is more obvious than that of the original image.

The multiscale sliding window only needs to scan the image area at the salient area, which reduces the generation of many candidate windows and improves the detection speed. Each window slides 20 pixels in turn. When using multiscale sliding window to detect pedestrians, there may be one pedestrian and multiple windows. The principle of merging is that if the ratio of the intersection area of two overlapping detection windows to the smaller one of the two windows is greater than a threshold, the threshold in this paper is set to 0.6, and then the window with higher output score is selected.

The experiment in this paper is mainly simulated on Caltech 7.5x and TUD-Brussels datasets, in which the deep network model is trained with Caffe deep learning framework. Because INRIA dataset contains a small number of pedestrian samples, it is not suitable for deep network training. This paper still adopts dollar’s open evaluation algorithm and uses the average log miss rate (MR) and the average false positives per image (FPPI) in the ROC curve to comprehensively evaluate the detection performance.

| Confidence threshold | Transformation of output layer | Improving the network |
|----------------------|-----------------------------|----------------------|
|                      | Accuracy | Recall | F1 score | Accuracy | Recall | F1 score |
| 0.1                  | 0.87     | 0.88   | 0.87     | 0.93     | 0.93   | 0.93     |
| 0.2                  | 0.92     | 0.87   | 0.89     | 0.96     | 0.92   | 0.94     |
| 0.3                  | 0.94     | 0.85   | 0.89     | 0.97     | 0.91   | 0.94     |
| 0.4                  | 0.95     | 0.84   | 0.88     | 0.98     | 0.89   | 0.93     |
| 0.5                  | 0.96     | 0.82   | 0.89     | 0.98     | 0.86   | 0.92     |
| 0.6                  | 0.97     | 0.81   | 0.88     | 0.99     | 0.87   | 0.92     |
| 0.7                  | 0.98     | 0.78   | 0.86     | 0.98     | 0.84   | 0.91     |
| 0.8                  | 0.98     | 0.73   | 0.84     | 0.99     | 0.81   | 0.89     |
| 0.9                  | 0.99     | 0.63   | 0.77     | 0.98     | 0.73   | 0.85     |
| Average accuracy     | 0.8713   |         |          | 0.9131   |         |          |

Table 2: Performance of fusion detection algorithm under low confidence threshold.

| Confidence threshold | Accuracy | Recall | F1 score |
|----------------------|----------|--------|----------|
| 0.02                 | 0.92     | 0.92   | 0.92     |
| 0.06                 | 0.93     | 0.91   | 0.92     |
| 0.11                 | 0.95     | 0.90   | 0.92     |
| 0.16                 | 0.96     | 0.89   | 0.92     |
| 0.21                 | 0.97     | 0.89   | 0.93     |

Figure 6: Performance comparison of models with different confidence thresholds.
In order to prove the effectiveness of the depth network and regression discriminant algorithm proposed in this paper, we evaluate them on the Caltech 7.5x dataset. After adding the regression discriminant algorithm, the average logarithm miss detection rate decreased by 0.93%, reflecting the effectiveness of the algorithm. After replacing convolution channel features with fine-tuning depth features, the average logarithm miss detection rate decreased from 14.84% to 13.20%, decreased by 1.64%, which fully shows the great advantages of depth network feature extraction (Figure 10).
5. Conclusion

This paper mainly analyzes the advantages and disadvantages of moving object detection algorithm based on background difference method and pedestrian detection algorithm based on improved YOLOv3 tiny network. Based on these advantages and disadvantages, the feasibility of fusing the detection results of the two algorithms is analyzed, and the overall process and specific implementation of the fusion detection algorithm are given. Firstly, the possible positions of two kinds of detection bounding boxes in the fusion process are analyzed, and the process of screening and fusion of the two kinds of bounding boxes is introduced. Then, it introduces the process of caviar dataset and the training process of pedestrian detection algorithm based on improved YOLOv3 tiny network and gives the performance of pedestrian detection algorithm based on improved YOLOv3 tiny network on the caviar dataset. Using ROI pooling operation, pedestrian detection network and pedestrian reidentification network are combined into an end-to-end system network. Finally, the performance of the fusion detection algorithm on the caviar dataset is given, which proves the effectiveness of the proposed fusion detection algorithm. Experiments show that our proposed end-to-end pedestrian detection and recognition network based on feature fusion not only improves the applicability of pedestrian recognition in practical applications but also improves the recognition rate of pedestrian recognition.

Data Availability

No datasets were generated or analyzed during the current study.

Consent

All authors approved the publication of the paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the 2020 Characteristic Innovation Project of Ordinary Universities in Guangdong Province: Research and Implementation of Image Recognition Algorithm Based on Artificial Intelligence Technology, Project no. 2020KTSCX399.

References

[1] Y. Zhang, “Pedestrian detection method of texture feature and deep learning,” Liaoning Gongcheng Jishu Daxue Xuebao (Ziran Kexue Ban)/Journal of Liaoning Technical University (Natural Science Edition), vol. 35, no. 2, pp. 206–210, 2016.
[2] Z. Yang, T. Wang, J. Deng et al., “Fast pedestrian detection based on motion information,” Journal of Computational Information Systems, vol. 11, no. 6, pp. 2303–2313, 2015.
[3] W. Zhou, S. i. Kaneko, M. Hashimoto, Y. Satoh, and D. Liang, “Foreground detection based on Co-occurrence background model with hypothesis on degradation modification in dynamic scenes,” Signal Processing, vol. 160, pp. 66–79, 2019.
[4] M. Qin, Y. Lu, H. Di, and W. Huang, “A background basis selection-based foreground detection method,” IEEE Transactions on Multimedia, vol. 18, no. 7, pp. 1283–1296, 2016.
[5] J. Chavat, S. Nesmachnow, A. Tchernykh, and V. Shepelev, “Active safety system for urban environments with detecting harmful pedestrian movement patterns using computational intelligence,” Applied Sciences, vol. 10, no. 24, pp. 9021–9042, 2020.
[6] Biao and L. Yang, “Robust foreground detection using block-based RPCA,” Optik: Zeitschrift Fur Licht- und Elektronenoptik: = Journal for Light-and Electronoptic, vol. 126, no. 23, pp. 4586–4590, 2015.
[7] Y. Yuan, F. Dai, L. An, D. Yin, Y. Zhu, and Y. Yan, “Research on fatigue detection method based on deep learning,”
[8] J. Q. Gui and L. U. Zhe-Ming, "Pedestrian detection based on modified dynamic background using Gaussian mixture models and HOG-SVM detection," International Journal of Innovative Computing Information and Control, vol. 14, no. 1, pp. 279–295, 2018.

[9] M. Qin and Y. Lu, "A foreground detection method based on cross-covariance subspace estimation," Beijing Ligong Daxue Xuebao/Transaction of Beijing Institute of Technology, vol. 38, no. 1, pp. 91–95, 2018.

[10] J. Dou, J. Li, Q. Qin, and Z. Tu, "Moving object detection based on incremental learning low rank representation and spatial constraint," Neurocomputing, vol. 168, pp. 382–400, 2015.

[11] M. B. Pont, G. Stavroulaki, and L. Marcus, "Development of urban types based on network centrality, built density and their impact on pedestrian movement," Environment and Planning, vol. 46, no. 8, pp. 1549–1564, 2019.

[12] S. Varadarajan, P. Miller, and H. Zhou, "Region-based mixture of Gaussians modelling for foreground detection in dynamic scenes," Pattern Recognition, vol. 48, no. 11, pp. 3488–3503, 2015.

[13] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," Applied Soft Computing, vol. 70, pp. 41–65, 2018.

[14] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Scherer, and M. Woźniak, "Accurate and fast URL phishing detector: a convolutional neural network approach," Computer Networks, vol. 178, Article ID 107275, 2020.

[15] H. Yousif, J. Yuan, R. Kays, and Z. He, "Object detection from dynamic scene using joint background modeling and fast deep learning classification," Journal of Visual Communication and Image Representation, vol. 55, pp. 802–815, 2018.

[16] Y. Wang, S. Pierard, S. Z. Su et al., "Improving pedestrian detection using motion-guided filtering," Pattern Recognition Letters, vol. 96, no. 1, pp. 106–112, 2016.

[17] B. Ashqar and S. S. Abu-Naser, "Image-based tomato leaves diseases detection using deep learning," International Journal of Engineering Research, vol. 2, no. 12, pp. 10–16, 2019.

[18] S. H. Ahmed, V. H. C. de Albuquerque, and W. Wei, "Guest editorial: special section on advanced deep learning algorithms for industrial internet of things," IEEE Transactions on Industrial Informatics, vol. 17, no. 4, pp. 2764–2766, 2021.

[19] J. Li, X. Shi, J. Wang, and M. Yan, "Adaptive road detection method combining lane line and obstacle boundary," IET Image Processing, vol. 14, no. 10, pp. 2216–2226, 2020.

[20] Q. Wu, Y. Zhou, X. Wu, G. Liang, Y. Ou, and T. Sun, "Real-time running detection system for UAV imagery based on optical flow and deep convolutional networks," IET Intelligent Transport Systems, vol. 14, no. 5, pp. 278–287, 2020.

[21] J.-H. Wu, W. Wei, L. Zhang et al., "Risk assessment of hypertension in steel workers based on LVQ and Fisher-SVM deep excavation," IEEE Access, vol. 7, pp. 23109–23119, 2019.

[22] J. Dou, Q. Qin, and Z. Tu, "Background subtraction based on deep convolutional neural networks features," Multimedia Tools and Applications, vol. 78, no. 11, pp. 14549–14571, 2019.

[23] L. Yang, G. Hu, Y. Song, G. Li, and L. Xie, "Intelligent video analysis: a Pedestrian trajectory extraction method for the whole indoor space without blind areas," Computer Vision and Image Understanding, vol. 196, Article ID 102968, 2020.